United Nations Educational,
Scientific and Cultural Organization

World Summit on the Information Society

Measuring Linguistic Diversity on the Internet

UNESCO Publications for the World Summit on the Information Society

United Nations
Educational,
Scientific and
Cultural Organization

2005

Measuring Linguistic Diversity on the Internet

A collection of papers by:

**John Paolillo, Daniel Pimienta,
Daniel Prado, et al.,**

Edited with an introduction by
UNESCO Institute for Statistics
Montreal, Canada

Published in 2005
by the United Nations Educational,
Scientific and Cultural Organisation
7, place de Fontenoy, 75352 PARIS 07 SP
Composed and printed in the workshops of UNESCO

Printed in France

(CI-2005/WS/06 CLD 24821)

Contents

1.

Introduction

UNESCO has been emphasizing the concept of "knowledge societies", which stresses plurality and diversity instead of a global uniformity in order to bridge the digital divide and to form an inclusive information society. An important theme of this concept is that of multilingualism for cultural diversity and participation for all the languages in cyberspace. There is a growing concern that in the efforts to bridge the digital divide, hundreds of local languages may be side-stepped, albeit unintentionally. Hence, the importance attached to linguistic diversity and local content as part of an Action Line of the WSIS Action Plan for which UNESCO has the responsibility of coordination.[1]

The issue of language diversity on the Internet proves to be central to the debate on the Information Society in a number of unexpected ways. At first glance the question seems to revolve around the communities using the Internet – allowing communities to talk to each other in their own mother tongues, but other questions quickly follow.

Through what channels does communication happen across the Internet? The World Wide Web is a series of information sources which provide little interactivity. Discussion fora and email provide more direct interchange. However there is insufficient information about the languages used in email or discussion fora (see some discussion of this by Paolillo's paper below Chapter 3 , including the work of Sue Wright).

For most language analysis researchers therefore turn to Web pages. Here, as in all communication, we must consider the character of the audience. A Web

[1] See Koïchiro Matsuura, Director General, UNESCO's address to Permanent Delegation on WSIS, 8 July 2005.

page is only read by people who have Internet access. Thus while linguistic diversity might be served by having Web pages in the ‘vanishing’ language of a very remote tribe, few people would read them as it is unlikely that tribal members have Internet access. Pages about the language of the tribe in a more international language would however serve an important role in drawing attention to the cultural value of the language concerned, and perhaps attract support for the linguistic group concerned. It is in addition, a contribution to the preservation of endangered languages.

The papers in this volume demonstrate that there are many technical problems in calculating language diversity on the Internet. We can easily produce a random count of Internet pages by using any number of commercial search engines, but we cannot judge how often Web pages are read or whether the reading of a page helped the reader in any way. Care has to be taken to ensure that the items searched for in different languages are equivalent in their value, meaning and usage (See Pimienta).

Languages and the Information Society

UNESCO Institute for Statistics is committed to an approach of measuring the Information Society which goes beyond a technocentric view to consider the social impact of the Internet and other channels of information dissemination. There are huge problems to be surmounted in terms of

— standardisation of definitions to achieve international comparability;

— identification of indicators relevant for developed and developing country policies;

— capacity building at national and international levels to allow quality data to be collected on a regular basis.

Language is the medium through which all information society exchanges occur. Language is a fundamental medium for all communication, the basis by which individuals and communities express themselves whether in oral tradition or in written text. For UNESCO, mother tongue teaching is a right of all children. UNESCO also supports language diversity ensuring that the richness of culture that diversity represents will be preserved for all countries and for the world as a whole.

The cultural issue of languages on the Internet serves as a counter to the perceived concentration of issues surrounding the information society on ICTs and their impact. UNESCO Institute for Statistics wants to present a view that raises issues about the importance of 'content' issues and the enabling environment, while at the same time indicating the technical problems in measuring culture and content in the information society.

The papers presented in this volume present a variety of different perspectives on this problem. The paper in this volume by Prof John Paolillo presents the view of a professional linguist working in the English speaking world. The report is divided into four main sections. The first section deals with the ethical framework for evaluating bias in computer systems and relates this framework to the status of the world's languages on the Internet. The second section addresses issues of pre-existing bias in the Internet's recent development, using statistics on the growth of the Internet and their relation to worldwide language diversity. The third section examines issues of linguistic biases that emerge in the context of the Internet. The fourth section looks at such biases in the technical systems of the Internet.

This has been complemented with a set of shorter papers from non-English contexts. They were organised and collected by Daniel Pimienta of FUNREDES a non-governmental project which has developed a system of enumerating language from the perspective of Latin languages. Pimienta sets out from the point of view of a civil society NGO, the barriers faced by local groups in accessing the Internet and a summary of currently available indicators. His paper is followed by a note from Daniel Prado presenting the reaction of the 'neo-Latin' language community to the perceived dominance of English. These shorter papers also include a very interesting Asian perspective from Mikami et al, as well as a note on the situation in Africa by Fantognan which summarises the situation from an African point of view.

The volume does not present any final answers as to how to measure languages on the Internet, but it seeks to dispel many of the myths surrounding existing published figures. It indicates that simply counts of Web pages are not sufficient and that much more further development is required from Internet Service Providers, and governments. Each author presents from his perspective a number of suggestions on how to address these required developments.

Language diversity on the Internet: an overview

Language diversity can itself be interpreted in a number of different ways. English is a tongue spoken relatively uniformly across the countries where it has dominance. Papua New Guinea has over 830 languages. Residents of English countries may have many other language skills, but few other countries can match Papua for diversity within one country. The numbers of speakers of neo-Latin languages, including those in the US, may be more than twice the numbers of people of English mother tongue (see Prado) but the US controls much of the machinery behind the World Wide Web (Paolillo, Mikami below). The relationship between languages on the Internet and diversity of language within a country indicates that even with a globalised network nation states have a role to play in encouraging language diversity in cyberspace. Language diversity can be viewed as much within a country as within the Internet as a whole.

It is a common assumption that English is the dominant force in the Internet. The papers in this volume differ in their views on this question. Paolillo agrees with the proposition and assumes, as do most others who see English as dominant, that this is a problem. Pimienta considers that English covers about half of all Web pages and its proportion of them is falling as other nations and linguistic groups expand their presence on the Web. Paolillo points to US dominance of the force behind the Web, both commercial and regulatory, to the extent that the latter exist. Mikami supports Paolillo on this point and emphasises the difficulties in reconciling US or western ICT and linguistic conventions with Asiatic scripts. There is a hint in Mikami though, as in Pimienta that change is coming as take up increases in India and China. This division of opinion on the dominance of English and the future of languages on the Web is one that cannot easily be resolved. At the end of the day it may be taken as illustrating the difficulty in measuring the use of languages on the Internet for which, despite the myriad ICT platforms, but partly because of the lack of regulation and the phenomenal growth, we have no good statistical indicators. Pimienta suggests that the area of

Internet indicators has largely been left to business marketing and that there is a need for high quality academic analysis.

Paolillo argues that telecommunication companies who profit from the demand for communication and technology services have a special responsibility to bear in mind the linguistic diversity of the countries whose markets they serve. Hardware and software companies have a similar influence on the linguistic make up of the Internet, by producing computers with keyboards, displays and operating systems that favour particular languages. The acts of computer companies locked in competition for market dominance have a detrimental effect on the climate of multilingual computing and on-line linguistic diversity. In such circumstances, the ethno-linguistic awareness of telecommunication companies, computer companies and Internet governing authorities will begin to broaden only if a critical mass of under-represented ethno-linguistic groups can command their attention. Hence, the general issue of emergent linguistic bias requires close monitoring on global, regional and local scales.

The measurement of languages on the Internet can be used as a paradigm for many issues of measuring content. To put it bluntly if we cannot measure this seemingly simply dimension of Web site content what can we measure? But we should not be so pessimistic. Mikami's project offers great potential for addressing many of the technical problems posed by the earlier papers, and avowedly adopting a non-English starting point.

We need to move to develop more intelligent indicators. Measuring the languages in the overall number of pages on the Web increasingly presents challenges caused by the sheer volume of Web content, but just because a page is on the Web does not mean it is used, or even 'visited'. If we are to truly measure the impact of the Information Society, we need to have statistics on how the Internet is used, and by whom. In this view Web pages are simply the supply side, in all its linguistic homogeneity or diversity, and not necessarily a reflection of use and demand. In an oversupplied market of say English language Web pages offering a variety of services, many poor quality sites may receive few or no visits. It is also a common observation that many Web sites remain without updates or modification for years.

From an economic perspective the Web has some aspects of a free market and some market failures (see Paolillo). Web sites are developed to meet the needs of a particular audience. If there is little domestic Internet access, business Web sites will be orientated to an external foreign market, and hence will be written in an international language like English. On the other hand low usage of an Internet site, and low maintenance costs of Web sites mean that they may continue to exist and be registered on search engines long after the last visit of a potential user. Ideally we need analysis of 'useful' sites and visitors who use them.

Even within the limitations of the present studies these problems indicate how little statistics on the percentage of people with computers, or the number of Internet subscriptions (both Millennium Development Goals indicators) say about the fundamental changes in information exchange brought forward by the Information Society. If we set aside the arguments for or against English dominance we can see in this volume the rapid expansion of Internet use in Asia and hence the growth in Asian language Web sites (Mikami), as well as the way in which the expansion of the World Wide Web has brought together the 'neo-Latin' communities to consider their place in a global knowledge society (Prado). It is important to underline that the digital world provides an enabling environment for as many languages as possible. This could ensure true language digital inclusion.

Next steps

It is hoped that this report indicates the need as suggested above for all agencies to come together at national and international levels. The World Summit on the Information Society presents a suitable context for discussions of both language policy and technological standards, as well as the policy objectives to be achieved by promoting freer exchanges of information.

The studies indicate how important it is to understand the cultural context for information exchange. Given this it seems unlikely that any global perspective is likely to provide comparable or policy relevant data that is sufficiently sensitive to the technical and policy issues involved. It may instead be preferable for monitoring to be led by regional initiatives, whose studies could then be combined in an overall global perspective. The FUNREDES project and Mikami's Observatory are two potential projects which indicate how such a regional network might operate.

To conclude, as Paolillo aptly states in his report, actions may be needed to ensure that the values of digital access and digital literacy are upheld especially on behalf of the many linguistically diverse and developing countries of the world.

UNESCO recommends to the national, regional and international levels to work together to provide the necessary resources and take the necessary measures to alleviate language barriers and promote human interaction on the Internet by encouraging the creation and processing of, and access to, educational, cultural and scientific content in digital form, so as to ensure that all cultures can express themselves and have access to cyberspace in all languages, including indigenous ones.[2]

[2] See: UNESCO Recommendation Concerning the promotion and use of multilingualism and universal access to cyberspace and 32 C/27 document, 2003, UNESCO Declaration on Cultural Diversity, Paris 02.11.2001.

2.

Models and Approaches

a) Linguistic Diversity in Cyberspace – Models for Development and Measurement

Daniel Pimienta, FUNREDES

Introduction

There is a word that actors in civil society use, especially those who think that the essence of the new paradigms of knowledge societies in participative democracy is to call upon an ethic of processes that can be used to translate our vision. This word is **consistency**.

Consistency between word and action is what allows us to believe in pronouncements and forgive errors, which in a process approach, become opportunities to learn, draw lessons from, and continue to believe. This process, specific to action-research, particularly adapted to addressing development issues, is a driving force in this document, the purpose of which, more than presenting solutions to a question as complex as linguistic diversity on the Internet, is to question false evidence, present provocative points of view, indicate paths for reflection and action that come out of well-trodden paths and preconceived ideas, that can take into account the complexity of the topic. All this occurs, of course, with the humility of a researcher who proceeds by trial and error and the confidence of a man of action who is actively working in the field.

Consistency will express itself in many ways in this document:

— Choice of communicating in the mother tongue, a fundamental right;

— The will to allow diversity to be expressed in the selection of people, knowledgeable in the field, invited to express themselves. As much as possible, we have tried to cover geographic regions, cultures, languages, backgrounds, sectors, age groups and both genders. It is obvious we have not been completely successful – for example, we regret that the space for female authors is not larger – but consistency is above all expressed in the sincerity of the intention;

— The decision not to write in a diplomatic language and to risk being provocative, never gratuitous, at times grateful, always based on experience in the field and with the idea of opening peoples' minds not just for the sake of it.

A Structured Approach to Integrate ICTs and Human Development

The "digital divide" is a concept that has become very fashionable and has spurred much reflection and many international meetings. The consensus vision of civil society (Pimienta, 2002, Communauté MISTICA, 2002) is that we cannot ignore other dimensions of the divide and we should avoid the simplification that places the blame on technology. What follows is a rather unique table of reading and analysis of the use of ICTs for development to illustrate the fact that resolving the digital divide is not simply a question of access to technology – far from it – and that the question of language also plays a key role.

The purpose of the table is to identify successive obstacles to overcome so that ICTs can be used for human development. The table implies a progression in the listing of obstacles, from infrastructure to infoculture by way of infostructure. It is likely that this progression does not correspond exactly to the reality experienced by each person or social group, and that the order of the factors depends on the individual context. Nevertheless, for practical and pedagogical reasons, we have simplified this complex reality as a series of obstacles to overcome to reach the next levels.

Table 1. ICT for Development, a Long Road Filled with Obstacles to Accessing Human Development Tools

Usage Level	Description of use and obstacles	Language Issues
ACCESS	*The possibility for a person or group of people to have the physical means of using ICT.*	
	The obstacles to be overcome in order to have access are many, and can present themselves in the form of successive stages:	
	– Existence of Infrastructure Service side: Suppliers of access to ICT and telecommunications networks with capacity large enough to serve a quantity of users with acceptable response times and congestion rates. User side: Computer equipment required for access with appropriate characteristics for providing acceptable performance. This can be on an individual (personal work station) or	**– Existence of the infrastructure.** Interfaces should allow access in the user's mother tongue and be adapted to one's culture. In terms of hardware, the language issue relates to computer keyboards, as well as to software – in managing characters associated with a language and that should be codified for computer processing. However, the operational aspect of software programs related to language does not stop

Usage Level	Description of use and obstacles	Language Issues
	collective basis (telecentres or internet kiosks).	at coding. For optimal functioning in a given language, publishing programs require a corpus and dictionaries for spell-checking and grammar verification. A more ambitious long-term vision would assume that machine translation programs be part of the operating layer, not of the application layer. A great deal of work remains to be done on translation programs to extend them beyond the so-called dominant languages. This is a perfectly suitable area for open-source software, but unfortunately there is practically no activity here, and developers should be encouraged to fill the gap. ICANN (Webopedia, 2005b), is now considering the introduction of internet domain names in all languages (Wikipedia, 2005a).
	– Affordable access to infrastructure The costs of access to the infrastructure should be	**– Affordable access to infrastructure** The principle of universal access should include

Usage Level	Description of use and obstacles	Language Issues
	affordable for users. There are clearly many direct or indirect factors in the price[3] equation, and individual or collective access present different parameters. It suffices to compare, for example, the cost of ADSL access (Webopedia, 2005a) (between 10 and 50 US$ a month) and salaries in the social pyramid to discover that this represents more than a year's salary for a sizeable proportion of humanity (those who live below the poverty line), a month's salary for another sizeable proportion (that of the people in the South), approximately 10% of the monthly salary of the middle class in developing countries, and in the order of 1% of the monthly salary of the middle class in developed countries.	consideration of an access cost that is consistent with the income level of the populations in question.

[3] Direct costs, such as the cost of a computer, the Internet Service Provider (ISP), in some cases, that of the information provider, that of hosting the server or Internet domain (because access means also producing content); or indirect costs, such as savings that allow access, for example telephone IP or travel costs saved, or costs of maintaining hardware and training personnel.

Usage Level	Description of use and obstacles	Language Issues
	The first divide is not digital but economic and social. Resolving the first two aforementioned problems should[4] represent what the ITU and telecommunications regulators call **"universal access"** (ITU, 2003). But if this is a prerequisite for closing the digital divide, it is far from sufficient.	
	– Functional literacy That the person who uses the infrastructure has the functional competency to read and write in his own language. This is probably the second divide that must be resolved when we assert that there is an "Internet for all."	**– Functional literacy** The multimedia component of ICTs should certainly be considered in adapting interfaces to help for people who are illiterate. However, we must look at the evidence if we are referring to access to knowledge and not simply access to technology. Functional literacy is a higher priority than access to technology for illiterate populations.

[4] We are using "should» because too often the economic aspect is neglected in universal access plans and the concept is understood as total physical coverage of access to infrastructure, which certainly creates business for sellers of hardware but not necessarily for users.

Usage Level	Description of use and obstacles	Language Issues
		There is also the issue of languages that are oral and have no written form. For these, the digital world can be a fatal barrier unless a written form can be codified.
	- Computerizing the alphabet The mother tongue of the person using the infrastructure must be put into digital form. To accomplish this it must have a written form and the characters of its alphabet be easily coded. Unfortunately, this is not the case for the majority of languages still being used.	**- Computerizing the alphabet** This is still a major obstacle for a large number of languages, and should be considered a major priority at the outset. Work is under way in the context of UNICODE (Wikepedia, 2005b) – it should be maintained and expanded.
USE	*The potential to make the use of ICTs effective (achieving a preset goal) and time-efficient.*	
	For this the person must have a large number of capabilities in managing digital tools as well as an understanding of the concepts, methods and culture associated with cyberspace. The sheer magnitude of competencies required for	**Digital literacy** The formidable effort needed for digital learning should be conceived and conducted in the mother tongue of the populations and in the context of their cultures. It is important to note that this imperative

Usage Level	Description of use and obstacles	Language Issues
	digital literacy should not be underestimated. Learning cyberspace, that should not just involve simple training in using certain computer programs, but should also include a holistic vision of societal implications and impacts[5] in using ICT for development, is without a doubt the most difficult nut to crack – the most important and yet most neglected aspect of the effort to reach consensus in closing the digital divide. Contrary to widespread belief, the three pillars of the information society we are building are not telecommunications, hardware and software, but the ethic of information, education and participation.	criterion should also be applied to the interface of government computer applications.

[5] Political, economic, social, cultural, linguistic, organisational, ethical, biological, psychological impact.

Usage Level	Description of use and obstacles	Language Issues
TECHNOLOGICAL APPROPRIATION	*When the user is skilled enough so that the technology is* ***invisible in his personal use***	
	An example is a pair of glasses – an optical technology that is put on the nose in the morning and is completely forgotten throughout the day, or in the field of ICT, the person who uses his telephone without his being constantly conscious of it acting as a tool for conversing at a distance. For ICTs, this appropriation obviously requires more sophisticated capabilities in terms of using a PC and computer applications that intervene in the process. Therefore, it is clear that a certain level of expertise is required to research information, communicate by email, and indeed behave in the virtual community. In addition to good digital education, a minimum level of **practice** is needed to get to this stage.	How can technology be made transparent if access to it requires using a language other than the mother tongue? This clearly reinforces the arguments put forward for attaining the previous levels.

Usage Level	Description of use and obstacles	Language Issues
CARRIER OF MEANING	*The capacity to make use of ICTs has a social significance for the person in his personal, work and community context.*	
	This means going beyond recreational use as a simple interpersonal communication tool and directing its use to human development goals. This is where fundamental capabilities should appear so as not to be a mere consumer but to become a **producer** of **content**, for example, or a **creator**, of **virtual communities**. It is clear that education is required in order to achieve this level of personal development.	Language is essential at this level to create the possibility and motivation for producing content and local virtual communities. It raises the issue of multilingualism and the need for navigation bridges between languages.
SOCIAL APPROPRIATION	*When the user is skilled enough so that the technology is* ***invisible in his social use.***	
	This level requires a lucid understanding of the social impacts of using ICTs for development and the **cultural and ethical implications** related to this use (culture/ethic of	The ethical and cultural aspects of networks are not entirely neutral and should pass through a kind of cross-cultural filter (indeed a certain form of syncretism) in terms of

Usage Level	Description of use and obstacles	Language Issues
	the network and of information, and knowledge of the methodological aspects related to productive use for development). In addition to good digital education, practical experience focused on development is needed to accomplish this stage.	local cultures and ethics. Language, being one of the carrier vectors of cultures, is not free from complex and sensitive issues.
EMPOWER-MENT[6]	*When the person and/or community is in a position of* ***transforming his social reality*** *thanks to the social appropriation of ICTs for development.*	
	This requires not just the capabilities themselves but their being **put into practice** at both an individual and a collective level. This putting into practice obviously requires the application of values associated with the culture of the Internet and the culture of information, a propensity for collaborative work, acquired invisibility	The closer we get to the end of the chain that leads from access to development, the clearer it is that culture becomes important, without losing sight of the fact that language is often an important issue. What does "empowerment" mean and how does it manifest itself in each culture?

[6] This word unites the sense of receiving and assuming the capability, as well as the notion of acquiring power by using it.

Usage Level	Description of use and obstacles	Language Issues
	of the ICT, and proactive participation.	
SOCIAL INNOVATION	*When the action of transforming social reality is a carrier of* ***original solutions*** *created by the person or community.*	
	The new paradigm of working in networks carries the seeds of innovation, in particular social (new forms of organisation, new responses to known problems, etc.)	What does "innovation" mean and how does it manifest itself in each culture?
HUMAN DEVELOPMENT	*When the options of individual and collective freedom are open to the person or community and can be exercised in the form of* ***"capabilities."***[7]	
	This is the end of the process, but it should remain clear that in any social process what is found in the end can only be what has been maintained during the whole process from its very beginning.	***Options of freedom*** *in the form of* ***"capabilities."*** What does "participation" mean and how does it manifest itself in each culture? Is real "participation" truly

[7] Development can be seen as a process of expanding real freedoms which people have. Taking into account human freedoms or capabilities differs from narrower views of development, such as those that identify it with growth in the GNP, increases in personal income, industrialization, technological advances or social modernization" (Sen, 2005).

Usage Level	Description of use and obstacles	Language Issues
	Therefore the choices that freedom provides cannot flourish unless the **participation** of the people and their communities has been a reality during the whole process.	possible in social processes and is it still possible if a language other than the mother tongue is imposed?

The Information Society – Issues at the Crossroads of Languages and Cultures

The discipline of information ethics was born in recent years and UNESCO has made numerous contributions to it. Linking this discipline to the question of cultural and linguistic diversity opens perspectives and avenues for reflection that are completely relevant to our debate. A conference in 2004[8] organized by the ICIE (International Centre for Information Ethics) was devoted to this theme, and a book containing papers from the conference will be published in late 2005 (Capuro, 2005). These papers are also relevant to the topic under discussion here.

Among these, Charles Ess (2004) has said that, contrary to the common hypothesis that ICT is culturally neutral, a large number of studies have shown that ICTs originating from Western cultures, especially North American, carry, and in a certain way promote, their cultural values and preferences in terms of communication. According to Charles Ess, this is apparent in the many ways these values and preferences enter into conflict with the cultures which receive the technologies, especially indigenous, Asian, Latin and Arab. The resulting

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

conflicts translate into sometimes spectacular failures in well-intentioned efforts to overcome poverty and marginalization (Postma, 2001). Ess goes even further by pointing out the danger of "computer-assisted colonization" that may be the product of a naïve plan to "connect the world" but does not acknowledge the risks proven to have a negative impact on domestic cultures and values by the careless introduction of ICTs.

However, Charles Ess reassures us by indicating that such conflicts are avoidable, first by adopting a conscious attitude towards cultural issues. He shows us the pathways to designing human-machine interactions that respond to this cultural criterion (Hall, 1976).

If we agree that digital education is an essential component of the transition to an inclusive information society, it is also clear that such education should fulfil the fundamental ethical criterion of respect for cultural and linguistic diversity, and therefore avoid the ethnocentrism and colonization implicit in technologies.

There is another essential and cross-sectional issue among those associated with the information society – a public domain of knowledge that should be free from the marketplace that derives from open content and software. This question also relates to language diversity in the information society.

José Antonio Millán (2001), the Spanish expert on languages and the Internet, reminds us that our languages remain the most complete interface that exists. In either oral or written form, they are increasingly used to interact with a variety of programs, such as information search engines. The linguistic knowledge incorporated in programs, such as auto-correction, outlining, and text-voice transformation, are not really visible to the user. However, its economic importance is enormous. Basic resources that are the substrate of programs most often have their source in research funded from public sources. However, these programs often benefit from commercial software which are not open-source and therefore cannot be improved or extended (e.g. to address minority variations in the most-used languages). Furthermore, they cannot be used as a base from which people using minority languages can create their own software. According to Millán, the democratisation of software in minority languages will occur by the freeing up of language resources (under GPL licenses or similar agreements – Wikipedia, 2005c) produced with public funds or that are simply in the public domain.

Whatever the case, open-source software programs, which by their very nature should be playing an important role in the language sector, have only a modest presence, and the community of developers needs to be engaged.

The theme of open content naturally leads us to consider the changes required in a system of scientific publishing that is considered by specialists who work in the field of the information society (Guédon, 1998) as obsolete because it hinders the sharing of scientific knowledge, particularly with the countries in the South. This system is beginning to be questionned by initiatives such as the Public Library of Science and the Berlin Declaration on Open Access to Knowledge in the Sciences (ZIM, 2003). Linguistic diversity has everything to gain from evolution in the system of scientific publishing towards models taken in large part from ICT and based on the notion of open content.

Underlying this situation, and given a certain lack of movement in the countries concerned, is the absence of language policies; in fact, the critical gap to close, as emphasized by José Antonio Millán, is that of a veritable policy on digital content that obviously includes a language policy for the digital world. In this regard, the role of international organizations such as UNESCO can be to sensitise member States on the importance of voluntary policies that promote multilingualism.

Measures and Indicators

Is it reasonable to define and direct linguistic policies in digital space without having sufficient, accurate and precise indicators on the situation of languages and their progress?

Quite paradoxically, the world of networks, born and developed in universities, in a way surrendered measuring the language situation to marketing companies, whose intentions are different from those of scientific publication, and who are therefore not very concerned with documenting their methods. Disorder and confusion regarding the state of languages on the Internet has been the result, which can lead to disinformation. Therefore, while the proportion of English-language speakers who use the Internet has gone from more than 80% in the year the Web was born to 35% today, the figures circulating in the media, against all evidence, are reported as stable between 70% and 80%!

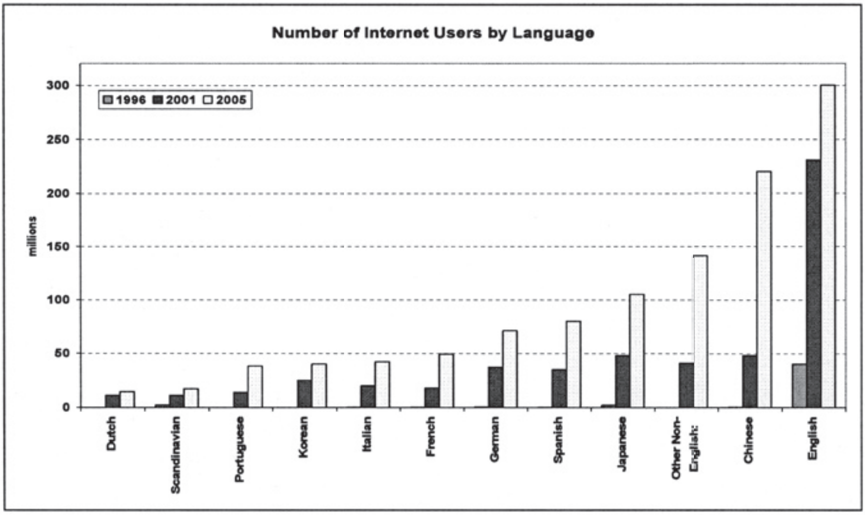
It is urgent that the academic world regains its role in this area along with public institutions, both national and international. There are clear signs that this change is finally occurring! For an update, consult the proceedings online of the meeting multilingualism in cyberspace[9] organized by UNESCO with ACALAN (Academy of African Languages) and AIF (Intergovernmental Agency for Francophone Countries and Regions) held in Bamako.

While waiting for this process to develop accurate documented indicators updated at the speed of the development of new media, gaining a clear perspective on this situation and its trends is extremely difficult.

I – For data on the proportion of Internet users in each language group, one source has distinguished itself for a number of years. With great regularity, Global Reach has supplied figures which certainly come from multiple sources and that are not consistent in terms of methodology, but at least they are known (Figure 1). The figures are not completely accurate but at least they exist and should be frequently updated. Even if we ascribe only relative confidence in them (20% margin of error), they provide a reasonable perspective on the growth of Internet users by language group.

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

Figure 1: Number of Internet Users by Language



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

II – data related to languages on the Web, there are a certain number of simultaneous approaches:

1) One consists of extrapolating figures from search engines by language. This is the easiest, and it gives acceptable orders of magnitude, but not figures reliable enough to constitute serious monitoring, given the weaknesses of the algorithms at recognizing languages and the erratic behaviours of the engines in arriving at totals.

2) Another was launched by one of the first studies on the topic that Alis Technologies conducted in June of 1997 with the support of the Internet Society. Their method was subsequently used by others, in particular the OCLC (Online Computer Library Centre), that conducted a study that seems to be the reference upon which numerous authors and media continue to base their figure of more than 70% of Web pages being in English (O'Neill, 2003). The method consists of creating a sample of several thousand websites through a random selection of

IP addresses (Wikipedia, 2005d), running language recognition engines on this sample of sites, and then generalizing the results.

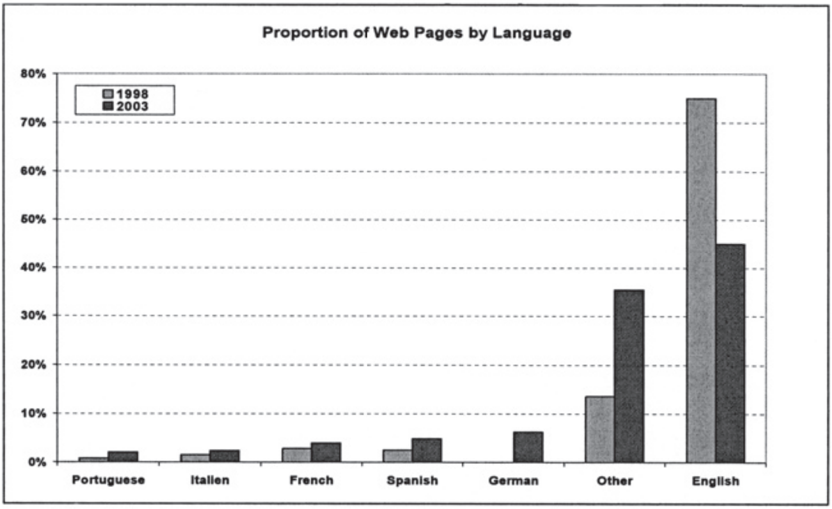
It shares the limitation of the first approach in terms of language recognition algorithms. Important progress has been made since 1997 but future methods will decisively increase the reliability of the results.

Another limitation is of major concern because it is statistical. The mathematical processing planned for a random variable (as is the case of the random sample of websites to which is applied language recognition) is to analyse the statistical distribution to extract the mean, the variance and deduce the confidence interval. Taking only one random sample cannot provide any credible result. What does 8,000 websites represent when there are 8 billion Web pages indexed by Google? From the little documentation published, it appears however that the language figures produced by OCLC use this method.

3) There is a large category in which figures are published and no methodology is revealed. It is therefore impossible to validate the results. This is the case of the Inktomi study in 2001, which was announced with a great marketing flourish. It had gross errors, such as presenting the worldwide percentages of Web pages in a limited number of languages, the total of which was 100%!

4) The last category comprises some rare methods that were documented, such as the original approach employed by Xerox researchers in 2001 (Grefenstette & Nioche, 2001). Among these methods, FUNREDES and the Latin Union has used a specific one since 1996 (see Figure 2).

Figure 2: Proportion of Web Pages by Language



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

The method consists of using search engines to obtain the number of occurrences of a given word in a given sector of cyberspace, such as Web pages or discussion groups. A sample of keywords in each of the languages under study is constructed with attention paid to providing the best semantic and syntactic equivalence among these. The values for the appearance of each word measured by the search engines are compiled for each concept in each language. These values are processed as a random variable, the mathematical distribution of which is analysed using traditional statistical tools, such as means, variance, confidence intervals, and Fisher's law. The result consists of an estimate of the weight for the presence of each language relative to English, which is considered the reference language. This estimate is then quantitatively validated by statistical instruments such as the confidence interval. Repeating this measurement at successive intervals provides an evolving picture of the presence of languages in the sectors of the Web under study. The value of this method is that it is providing consistent results that can indicate trends.

Although this methodology has not been subject to academic criticism since it began to be used, it does have certain limitations:

- It provides a value of the percentage of Web pages in a language (German, Spanish, French, Italian, Portuguese or Romanian) compared to English but not an absolute value. To obtain an absolute value, there must be an estimate of the absolute magnitude of the presence of English based on the increasing difficulty and uncertainty of checking for occurrence of key words given the multiplication of languages on the Internet;

- It is difficult linguistically and in terms of cost to add a new language;

- It gives a value that corresponds to the cyberspace of pages indexed by search engines but does not take into account the invisible Web (Bergman, 2001). But do unindexed pages really "exist"?

- The method is above all very dependent on the accuracy of search engine[10] counters, can be unreliable since they take increasing liberties with processing searches by word.[11]

On the positive side, the advantage of this method is that it provides a means of consistent monitoring over a long period, of examining the cyberspace sector other than just the Web,[12] and above all, of producing a series of unique and very significant indicators based on research by country and domain (Pimienta, 2001).

Perspectives on New Approaches

The Observatory of Languages project (see article by Yoshiki Mikami) promises to fill the void and provides the responses that policy makers need in order to develop strategies and measure their impact.

[10] The major part of the work today in conducting measurements consists of verifying the behaviour of the search engines, selecting the most accurate, and compensating for their erratic behaviours, especially in their processing of diacritic symbols.

[11] It is probable that very soon search engines will provide results involving documents with the translation of the search words into other languages.

[12] It has also provided a first approximation which is certainly large but interesting in terms of the growth of cultures on the Internet.

Our experience in the field has made us think that a promising approach that does not yet seem to be used would be a method similar to that used by Alexa to paint a portrait of the most visited sites and to provide other invaluable information. Alexa compiles data on the behaviour of a large number of users who have accepted to download spyware to their computers; this then provides extremely detailed statistics. Following the same method, we can imagine a programme that would be capable of measuring the languages used in a variety of contexts which would be relevant to indicators such as the composing and reading language of emails, languages of sites accessed, etc.

Bibliography

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Glossary

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

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

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

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

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

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

b) Political and Legal Context

Daniel Prado, Latin Union

As a general rule, the most prominent European languages are experiencing a significant decline in use in scientific and technical communication to the benefit of English. With the exception of certain languages with little distribution that have seen a resurgence in recent years, the great European languages such as German, Spanish, French, Italian, Portuguese, Russian and the Scandinavian languages have been affected (Hamel, 2002).

Among the European languages, the Latin ones have been particularly affected, whether in technical publishing, scientific conferences, international organizations, media or teaching, etc.

In November 2002, the first international conference on the place of Latin languages in scientific and technical communication was held (UNILAT, 2002a), gathering specialists on language policies in countries and regions of three major language groups, French, Portuguese and Spanish.

During this conference, statistics and observations showed the vertiginous loss of the vitality of Latin languages in many sectors related to scientific and technical matters. According to Calvet (2002), in spite of being the official languages in more than a quarter of the planet's countries (27,53%) and spoken by nearly one billion speakers, languages such as French, Spanish, Portuguese, Italian, Romanian and Catalan, as well as twenty others with fewer numbers, produce only a tenth of the scientific publications written in English, based on the most important international databases.[13] Indeed, Hamel suggests, that English is the language of 80% to 90% of publications in the natural sciences and 74% to 82% of those in the social sciences and humanities, whereas the three most

[13] It is often considered that English-language scientific journals are over-represented in these international databases, and that by the same token, journals in countries other than those in the OECD are under-represented (UIS).

common Latin languages account for 12% of publications in the social sciences and 18% in the humanities. However, Hamel nuances his observations, by indicating that the statistics come from databases of scientific publications and that book publication is as vigorous as scientific journals. It is interesting to note that the publishing world of the Latin languages is doing relatively well, with 18.9% of world production (Rousseau, 2002), but that this figure is mainly comprised of fiction (Leáñez Aristimuño, 2002).

It is well understood that compared to most languages on the planet, the Latin ones are not the worst in terms of the distribution of knowledge. Indeed, for every 100 pages measured in English, there are close to 38 pages (UNILAT, 2005) in Latin languages.[14] French is the second in international use, Spanish a comfortable third, and the teaching of Spanish is increasing around the world. Portuguese has good demographics and is on more than one continent, and Italian remains a prestige language in spite of its low demographics and geographic concentration (Italy, Switzerland and San Marino).

It should not be forgotten that English, with two and a half times the number of speakers than all the Latin languages combined, has two and a half times the number of Web pages than that of all Latin languages combined. It also must not be forgotten that scientific publications written in English represent two-thirds of all scientific publications in the world, whereas all the Latin languages combined only represent one tenth.

It is far from our intention here to ignore the decline of the scientific and technical use other languages are experiencing such as those of northern Europe, especially Scandinavian, in which a huge proportion of scientific vocabulary is disappearing because of English unilingualism in practice among specialists in certain disciplines (Nilsson, 2005). It is also far from our intention to want to dramatize the situation of European languages, when, as Leáñez tells us, 98% of languages on the planet do not even have basic special-term vocabularies, whether administrative, scientific, technical, legal or business. The alarm bell must

[14] The study was conducted on the top five Latin languages in terms of number of speakers, namely Spanish, French, Italian, Portuguese and Romanian.

be sounded on this worrying situation, since no language other than English is free from it.

To return to languages on the Internet, even if the FUNREDES/Latin Union statistics show that in 2003 close to 14% of Web pages were published in at least one Latin language, close to 45% were in English. Even German, with ten times fewer speakers, had only two times fewer pages than all the romance languages combined. But the most worrisome in terms of Latin languages on the Internet are the unpublished data, the invisible Internet – the Intranets, databases, distribution lists, fora, etc. We do not have any statistics on these, but simple everyday practice shows the overwhelming predominance of English as soon as an international technical discussion begins in an electronic discussion group, or a scientific database has international implication, or even a chat room of young people start conversing about their favourite star. This phenomenon can be easily explained as telematic networks serve a population of international researchers, and it is pointless to repeat that English is perceived in the scientific world as the main language of communication. But what is regrettable is that this model has not evolved, and therefore populations and communities who are less skilled at handling the English language are excluded.

Leáñez reminds us that "a language that has little value is little used, and a little-used language has little value," affirming that if our own languages do not meet our needs, we will learn and teach another.

Given the above, UNESCO's Action Plan (2005) for the WSIS falls at an opportune time. Indeed, in the first chapter, one of the recommendations concerns cultural and linguistic diversity. It is recommended to "Create policies that support the respect, preservation, promotion and enhancement of cultural and linguistic diversity and cultural heritage within the Information Society." Currently, not one country with a Latin language has a policy which allows full use of Latin languages, notably in the Knowledge Society and in the Sharing of Knowledge.

Furthermore, with regard to language policies, countries with Latin languages, aside from rare exceptions, are too exclusively focused on administrative aspects on the one hand, and on the protection of endogenous languages on the other (and in rare cases, on consumer protection). Not creating the necessary control apparatus and not giving themselves the means to put into practice what the laws on the books extol, they do not have sufficient resources to develop their language, and therefore leave an open space that is quickly filled by English, notably in scientific discourse, technical documentation, higher education, the Internet, etc.

With the exception of Québec, Catalonia and France, no state-run organization in provinces or countries with Latin languages has taken charge of all the components that lead to an overall policy of development, enrichment, modernization and spread of a language. In Belgium, Switzerland, Spain and Portugal, institutions exist but only partially carry out this task. Furthermore, in regions or countries that are more advanced in terms of language policies, a policy that supports digital multilingualism is absent. Too often, it is private associations with little means, or intergovernmental organizations with no clear mandate, that have to come in and do the work.

The good news is that many minority or "minoritized" languages, contrary to what is happening with the larger languages, are participating in scientific and technical communication at a level not seen before. This is notably the case with Catalan, but also with Galician, Basque, indeed Sardinian, and others. However, much remains to be done, so and that they are able to cover all the spheres needed for their populations to flourish.

There rests the thorny issue of access to information when it is produced in a language in which we are not fluent. Translation, as we know, is expensive. For certain processes, such as the translation of a tender from an IGO, translation is a slow process. Machine translation will never replace human translation, but will only help in improving performance, speed and affordability. However, it is an indispensable instrument for the transformation needed in the world of digital and paper publishing. No current system does satisfactory translations for the most common language pairs. All machine translations need revision. But the most serious point is that most machine translation systems or computer-assisted translations cover only a small number of pairs of languages.[15]

[15] Indeed, they counted far fewer than 100 languages that were processed by machine or computer-assisted translation systems, out of close to 6,000 existing languages.

The quality of existing systems should be improved, and given their technological evolution, this will no doubt occur, but nothing leads us to believe that the crucial percentage of less than 1% of languages linked through computer translations can be increased soon. Voluntary initiatives should point the way for translating languages that have no market interest for commercial enterprises. The Latin Union has initiated certain processes in this direction,[16] as well as United Nations University. It remains to be seen if others can come forward to address the least promoted languages.

What is to be done therefore to accomplish a multilingual digital world? The recent discussion in France on a European Google, that was picked up by the international press, has inspired some ideas (Millán, 2005), and UNESCO has been emphasizing the role of libraries and the format of their collections. One idea is of implementing large-scale projects of computerizing collections, calling on countries as well as inter- and non-governmental organizations or private Internet providers to do so, but only those who commit to respecting a code of ethics regarding the use of this information. Obviously appropriating this digital information for commercial purposes or demanding distribution rights or licenses must be impeded. The objective is to widely distribute digital contents free of charge, the only means of guaranteeing veritable linguistic diversity.

In its everyday use, the Internet spontaneously shows us new types of communication – independent and autonomous press services, blogs, citizen initiatives that are born every day. These demonstrate that voices other than unilingual monopolies exist. Perhaps we should be observing these more closely, to support and inspire them.

As a general rule, countries with Latin languages are behind in issues related to the presence of their languages in the digital world. Given this, many actions are needed – the creation of a voluntary policy of digitizing information and catalogues which at present only exist on paper, and an ongoing policy of scientific publishing in national languages, or at least translation of publications if they are in English, and their immediate distribution on the Internet; the implementation of a charter of the rights of citizens to be informed in their language

[16] Notably by introducing Romanian into the Atamiri project.
(http://lux0.atamiri.cc/forum/init.do)

and therefore the obligation to respect multilingualism on the websites of international organizations, international companies, and obviously an obligation on the part of national corporations to distribute materials in the local languages; and, expansion of machine translations, especially for languages not yet covered by the software.

The Latin Union is preparing a second meeting on the place of Latin languages in professional and technical communication to put into practice the recommendations that were proposed in the first meeting (UNILAT, 2002b). The plan is to have mechanisms for consultation, monitoring, gathering and disseminating statistics, and actions that encourage publishing and research in the Latin languages, and the development of effective linguistic tools. This meeting will be held in 2006 in Spain, in close collaboration with the *Trois Espaces Linguistiques* (representing the French-speaking, Spanish-speaking and Portuguese-speaking regions of the world), and it is hoped that solutions to the above-mentioned problems will be found.

Bibliography

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

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

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

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

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

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

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

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

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

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

3.

Language Diversity on the Internet

John Paolillo, School of Informatics,
Indiana University

More than two decades since the Internet arose in the English-speaking world, the representation of different languages on the Internet remains highly skewed toward English. English remains the most prevalent language on the Internet, and some very populous languages have little or no representation. To what extent does this situation represent a bias toward English, and away from other languages? This report[17] addresses this question by introducing the ethical framework of Friedman and Nissenbaum (1997) for evaluating bias in computer systems, and relating it to the status of the world's languages on the Internet. This framework helps us to interpret the probable causes and remedies for potential bias. Current claims about the status of the world's languages on the Internet are also presented and re-stated in terms of their meaning with respect to this framework, which guides us to considering not only the distribution and use of languages on the Internet, but also the social institutions guiding governance and development of the Internet that may lead to what Friedman and Nissenbaum term "emergent bias". Finally, we consider issues of linguistic bias in the technical systems of the Internet.

[17] This report was assisted by : ELIJAH WRIGHT and HONG ZHANG, Indiana University, Baskaran, S., G. V., Ramanan, S. V., Rameshkumar, S., SHOBA NAIR, L., VINOSHBABU JAMES, VISWANATHAN, S. Anna University, Chennai, India. The complete version of the report can be accessed from: http://ella.slis.indiana.edu/~paolillo/paolillo.diversity041027.pdf.

Bias, multilingualism and computer systems

The "digital divide", that is, the unequal distribution of access to digital information sources and services, stands out as one of the key policy issues of the present digital information era. Governments, international agencies, citizens' groups, corporations and others all seek to take advantage of the promises of lower cost and instantaneous information access by moving many of their communication functions to networked computer media. But if traditional social barriers, such as socio-economic status, education, ethnicity, gender, etc., hamper access to digital information, then policies must be directed to equalizing access for these benefits to be realized.

The questions of the status of the world's languages online may be framed in terms of the digital divide. Some languages have large amounts of readily accessible digital content. Internet users who speak, read and write such languages have far less difficulty accessing and sharing useful information than speakers of less well-represented languages. This situation naturally raises the question of whether the digital information systems, their configuration, or their use constitutes a form of bias against the less well-represented languages. Has linguistic difference become a barrier to information access, that provides unfair advantages to some, and disadvantages to others? Questions of this nature are fundamentally ethical and moral questions, requiring a framework that takes these questions into account.

UNESCO and cultural diversity

In 2001, UNESCO's member states adopted a Universal Declaration on Cultural Diversity.[18] Article 6, "Towards access for all to cultural diversity", states:

> While ensuring the free flow of ideas by word and image, care should be exercised that all cultures can express themselves and make themselves

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

> known. Freedom of expression, media pluralism, multilingualism, equal access to art and to scientific and technological knowledge, including in digital form, and the possibility for all cultures to have access to the means of expression and dissemination are the guarantees of cultural diversity.

Hence, UNESCO unambiguously favors the provision of equal access to digital information, both in production and consumption, for all linguistic and cultural groups. The declaration elaborates this position by enumerating a number of lines of action for the implementation of the declaration. Three of the points pertain directly to questions of digital media and information technology.

> 9. Encouraging "digital literacy" and ensuring greater mastery of the new information and communication technologies, which should be seen both as educational discipline and as pedagogical tools capable of enhancing the effectiveness of educational services;
>
> 10. Promoting linguistic diversity in cyberspace and encouraging universal access through the global network to all information in the public domain;
>
> 11. Countering the digital divide, in close cooperation in relevant United Nations system organizations, by fostering access by the developing countries to the new technologies, by helping them to master information technologies and by facilitating the digital dissemination of endogenous cultural products and access by those countries to the educational, cultural and scientific digital resources available worldwide (UNESCO, 2001, pp.6).

These principles and lines of action establish values for evaluating the attributes of the information society in ethical terms and goals for its development. However, they do not provide sufficient insight into the possible causes for any biases that might be shown to exist. Without this, it is difficult to make appropriate recommendations for action in specific cases.

For example, digital libraries have not been well accepted among the Maori people of New Zealand. Rather than this simply being a problem of digital literacy, careful study has revealed a number of cultural issues preventing successful use of the resource, including the fact that the metaphor of the library belongs to a “Pakeha” (white, western European) form of institution with assumptions about access to information alien to Maori culture (Dunker, 2002). A key locus of conflict for the Maori is the open availability of information that is traditionally protected in their culture, such as genealogical information. Libraries, which typically apply a principle of open access to information irrespective of its content, run foul of this value. Thus, the information access model for digital libraries need to be reconsidered before a digital library can be created that gains acceptance among the Maori.[19]

An ethical framework

Friedman and Nissenbaum (1995, 1997) provide a useful framework for analyzing bias in computer systems that helps focus attention on the causes of bias. They identify three main categories of bias: pre-existing, technical, and emergent. Pre-existing bias is rooted in social institutions, practices and attitudes, and exists independent of the computer systems themselves. Technical bias arises from technical properties of the systems used when assumptions are made that do not fit all aspects of the world to which they are applied. Emergent bias arises in a context of use with real users; the bias is not an intentional part of the system design or inherent in the social context, but emerges from an interaction of the two in a specific situation.

With respect to language, we can find examples of all three forms of bias. Pre-existing bias is evident when a government, an industry or a powerful corporation refuses to make information, technologies or products available for speakers of one or more languages. For example, in the mid 1990s, Microsoft Inc. refused to make versions of its products that would work with non-roman writing systems, despite the availability of already commercialized technological solutions such as

[19] The situation is not unlike the problems caused by personal medical records accidentally becoming public via the Internet.

WorldScript, from Apple Computer Inc. The reason offered by Microsoft was that the non-roman markets were too small to justify creating a new version of their product; hence, this example of pre-existing bias had an economic motivation.[20] Technical bias arises in encoding schemes for text such as Unicode UTF-8, which causes text in a non-roman script to require two to three times more space than comparable text in a roman script. Here, the motivation stems from issues of compatibility between older roman-based systems and more recent Unicode systems. Finally, emergent bias arises when computer systems created for one purpose are deployed for another, such as the digital library system developed for a white New Zealand urban context that met with poor acceptance in a rural Maori context.

The three types of bias need to be addressed in different ways. Pre-existing bias needs to be addressed through the educational, legal and institutional resources of countries, industries or corporations. Technical bias can be addressed in the design of the underlying principles of the computer systems themselves. Emergent bias needs to be addressed through a combination of education and design, based on observations on the use of the computer systems in actual use.

Because the development of the Internet involves the interaction of technologies, pre-conditions, purposes, industries, and actors, all three forms of bias are involved in the development of languages on the Internet, in many different places and at many different times.

Internationalization and the Internet: popular conceptions

Popular media discourse about the potential for linguistic bias on the Internet tends to fall into two opposing perspectives. This opposition is described by Wasserman in the following terms:

[20] Since that time Microsoft has changed its stance, and created versions of its products for other language markets.

> Because the Internet contributes to... the intensification of the consciousness of the world as interconnected and interdependent, it could be seen as one of the most recent developments in the acceleration of globalisation... Because globalisation is seen as a force emanating from the so-called developed world, some critics envisage the destruction of localities and cultural specificities within minority countries and communities. On the other hand, some critics argue that global and local forces interact in the process of globalisation, making it a multidirectional process from which local cultures and languages can benefit and even draw empowerment. (Wasserman, 2002:2)

Those taking the former view tend to be advocates of minority rights, while those taking the latter view tend to be proponents of the new, networked information technology. The former view is something of a reaction to rapid and far-reaching changes resulting from the spread of the Internet, while the latter view has been heavily promoted by the creators of the technology from the outset.

It is quite easy to find popular accounts representing the engineering teams working on the early ARPANET (the first computer network) as an idealized, decentralized democratic organization (e.g., Hafner and Lyon, 1996), or the Whole Earth 'Lectronic Link (otherwise known as the WELL) spreading virtual communities to the world through the Internet (Rheingold, 2000). It is a short jump from this perspective to one that views linguistic dominance as one more form of inequality that the technology of the Internet will rapidly eradicate. First (so the argument goes), the Internet is global and decentralized; no user or group of users can exercise hierarchical control over any other user or group of users, because the Internet allows complete freedom of association. Hence, anyone can use any language, as long as at least one other person is willing to use it as well. Second, the growth of non-English users, especially Chinese-speaking users, is expected to exceed the current rate of growth for English-speaking users. In other words, eventually, English will no longer dominate the Internet, because other languages are spoken by many more people. The question regarding which language is most prevalent online is simply a matter of demographic distribution. Finally, proponents argue that the technical affordances of the Internet, such as Unicode for multilingual text, and systems like BabelFish for instance translation of Web-based documents, can solve any problems that speakers of other languages might have in using the information on the Internet. Notably, this perspective largely characterizes the position of the document *Cultural and Linguistic Diversity in the Information Society*, a UNESCO publication written for the World Summit on the Information Society (UNESCO, 2003).

Each of these arguments has a counter-argument from the alternative perspective which, in its more specific form, holds that the English language and to some extent other European languages are dominant in communication on the Internet. The reasons given are partly social as well as technical. First, it is argued that the Internet uses a telecommunications infrastructure which is economically dominated by US companies. The geographic center of connectivity in the global telecommunications network is the US, so anything that promotes this will disproportionately benefit the US, through lower communications costs and a greater number of reachable destinations. Second, in spite of any recent trends, English users are still the largest group of Internet users. At the very least, the representation of English speakers on the Internet is disproportionate with respect to that of global language populations. Finally, most of the technologies used on the Internet are best adapted to English. Interfaces for non-Roman alphabets are cumbersome or do not yet exist for some languages. Even such systems as Unicode incorporate technical biases which favor English, and translation systems are not good enough to work on the required scale.[21]

These perspectives differ in ways that underscore the three types of bias identified by Friedman and Nissenbaum (1997). The language demographics of Internet users raise the question of pre-existing biases. The issue of the availability of technical affordances for different languages raises questions of technical bias. In addition, the issues of decentralization versus de-facto central control raises the question of emergent biases in a system that has expanded beyond its original, national boundaries.

In spite of the divergent opinions and the sometimes heated debate they generate, there is a dearth of empirical research that directly addresses these questions of pre-existing, technical and emergent language bias on the Internet. Part of the reason for this is that the Internet is both vast in scale and rapidly

[21] Variants of these two positions and their relation to similar perspectives on global-ization are discussed in Block (2004).

changing. Both conditions make it difficult to obtain reliable data. And while linguistic surveys are sometimes collected by marketing organizations such as Jupiter Research (http://www.jupiterresearch.com/), and Global Reach (http://www.glreach.com/), such data has dubious value with regard to addressing questions of linguistic bias, because of the economic agendas of the marketers and their clients. In addition, a reliable, large-scale survey of online multilingualism would also be expensive, putting it beyond the reach of small budget or unfunded research scholarship.

Sources of Pre-Existing Bias

Pre-existing bias concerns the social institutions, practices and attitudes independent of technologies. The sources of pre-existing bias include the historical distribution of language populations, economic arrangements favoring larger languages, and institutional policies of nation states. As concerns language diversity on the Internet, pre-existing biases are found in the dispositions of governments, institutions and companies toward people of different language backgrounds in the implementation of information technology policy. Understanding such biases is complex, but most fundamentally, since the Internet is a global phenomenon, they need to be understood in the context of global linguistic diversity.

Global Linguistic Diversity

Coherent discussion of linguistic diversity on global or regional scales requires a quantitative index of diversity. Unfortunately, quantitative measures of linguistic diversity are rarely employed in current linguistic research, and no established measure is widely used. Existing measures tend to be somewhat simplistic, such as numbers of languages or numbers of language groups, as used in Barrera-Brassols and Zenck (2002) and Smith (2001). More sophisticated measures of diversity were proposed in the past (e.g. Greenberg, 1956; Lieberson, 1964), but these measures were not always statistically well-founded, and have fallen out of use. The approach adopted in this report follows that of Nettle (1999) in using a measure of variance as a diversity index.

A satisfactory linguistic diversity index must take into account several factors. Firstly, it must address some unit of analysis, such as a country, a continent or the Internet. Secondly, linguistic diversity should take into account the probabilities of finding speakers of any particular language. It should have a natural minimum of zero, for a completely homogeneous population, and no fixed maximum value. A greater variety of languages should increase the value of the index, but as the proportion of a language group decreases, its contribution to diversity should also decrease. This way, countries with many language groups of roughly equal size (e.g. Tanzania; Mafu, 2004) will show relatively high linguistic diversity, whereas countries with comparable numbers of languages, but with one or two dominant languages (e.g. the US) will show relatively lower linguistic diversity. A measure that has these properties is the information-theoretic construct *entropy*, on which we base our linguistic diversity measure. In statistical terms, entropy is a measure of variance. Entropy is calculated from the estimated proportion of the country population for each language by multiplied it by its natural logarithm and summing all the entries for a given unit (country, region). The final index value is -2 times this sum.

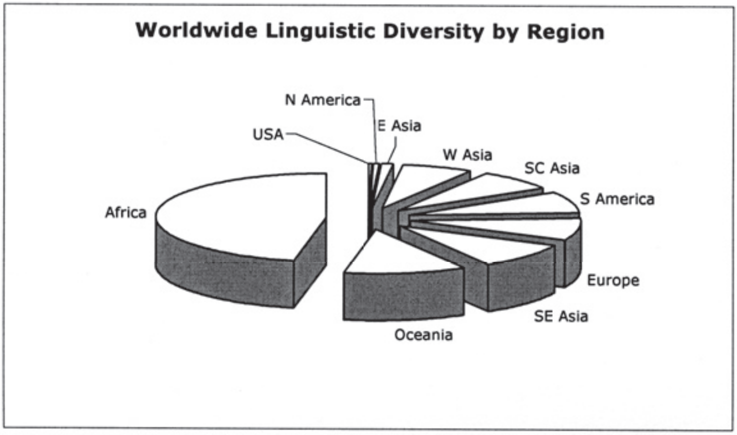
Table 1 and Figure 1 present figures for this entropy-based diversity measure for different regions of the world, based on the 7,639 language population figures presented in the Ethnologue (www.ethnologue.com), and ordered from lowest to greatest linguistic diversity. The USA, the birthplace of the Internet, has been split out in the first row for comparison. Regions that are well-known reservoirs of linguistic diversity (e.g. Africa, Oceania) show the highest linguistic diversity, while regions with large national languages (East Asia, North America) show the lowest. These last two regions are especially important in understanding linguistic diversity on the Internet: the US and China are arguably the two largest players in the Internet (some projections show China overtaking the US in numbers of users in the next few years); neither is very linguistically diverse, compares to the countries of Oceania or Africa. To the extent that these two countries dominate the Internet (or by extension, the discussion of linguistic diversity on the Internet), the Internet cannot be expected to reflect the world's linguistic diversity.

Table 1. Linguistic diversity index scores by region

Region	Languages	Diversity index	Prop. of world total
USA	170	0.7809	0.0020
N Am. (incl. USA)	248	3.3843	0.0086
E Asia	200	4.4514	0.0112
W Asia	159	26.1539	0.0659
SC Asia	661	29.8093	0.0752
S America	930	30.5007	0.0769
Europe	364	32.4369	0.0818
SE Asia	1317	37.6615	0.0949
Oceania	1322	46.5653	0.1174
Africa	2390	185.6836	0.4681

Source: Ethnologue.

Figure 1. Linguistic diversity index by region

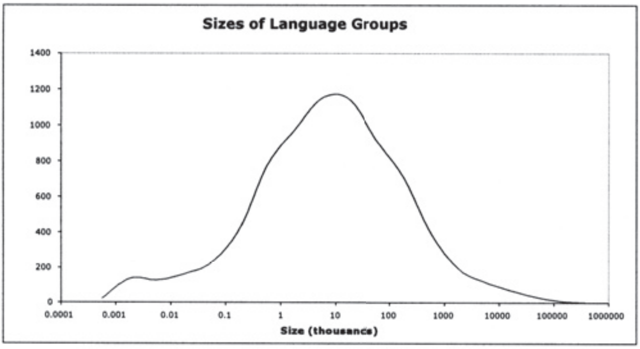


Source: Ethnologue.

The Evolution of Linguistic Diversity

One perspective on the meaning of linguistic diversity can be obtained from examining the sizes of language populations. Figure 2 shows a display of the number of different language groups at different population sizes, again from the Ethnologue data. The horizontal axis is on a logarithmic scale, meaning that the bell-shaped curve reflects a log-normal distribution (Grimes, 1986). The typical size of a language group is in the tens of thousands of people — about the size of a small urban community. Languages with hundreds of millions of speakers, such as English, Chinese, Spanish, French, etc., are rather atypical, as are smaller language groups with populations in the hundreds. In terms of human experience, the situation is different: nearly half of the people in the world speak a language spoken by hundreds of millions of other people.

Figure 2. Sizes of language groups



(Source: Ethnologue, UNPD).

Global and regional linguistic diversity is not static but evolves over time. It is influenced by socio-historical events such as mass migration, colonization, war, disease epidemics, and the like. Global linguistic diversity is currently in decline, and has been for a long time. For linguists, who make a study of the diversity of human speech, the situation is a crisis. The extinction of hundreds of languages in recent historical time has meant that large bodies of knowledge about the distinctly human capacity for speech are forever lost, as are the literatures, histories and cultures of the peoples who spoke those languages. For the communities whose languages, histories and cultures are lost, the situation is catastrophic.

By some estimates, nearly half of the world's languages will be extinct by the year 2050 (Dalby, 2003; Krauss, 1992; Nettle and Romaine, 2000). As linguistic diversity is lost through extinction of smaller language groups, the proportion of people in the world belonging to large language groups increases.

Loss of linguistic diversity is not localized in any particular region of the world: languages have been lost in large numbers in Europe since the rise of nation-states; in North America, South America, and Australia, dramatic losses follow European colonization and continue to the present; in the Pacific Islands and Indonesia, English and Indonesian are replacing indigenous languages; and in Asia, the major languages in China, Japan, India, and Russia have expanded for centuries at the expense of other languages (Crystal, 2000; Muhlhausler, 1996).

Some causes of language death are obvious. For example, the accidental or deliberate extermination of a group pf people can lead to language death (Wurm, 1991). Much of North America's linguistic diversity was lost in this way: wars with European settlers and foreign diseases spread by European contact left many indigenous populations diminished to the point that their languages could not be maintained. Other causes of language death are less obvious, particularly when changes in cultural ecology are involved.

Global Linguistic Diversity and the Internet

Low linguistic diversity, such as in North America, Latin America and the Caribbean, Europe and East Asia, facilitates the provision of Internet access using a small number of standardized technological solutions targeting each of the major language populations. Regions and countries with greater linguistic diversity typically require more complex arrangements for Internet access which may require customization of resources for each of a large number of minority languages. Hence, the Internet can be said to be biased in favor of larger languages, at the very outset. But even large linguistic groups often lack consistent technical standards. For example the speakers of Hindi number in the hundreds of millions, but according to one University of Southern California researcher, almost every Hindi website has its own Hindi font set, mutually incompatible with all the other Hindi font sets. People who wish to read Hindi material on these websites must install the fonts required by each site separately, and searching across these different sites is extremely difficult, as words do not match in the different representations (Information Sciences Institute, 2003). In other words, not all large language groups are equally favored by the Internet. Regions such as Africa, Oceania and Southeastern Asia face even more serious challenges, because of the large number of languages they have that are not yet used on the Internet. Hence, significant technical development work may remain before the goal of reaching these language groups can be realized.

The evolutionary perspective on language diversity is important to keep in mind when we consider the effects of the Internet. While the Internet may well have a long-term impact on language diversity, it is unclear what or how large that impact is likely to be in historical terms. Since the Internet extends the reach of individual languages, it potentially strengthens them, but since it does the same for larger languages, while also facilitating language contact among, it potentially weakens them. Both of these effects could be far smaller than influences of other equally pervasive social causes on linguistic diversity, such as the development of agriculture, the urbanization of populations, geopolitical events, etc., and these could be well beyond the means of any human government or collective agency such as the United Nations to prevent. At the same time, the world faces a real decline in linguistic diversity, and the historical and cultural continuity of hundreds of communities around the world are directly at stake. It is important that these concerns are understood in any policy that is framed to address language diversity on the Internet.

Sources of Emergent Bias

Emergent bias concerns the effects of bias that arise in the actual use of the Internet technologies. For linguistic diversity on the Internet, emergent bias is constituted in the experience of users of information technology when their language background becomes relevant to their ability to use the technology or information provided. This bias is manifest in two major ways: first in the distribution of languages on the Internet, and second in the economic control of the telecommunications and information technology markets. In this section, we examine the sources of such emergent bias. The findings presented here suggest a substantial bias in favor of English at the present time.

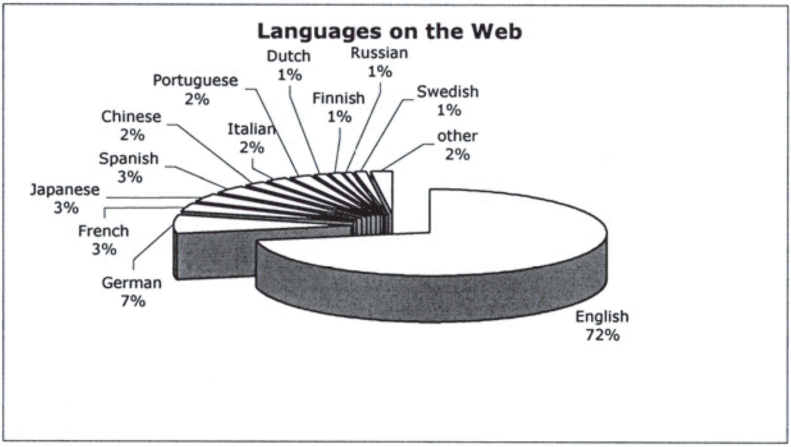
Linguistic Diversity of Internet Information Sources

A few studies undertake large-scale quantitative analysis of the languages used on the Internet. Generally, these focus on the World-Wide Web, to the exclusion of other communications modes like email and chat, because the Web is more directly observable and easier to survey than other forms of Internet communication. Two noteworthy pieces of research have generated interesting results in this area: a series of studies by Lavoie. O'Neill and colleagues at the Online Computer Library Center (OCLC), and a study by Nunberg (1998) at PARC of non-English websites.

The OCLC studies (Lavoie and O'Neill, 1999; O'Neill, Lavoie and Bennett, 2003) used a random sample of available websites on the Internet. They accomplished this by generating random IP numbers and attempting to connect to a website on each such address. If a Web server answered, they downloaded its main home page and ran an automated language classification system on it (O'Neill, McClain and Lavoie, 1997). This method of sampling has the advantage of being unbiased. All other methods of sampling rely directly or indirectly on search engines or "Web spiders", programs which discover new Web pages by following all the links in a known set of Web pages. Spiders produce what is known as a "snowball sample", a sample which is biased by its proximity to an arbitrary starting point. Search engines depend on spiders to build their indexes, and so samples drawn from them are similarly biased. If a reliable estimate of the prevalence of different languages on the Web is to be obtained, such biased samples must be avoided.

The initial survey was conducted at two different times, one year apart, to assess any trends in the use of these different languages. A later study in 2002 sought to confirm these observations. The 1998-1999 survey suggested that some international expansion of the Web was taking place, and that the use of different languages was closely correlated with the domain in which each website originated. The 1999 sample of 2229 random websites, for example, provided 29 identifiable languages with the distribution presented in Figure 3. As can be expected, English is clearly dominant with 72% of the total websites surveyed. The diversity index for this sample of Web pages is 2.47, less than that of a typical Southeast Asian country and more than a typical country in South Central Asia. It is also hundreds of times smaller than the global linguistic diversity. Hence, linguistic diversity of the worldwide Web, while it approaches that of many multilingual countries, is a poor representation of linguistic diversity worldwide.

Figure 3. Proportion of languages on the Web in a random sample of Web pages



Source: O'Neill, Lavoie and Bennett, 2003.

The follow-up survey conducted in 2002 shows the proportion of English on the Web to be fairly constant, in relation to the previous study, although small differences appear among the other languages (O'Neill, Lavoie and Bennett, 2003). The diversity index in 2002 is 2.44, changing little from the earlier survey. However, this may be partly a consequence of the methodology. The 29 languages they identify in their sample of Web pages is effectively the limit of the language identification program they use (http://www-rali.iro.umontreal.ca/SILC/SILC.en.cgi), and new languages coming onto the Web cannot be discovered by this method. Even if the language identification program encompassed more languages, they represent small proportions and so would not substantially change the calculated diversity of the World-Wide Web.

The 1999 OCLC survey also identified the proportions of Web pages that are multilingual from each domain of origin, and which language pairs are used. If a website used more than one language, English was always one of these: fully

100% of the 156 multilingual sites identified used English. French, German, Italian and Spanish appeared on about 30% of multilingual sites each, while other languages had much smaller shares. Furthermore, 87% of the multilingual websites originated from domains outside the major English-speaking countries (Australia, Canada, the United Kingdom, and the United States). Within any given domain, rates of multilingualism ranged from 6 out of 13 (42%) of Russian sites, to 16 out of 1103 (1.5%) for US domains. Hence, the World-Wide Web tends heavily toward monolingualism, and most displays of multilingualism are merely a tip of the hat to English dominance. This finding directly contradicts the popular notion that the Web somehow promotes linguistic diversity.

The tendencies observed in the OCLC surveys are confirmed in Nunberg's (1998) study, which adopted a different methodology. In this study, Web crawl of 2.5 million pages collected in 1997 by Alexa, an Internet service company, was analyzed using an automatic language identifier written by Nunberg's colleague Heinrich Schütze. While this sample is a biased snowball sample, it is over a thousand times larger than the OCLC sample. Nunberg's main finding is that countries with low Internet penetration used mostly English on their websites, while countries with greater levels of penetration used greater proportions of non-English languages. Latin America stands out in contrast to this pattern, having a very low level of Internet penetration in 1997, and an overwhelming predominance of non-English websites. Hence, the extent of English bilingualism in a non-English-speaking country may influence the expression of linguistic diversity on its websites.

Apart from the above studies, there are a few other attempts to study the distributions of languages based on statistics obtained from search engines. For various reasons, these do not provide as much usable information. For example, FUNREDES, an NGO promoting the adoption of information and communication technologies in Latin America, has conducted a series of studies since 1995 to assess the distribution of language and national influences on the Internet (Pimienta and Lamey 2001; Pimienta, et al., 1995-2003). These studies count the number of Web pages indexed by popular search engines containing selected words from different languages and national groups. Notably, they claim a much smaller proportion of English pages (52% in 2001, 45% in 2003), than that observed in the studies by Lavoie and O'Neill and Nunberg.

Page counts derived from search engines, however, are an unreliable methodology for determining language representation on the Web. Apart from the biased samples that provide pages to search engines, there are numerous other confounding influences. Search engines typically employ a variety of proprietary indexing methods that are not open to inspection, and these may bias the page counts returned in ways that cannot be corrected or even reckoned. A word need not appear in the page at all for it to be included in the count, and pages containing the word might also be dropped from the count. In addition, the method assumes that word frequencies of related "culturally neutral" concepts are stable across languages. Cultural neutrality is unattainable, however. Many of the words whose frequencies they observe represent culture-bound concepts, such as "cheese": American English-speaking culture and continental French culture assign quite different dietary significance to *cheese* and *fromage* respectively. These facts will be represented in the frequency of the corresponding terms. In addition, since page counts are returned, not word counts, the counts returned for different language forms may include bilingual or multilingual pages, which are multiply counted.

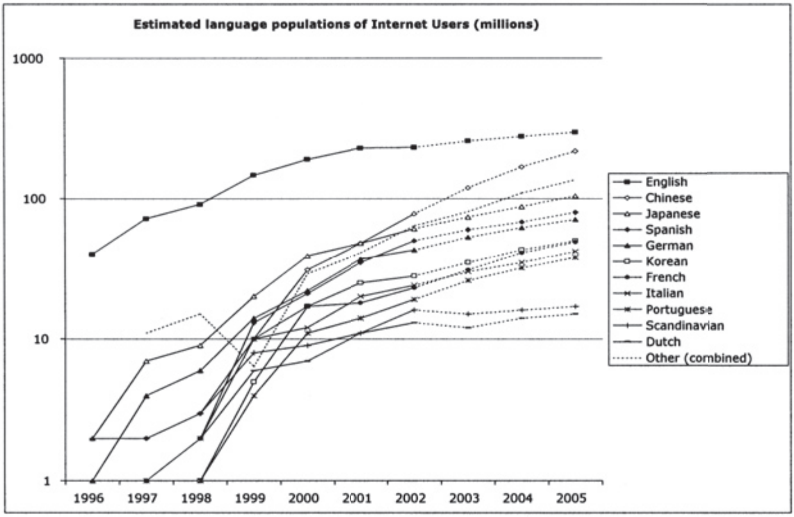
Linguistic Diversity among Internet Users

The most direct effort to estimate the linguistic diversity of Internet users comes from the translation services company Global Reach. These estimates, produced every year from 1996 to 2002, are widely cited as projecting an Internet of ever-increasing linguistic diversity.[22] These estimates are based on ITU estimates of user populations in each country, hence a user is defined as someone who has used the Internet in the past three months. The user populations are divided into language populations calculated from Ethnologue estimates and adjusted with UN population data, much as we have done in calculating linguistic diversity above. In some cases, they have supplemented these sources with marketing statistics from companies such as Nielsen Net Ratings. Absent from the data is any kind of actual survey of Internet users, so the Global Reach data do not represent the languages Internet users actually speak. Because these figures are cited so often as

[22] These data are available from http://global-reach.biz/globstats/evol.html.

evidence of the linguistic diversity of internet users, it is worth examining them in more detail.

Figure 4. Estimated language populations of Internet users (logarithmic y-axis)



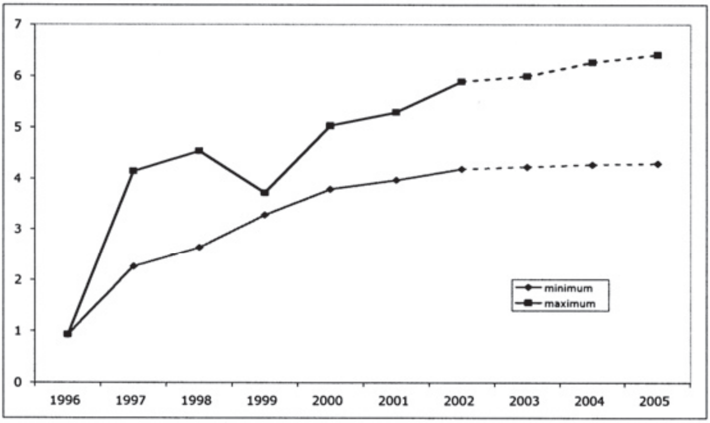
Source: Global Reach.

Figure 4 presents Global Reach's estimated user populations for different languages. The period from 2003 to 2005 is shown in dashed lines, as these are projected estimates. The languages identified are consistent with the languages found in the OCLC studies. As expected, English, with an estimated 230 million users, had nearly three times as many users in 2001 as the nearest language, namely Chinese, with approximately 60 million users.[23] Figure 4 shows that all of these user groups appear to be growing exponentially except for English and Japanese, which appear to be slowing. Both language groups are estimated to have about 50% of their available populations as Internet users already.

[23] These estimates appear to treat all varieties of Chinese equivalently, even though linguists consider Chinese a family of nine different languages (often called "dialects" by non-linguists).

From the Global Reach estimates one can calculate linguistic diversity indices for the global population of Internet users; these values are presented in Figure 5. Because the composition of the "other" language group is left unexplained in the Global Reach data, we have calculated minimum and maximum values for the index, based on the assumption of "other" representing a single language (the minimum diversity) or a uniform distribution across 6,000 languages (the maximum diversity). It is striking that although there are initially large gains in the diversity index from 1996 to 1999, linguistic diversity appears to be leveling off after 2000, in spite of the exponential growth of many of the languages. Additionally, the 2003-2005 projections continue this leveling trend; the projected increase in the number of Chinese speakers, because it is so large, actually mitigates the increase in diversity. The end result is a linguistic diversity index between that of a typical African country, and the North American and European regional indexes combined. This is perhaps not surprising, given that Internet hosts remain concentrated in North America and Europe. Yet Internet linguistic diversity is nowhere near as large as the index of any other region or of the world as a whole. Hence, contrary to popular belief, the Internet cannot be said to embrace linguistic diversity in this sense.

Figure 5. Estimated diversity of Internet users



Source: Global Reach.

Hence, the Internet has not become linguistically diverse merely by being global and interconnecting large numbers of people. Other issues need to be addressed in order to guarantee that languages of the connected peoples are represented online, and as we see below, these may be highly particular to the contexts of the connected communities.

The Internet and the Practice of Multilingualism

Access to the Internet is a pre-requisite to using the information it provides. So far, we have considered what such access might mean in global terms. However, no such effort will be successful if the speakers of the world's many languages simply opt for one of the few dominant languages. What then governs users' choice of languages on the Internet?

Languages are more than mere vehicles for the conveyance of information — they are complex systems of symbols bearing rich and subtle evaluations of their context of use. Sociolinguistic studies of multilingualism have illuminated in great detail the sensitive and turbulent ecologies of languages in contact; recent research on Internet multilingualism underscores the relevance of these lessons in relation to the Internet. Moreover, the keen global interest in the Internet centers on the economic advantages it offers. Does the Internet also favor larger languages in the same way?

It is not easy to identify in general terms what languages are used online and how they are used. A range of issues are involved, from individuating language communities, to differential Internet access, to different writing systems and computer encodings to different communication modes. Much of the available research addressing the Internet's potential effects on language and culture examines case studies of particular linguistic groups using the Internet in specific contexts, rather than from a macro-social perspective. These case studies suggest that language contact on the Internet favors powerful languages, much as off-line contacts do. For example, Wright (2004) and Holmes (2004) report on a survey of the online linguistic behaviors of college students in eight countries. The results indicate that the extent to which people use their native languages online varies tremendously with the context examined. At the same time, none of the populations surveyed shows evidence of using their full linguistic repertoires online.

Lesser-used languages appear to be used not at all on the Internet. Hence the questions around this issue are subtle and complex.

In early research, Paolillo (1996) found that English is heavily favored over Punjabi in Usenet discussion groups that have primarily Punjabi audiences. This behavior is partly expected from the predominantly expatriate and English-educated status of the participants, but the tendencies observed marginalize online use of Punjabi to the point that it becomes specialized for use in highly ritualistic or nationalistic communicative functions, and serves more as a stamp of identity than as a vehicle of informative communication. In a later paper, Paolillo (2006) compared interactivity and linguistic homogeneity among South Asians interacting in chat rooms and discussion groups on the Internet, and found that both favored the use of the minority language (Hindi or Punjabi, depending on the forum). These tendencies are echoed in Peel (2004), which reports that interactive chat rooms in the United Arab Emirates favor Arabic, whereas email favors English. In another paper, Paolillo (2001) found that central participants on a chat channel were more likely to use minority languages than peripheral participants. Because chat systems make it easy for participants to come and go, peripheral participants and their linguistic preferences predominate. Hence, technological and social aspects of Internet communication interact in complex ways that nonetheless favor majority over minority languages. Technological variables might be manipulated so as to mitigate the effects of linguistic dominance somewhat, but it is unknown how effective this might be.

Studies of Greek in Internet communication by Koutsogiannis and Mitsakopolou (2004), Georgakopoulou (2004, Forthcoming) and Androtsopolous (1998) explore a range of issues overlapping with the studies cited above. Like the Gurmukhi script of Punjabi, the Greek alphabet is not easily used on the Internet, and a romanized form of Greek adapted from an offline form known as "Greeklish" is favored, especially by expatriates in multilingual contexts which favor English (Georgakopoulou, 2004,) or German (Androtsopolous, 1998). This in turn subverts the Greek norm of diglossia (Ferguson, 1959), where speakers employ a distinct vernacular variety for informal speech and a classical language for writing. At various times in the past, government of Greece has expended considerable effort on maintaining literacy in Katharevousa, the classical language for formal writing; the erosion of Greek diglossia on the Internet potentially undermines these efforts. In a second diglossic context, that of Arabic, Warschauer, et al. (2002) observe that vernacular Egyptian, Arabic and English are encroaching on traditional functions of Classical Arabic. Such encroachment tends to destabilize diglossic situations, ultimately leading to language shift toward an outside, dominant language. Hence, when linguistic norms are eroded on the Internet, universal provision of Internet access could have a potentially damaging effect on such linguistic diversity.

Influence from English is both widespread and subtle. Sharply contrasting situations involve email in Switzerland (Durham, 2004) and Internet use in Tanzania (Mafu, 2004), where bilinguals in both countries favor the use of English over more obvious local languages. While there is a colonial precedent of English use by elites in Tanzania, this is not at all the case in Switzerland. Only in the international status of English (Crystal 2003; Phillipson, 1992, 2003) do we find an explanation for this phenomenon. Another example of English influence in Internet involves the spread of certain oral language features into writing through short message services (SMS) messages, Instant Messages (IM) and Web-based chat in Swedish (Hård af Segerstad, 2002). Similarly, Torres (1999, 2001), observes many pragmatic functions of emoticons ("smileys") in Catalan chat. These forms originated in English-speaking contexts, and hence indicate contact influence from English to Catalan through the medium of the Internet.

These studies and others collectively illuminate the richness and complexity of factors bearing on the use of minority languages by multilingual Internet users. A point that re-emerges in many is the fragility of the use of the non-dominant languages in Internet communication contexts.

Institutions and Interests Governing the Internet

Contrary to popular belief, the Internet is not an open and democratic (or anarchic) institution. Rather, it is an institution with a complex network of powerful interests, many of which are highly centralized. These powerful interests are not often concerned with the actions of individual users, leaving the impression that the Internet is free of constraints from civic, governmental, or corporate interests. Each level of interest nonetheless represents a locus of opportunity for linguistic biases to determine what languages are used on the Internet.

There are several different major actors involved in regulating the Internet. Firstly, there are the telecommunications monopolies and oligopolies of different regions in the world. These companies maintain the infrastructure that permits individuals to connect to the Internet, and for Internet sites to connect to one another. Secondly, there are the computer and software manufacturing companies, such as Intel, IBM, Hewlett-Packard, Cisco Systems, Sun Microsystems, Microsoft, Adobe, among others. These companies create and market the hardware and software that constitutes the Internet's infrastructure. In addition, there are Internet-specific governing bodies, such as the Internet Corporation for Assigned Names and Numbers, or ICANN, and the Network Information Centers such as the American Registry for Internet Numbers (ARIN), Réseaux IP Européens (RIPE) and the Asia Pacific Networking Information Centre (APNIC), which make decisions regarding Internet connectivity. National governments also play a role, both in administering Internet resources at the country level, and in implementing other forms of information policy. Finally, there are other organizations and consortia, such as the World-Wide Web Consortium (W3C), the Unicode Consortium, and the International Standards Organization (ISO), which develop standards for implementing Internet technologies.

The telephone network has always been important to the Internet, from its earliest days. When an Internet host connects to another host, modems, leased lines, Digital Subscriber Lines, fiber-optic backbones and geosynchronous satellites may all be engaged in some phase of the digital communication, physically conveying the data on the telephone network. More recently, other forms of telecommunications networks, such as the television cable networks, have been adapted for carrying Internet traffic. Historically and today, the economic control of these resources has been in the hands of large companies, often private or state-owned monopolies. Internationally, these concerns are most developed in the United States. For example, through its subsidiary UUNET, MCI runs a network that carries an overwhelming majority of international Internet traffic (see Mapnet, http://www.caida.org/tools/visualization/mapnet). The fiber-optic backbone that MCI put in place several years ago is central in this network. While companies such as MCI are relatively uninterested in the languages used on their data lines by Internet users, the centrality of the United States in the distribution of data traffic guarantees that high-level administrative tasks concerning backbone traffic will take place in English. Hence, regional networks connecting to these central networks must necessarily engage people with high levels of English skill.

While this might not seem a terrific burden, given that computer professionals the world over tend to be highly proficient in English, these two tendencies feed and reinforce each other. If regional network authorities cannot communicate with their providers in a language of their choice, then English will remain the dominant language of network administration, by default. Telecommunications companies, who reap bountiful profits from the demand for communication and technology services, have a special responsibility to take into consideration the linguistic diversity of the countries whose markets they serve.

Hardware and software companies have a similar influence on the linguistic makeup of the Internet, by producing computers with keyboards, displays and operating systems that favor particular languages. These products are produced at a low cost by achieving economies of scale, which entail marketing a standardized product to the broadest available market. Computer technology, with its offshore chip factories, outsourced software development (and even management), and commodity markets, is one of the original globalized sectors of industry. Because of this, and because of the prominence of US-based companies in developing new systems and standards, computer systems that make their way into linguistically diverse regions such as Africa are overwhelmingly designed for English or European language use, and have little if any tailoring to local languages. Such circumstances constitute another form of emergent bias toward European languages on the Internet, and away from the languages of less industrialized countries. As with the telecommunications companies, hardware and software companies have a special responsibility to the linguistic diversity of the countries whose markets they serve.

Thus, the acts of computer companies locked in competition for market dominance have a detrimental effect on the climate of multilingual computing and online linguistic diversity. If multilingual computing is to be promoted, then arrangements are needed where international interests can assert precedence over the competitive goals of private companies. Some of these tendencies are ameliorated by the activities of international organizations and consortia, such as the International Standards Organization, the Unicode Consortium, and the World-Wide Web Consortium, which oversee different aspects of Internet technology development. Many major computer companies (including Apple and Microsoft) work through these organizations. While some technologists complain that these organizations impede innovation, their international character helps to ensure that the interests of different national and language groups are considered. On the other hand, these standards organizations have no real enforcement mechanisms. Hence, a number of Internet technologies have standards which are not widely followed in practice. These include the HTML used in Web pages, and the ECMAScript programming language for Web browser interactivity. The incompatibilities that result when standards are not followed are detrimental to the progress of multilingual computing. If language diversity is to be promoted and protected through these organizations, their enforcement mechanisms need to be strengthened.

Another actor governing the Internet that has a large impact on Internet language diversity is ICANN, which administers the protocol known as the Domain Name System (DNS), under contract with the US Commerce Department. The DNS performs the function of associating unique mnemonic names with all of the Internet's hosts, a function which is fundamentally linguistic. Unfortunately, the DNS is awkward to use with languages other than US English, and furthermore is at odds with the way naming actually works in human language. The DNS is deeply integrated into the functioning of the Internet, as most other Internet application protocols depend on it to locate Internet hosts. It is also the only protocol actually administered, rather than merely codified, by a central authority. ICANN regulates the DNS primarily by delegation, but its administrative structure, its network of contracts with the US government and other parties, and its various policies have all worked to restrict multilingualism in naming Internet hosts. The resulting effect is that the DNS cannot fulfill its original goal of providing useful mnemonics for Internet hosts. Changes to ICANN, the DNS itself and the policies of domain name administration are all needed to improve this situation.

Internet users look at Internet host names much as any other names. In fact, they are very different. The DNS requires internet hostnames to be globally unique, whereas in natural language, metaphor, symbolism and acronyms make it unlikely that any particular name will be unique. Once a domain "acl.org" is registered to the Association of Christian Librarians, it is unavailable to the Association for Computational Linguistics, or any other organization in the world that would like to refer to itself with the same acronym.

To enforce uniqueness while allowing limited flexibility, the DNS uses hierarchically-structured names: individual host names are composed of strings of names, ordered from greater to lesser specificity. The top-level of the hierarchy as the last field in the name; this will be a generic or country-code top-level domain (gTLD or ccTLD), which functions as a general classifier. However, it is often unclear which classifier is relevant to a particular purpose. Under their agreements with ICANN, TLDs are supposed to be administered for different functions: .com is for commercial sites, .net is for networks, .org is for non-profit and not-for-profit organizations, and country codes are to be administered by the associated countries for their own purposes. Domain names in gTLDs are more desirable, however, because they tend to be short and more easily remembered. Since there are only a small number of gTLDs and hundreds of millions of hosts, conflicts in the assignment of domain names are inevitable.

ICANN's disposition toward such conflicts and their resolution is to favor holders of legally recognized trademarks. Otherwise, the first party to register a domain name keeps it, as long as the registration is kept up-to-date. This does not help registrants who are not trademark holders, or whose provenance is a small locale or minority language. It particularly does not help international registrants if their natural identities happen to be homographs of some already-registered domain name. Once a domain is registered, it requires expensive negotiation and/or legal action to change it. The prior registration of hundreds of millions of hosts in English thus entails a manifest bias against non-English host registrations, as many thousands of desirable hostnames in other languages will be homographs of already-registered hosts in the gTLDs. Hence, in the DNS, trademarking, a trade-related US legal issue, is given precedence over transparent multilingual naming, a language and communication-related international issue. This skewed ordering of priorities will not change until the DNS is governed by a fully international authority, rather than a private body with contractual ties with the US (or any other) government.

The original design of the DNS had a strong technical bias toward English, in that it could only use 7-bit US-ASCII encodings. Hence, even European languages such as French, Spanish, and German, which use diacritics that are not in US-ASCII, are at a disadvantage when it comes to selecting suitable names for Internet hosts. A number of organizations, such as the Multilingual Internet Names Consortium (MINC), New.net and RealNames, have labored for years to persuade ICANN to develop alternatives to the current DNS with better multilingual support. Although these groups have made many constructive proposals deserving greater consideration, their efforts have been received with much resistance on the part of ICANN. Only recently did ICANN adopt a variant of Unicode known as punycode, to allow multilingual domain names, but its deployment has been unsatisfactorily slow and politically fraught.

The domain name issue is principally a symbolic one. Nonetheless, the symbolism is powerful, and ICANN's intransigence over multilingual domain names has led to a worldwide perception that it does not really care about internationalism or linguistic diversity. While ICANN has recently undergone major reform, and now claims a more international board, it has lost much of its public trust over the multilingual domain names issue, and it is not clear whether these changes will lead to a fair, functioning and international DNS, or if the lost trust can be restored.

The role of the organizations ARIN, RIPE and APNIC (as well as other Network Information Centers, or NICs) in emergent linguistic bias is more subtle than that of ICANN. These organizations, whose membership is relatively open, govern the physical inter-connection of regional and local networks. One of their chief functions is in the maintenance of the Internet Protocol (IP) address space. IP numbers are 32-bit numbers that are used to uniquely identify hosts. Like domain names, IP numbers are assigned through a process of delegation to intermediaries, who may further delegate authority. Unlike domain names, each assigned range corresponds to a physical branch of the network, whose associated equipment is operated by a single authority. IP numbers are assigned in ranges, and because address space is ultimately limited, each such assignment has an opportunity cost — the same numbers cannot be assigned later to somewhere else, unless that part of the network is taken down.

The intersection of the NICs' role with language diversity issues comes from their function as regional authorities. Network resources available to a particular country or linguistic group depend on the ranges of IP numbers available to the relevant regional authority, and its allocation of them to other groups and countries. Poor allocation of addresses or a small range of available space to start with are two conditions that could lead to a shortage of addresses for new hosts. Controversy has raged about whether APNIC, whose regional responsibilities include Oceania, Eastern and South-eastern Asia, has enough address space to continue assigning IP ranges at the necessary rate. APNIC denies that there is a problem, but the specter of a looming crisis causes concern. The address space problems are expected to be ameliorated by upgrading from IP version 4 (IPv4) to IP version 6 (IPv6), which uses a larger range of address numbers, but that conversion is several years off because of technical incompatibilities with IPv4.

Nonetheless, the assignment of IPv4 address space is very inefficient. Large ranges of address space are designated as special-purpose or unusable entirely; these are known as "bogons", and careful records of these ranges are kept so that system administrators can monitor them for security reasons (see http://www.cymru.com/Bogons/). Even when the bogon ranges are masked out, a random sample of 1,107 IP numbers returned 203 IP numbers (18%) apparently allocated for a testing a rarely-used "multicast" protocol. In other words, 18% of the globally available IP address space was blocked off and made unusable because of inefficiency in its allocation. To the extent that such inefficiencies are allowed to occur, and to the extent that they impact the address space available to regional authorities, local linguistic groups could be denied Internet resources. For different languages to have a fair chance of being used online, the administration and allocation of Internet address spaces must also be fair.

National governments can play both positive role and negative role in influencing linguistic biases on the Internet. To the extent that national governments implement policies within their borders that protect and promote linguistic human rights of their multilingual citizens (Skutnabb-Kanngas and Phillipson, 1995), pre-existing linguistic biases in those countries are held in check. To the extent that their language policies are carried over into relevant areas of Information policy, they promote linguistic diversity on the Internet. But governments are typically more concerned with administrative efficiency and the perils of separatism and many of the world's people live without guarantee of even their most basic linguistic rights. When countries interface with the global Internet and demand accommodation to their national languages, they facilitate emergent biases against their own constituent ethno-linguistic minorities, ultimately doing little to advance the cause of linguistic diversity online. If national language groups hope to secure their own niche in the global telecommunications ethnosphere, then they must acknowledge and address linguistic diversity within their national borders. They must specifically strive to educate citizens from all their linguistic groups in the digital literacies needed to fully participate in the Internet. The ethno-linguistic awareness of telecommunications companies, computer companies and Internet governing authorities will only begin to broaden if a critical mass of under-represented ethno-linguistic groups can command their attention. This is not likely to happen if the true extent of international linguistic diversity remains concealed.

Emergent linguistic bias is a significant area of concern for linguistic diversity on the Internet. The areas discussed here are merely representative, and not a comprehensive list of possible emergent biases. As telecommunications, computer hardware and software markets change, and as Internet governing authorities evolve, new linguistic biases may emerge. Emergent linguistic biases, because they arise in particular contexts of technology and language use, may also be highly local, being manifest in a particular way only within a particular country. Hence, the general issue of emergent linguistic bias requires close monitoring on global, regional and local scales.

Sources of Technical Bias

Three areas of technical bias, having different relationships to linguistic diversity, are relevant to current internationalization efforts, under the three UNESCO lines of action mentioned earlier. First, there is the issue of encoding standards, which relates directly to line of action number 10, promoting linguistic and cultural diversity on the Internet. Text encodings are the primary technical means for achieving linguistic diversity in this primarily text-based communications medium. Second, there is the issue of the markup and programming languages used for creating and maintaining Internet applications and content. These technical systems bear directly on line of action number 9, promoting digital literacy. If digital literacy requires literacy in another language as a pre-requisite, openness and universal access cannot be assured. Finally, there are issues of technical linguistic bias in the applications protocols of the Internet, which relates to lines of action 9 and 10. In order to foster access to information technologies among the developing countries, the major Internet applications (electronic mail, hypertext browsing, instant messaging, etc.) should permit use of the languages of those countries. Without this, barriers to the acceptance of the technology may be prohibitive. These three areas of technical bias are treated in the sections below.

Encoding

Encodings specify the arbitrary assignment of numbers to the symbols of the world's written languages. Two different encodings can be incompatible by assigning the same number to two distinct symbols, or vice versa. In order to take advantage of the computer's ability to manipulate text (e.g., displaying, editing, sorting, searching and efficiently transmitting it), communications in a given language need to be represented in some kind of encoding. Thus, much of what the Internet offers with respect to linguistic diversity comes down to the encodings available for text.

The most widely used encoding is the American Standard Code for Information Interchange (ASCII), a code devised during the 1950s and 1960s under the auspices of the American National Standards Institute (ANSI) to standardize teletype technology. This encoding comprises 128 character assignments and is suitable primarily for North American English. Because of its early spread and widespread adoption, most subsequent encodings have been defined around ASCII, for example the International Standards Organization's ISO-8859-1, or Latin-1 encoding, specifies 256 codes, the first 128 of which are the same as ASCII. Unicode, an effort to provide compatible encodings for all of the world's languages (Unicode Consortium 1991, 1996, 2000, 2003), adopts a similar strategy by making the first 256 characters of the 65,536 characters in the Basic Multilingual Plane (BMP) the same as ISO-8859-1. Most of the supporting technologies of the Internet rely on ASCII or its derivatives. Systems such as the DNS, Usenet news and Internet Relay Chat permit only a subset of ASCII characters to be used. Operating systems such as Linux rely extensively on "flat ASCII-text files" for some of their most basic functions. All of these systems enforce a technical bias toward English.

Most hopes for internationalization of the Internet's infrastructure pivot on the eventual acceptance of Unicode, a standardization effort undertaken by the Unicode Consortium in cooperation with the ISO. The membership of the Unicode Consortium is composed of major software vendors, international religions, regional and educational organizations and national governments. The Unicode standard (now in version 4.0) provides over a million possible character codes, permitting all modern and historical scripts to be used in a single text. Sixty-five thousand characters form the basic multilingual plane (BMP), which is expected to suffice for most written communication. Such versatility comes with a cost. In its most basic form, UTF-32, Unicode text occupies four times as much space as the same text in ASCII. Many software developers have assumed that users would not want this penalty for multilingual text, particularly if computer use occurs mainly in monolingual contexts.[24] Unicode offers other variable-length encodings that are more efficient, but the space costs are passed on to non-roman scripts which are forced to consume more space. Although data storage costs have dropped considerably in the last decade, enough to make Unicode less of a problem, handling Unicode still substantially complicates the software developer's task, since most applications require inter-operability with ASCII. In addition, the larger sizes of Unicode documents carry costs for transmission, compression and decompression, and these costs are enough of a penalty to discourage use of Unicode in some contexts.

Although major strides have been made in the internationalization of computing through Unicode, the problems of using multilingual text on the Internet are far from solved. For a variety of technical, economic and organizational reasons, development of an acceptable technical standard has been slower than the pace of development of the Internet itself. Consequently, international use of the Internet has favored languages based on Roman scripts and especially English, which has benefited from having a widely adopted standard encoding since before the spread of the Internet. For the Internet to allow equivalent use of all of the world's languages, Unicode needs to be more widely adopted. As in the case of the DNS, this may require updating certain Internet protocols, so that they can work with Unicode.

Markup languages and programming languages

Another way that technical biases favoring English are perpetuated on the Internet is through the computer "codes" — the markup and programming languages — that are used to configure Internet content and services. The first and most

[24] Whether this is true is an important question that is not satisfactorily addressed in the research literature.

obvious way that technical bias arises is in the support they provide for multilingual content. Markup languages such as Hypertext Markup Language (HTML) and eXtensible Markup Language (XML) need to be able to describe text in the full range of human languages. The World-Wide Web Consortium has provided for this by requiring Unicode support as part of its standards. This means that where Unicode support lags, as with the majority of Western, South Central and Southeast Asian languages, HTML and XML support also lag. Thus, there is uniformity in the bias toward certain languages because of this. Programming languages must also be made compatible with multilingual text. Unfortunately, many commonly-used programming languages such as C do not yet offer standard support for Unicode.[25] A growing number of languages designed for Web-based applications do (examples include Java, JavaScript, Perl, PHP, Python, and Ruby, all of which are widely adopted), but other systems, such as database software, vary more in their support for Unicode. The promise of electronic commerce in languages other than English assumes that Unicode-compliant databases will become widely available.

The second way that English bias is present is in the design of the markup and programming languages themselves. Programming languages offer the most basic human interface available for the control of computers, mediating between the cognitive processes of programmers and the logical capacities of the computers themselves. A plethora of programming languages exist; estimates range from 2,500 to more than the number of human languages. In spite of this apparent diversity, the vast majority of these languages ultimately trace their lineage to FORTRAN, the earliest high-level programming language, released by IBM in 1957 (Lévénez, 2003). These languages make extensive use of English keywords to define important programming constructs, such as conditionals (*if*, *then*, *else*, *case*, etc.) and iterative looping (*while*, *for*, *until*, etc.). Even though many human languages have equivalents for these keywords, they never appear to be substituted for the English ones in executable code. For example, Ruby, authored by Japanese programmer Yukihiro Matsumoto and designed with specific attention to internationalization, also uses English keywords.

[25] The International Components for Unicode website offers an open-source C library that assists in Unicode support (http://oss.software.ibm.com/icu/).

HTML and XML are similar in this regard. HTML tags are generally mnemonic abbreviations for English words (e.g. *b* “bold”, *ul* “unordered list”, *li* “list item”, etc.). Although XML is not a markup language per se, it is a syntax for defining markup languages, and all of the XML-based markup languages with any acceptance are based on English (e.g. MathML, for mathematical expressions, and XML:FO for formatting text documents), in spite of the XML standard being based on Unicode. This trend deepens with the Semantic Web development project, which aims to bring “common knowledge” reasoning to the World-Wide Web. Large Artificial Intelligence databases such as Cyc (Reed and Lenat 2002) and WordNet (Fellbaum and Miller, 1998) are expected to be used to develop new markup that will assist Internet programs in finding and processing information for users. These databases have already been critiqued from within the Northern Hemispheric cultural perspective as bearing sexist, androcentric biases (Adam, 1998). They surely have inherent cultural biases as well. Hence, projects like the Semantic Web, which are promised to bring the “next generation” of Internet information services, threaten to further reinforce already existing linguistic and cultural biases.

The potential for linguistic bias in programming and markup languages must be considered alongside the cultural nature of computation. Modern computation arose out of centuries of mathematical learning, and its current spread can be compared to that of the decimal number system, in nature and in importance. Decimal numbers were originally invented in northern India sometime around the 7th century A.D. and spread globally, replacing most other number systems. Cultural transmission of decimal numbers did not require importing vocabulary, however; many languages modified their existing number vocabularies to accommodate the new practice. The computer develops the principles of decimal numbers further by automating their manipulation. Unlike the spread of decimal numbers, however, the spread of computers has brought with it large and complex English vocabularies – the programming languages.

No doubt, the computer as a physical artifact, by coupling symbols to actions, has a role in this relationship. The exact pairing of symbol and action is arbitrary, and hence any language could be accommodated, but at the same time is complex enough that doing so is not trivial. Hence, a large question for linguistic diversity has not been adequately addressed in the research literature: to what extent do various features of programming languages assist their acquisition and use by speakers of different languages?[27] Transfer effects for speakers of one human language learning another are well known. It stands to reason that programming languages, being formal linguistic systems, could exhibit native language transfer, leading to systematic difficulties or errors for speakers of particular language backgrounds. Design properties of programming languages vary greatly. Is it possible that speakers of a given language are better served by programming languages with properties that match their own language? Perhaps programming languages could be designed to reflect the systems of reasoning of different cultural and linguistic traditions. Would such adaptations help people from these backgrounds take mastery of their own information technology resources?

UNESCO and the other UN agencies have a compelling need to see these questions answered, if the educational goals required for promoting linguistic diversity are to be achieved. Through computer programming, language becomes powerful and animate, with a potential to reshape cultures. Sadly, at present it is primarily the English language that is animated in this way. If digital literacy in computer programming languages requires linguistic and/or cultural knowledge in English, then speakers of other languages ultimately must bear heavy educational and perhaps cultural costs in order to claim ownership of their Internet information resources.

Communication modes

While the Internet is known to most people through the World-Wide Web (some assume they are synonymous), it is actually a more heterogeneous environment offering a variety of modes of communication. Moreover, the design of the Internet means that new modes can always be created and deployed inexpensively. While we make use of electronic mail, the Web and instant messages on the Internet today, it is entirely unknown what we might use in the near future. Certain communications modes have nonetheless become widely adopted, and at times these modes incorporate technical forms of linguistic bias.

[27] See Anis (1997) for suggestions in this direction.

Usenet News, first created as a network of three university computer systems in 1978 (Spencer and Lawrence, 1998), is one such communication mode. Usenet is a collection of thousands of "newsgroups", public message spaces having names suggesting a topical focus. Usenet server and client software is freely available, and its administration is relatively open. Usenet administrators can individually set the amount, rate and frequency of sharing messages with other servers, so in areas where connectivity is poor, they can readily optimize use of the network. Thus, the barriers for entry into Usenet are relatively low. Usenet is an extremely important resource internationally. As of 1999, 205 countries in the world had Usenet access (Smith, 1999).

As a technical system, Usenet is like a microcosm of the Internet. Its naming scheme for newsgroups is hierarchical, and uses a subset of ASCII, much as the DNS does. It has top-level hierarchies and local, regional and country-code hierarchies.[28] Message text needs to maintain compatibility with ASCII. Chinese and Japanese text on Usenet uses special encodings. As in the rest of the Internet, English predominates in the generic top-level hierarchies. For example, the comp. hierarchy, the generic category for postings related to computer systems, there are few if any Japanese postings, even on comp.lang.ruby. Only on the fj.comp hierarchy does one find discussions of technical computer science topics in Japanese. The soc.culture sub-hierarchy also provides space for multilingual traffic, but primarily in European languages. Thus, in spite of its low barrier costs for countries with very limited resources, Usenet is poorly internationalized at best, and has many technical biases that favor English. Some of these lead to additional emergent biases.

A second communication mode that became popular in the early 1990s is Internet Relay Chat (IRC), a multi-party real-time synchronous communications mode. Participants on a chat channel communicate in real-time with all other participants, much as if they were in a telephone conference call, except that the conversation is typed. IRC servers networked together may host thousands of channels and it is common on IRC networks such as the EFNet or the Under-Net for chat channels to have cultural, regional or national themes, and to draw participants from around the world (Paolillo, 2001). IRC originated in Northern

[28] Usenet name space, like the DNS the name space, has also been badly abused.

Europe, so some features, such as the allowable characters in text messages and participant names differ from that of Usenet. However, the support of multilingual text is not any better in IRC than in Usenet. In fact, the display differences between American English and Northern European computers cause glaring problems (e.g. substitution of punctuation characters for diacritic vowel characters in Scandanavian names and words).

Thus, in spite of the appeal of these two systems for international use, they have shortcomings stemming from linguistic biases that are part of their design. Of course, new communications modes such as Instant Messaging, Web-logging, Web-chat and others are constantly emerging. Although some of these specifically incorporate design features such as XML and Unicode, the state of development of those standards is such that only a small proportion of the world's population and the world's languages can be readily served by the technologies. Some technology proponents may hold out hope for yet other communication protocols, such as voice-over-internet, or multi-modal interfaces. Even if these manage to solve certain linguistic issues, others may remain, such as support for hearing or visually impaired. Furthermore, existing technical biases reinforce emergent biases that arise through demographics, economics and other means. In order that linguistic biases be minimized on the Internet, new communications modes should be scrutinized for latent technical bias before they are allowed to become widely adopted.

Many technophiles have placed hope in machine translation as an answer to problems of multilingual communication on the Internet. Already there is a high demand for translation services offered by such companies as Systran, the provider of the BabelFish translation system, and in certain situations, such as Catalan-Spanish, machine translation has been proposed as a serious solution to communication problems (Climent et al., 2004). Will it be possible for people to access the Internet in their own languages by simply using one of the online translation systems? This question is too optimistic for several reasons.

Firstly, having a machine translation system presupposes that the more mundane problems of representing and rendering text in the language are already solved, when for a large number of languages, they are not. Secondly, the design of machine translation systems is extremely labor-intensive. Special problems may arise in translation between any pair of languages that have to be solved for that language pair alone. Translation among all of the world's languages is therefore a challenge that is not likely to be accomplished in the near future. Third, the design of machine translation systems requires large amounts of material in the languages to be translated;[29] today, these materials are harvested from the websites of the languages to be translated (Grefenstette, 1999; Resnik, 1999), and so they need to be created by native speakers. This cannot happen unless there is adequate technical support for the language. Finally, machine translation is never as good as that produced by a human translator (Kay et al., 1993). Users of machine translation systems have to adapt to strange residues of vocabulary and word order that merely represent a covert form of the linguistic bias that led to the translation need in the first place. Consequently, we cannot expect technological approaches such as machine translation to diminish problems of linguistic bias on the Internet in a substantial way.

Conclusions

The exploration of potential sources of bias conducted in the foregoing discussion finds many sources of linguistic bias in the Internet, whether pre-existing, technical or emergent. Consequently, the answer to the question, is there linguistic bias on the Internet, can only be affirmative. The principal effects of bias are to favor large languages with well established technical standards. Notably, English is perhaps foremost among these languages, being not coincidentally, the language of the founders of the Internet and its precursor research projects. However, it is also evident that the causes and effects of bias are subtle, diverse, and in many places, unanticipated. If UNESCO seriously hopes to address linguistic bias on the Internet, we must do more to educate ourselves and the principal agents of Internet development about both the manifestations of linguistic bias and the importance and value of linguistic diversity.

[29] Not just any text can be used. Typically, bilingual texts aligned sentence-by-sentence are required. These are expensive to produce, and are not available for all language pairs.

Glossary

ACM. Association for Computing Machinery. The largest international society of computer professionals. The ACM has many special Interest groups that address technical, social and policy-oriented issues around computing and computer networks.

APNIC. Asia-Pacific Network Information Center. The NIC that oversees the operation of the Internet in Asia and the Pacific. Its operations cover Australia, China, Japan, Korea, Indonesia, Malaysia and all of the Pacific island nations.

Application protocol. A network protocol that a computer user normally operates. Application protocols are typically organized around a particular purpose for network use, for example, exchanging files or mail between computers.

ARIN. American Registry for Internet Numbers. The NIC which oversees the technical operation of the Internet in North America.

ASCII. American Standard Code for Information Interchange. An early seven-bit standard for computerized text encoding that has become pervasively supported in most computing applications. Most modern text encodings, including Unicode, are designed to be backward-compatible with ASCII, whose seven bits allow 128 distinct characters to be encoded. Extended ASCII is an eight-bit extension of ASCII which has no standard; different vendors support different versions of extended ASCII which are mutually incompatible.

BMP. Basic Multilingual Plane. The region of Unicode code values that comprises the codes for all the scripts most commonly used around the world.

ccTLD. Country-Code Top-Level Domain. Top-level domains that are associated with specific countries. ccTLDs are identical to ISO-3166 country codes. Examples are .uk (United Kingdom) and .za (South Africa). See appendix B for additional ccTLDs.

CMC. Computer-Mediated Communication. Any form of human-to-human communication that takes place using networked computers as a medium.

CNNIC. China Network Information Center. The NIC that oversees the technical operation of the Internet in China.

DNS. Domain-Name System. The technical system administered by ICANN that allows the assignment of mnemonic codes to networked computers.

Domain (name). A name that is registered in the DNS, and is used to refer to an Internet host computer. Domain names are assigned to organizations, who may assign them to specific computers or sets of computers, with the cooperation of the network service providers they contract with.

Domain-name registry. An organization that under contract with ICANN administers some portion of the DNS name space. Generally a registry is responsible for the maintenance of one or more TLDs. Verisign and Educause are examples of domain-name registries.

Emergent bias. In Friedman and Nissenbaum (1997), bias which emerges out of the interactions of technical systems in specific social contexts.

Ethno-linguistic vitality. The potential of an ethno-linguistic community to survive.

Ethnologue. The database maintained by SIL International (Barbara Grimes, ed.) which records general descriptive information for all known language populations in the world.

GPL. Gnu Public License. A legal licensing arrangement used in some open-source software intended to protect copyright while allowing open access to the source code of the software to developers.

gTLD. Generic Top-Level Domain. Top-level domains that are assigned to "generic" purposes, without necessarily referencing a specific country. Common gTLDs are .com, (commercial) .edu (US-accredited higher education), .mil (US military), .net (network providers), .org (non-profit or not-for-profit organizations), etc.

Host, **Internet host**. Any computer that is connected to the Internet.

HTML. Hypertext Markup Language. The markup language that is used to format pages on the World-Wide Web. It is a simple markup language that is widely understood by Web browsers and other software, and whose standard is now maintained by the W3C.

IANA. Internet Assigned Numbers Authority.

ICANN. Internet Corporation for Assigned Names and Numbers. The organization, run as a public-private partnership, which oversees the DNS.

ICT. Information and Communications Technology. Any technology that is used for information processing or communications.

Internet. The global computer network that emerged from linking the ARPANET and other regional computer networks.

IP. Internet Protocol. See TCP/IP.

IPv4. IP version 4. The version of IP in most common use today, characterized by 32-bit address numbers for each Internet host. Address space under IPv4 is limited, hence the Internet is currently in transition from IPv4 to IPv6.

IPv6. IP version 6. The "next generation" version of the Internet Protocol, which uses 128-bit addresses. IPv6 support is expanding in a number of networking applications, but it is not widely deployed yet, as IPv4 applications are not compatible with IPv6 hosts.

IRC. Internet Relay Chat, an application protocol for simultaneous, real-time multiparty communications on the Internet. Most "chat" programs, including many proprietary ones, borrow heavily from IRC. There are many IRC networks used by millions of people worldwide, primarily for recreational purposes.

ISO-8859-1, Latin-1. This is the eight-bit standard text encoding that supports most European languages using the roman alphabet.

Markup Language. A system for introducing formatting or other codes ("markup") into text documents so that the text can be formatted or interpreted by a device that understands the markup. HTML is an example of a markup language, but other markup languages such as SVG (Scalable Vector Graphics) work similarly yet perform quite distinct functions. See XML.

NIC. Network Information Center. The technical organization responsible for overseeing the technical operation of the Internet in a region or locale. There are three main regional NICs: ARIN, RIPE and APNIC, for North America, Europe and Asia and the Pacific, respectively.

Pre-existing bias. In Friedman and Nissenbaum (1997), any bias that arises from strictly social causes, which pre-dates a particular application of technology in which it becomes manifest.

Protocol. A standardized set of messages and rules for exchanging messages between computers on a network. Protocols are complex, and usually discussed in terms of "layers": the application layer, the link layer, etc.

RIPE. Réseaux IP Européens. The NIC that oversees the technical operation of the Internet in Europe.

SGML. Standard Generalized Markup Language. A markup language definition language that is in standard use in the domain of print publishing. HTML was originally developed as an SGML application.

TCP/IP. Transmission Control Protocol/Internet Protocol. The primary protocol suite used to operate the Internet. TCP and IP are actually independent "layers" of the Internet networking protocols, as they concern different aspects of the network operation, but most often they are used in conjunction.

Technical bias. In Friedman and Nissenbaum (1997), any bias that is directly inscribed into a technical system. The bias of ASCII toward American English is an example of technical bias.

Techno-linguistic vitality. The potential of an ethno-linguistic community to avail itself of technologies, especially information technologies, and for its language to be used in connection with those functions. By analogy with ethno-linguistic vitality.

TLD. Top-Level Domain. A domain name directly assigned by ICANN to a domain-name registry, which groups together a number of related hosts, typically by country or organizational purpose.

Unicode Consortium. The consortium overseeing the development of Unicode.

Unicode. The 64-bit character encoding currently in development which aims to provide a standard technical means for representing the characters of all of the world's written languages. Unicode is being developed in cooperation with the International Standards Organization and the W3C, to ensure that the standards of all three organizations will be compatible.

Usenet (news). An application for exchanging publicly posted and widely distributed messages ("news"), among networked computers. Also any news, or the totality of the news that is exchanged this way. Usenet is important in the Internet because it is a low-cost easily-implemented protocol that can serve the functions of mail, and which does not require a dedicated network connection to operate. Hence, it is often the first Internet application to reach a new location.

UTF-8, UTF-16, UTF-32. Encodings for Unicode characters that use units of 8, 16 and 32 characters respectively. UTF-8 and UTF-16 are variable-width codes, meaning that some characters require more than one 8 or 16-bit unit to encode. UTF-32 is a fixed width code, meaning that all characters require 32 bits to encode.

W3C. World-Wide Web Consortium. The consortium of which oversees the development of protocols, markup languages and other technical standards that pertain to the World-Wide Web.

World-Wide Web ("the Web"). An application for exchanging formatted documents, programs and multimedia content over the Internet. Also the entire connected set of documents and content that is available via the Web. The Web is most common application on the Internet, owing to the ease with which a Web browser is operated to perform requests for documents and other content.

XML. Extensible Markup Language. The markup language definition language, a simplified version of SGML, which was intended as a format for delivering information on the World-Wide Web that is more flexible than HTML, as it permits many different kinds of markup to be defined. Current Markup Languages defined in XML include ones for Web content (XHTML), graphics (Scalable Vector Graphics [SVG]), mathematical equations (MathML), music (MML, MusicML), and many other applications.

Bibliography

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4.

Alternative Perspectives

Language Diversity On The Internet: An Asian View

Yoshiki Mikami[*], Ahamed Zaki abu Bakar[•], Virach Sonlert-lamvanich[○], 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)

"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 (Priolkar, 1958).

[*] Nagaoka University of Technology, Japan: [•] Universiti Tekmologi Malaysia, Malaysia: [○] Thai Computational Linguistic Laboratory, Thailand: [■] Technology Depart-ment of Indian Languages (TDIL), Ministry of Information Technology, India: [▲] Miskolc University, Hungary. Authors can be contacted by mail to mi-kami@kjs.nagaokaut.ac.jp.

"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 (Lunde, 1981)

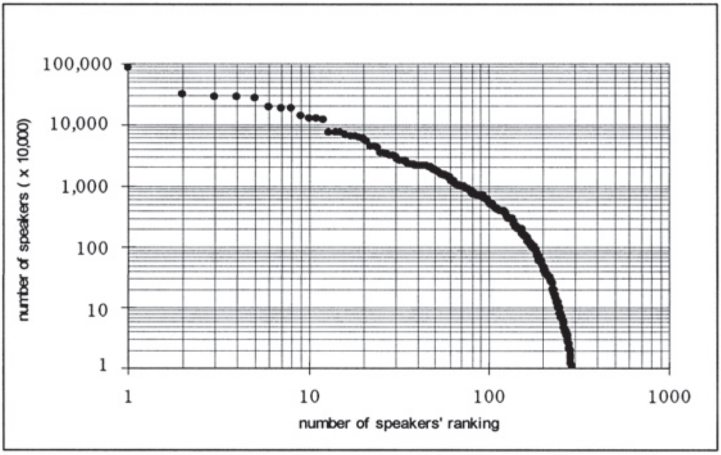
Language and Script Diversity in Asia

Language experts estimate that nearly 7,000 languages are spoken in the world today (Gordon, 2005). 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. (UNHCHR, 2005).

Among all the languages appearing in this site, Chinese has the largest speaking population of almost a billion, and is followed by English, Russian, Arabic, Spanish, Bengali, Hindi, Portuguese, Indonesian and Japanese. The language list continues to cover those with less than a hundred thousand speakers. Asian languages occupy six 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 and plot each language in a logarithmic scale chart, 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 a 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 di-versity of scripts used to represent languages. From the view-point of complexity in localization, diversity of scripts is a 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 as well as its extensions used for various European languages, Vietnamese, Philippino, etc. as one category. We treat Cyrillic based scripts and Arabic based scripts as one category. In the same manner, we treat Chinese ideograms, Japanese syllabics and Korean Hangul script as forming one category. The remaining scripts are comprised of various kinds of differing scripts. Here, we take the "Indic scripts" to constitute the fifth category. This category includes not only Indian language scripts such as Devanagari, Bengali, Tamil, Gujarati, etc., and also four Southeast Asian major lan-guage scripts, i.e., Thai, Lao, Cambodian (Khmer) and Myan-mar. In spite of the differences in their shapes, these scripts have the same origin (the ancient Brahmi script) and have the same type of behavior in formulation. When we sum 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	451	462	1,085	807	129
[% of total]	[43.28%]	[8.71%]	[8.93%]	[20.98%]	[15.61%]	[2.49%]

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

Current status of language coverage - the 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 Afri-an 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. Although this figure may not be construed to be bad, the figure seems to be an overestimate that does not tally well with reality as we will see in this paper.

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	Gujarathi Tamil Telugu Kannada Bengali Malayalam Punjabi Hindi Marathi Sanskrit Konkani Oriya 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.

The 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 engines, Google, is found to have indexed more than 8 billion pages written in various languages as of April 2005. However, the languages covered so far are limited to approximately 35 languages. Among these languages, there are only seven Asian languages covered by Google. They include: Indonesian, Arabic, Chinese Traditional, Chinese Simplified, Japanese, Korean and Hebrew (Table 3). If we calculate the population-based coverage, it 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.

The case of the 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, posted as "GIF" or "PDF" files, not in the form of encoded text. Again, we summarize the situation by major script grouping like 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 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
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 Magadi 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 texts using local script typefaces. According to one Indian historian, the first printed text to use local letters in Asia was Doctrina Christiana in Tamil. The second page of the text tells us what kind of approach was employed in the localization of type-printing technology into Tamil language. Although Tamil language has some 246 syllables in all , 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" (Priol-kar, 1958).

In Manila, the central place of Spanish colonial activities at that time, Doctrina was translated into Tagalog language in 1593. However, 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 (Hernandez, 1996). 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, the emergence of new technology would even destroy the writing system or the culture itself.

Encoding standards as a cornerstone of localization

There are certainly many factors behind this divide; economical, political, social etc. But among these, from a 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 even as today's computer engineers do. Especially the lack or non-availability of appropriate encoding standards is the major obstacle in non-Latin script using languages. 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 (IANA, 2005) or ISO International Registry of Escape Sequences (IPSJ/ITSCJ, 2004), we can not find any encoding schemes for these languages which we term as having '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 Standards. 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 associations and government standard bodies is another reason for this chaotic situation. Aruna Rohra and Ananda of Saora Inc., has produced an interesting study (see:

http://www.gse.uci.edu/markw/languages.html), which collected the language corpora of Indian languages. It found 15 different encoding schemes from 49 Tamil Web sites visited (Aruna & Ananda, 2005).

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 it has not acquired a prevailing status at least in the Asian part of the world. Our most recent study has disclosed that penetration of UTF-8 encoding is limited to only 8.35% of all Web pages under Asian ccTLDs (Mikami, et al. 2005). 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

The Language Observatory Project - Objectives

Recognizing the importance of monitoring language activities level in cyberspace, the Language Observatory Project (LOP) was launched in 2003 (UNESCO, 2004). 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 is funded by Japan Science and Technology Agency under the RISTEX (RISTEX, 2005) program. UNESCO has given official support to the project since its inception. Major technical components of the Language Observatory consist of a powerful crawler technology and a language property identification technology (Suzuki, et al. 2002). As for crawler technology, the UbiCrawler (Boldi, et al. 2004), 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 Instituto 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 (UNESCO, 2004).

Conclusion

In this paper, we stressed the importance of monitoring the behavior and activities of world languages in cyberspace. The Language Observatory Project allows for a sophisticated method to understand and to monitor the languages. The LOP consortium hopes to make the world more aware of its living and dying languages. Steps to assist endangered languages can then be made before language extinction. For this effort to bear fruits, the observatory is also designed to be the focal point for human capital development as well as to be a depository for various language resources. The accumulation of these digital resources through research and development will assist developing countries and communities in regions to acquire the ability and capacity to get their indigenous languages into cyber-space and hence preserve their national heritage from extinction.

Bibliography

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

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

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

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

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

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

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

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

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

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

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

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

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

A Note on African Languages on the Worldwide Web

Xavier Fantognan

Overview

The *Cahiers du RFAL n°23* "Traitement informatique des langues africaines" (Computerization of African Languages) indicates that the number of African languages is estimated at approximately 2,000, which represents two-thirds of the languages in the world. This constitutes a wealth of heritage and culture that merits close attention. Today cyberspace provides a participatory means by which all languages can be veritable instruments of large-scale communication. However, all the languages in the world are not being used and are not taking advantage of the opportunity that cyberspace provides. It is very clear that to achieve multilingualism, there must be a process of computerizing languages that begins with first codifying them. The first question to be asked is the extent of the use of African languages in cyberspace. Marcel Diki-Kidiri and Edema Atibakwa, in "*Les langues africaines sur la Toile*" (African Languages on the Web), reported having accessed 3,000 websites and then retaining only those which cover African languages. Their analysis shows that there is abundant documentation on African languages on the Web, but very few sites use an African language as the language of communication. Although numerous factors can account for this, two major ones could be the absence of cybercommunities capable of expanding communication in their languages, and the proper tools to computerize and process them.

However, the findings of a study conducted by Gilles Maurice de Schryver and Anneleen Van der Veken, "*Les langues africaines sur la Toile: étude des cas haoussa, somali, lingala et isixhosa*" (African Languages on the Web: A Case Study of Hausa, Somali, Lingala et Isixhosa) seem to balance, nuance and indeed correct those of the aforementioned study. These authors examined discussion fora and discovered a very satisfactory level of using three widely-spoken African languages – Kiswahili, Hausa and Lingala.

The major findings of the RIFAL study are the following:

- African languages appear on the Web more as topics of study (referenced, documentation, description, samples, texts, courses) than as vehicles of communication;
- The language of communication used to talk about African languages is by far English, even for languages in francophone regions;
- African language courses are much too rare on the Web. This suggests the possibility of developing cybercommunities of African language speakers who use their language as a means of communication on the Internet;
- Software products or computer solutions standardizing fonts for all African languages are rarely suggested on websites.

To correct this situation, the following should be promoted:

- Increasing the number of bilingual or multilingual sites involving French or English and at least one African language as languages of communication;
- Greater distribution of documentation on African languages, since this exists but is not being systematically disseminated on the Web;
- Quality African language courses given on the Web;
- Development and distribution of software or computer solutions facilitating the writing of African languages and their regular and ongoing use in cyberspace.

Today African languages do have a presence on the worldwide Web. There is much documentation on African languages on the Web, but very few documents are actually written in African languages. Why? The lack of motivation among Africans to write in their own language is one of the reasons we can cite to explain the relative lack of success of African languages on the Web. Internet users who communicate on the Web want to be read and understood, and therefore they write in a language used by greatest number of people.

Furthermore, a large number of African documents found on the Web were not written by Africans, such as numerous religious documents or those destined for teaching purposes. Fora where Africans communicate with other Africans in African languages, are the exception, not the rule.

Microsoft has announced that Windows and Office will soon be translated into Swahili. Kiswahili is without a doubt the most spoken language in Africa. Approximately 100 million people speak this language on the continent and islands in the Indian Ocean. Before doing the actual translation, Microsoft linguists should create a common glossary of the various Kiswahili dialects. Microsoft is also planning to translate programs into other African languages, such as Hausa and Yoruba.

Though Microsoft's intentions seem good, it is still worrisome to note that Microsoft software programs will be the only alternative for Swahili speakers who do not speak another language. Indeed, open-source software programs translated into Kiswahili are not legion. Let us hope that Microsoft's efforts to standardize African languages will also be of benefit to Linux and other open-source software.

In the case of open-source software, considerable work is under way in Africa. In Burkina-Faso, languages such as Mooré and Dioula are seeing localization with Open Office. The same type of work is under way in Mali with Bambara, in Bénin with Fongbé, Yoruba, Mina and Dendi. The great work done with Amharic and its alphabet demonstrates the possibility of making research on computerizing African languages more effective. Steps taken by UNICODE to standardize N'ko are great news for many people.

However, real questions such as those regarding orthography and the standardisation of African languages remain to be resolved. Many languages are still transcribed phonetically, and the risk of seeing each language lose its alphabet cannot be ignored.

Though Africa has approximately 2,000 languages, only 400 have been studied. There remain 1,600 that have not benefited from serious examination. None of these languages have an audience on the Web today, and even the 400 which have been studied lack technological adaptation in terms of becoming living languages on the worldwide Web.

Bibliography

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

Meloni H., 1996 . Fondements et Perspectives en traitement automatique de la parole. AUPELF/UREF.Morvan P., 2000. Dictionnaire de l'Informatique : Acteurs concepts, réseaux, Larousse, Paris.

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

Rint-Riofil, C., Chanard, et Diki-Kidiri, M. (undated) Stage de formation niveau 1 et 3, Document de travail : Introduction aux inforoutes par le développement de la terminologie et des contenus textuels pour le français et les langues partenaires, Lumigny, Marseilles.

The Authors

Xavier Fantognon is a student in Togolese linguistics at the University of Benin (xavier@bj.refer.org) who has decided to devote himself to promoting African languages on the Internet. He has translated the interface of the SPIP platform into the Fongbè language (http://www.spip.net/fon) and is also actively engaged in traditional and multimedia cultural activities.

Yoshiki Mikami is Professor of Management and Information Sciences at the University of Nagaoka. He has held management positions at MITI (Information Policies and Standards). He is in charge of the Observatory of Languages on the Internet (http://www.language-observatory.org/

- http://gii.nagaokaut.ac.jp/gii/ -
- http://kjs.nagaokaut.ac.jp/mikami/)

John Paolillo, Associate Professor of Information Science and Informatics; Adjunct Associate Professor of Linguistics, School of Library and Information Science. Ph.D., Linguistics, Stanford University, 1992, B.A., Linguistics, Cornell University, 1986 Research Interests: Computational linguistics, information retrieval, computer-mediated communication, statistical models and quantitative research methods, sociolinguistics and language acquisition, second language acquisition, and South Asian languages.

Daniel Pimienta, French of Moroccan origin who lives in Santo Domingo, is the President of the *Association Réseaux & Développement* (FUNREDES – http://funredes.org) (Networks and Development Association), a NGO that has worked in the field of ICT and development since 1988. FUNREDES has conducted a certain number of experiments in the field on languages and culture, some in collaboration with the Latin Union and/or the support of the *Agence de la Francophonie*. (http://funredes.org/tradauto/index.htm/bamaktxt - http://funredes.org/lc).

Daniel Prado, an Argentinean who lives in Paris, is the Director of the *Programme de Terminologie et Industries de la Langue de l'Union Latine* (http://unilat.org/dtil/), an intergovernmental organization that promotes Latin languages. He manages statistics on the dynamic reality of languages in our society and information on language and terminology polices.

World Summit on the Information Society

Full texts of the studies at:
http://www.unesco.org/wsis

UNESCO