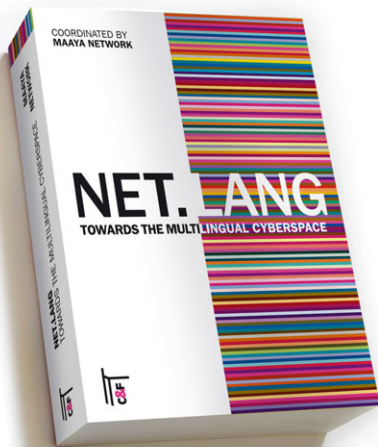


COORDINATED BY
MAAYA NETWORK

NET. LANG

TOWARDS THE MULTILINGUAL CYBERSPACE





NET.LANG IS AVAILABLE IN VARIOUS LANGUAGES AND FORMATS :

In print (French or English)

446 pages, 17 × 22.5 cm, softcover

Price : 34 euros

French : ISBN 978-2-915825-08-4

English : ISBN 978-2-915825-09-1

In bookstores or at <http://cfeditions.com>

eBook (French or English)

DRM free

French : ISBN 978-2-915825-25-1

English : ISBN 978-2-915825-26-8

<http://net-lang.net>

PDF (French or English)

DRM free

French : ISBN 978-2-915825-23-7

English : ISBN 978-2-915825-24-4

<http://net-lang.net>

**FOR OTHER VERSIONS AND TRANSLATIONS
VISIT [HTTP://NET-LANG.NET](http://NET-LANG.NET)**

NET.LANG

TOWARDS THE MULTILINGUAL CYBERSPACE

Associate website :

<http://net-lang.net>

Other books (in french) from C&F éditions :

Libres savoirs,

les biens communs de la connaissance

Ouvrage collectif coordonné par l'association VECAM

Mai 2011, ISBN 978-2-915825-06-0

Aux sources de l'utopie numérique

De la contre-culture à la cyberculture : Steward Brand, un homme d'influence

Par Fred Turner

Avril 2012, ISBN 2-915825-10-6

Dans le labyrinthe

Évaluer l'information sur internet

Par Alexandre Serres

Avril 2012, ISBN 978-2-915825-22-0

Complete catalog and online bookstore :

<http://cfeditions.com>

ISBN PDF edition 978-2-915825-24-4

C&F éditions, mars 2012

35C rue des Rosiers – 14000 Caen, France

<http://cfeditions.com>

This book is published under a Creative Commons license :

attribution, share alike (<http://creativecommons.org/licenses/by-sa/3.0/fr/>).

MAAYA NETWORK

NET.LANG

TOWARDS THE MULTILINGUAL CYBERSPACE

**EDITION AND COORDINATION:
LAURENT VANNINI
HERVÉ LE CROSNIER**

**C&F ÉDITIONS
2012**

MAAYA NETWORK

NET.LANG

TOWARDS THE MULTILINGUAL CYBERSPACE

HERVÉ LE GROSNIER
LAURENT VANNINI
: EDITION AND COORDINATION

2013
G&F ÉDITIONS

CONTENTS

Credits 9

FOREWORDS 10

Irina Bokova *General Director, UNESCO* 13

Abdou Diouf *General Secretary, La Francophonie* 17

José Luis Dicenta *General Secretary, Union Latine* 21

Dwayne Bailey *Research Director, ANLoc* 23

Daniel Prado *Executive Secretary, Maaya Network* 27

PART 1 33

WHEN TECHNOLOGY MEETS MULTILINGUALISM

Daniel Prado
Language Presence in the Real World and Cyberspace 35

Michaël Oustinoff
English Won't Be the Internet's *Lingua Franca* 53

Éric Poncet
Technological Innovation and Language Preservation 69

Maik Gibson
Preserving the Heritage of Extinct or Endangered Languages 75

Marcel Diki-Kidiri
Cyberspace and Mother Tongue Education 89

PART 2 **102** **DIGITAL SPACES**

Stéphane Bortzmeyer Multilingualism and the Internet's Standardisation	105
Mikami Yoshiki & Shigeaki Kodama Measuring Linguistic Diversity on the Web	119
Joseph Mariani How Language Technologies Support Multilingualism	141
Vassili Rivron The Use of Facebook by the Eton of Cameroon	161
Pann Yu Mon & Madhukara Phatak Search Engines and Asian Languages	169
Hervé Le Crosnier Digital Libraries	185
Dwayne Bailey Software Localization: Open Source as a Major Tool for Digital Multilingualism	205
Mélanie Dulong De Rosnay Translation and Localization of Creative Commons Licenses	221

PART 3 **227** **DIGITAL MULTILINGUALISM: BUILDING INCLUSIVE SOCIETIES**

Viola Krebs & Vicent Climent-Ferrando Languages, Cyberspace, Migrations	229
Annelies Braffort & Patrice Dalle Accessibility in Cyberspace: Sign Languages	249
Tjeerd de Graaf How Oral Archives Benefit Endangered Languages	269
Evgeny Kuzmin Linguistic Policies to Counter Languages Marginalization	287

Tunde Adegbola Multimedia and Signed, Written or Oral Languages	311
Adel El Zaim Cyberactivism and Regional Languages in the 2011 Arab Spring	325
Adama Samassékou Multilingualism, the Millenium Development Goals, and Cyberspace	337

PART 4 **348** **MULTILINGUALISM ON THE INTERNET: A MULTILATERAL ISSUE**

Isabella Pierangeli Borletti Describing the World: Multilingualism, the Internet, and Human Rights	351
--	-----

Stéphane Bortzmeyer Multilingualism and Internet Governance	373
--	-----

Marcel Diki-Kidiri Ethical Principles Required for an Equitable Language Presence in the Information Society	387
--	-----

Stéphane Grumbach The Internet in China	401
--	-----

Michaël Oustinoff The Economy of Languages	407
---	-----

Daniel Prado & Daniel Pimienta Public Policies for Languages in Cyberspace	423
---	-----

CONCLUSION **437** **THE FUTURE SPEAKS, READS AND WRITES IN ALL LANGUAGES**

Adama Samassékou *President of Maaya*



LAURA KRAFTOWITZ born in 1982 at Pittsburgh, splits her time between Paris and the Middle East. She is the author of numerous articles about the Israeli-Palestinian conflict, and the forthcoming memoir *The End of Abu Jameel Street*, regarding her time as a human rights activist in the Gaza Strip. She is currently completing her Master's Degree in Political Science and History at the École des Hautes Études en Sciences Sociales in Paris, where she is researching the past century of Israeli and Palestinian binational ideation.



HERVÉ LE CROSNIER is a senior lecturer at the University of Caen Basse-Normandie, where he teaches Internet technologies and digital culture. He is currently working with ISCC, the Institute for Communication Sciences of the CNRS. His research focuses on the impact of the Internet on social and cultural organization, and extending knowledge in the public domain. He is one of the founders of C&F éditions.



JOHN ROSBOTTOM was Principal Lecturer in Computer Science at the University of Portsmouth, UK, until retirement in 2010. He now pursues interests and activities in several areas that were precluded during a busy working life, but continues to enjoy working on computing projects and using his language skills in the translation of technical books and documents.



LAURENT VANNINI after working as a linguistic diversity activist for the Babels network, a journalist, and a telecommunications consultant, decided to travel the world and return to academe, where he is currently studying in Arts and Languages at EHESS. His work probes “the disappearance of the animal and oblivion”. In addition to his writing on ICTs, he penned the French translation of the book *From Counterculture to Cyberculture: Stewart Brand, the Whole Earth Network, and the Rise of Digital Utopianism*, by Fred Turner (C&F éditions, 2012).

CREDITS

Maaya Scientific Committee

Adama Samassékou
Daniel Prado
Daniel Pimienta
Marcel Diki-Kidiri
Louis Pouzin
Yoshiki Mikami
Evgeny Kuzmin

Editing and Coordination

Laurent Vannini
Hervé Le Crosnier

Translation

John Rosbottom
Laura Kraftowitz
Laurent Vannini

Graphic Design and Page Layout

Nicolas Taffin
Kathleen Ponsard

With the support of

Unesco, communication and information sector
La Francophonie
Union Latine
ANLoc, South Africa
IDRC / CRDI, Canada

FOREWORDS



United Nations
Educational, Scientific and
Cultural Organization

With the support of
**Communication and
Information Sector**

UNESCO works to create the conditions for dialogue among civilisations, cultures and peoples, based upon respect for commonly shared values. It is through this dialogue that the world can achieve global visions of sustainable development encompassing observance of human rights, mutual respect and the alleviation of poverty, all of which are at the heart of UNESCO's mission and activities.

UNESCO's mission is to contribute to the building of peace, the eradication of poverty, sustainable development and intercultural dialogue through education, the sciences, culture, communication and information.

<http://unesco.org>

PRESERVING AND PROMOTING LINGUISTIC DIVERSITY

BY IRINA BOKOVA
General Director, UNESCO

Languages are essential components of individual and common human heritage. They are the first and foremost vehicle for expressing identity, communicating ideas, attaining educational, economic and political autonomy, and promoting peace and sustainable human development.

Languages are important for sharing information and knowledge and for transmitting unique cultural wisdom, including across generations and nations. They form an intrinsic part of the identity of individuals and people, and they are of vital importance to manage the cultural diversity of our world. They open opportunities for dialogue, cooperation and mutual understanding. In this perspective, a plural and diverse linguistic space can expand the conditions for such dialogue by allowing each and every individual to contribute freely in the languages of their choice.

At the same time, languages are a fragile resource which requires endorsement, revitalization and promotion. Today, a significant number of languages are at risk of disappearing. About 97 per cent of the world's population speaks about 4 per cent of the world's languages. Conversely, about 96 per cent of the world's languages are spoken by only about 3 per cent of people around the world. This means that at least half of more than 6,000 languages worldwide are losing their speakers. It is estimated that about 90 per cent of the languages may be replaced by dominant languages by the end of the twenty-first century.

As the information and communication technologies (ICTs) became central to all aspects of social, cultural, economic and political life, it is important to ensure that everyone has access and can contribute with their own content to the multilingual internet. ICTs can be a powerful tool for the safeguarding and promotion of linguistic diversity. In principle, the internet is open to all languages of the world, but only when certain conditions are met, and when the necessary human and financial resources are in place. A multilingual internet is essential for nations, communities and individuals to access, share and use information and resources which are critical for sustainable development and for managing innovation and change.

UNESCO is strongly committed to promoting multilingualism on the internet. A plural linguistic cyberspace allows the wealth of diversity to be put in common. These goals guide the Organization in its work with the Internet Corporation of Assigned Names and Numbers (ICANN). They are also debated in the various meetings of the Internet Governance Forum (IGF) as well as in the World Summit on Information Society Forums. UNESCO's partnership with the International Telecommunications Union (ITU) in the Broadband Commission for Digital Development has also placed emphasis on the need for rich, culturally and linguistically diverse local content and applications as a key target in world leaders' commitment to broadband inclusion for all.

The importance of cultural and linguistic diversity is also echoed in the *Recommendation concerning the Promotion and Use of Multilingualism and Universal Access to Cyberspace*, adopted by the General Conference of UNESCO in 2003. This Recommendation calls for urgent and concrete measures to promote linguistic diversity, especially through a multilingual internet, and the preservation of languages, including endangered ones. It encourages pilot projects and the development of multilingual content management tools and resources. It calls also for a wider and more equitable access to information networks and services, while reaffirming the need for a balance between the interests of the rights-holders and the public interest.

Since a decade, UNESCO has promoted the concept of knowledge societies that are open, pluralistic, equitable, and participatory. Internet and social networks have a key role to play in fostering such inclusive societies. In order for Internet to be such an open and equitable global platform, it

must be guided by principles of openness, freedom of expression, cultural diversity and multilingualism.

I hope that this publication, supported by UNESCO, will contribute to interdisciplinary dialogue among various stakeholders, open up new horizons for better understanding the importance of linguistic diversity, raise awareness about new developments linked to the Internet, and most importantly, convey the message that multilingualism contributes to wealth creation, social transformation and human development.



ORGANISATION INTERNATIONALE DE LA FRANCOPHONIE

With a population of over 890 million inhabitants, and 220 million French speakers worldwide, the Organisation Internationale de la Francophonie's (OIF) mission is to embody an active solidarity between the 75 countries and governments that compose it (56 members and 19 observers) – which represents over a third of UN Member States.

It operates with respect for cultural and linguistic diversity and to the service of promoting the French language, peace, and sustainable development.

<http://www.francophonie.org/>

PROMOTING CULTURAL DIVERSITY IN THE DIGIVERSE

BY ABDOU DIOUF

General Secretary, La Francophonie

Much of our common destiny plays out in the Information Society. In every direction, not only are our modes of information, communication, production and consumption changing form, but also our ways of thinking, being creative and accessing knowledge. Digital technology is opening up all of us, especially our youth, to new perspectives.

Faced with today's global challenges, La Francophonie, an international organization with 75 states and governments members and observers, representing over a third of UN members, grasped early the need to mobilize in coordination to meet the challenges of the Information Society. Taking a concerted activist presence in international decision-making has lent resonance to its positions and force to its proposals, and has allowed the organization and its member states and governments to weigh in on international decisions.

Today, the appropriation of digital culture is fomenting an intense proliferation of creativity in all fields. The digital switchover poses as much a risk as it does opportunity for French and other linguistic expressions on the web. Furthermore, the question of languages' respective weight in the web's various formats is quite poorly documented. Since 2002, our organization, through the *Direction de la Francophonie Numérique* (Direction for a Digital Francophonie) and the *Observatoire de la Langue Française* (French Language Observatory), has supported frequent studies by Funredes about the French's place on the internet. Taking into account cyberspace's constant evolution, OIF (*Organization internationale de la Francophonie*) also supports all ongoing initiatives to establish new indicators for measuring linguistic diversity in the digital world.

The Francophone commitment aims to ensure the necessary preconditions for cultural diversity and linguistic pluralism in the information society. Our objective is for the Francophone community to express its specificity, thereby appropriating digital culture in all its diversity. In this context, free access to digital content and innovative technologies is a priority.

With the cultural domain as with the digital, La Francophonie works to oppose uniformity and preserve the values that allow a society to maintain both its identity and the seeds of its development and rejuvenation.

OIF's support to this publication falls under this work, demonstrating the synergism of cultural and linguistic diversity that La Francophonie has been cultivating for years, together with intergovernmental and international civil society actors, notably Unesco, the Maaya Network and the Union Latine.



UNION LATINE For a Plural Latinity

The peoples of the Romance languages have very diverse origins; however, they share the same linguistic heritage and the same system of historical, legal and cultural references. So it is natural that this family, as dispersed and broad as it is, should have an institution dedicated to the promotion and dissemination of the Latin world's shared heritage and identities. Such action is particularly necessary at a time when the preservation of cultural diversity constitutes a major concern of the contemporary world.

<http://www.unilat.org/>

FOSTERING THE PRESENCE OF LANGUAGES IN CYBERSPACE

BY JOSÉ LUIS DICENTA
General Secretary, Union Latine

The Union Latine is an intergovernmental organization that was founded 57 years ago. Since then, it has been a leading proponent of culture (meetings, heritage, film festivals, exhibitions, etc.) and language.

Concretely, it has become a pioneer in the field of linguistics, particularly with regards to the measure of languages in cyberspace. It is currently one of the few entities worldwide devoted to this task, a task of undeniable importance given the parameters of our contemporary world. Cyberspace has become a real territory necessitating comprehensive understanding.

The Union Latine, eager to enrich the terminologies of Romance and other languages of its member countries, is contributing to the creation of networks, associations and other entities aiming to revitalize and modernize these languages.

Intercomprehension between Romance languages is another of the Union Latine's pursuits, to which end it has organized meetings and trainings for trainers who it identifies as potential future advocates of this crucial field. Graduates of our programs will hopefully go on to facilitate understanding between countries.

Promoting policies to support language development is yet another objective of the Union Latine, which has helped create indicators to help orient policies in the language sector.

Through its support of *Net.lang*, the Union Latine strives to promote the modernization of all languages that do not have adequate online presence, by providing decision makers with a practical guide for making correct choices. We are committed to this project, and wish it great success.



ANLOC
The African Network for Localization

ANLoc is an Africa wide network of researchers from academia, civil society and business. Partners from over nine African countries assist each other in addressing the needs of local languages. Our vision in ANLoc is that every African can participate in the digital age. There are over a billion African's and we want their voices to be heard.

<http://www.africanlocalization.net/>



IDRC / CRDI
International Development
Research Centre, Canada

A key part of Canada's aid program since 1970, IDRC / CRDI supports research in developing countries to promote growth and development. IDRC / CRDI also encourages sharing this knowledge with policymakers, other researchers, and communities around the world. The result is innovative, lasting local solutions that aim to bring choice and change to those who need it most.

<http://www.idrc.ca/>

LANGUAGE UNLOCKS HUMAN POTENTIAL

BY DWAYNE BAILEY
Research Director, ANLoc

The African Network for Localization – ANLoc, pronounced Unlock– chose its name because of the vast human potential locked out by monolingualism. It is thus exciting for us to be part of this book on multilingualism in cyberspace. Like all users of the internet we get excited by the latest developments on the internet and in the digital world, however we are even more excited when this happens for all languages.

ANLoc’s efforts and our support of this book are made possible by the International Development Research Centre, Canada (IDRC/CRDI) who have funded much of our work to empower African languages.

At ANLoc we have focused quite extensively on the technology aspect of multilingualism. In our work we have seen that many people shy away from multilingualism for fear of its technical complexity. The resultant monolingual content is usually thought of as the traditional English only website, but in Africa monolingualism appears in French, Portuguese and Arabic. Monolingualism chooses official languages while excluding national and minority languages.

Fixing broken monolingual technology can take some effort but we’ve found that many requirements are simple or almost effortless compared to the impact. As simple as adding characters to fonts, creating a keyboard or translating a piece of open source software. Thus allowing someone to write and read digital content. While those interventions have some level of technical expertise ANLoc has been able to empower others by taking on the technical burden and allowing language experts to do the rest.

With these interventions ANLoc has not changed the landscape, but we’ve eliminated a number of technical hurdles that prevented mother tongue speakers from participating in the digital age. With an empowered community the content of the internet can be shaped by mother tongue speakers. We have simply removed the technical obstacles put in place by the shortcomings of technology.

The future of a language is in the hands of people. This is as true in the digital age as it has been in any age. Once those technical barriers have been removed a linguist or computer programmer cannot force people

to participate in the digital world. And fortunately it does not seem that many people need to be forced to Google, Facebook or Twitter. ANLoc has looked at this store of human potential and we worked hard to see how we could harness those human volunteers. Part of our work led to web technology to allow community translation, resulted in community built ICT terminology and resulted in Firefox in 10 African languages. This trend of crowd sourced mother tongue information is a trend that is leading to vibrant community participation. Volunteers are translating software, social network sites, search engines, making spell checkers and translating Wikipedia articles. The digital age could potentially be the most vibrant for any language.

ANLoc helped tackle the technical challenges of multilingualism. We are excited about this book because it takes our message, and the messages of all of the contributors to this book, to policy makers showing that multilingualism is possible and powerful. With a selection of worldwide experts and the backing of Unesco we believe that this book can act as a catalyst to help policy makers move forward to the advantage of all humanity. To unlock the human potential limited by a monolingual world.



Maaya is a multilateral network created to contribute to the development and promotion of linguistic diversity worldwide. In terms of the Bambara language, Maaya could mean the neologism “humanitude”. The Maaya Network was established following the World Summit on the Information Society (WSIS), in which the cultural and linguistic diversity in cyberspace was identified as a priority. Maaya was founded by the African Academy of Languages (ACALAN), under the auspices of the African Union.

<http://www.maayajo.org/>

A BRIEF RETURN TO THE GENESIS OF *NET.LANG*

BY DANIEL PRADO

Executive Secretary, Maaya Network

The book *Net.lang* that you currently have in hand, either in print or in digital version (in PDF or ePub format, or through direct online access) was designed and coordinated by the Maaya Network.

Maaya, the World Network for Linguistic Diversity, came into being during the World Summit on the Information Society (WSIS). Its objectives focus on achieving a truly multilingual world, where each language, and the wealth of culture and knowledge contained within it, enjoys the right to citizenship, receives respect, and contributes to the expansion of shared knowledge. Such an objective certainly requires the mobilisation of individuals and associations, but it also must be taken into account by governments and institutions. With this in mind, Maaya undertook the collective writing of a book enabling all persons and organizations concerned to act to develop a multilingual cyberspace.

I am aware that cyberspace presents both a threat and an opportunity for multilingualism. A threat, because the most highly equipped languages, and those spoken in dominant states, impose themselves over others and are supported by the network's technicality. An opportunity, because cyberspace's accessibility and universality allows it to give voice to languages that have been unable to make themselves heard via other recording and knowledge dissemination tools. We believe that this ease of access, the internet's ability to mobilise and coordinate many people, and its multimedia capabilities, will assure the rescue and revitalisation of minority languages.

Net.lang attempts to equip a concerned and informed public with the necessary analysis and methodology to extract the maximum benefit from the internet, to catalyse an equitable representation of languages.

Using simple, accessible language, this book strives for comprehensiveness alongside concision and clarity. It presents state-of-the-art thinking on its topic, alongside the desirable and necessary actions for using Information and Communication Technologies (ICTs) to promote linguistic diversity. *Net.lang* aims to address both content-oriented and technical topics as they relate to the presence of the world's languages

in Information and Communication Technologies, especially online. It should thus show the issues of multilingualism in cyberspace.

The capacity to analyse, measure, and report on language status in cyberspace, in order to help define public policy, advocacy, revitalisation and cultural sharing, is a constant concern of the institutions that have joined the Maaya Network in completing this book. In the framework of a 2007 study on multilingualism in cyberspace, assigned to the Union Latine by Unesco, the former asked Marcel Diki-Kidiri to compose a special leaflet containing concrete strategies to provide all languages, even those considered poorly equipped or “oral”, with the tools necessary for online presence. The resulting study, *How to ensure a language’s presence in cyberspace*?¹, quickly became a reference point, summarising in only a few pages the essential steps for enabling a language to embark on the digital path.

With technology’s rapid evolution and the internet’s swift development as a geopolitical and cultural player, it became a pressing issue to follow this work with a second that would give priority both to analysis and to a variety of perspectives. At the Bamako International Forum on Multilingualism (February 2009) organized on the initiative of the African Academy of Languages and the Maaya World Network for Linguistic Diversity, all within the framework of International Year of Languages, the recommendation was made to publish a “didactic manual in several languages to educate the uninformed public about issues relating to the presence of languages in cyberspace”².

One year later, Maaya, with support from Unesco and coordination from a first Scientific Committee³, convened a meeting in Paris with twenty of the field’s leading experts. The group sought to define the issues to be addressed and the experts to explicate them. This two-day seminar sketched out the framework and objectives, and assigned experts, reaching for representation from across the world, to ensure the best possible balance of gender and age, as well as cultural, political, and technical approaches.

1 <http://unesdoc.unesco.org/images/0014/001497/149786f.pdf>

2 http://www.acalan.org/eng/confeven/forum/plan_action.pdf

3 Consisting at the time of Daniel Prado (Union Latine), Louis Pouzin (Eurolinc), Marcel Diki-Kidiri (YSB Sängö Association), Daniel Pimienta (Funredes) and Yoshiki Mikami (The Language Observatory).

Writing a book of this magnitude required the contribution of specific editorial knowledge, together with regular monitoring of the authors and translators, to say nothing of technical typographic and digital document development skills. For this reason, the Union Latine decided to solicit the expertise of *C&F éditions*. Nicolas Taffin and Hervé Le Crosnier, together with Project Coordinator Laurent Vannini, with John Rosbottom and Laura Kraftowitz for editing English text and translation, and Kathleen Ponsard for graphic design and page layout, meet the challenge. They not only gave our vision a body, but also a title, image, depth and rigor that will make *Net.lang* a major work on language policy.

In addition to the institutional support provided by Unesco and the Union Latine, which led to the book's genesis, *Net.lang* also benefited from the enthusiasm of ANLoc (African Network for Localization), in partnership with the International Development Research Centre, Canada (IDRC/CRDI) and the international organization La Francophonie.

The project's launch seminar also allowed us to reflect on what a collective book is. Is it simply a printed object, or rather, a collection of texts that insist on living, on being spread, and finding their own way in cyberspace? While the printed version permits it to find a place in the library, assures its bibliographic representation, gives status and integrates its texts, the digital version emerges as the work's essential element. Publishing a book above all means gathering writers, defining a coherent topic, and ensuring text readability and reader interest. The vehicle is then of no importance. For this reason, the book is available in multiple digital formats in addition to print.

A second problem we had to address was to define how many languages are necessary to render a book truly "multilingual". Or rather, how the plasticity of a digital version allows us to start with two working languages, while opening the door to full or partial translations in all languages by all the structures involved. We chose French, the language of the publishers, as well as English, for obvious reasons. But our real goal is to see versions in multiple languages appear as quickly as possible.

To this end, we are using the legal provisions of Creative Commons BY-SA licensing, allowing derivative works (i.e. translations) as long as the book and the authors of the articles are attributed, and these translations are published under the same conditions. We deeply thank Evgeny Kuzmin,

who is already preparing a Russian edition, even before the original version appears. We hope that others, either privately or on behalf of institutions, will take up the same work for other languages. The website set up in conjunction with the release of *Net.lang* (<http://net-lang.net>) allows for the translation and publication of each article separately, a structure that facilitates organization and collaboration to obtain a full translation, which can then be transformed into a book, either digital or printed, in the relevant language and country.

Perhaps even now, you are reading the story of this project in Castilian, Chinese, or Malay, and we would be delighted to find it in Wolof, Quechua or Tagalog. Anyway, we hope this is the case, and we have taken legal and technical precautions to make that possible, based on the initiative of readers wishing to make the work gathered here available to those around them in their own language.

Once it is translated into multiple languages, we hope above all that *Net.lang* will become a tool that can assist and support all those responsible for language planning policies who are eager to provide their language with the needed tools to be fully present in the digital world. We hope the seeds sown by this book will bear the fruit of a multilingual cyberspace that is open to all languages, all peoples, and all the knowledge of the world.

**WHEN
TECHNOLOGY
MEETS MULTI-
LINGUALISM**

PART 1

Human societies are rich in their linguistic diversity. What does this mean for cyberspace? What language(s) are current and enrich cyberspace? What endangered languages may find a refuge, or a second life there? How can 6000 languages and as many human cultures find their place in this cultural space open to the winds? Can you imagine a digital world dominated by only a few languages?

LANGUAGE PRESENCE IN THE REAL WORLD AND CYBERSPACE

Barely 5% of the world's languages have a presence in cyberspace, and among those few, there are still considerable differences. Only a tiny handful of privileged languages offers a genuine production of content. This article attempts to compare the actual weight of a language (demography, economy, vitality, officialdom, literature, translation, etc.) and presence (or ability to be present for those still missing) on the internet.

Original article in French.
Translated by Laura Kraftowitz.



DANIEL PRADO is the former head of the linguistic unit of Union latine, an intergovernmental organization composed of 35 states whose mission is to disseminate and promote the Latin languages and cultures. He is the current Executive Secretary of Maaya, World Network for Linguistic Diversity.

DANIEL PRADO

LANGUAGE
PRESENCE IN THE
REAL WORLD AND
CYBERSPACE

Existing sources agree that linguistic diversity is vanishing. According to Unesco, nearly half the world's languages could disappear by the end of the century [LANGUES 2006]. Claude Hagège [HAGÈGE 2000] estimates that at current rates, one language disappears on average every two weeks, while Louis-Jean Calvet [CALVET 2002] believes the extinction to be a bit more gradual, at about ten per year. Is this process inevitable?

Speaker Population

The crux of language disappearance lies in a decrease in speaker numbers. Therein lies the significance of the figure 50% – the portion of the world's languages spoken by fewer than 10,000 individuals [CRYSTAL 2002]. Of course, this reduction is not in the hands of destiny alone. Dead and dying languages occasionally experience renewed vitality, as with Hebrew, which is today an official language after being considered dead for centuries; and Ainu, which is now being taught after counting “*no more than eight speakers on Hokkaido Island in the late 1980s*” [DIVERSITÉ LINGUISTIQUE 2005].

Political Will

The political will of a language's speakers, or at least of its representatives, can bring a vitality to a language that endogenous or exogenous factors have reduced for some time (some examples are Hebrew, Ainu, Catalan, Basque, and French in Quebec). The languages that have been able to recover a place in society and evolve both quantitatively and qualitatively are above all those with institutional support (public or private). But fewer than 3% of languages are publicly protected; while barely a hundred enjoy official status (*de facto* or *de jure*) in a country or region

[LECLERC 2011]. The chances of survival for lesser-used languages with no such protection is alarming.

Socioeconomic Factors

Let's not forget that the planet's linguistic diversity is far from homogenous. The world's 74 top languages are spoken by 94 % of its population [LECLERC 2010], while 70 % of languages are concentrated into twenty countries [EDUCATION 2003], most of which are among the most poor and therefore the least able to support linguistic diversity projects. The *Globalization Group* (2010) suggests that 90 % of total international GDP is produced by the speakers of only 14 languages¹.

Of course, these statistics somewhat reductively take into account only the official languages of the states surveyed. While their authors accordingly adopt a cautious stance in their statistical interpretation, the numbers nevertheless underline the extreme poverty of reliable indicators for measuring linguistic diversity – a phenomenon we will further explore in our discussion of internet user statistics.

Written Language, Oral Language

At a time when one of the *Millennium Development Goals* is the eradication of illiteracy [OBJECTIVES 2005], and in a contemporary society dominated by writing, we must urgently address the child and adult education for those who speak so-called “oral” languages. While most teaching materials are based on the written word, between 90 and 95 % of the world's languages have no alphabet.

Globalisation, Urbanization and the Knowledge Society

The phenomenon of language extinction, brought on by various factors in the recent and distant past (including colonization, genocide, epidemics, war, displacement, and language bans) is now being amplified by a globalisation process evolving on multiple levels (economic, technological,

1 That is, English, Chinese, Japanese, German, Spanish, French, Italian, Russian, Portuguese, Arabic, Dutch, Korean, Turkish and Polish.

social and political) alongside urbanization. The crucial role played by communications in linguistic power relations means an increasing extinction rate in the information age, as the ICT industry promotes the better equipped or more “prestigious” languages to the detriment of others.

According to Carlos Leañez, “[T]he less a language has value [in the eyes of its speakers], the less it is used, and the more it loses value” [LEAÑEZ 2005]. It is use in professional, administrative, educational and legal contexts that allows a language to persevere, because speakers who switch languages depending on context will gradually lean toward the language that allows them the widest range of expression. However, the vast majority of languages are used solely in emotional and local contexts.

In our knowledge society, language loses value for its speakers if they can’t use it to acquire knowledge. Differently said, if a language is absent from cyberspace, its speakers are likely to turn to the use of other languages.

Languages on the Web

Despite a significant increase in online multilingualism since the 90s, only a handful of languages² maintain a significant online presence. English³ certainly remains the most commonly used, although its relative presence has decreased from 75 % in 1998 to 45 % in 2007 [UNION LATINA - FUNREDES 2007], and according to various crossover studies (because as we mentioned, we have no reliable indicators), to almost 30 % today. Note that we are referring here to quantity of content, not of internet users.

Online Language Deficit: The case of African languages

While the few major languages of communication enjoy a decent web presence, the presence of the majority is highly symbolic, with only a few pages dedicated to them. A 2003 study by Marcel Diki-Kidiri showed that in a sample of 1,374 African sites, only 3.22 % used an African language

2 Sources vary: the Unesco *B@bel Initiative* brochure estimates up to 10% of languages; however, just over a hundred languages seem to constitute commonly accepted media of communication.

3 Or should we say “English and *Globish*”, as the latter is increasingly identified as a variant in its own right?

as the language of communication [DIKI-KIDIRI 2003]. The *Language Observatory Project* [LOP 2011] announced in 2009 a decline of autochthonous languages in the continent after a brief recovery period that lasted until 2005. The reason for this is clear: while Africa is considered, along with Asia, one of the two continents with the highest linguistic diversity (around 2,100 languages according to the website *Ethnologue*, 2011), Africa also has a long history of seeing colonial languages establish themselves as the *lingua franca*, and is only now returning to its original languages as media of expression in educational and professional contexts. Additionally, very few African languages have a graphic system and so cannot be represented online except on a multimedia platform, a subject to be further explored later.

Search Engines, Social Networks

As of March 2011, Google, the most widely used search engine, and the most sophisticated in terms of linguistic tools, offered language recognition in fifty languages. While Icelandic, with 240,000 speakers, has long been recognized, other languages with between 10 and 200 million speakers (including Bengali, Javanese, Tamil, Malay, Hausa, Yoruba, Fulani, and Quechua) remain excluded. The famous engine that recognizes⁴ thirty European languages recognizes only one African language and no indigenous American or Pacific languages.

Yahoo!, for its part, does no better, with just under forty languages recognized, of which eight are Asian and none are African, indigenous American, or Oceanic.

Few tools on the internet are as linguistically rich as Wikipedia, which counts almost 19 million entries in nearly 300 languages. Despite population outreach attempts by Twitter, Facebook, YouTube and other widely used internet services, through localized language versions, automatic translation services or subtitling, these companies remain far from meeting the needs of users of more than fifty languages.

⁴ Note that our use of the term “recognition” refers to an engine’s ability to search for a language and produce results. Google offers interfaces in 150 languages, but this is distinct from saying the engine recognizes all of them.

Machine Translation

Many analysts, including Graddol [GRADDOL 2007], see machine translation as a panacea with the potential to evade the need for a *lingua franca* by allowing all to speak their own language.

Keep this in mind: only sixty languages have access to such systems. Additionally, most systems encourage pairing the speaker's native language with English, or translating between a dozen major languages (French, Chinese, Spanish, German, Japanese, Russian, etc.). All other language coupling systems either have rudimentary technology or use English as a pivot language. In all, only 1 % of the world's languages have an automated translation system at their disposal⁵. Therefore, it seems that the *linguas francas* have some good times ahead [PRADO 2010], while a Bengali speaker who wants to communicate with a Yoruba or Quechua speaker will have to continue relying on an intermediary language.

The bottom line is that the most effective translation systems are those with a sufficient bilingual corpus (including "statistical" systems like *Google Translate*). The corresponding reality, if we are to believe the *Unesco Index Translationum*, is that only around fifty languages possess a sufficient number of translated texts⁶.

Languages of the Connected

Low productivity is a key risk faced by languages in cyberspace. It causes their speakers turn to better equipped languages, triggering a negative feedback loop: less productivity, less audience; less audience, less productivity.

Studies on the major international languages of communication indicate first and foremost that a language's online productivity is linked to its number of internet users and their level of computer literacy [PIMIENTA 2007]. But we cannot be as sure about the factors determining the productivity of lesser-used languages, and there are very strong reasons to even doubt about it.

⁵ See in this book: Joseph Mariani, *How Language Technologies Support Multilingualism*.

⁶ To see statistics of languages toward which are most often translated, see the *Unesco Index Translationum*: <http://databases.unesco.org/xtrans/stat/xTransStat.a?VL1=L&top=50&lg=1>.

Indeed, most statistics concerning online language use take into account only a few dozen languages [INTERNET WORLD STATS 2011] – a number close to Google’s – and never include those of African, American and Oceanic origin. The number of users/speakers of other languages is so insignificant that they are not even listed, a circumstance that makes it challenging to establish reliable rates of productivity.

However, our reservations concerning *Internet World Stats* numbers on internet users by language aside (*ibid.*)⁷, the service nevertheless permits us to grasp the evolution over time of internet penetration by geolinguistic area. In March 2011, the speakers of the following ten languages, in order of most likely to use the internet, were: English, Chinese, Spanish, Japanese, Portuguese, German, Arabic, French, Russian and Korean. It is interesting to note the progress of the Spanish (fourth in 2006, third in 2011), Portuguese (eighth in 2006, fifth in 2011), Arabic (absent in 2006, seventh in 2011), Russian (from tenth place to ninth) and especially Chinese, which is the main cause of the reduction in percentage English-speaking users from 30 % in 2006 to 25 % in 2011.

The Severe Lack of Indicators for Measuring Linguistic Diversity in Cyberspace

The reservations regarding statistics expressed in the preceding section brings us back to the debate surrounding the lack of indicators for measuring linguistic diversity online, particularly the lack of detail provided about these speakers. While many users resort to a dominant language because their language is not sufficiently equipped for online representation or simply because it has little “value” to them, the number of users of other languages is far from negligible, even if it is as yet difficult to quantify.

Languages Maladapted to the Web, or a Web Maladapted to Languages ?

In terms of tools, we know that a language’s online representation is not simply cultural or quantitative. It is most crucially a question of

⁷ The explanations provided by the site do not explain the method for deducing what language is used by a given internet user, and the number of potential users of a number of languages (English, Arabic, Chinese, French, Portuguese, etc.) seems wrong.

technology. The internet, recalls Paolillo [PAOLILLO 2005], is an instrument originally conceived primarily for the English language. By extension, languages sharing the Latin alphabet and Western cultures were able to find a comfortable place for expression more quickly than others, although let's not forget that European-specific diacritics still don't have a place everywhere online, despite advances are that sometimes given an excessively high profile, as with actions advocating the acceptance of domain names in different alphabets and diacritics. English remains the language of programming, markup, coding, communication between servers and most importantly, the bases of computer languages. Computer languages are based on English, and computer scientists are professionally required to know it.

But how many languages encounter more significant constraints related both to the technical problems of representation and to cyberspace-specific cultural media use [DIKI-KIDIRI 2007]?

Crossover between the Internet and Traditional Publishing

The publishing world and the world of the web display a certain degree of statistical similarity. Numbers show that only thirty languages publish more than a thousand volumes per year; six of these produce two thirds of world's literature, and English alone predominates considerably with 28 % of all volumes [LECLERC 2011].

That these dominant thirty correspond by and large with those receiving Google recognition seems to indicate a parallel between the world of the web and that of paper publishing.

Should we therefore conclude that online production must be preceded by paper?

Google Books, the most publicized online library initiative, and similar efforts by public and private operators to digitize library collections⁸, may at first sight seem only to reproduce the *status quo* of linguistic diversity, in the online context⁹. Yet we know that the internet has opened the door to forms of expression outside the interests of traditional publishers. For

⁸ See in this book: Hervé Le Crosnier, *Digital Libraries*.

⁹ See the portal on the challenges of digitizing works (*H)ex-Libris* <http://www.hex-libris.info>

example, scientific publishing in languages other than English has seen a revival thanks to the ease and low cost of online publishing, in contrast to traditional publishers who prefer to avoid the financial risk of editing articles with low reader appeal¹⁰.

However, this specific case is not reflected in international indexes indicating precisely the opposite – a steady increase in English. Thus the web, while providing a voice for minority languages traditionally absent from paper publishing, still does not adequately reflect the prism of cultural-linguistic diversity, even according to the most basic indicator of speaker numbers. There is no indication that the situation will change in coming years.

But it's not Just the Web

As of yet, no study has provided much insight into the role of linguistic diversity in “informal” media such as e-mail, instant messaging, chat forums, mailing lists, blogs and social networks. Surveys, statistics and analysis remain quite limited. However, a cross analysis of such texts permits us to assess general trends, finding that production via such informal channels is much higher than the production of web pages, even if it is often fleeting, and that in these contexts the place of lesser-used languages has increased significance.

Blogs

Even though Web 2.0 is still new, by the end of 2010 there already existed 152 million blogs and 600 million Facebook pages, compared to 255 million traditional websites [PINGDOM 2011]. These interactive spaces behave very differently from traditional web pages.

In 2006, Funredes experimentally applied its method of analysis to blogs, finding that :

... [T]he blog has a different logic of productivity compared to traditional websites according to various populations. Indeed, Hispanics, for example, produce proportionately as many blogs web pages as

10 But this is another debate that is more concerned with the commercial interests at play in scientific publishing.

Anglophones¹¹, while this proportion diminishes by a third for Francophones and Lusophones and by a tenth for Germanophones.

Can we conclude that there exists a cultural reticence on the part of Germanophones, Francophones and Lusophones regarding blog use, since it is difficult to imagine that such reticence could be due to technical or economic constraints, given that the socio-economic development of these three groups is inversely proportional to their comparative presence?

In fact, the exponential growth of the blogosphere (for which statistics are as ephemeral as they are contradictory) suggests that there is still some time left before its numbers stabilize. Some languages will probably catch up as they did with the web (for example, France's Minitel, which blocked the emergence of a French internet until 2000), at which point we will see these statistics evolve.

However, one can also see the blog as a cultural phenomenon (as well as a political phenomenon for certain countries) that certain cultures will resist adopting, as has been the case for chats and forums. Others on the other hand will find in these formats a more flexible and open means of expression.

The Web, by nature more institutional than blogs, chat rooms or social networks, serves to convey messages to large communities. Messages are therefore most often circulated in a language that will be best understood by the highest number of people¹².

However, the blog is usually the result of individual and local initiatives to enable one or more persons to express ideas, feelings, points of view, or, simply, to be known. In general, the creators of blogs are more interested in free expression than in the extent to which they use their own language as a means of expression.

11 The same proportion of blogs and websites in Spanish compared to English, that is, 10 pages in Spanish for 100 in English, for both blogs and web pages.

12 Unless it comes from the decisions of leaders who, belonging to professional categories that allow them to control the major languages of communication, are often tempted to use them to the detriment of others. It is common to find that international agencies and multinational corporations use only one or two major languages to communicate while their public mainly uses other languages.

How many blogs are written in marginal languages? While we can't say for certain, a simple web search¹³ reveals the blog as an emerging medium for these languages.

Interactive Cyberspace

E-mail, chat rooms, forums, and discussion boards represent a refuge for minority languages, as users sharing a virtual community also share a common language of proficiency, provided no speaker of another language intrudes [PRADO 2005]. Indeed, as soon as one of the interlocutors doesn't understand the local language spoken within a community, the more "prestigious" languages are imposed. This reality, common in scientific discussion forums where English enjoys primacy¹⁴, is also present in less formal discussion forums in diglot or polyglot areas where a "prestigious" language is either the official language or the *lingua franca*. But attitudes seem to differ from medium to medium. For Paolillo [PAOLILLO 2005], some people (speakers of Punjabi or the Arabic Gulf, for example) are more likely to use their mother tongue for chat sessions than for writing emails, including bilingual individuals who speak both their native language and their country's official language.

Various international forums and studies have highlighted the role of informal media against brain drain, a phenomenon that has been decried by emerging countries. Today, diaspora inhabitants maintain contact with the homeland and contribute to its development online and in their own language.

Many regional, national, ethnic and linguistic forums and discussion groups unite specialists living in their country with expatriates who want to stay connected. Assuming participants belong to the same linguistic community, these conversations do not take place in English, French or Spanish, (or any other official language or *lingua franca* of their country), but in Punjabi, Creole and Guarani.

13 For example: <http://blogsearch.google.fr/>

14 English often ends up imposing itself in a forum as soon as allophone enters, even if the majority of speakers use another language. This phenomenon also occurs with other major languages of communication dominating linguistic areas (for example in Francophone, Lusophone, Russian-speaking, and Arabic-speaking regions).

It's Not Just the Content, It's the Container

The internet is not culturally neutral. Its size, its way of representing reality, its topography, its governance, its protocols and norms, remain tied to the English milieu of its birth. The internet thus remains a place where Anglo-Saxon culture reigns over familiar territory, but not only because of linguistic dominance. The formats used, the flow of messages, methods of text combining, image and sound, screen size, the use of keyboards, the predominance of written over oral communication, and so on, are all factors that may not always correspond to cultures wishing to appropriate it.

Paolillo (*ibid.*) reminds us that the Maori cited solely cultural issues in their refusal to accept digital libraries. Specifically, “[T]he availability of information [is] protected in Maori culture”. This begs the question of whether the web, forums, blogs, and mailing lists do not at times go against the principles or cultural values of a people, leading to less or no use in a given culture.

Internet-specific formats are far from suitable for unwritten languages. Are these languages absent from cyberspace? In the booklet *How Does One Ensure the Presence of Language in Cyberspace?* (Comment assurer la présence d'une langue dans le cyberespace?), Marcel Diki-Kidiri shows how a language without a writing system may enter cyberspace [DIKI-KIDIRI 2007]. But what of languages whose speakers cannot or will not access the web through these channels?

Beyond Writing

Today, all forms of information or communication can employ electronic channels previously reserved for writing. IP telephony, digital radio and television, audio and video downloads, video hosting sites such as YouTube, streaming and others, are now a part of everyday life, at least in countries and regions with an ICT park and with easy, inexpensive and high-speed internet access. Many possibilities have opened for the representation of languages – including the unwritten ones – in cyberspace.

Keep in mind that the mobile phone is enjoying great success in Africa and that the radio is the communication tool of choice in a continent where the Press (especially sub-Saharan Africa) has a weak presence.

Non-text-based internet can offer an alternative to people with no written language¹⁵ or for languages with little or poor computer system recognition (i.e. encoding problems, fonts, keyboards, and software).

The Digital Divide

It is also necessary that populations have the access and capacity to produce audiovisual material and especially to be able to find this information. However, a global mapping of access to cyberspace illustrates a link between the digital divide and the socioeconomic divide. The internet has certainly become a tool of daily life for urban populations in industrialized countries, but it remains internationally inaccessible for five out of seven individuals. More than five billion people lacked internet access at the end of 2010 [PINGDOM 2011]; distribution itself is uneven: at most 10% of Africans are connected (and the vast majority of African users are concentrated in South Africa and on the Mediterranean); 25% of Asians are connected, while 80% of North Americans (excluding Mexico) and 65% of Europeans are online.

Audio and Video

Possible solutions that could reduce unequal access are financial, political and educational, rather than technical. IP telephony services like Skype¹⁶ and Messenger¹⁷ (which accept both voice and video) are currently within the technical reach of most users, because the linguistic constraints are minor. Similarly, the webcast – digital radio and television, podcasting, etc. – has become more user-friendly and less bandwidth-intensive.

Wishes

The barrier that prevents 95% of the world's languages from being present in cyberspace, namely the absence of writing or maladaptedness of a language for ICTs, could potentially disappear. For this to happen, of course, we must first have adapted computers and a high-speed connection. But

15 See in this book: Tunde Adegbola, *Multimedia and Signed, Written and Oral Languages*.

16 <http://www.skype.com>

17 <http://messenger.msn.fr/>

above all, target populations must appropriate the technology to develop tools for themselves.

There is no doubt that if the Tunis Agenda for the Information Society [SMSI 2005] were followed¹⁸, the danger of languages disappearing would fade away. By becoming instruments of communication, they would regain their value. However, ICT access isn't everything. Technology must be appropriated; technical, cultural and financial hurdles confronted [PIMIENTA, BLANCO 2005].

The planet's fabulous linguistic diversity is mostly absent from cyberspace. Cyberspace could be the medium that either gives dying languages a second chance... or kills them for good.

We must urgently lay out all missing information using reliable and comprehensive indicators; and we must urgently propose awareness raising policies among key players of endangered languages. It is vital that we provide linguistic diversity with the tools for its protection, in this emerging twenty-first century, through cyberspace above all.

BIBLIOGRAPHY

[CALVET 2002] CALVET, Louis-Jean. *Le marché aux langues*. Paris: Plon, 2002. ISBN 2-259-19660-8.

[CRYSTAL 2002] CRYSTAL, David. *Language Death*. Cambridge: Cambridge University Press, 2002. ISBN 0521012716.

[DIKI-KIDIRI 2003] DIKI KIDIRI, Marcel et BABOYA EDEMA, Atibakwa. Les langues africaines sur la Toile dans Les Cahiers du Rifal, n° 23, *Le traitement automatique des langues africaines*, Bruxelles, Agence intergouvernementale de la Francophonie et Communauté française de Belgique, novembre 2003, ISSN 1015-5716. <http://www.rifal.org/cahiers/rifal23/rifal23.pdf>

[DIKI-KIDIRI 2007] DIKI-KIDIRI, Marcel. *Comment assurer la présence d'une langue dans le cyberspace?* Paris, Unesco, 2007, CI-2007/WS/1 – CLD 31084. <http://unesdoc.unesco.org/images/0014/001497/149786f.pdf>

Ethnologue, languages of the world, 2011. http://www.ethnologue.com/ethno_docs/distribution.asp?by=area

Globalization Group inc. Top Languages by GDP 2010. <http://www.globalization-group.com/edge/2010/03/top-languages-by-gdp/>

18 Notably its recommendations for providing universal access to ICT, reducing connection costs, and expanding bandwidth.

- [GRADDOL 2007] GRADDOL, David. *English next*, Royaume-Uni, 2007. <http://www.britishcouncil.org/learning-research-english-next.pdf>
- [HAGÈGE 2000] HAGÈGE, Claude. *Halte à la mort des Langues*. Paris: Odile Jacob, 2000. ISBN 2-7381-0897-0
- [INTERNET WORLD STATS 2011] Internet World Stats. *Internet World Users by Language*, 2011. <http://www.internetworldstats.com/stats7.htm>
- [LANGUE 2006] *Langues en danger*, Paris, Unesco, 2006. <http://www.unesco.org/new/fr/unesco/themes/languages-and-multilingualism/endangered-languages/>
- [L'ÉDUCATION 2003] *L'éducation dans un monde multilingue: Les contextes multilingues: un défi pour les systèmes éducatifs*, Paris: Unesco, 2003. ED-2003/WS/2. <http://unesdoc.unesco.org/images/0012/001297/129728f.pdf>.
- [LEAÑES 2005] LEAÑEZ, Carlos. «Español, francés, portugués: ¿equipamiento o merma?» dans *Congreso internacional sobre lenguas neolatinas en la comunicación especializada*, México, Colegio de México, 2005. ISBN 968-12-1179-0. http://dtil.unilat.org/cong_com_esp/comunicaciones_es/leanez.htm#a
- [LECLERC 2011] LECLERC, Jacques. «L'expansion des langues» dans *L'aménagement linguistique dans le monde*, Québec, TLFQ, Université Laval, 24 janvier 2011. http://www.tlfq.ulaval.ca/axl/langues/2vital_expansion.htm
- [LECLERC 2010] LECLERC, Jacques. «L'inégalité des langues» dans *L'aménagement linguistique dans le monde*, Québec, TLFQ, Université Laval, 2 mai 2010. http://www.tlfq.ulaval.ca/axl/langues/1div_inegalite.htm
- [LOP 2011] LOP, *Language observatory project*, 2011. <http://www.language-observatory.org/>
- [OBJECTIFS 2005] *Objectifs du Millénaire pour le développement*, New York, Nations unies, 2005. <http://mdgs.un.org/unsd/mdg/Home.aspx>
- [PAOLILLO, PIMIENTA 2005] PAOLILLO, John, PIMIENTA, Daniel et PRADO, Daniel. *Mesurer la diversité linguistique sur Internet*. Paris, Unesco, 2005, CI.2005/WS/06. <http://unesdoc.unesco.org/images/0014/001421/142186f.pdf>
- [PAOLILLO 2005] PAOLILLO, John. Diversité linguistique sur Internet: examen des biais linguistiques dans *Mesurer la diversité linguistique sur Internet*. Paris, Unesco, 2005, CI-2005/WS/06 CLD 24822. <http://unesdoc.unesco.org/images/0014/001421/142186f.pdf>
- [PIMIENTA, BLANCO 2005] PIMIENTA, Daniel, BLANCO Alvaro. *Le chemin parsemé d'obstacles des technologies de l'information et de communication (TIC) vers les TIC pour le développement humain (DH) et l'approche par processus*, Saint-Domingue, Funredes, 2005. <http://www.funredes.org/presentation/TICpDHF.ppt>
- [PIMIENTA 2005] PIMIENTA, Daniel. Diversité linguistique dans le cyberspace: modèles de développement et de mesure dans *Mesurer la diversité linguistique sur Internet*. Paris, Unesco, 2005, CI-2005/WS/06 CLD 24822. <http://unesdoc.unesco.org/images/0014/001421/142186f.pdf>

[PIMIENTA 2007] PIMIENTA, Daniel. *Fracture numérique, fracture sociale, fracture paradigmatique*, Saint-Domingue, Funredes, juillet 2007. http://funredes.org/mistica/francais/cyberotheque/thematique/fracture_paradigmatique.pdf

[PINGDOM 2011] Pingdom. *Internet 2010 in numbers*, janvier 2011. <http://royal.pingdom.com/2011/01/12/internet-2010-in-numbers>

[PRADO 2010] PRADO, Daniel. « La traduction automatisée : le cas des langues romanes » dans *Traduction et mondialisation*, revue Hermès n°56, CNRS éditions, Paris, 2010, ISBN : 978-2-271-06992-4, ISSN : 0767-9513.

[DIVERSITÉ LINGUISTIQUE 2005] PRADO, Daniel. Diversité linguistique dans le cyberspace. Le contexte politique et juridique dans *Mesurer la diversité linguistique sur Internet*. Paris, Unesco, 2005, CI-2005/WS/06 CLD 24822. <http://unesdoc.unesco.org/images/0014/001421/142186f.pdf>

[SMIS 2005] Sommet mondial sur la société de l'information. *Agenda de Tunis pour la société de l'information*, UIT, 18 novembre 2005, WSIS-05/TUNIS/DOC/6(Rév.1)-F. <http://www.itu.int/wsis/docs2/tunis/off/6rev1-fr.html>

[UNION LATINE - FUNREDES 2007] Union Latine-Funredes. *Langues et cultures sur la Toile 2007*, Paris, 2007. http://dtl.unilat.org/LI/2007/index_fr.htm

Vitalité et disparition des langues, Paris, Unesco, 2003. <http://www.unesco.org/culture/ich/doc/src/00120-FR.pdf>

ENGLISH WON'T BE THE INTERNET'S *LINGUA FRANCA*

In the 1990s, English was so dominant on the web that many already saw it as the undisputed lingua franca of globalisation. Today, the share of English has fallen below the symbolic mark of 50% due to the rise of the languages of the BRIC countries (Brazil, Russia, India and China) and the gradual de-Westernization of the linguistic center of gravity of the planet. The question, therefore, changes in nature: what is the significance of the emergence of a multilingual cyberspace?

Original article in French.
Translated by Laura Kraftowitz.



MICHAËL OUSTINOFF is Associate Professor (Habil.) in Translation Studies at the Institute of the Anglophone World, University of Paris 3 Sorbonne Nouvelle and currently on sabbatical leave at the ISCC, the Institute for Communication Sciences of the CNRS. His third book *Traduire et communiquer à l'heure de la mondialisation* (Translating and Communicating in a Globalized World) was published by CNRS Éditions in 2011.

MICHAËL OUSTINOFF

ENGLISH
WON'T BE
THE INTERNET'S
LINGUA FRANCA

As the internet began to spread across the world in the early 1990s, English overwhelmed the Web. Many saw their predictions confirmed of English's inexorable rise as a global language, first among them David Crystal in his book *English as a Global Language* [CRYSTAL 1997].

That English has become global is undeniable, but to infer that it will become the single vector of international communication in the era of globalisation is an entirely different matter, and this argument holds less and less water.

It is a sign of the times that within the English speaking world itself, criticism against this claim is the most reasoned and radical, starting with a 2009 report by the *British Academy* [BRITISH ACADEMY 2009], and continuing with the no-less-serious United States *Department of Energy* (DoE), which in June 2010 launched the multilingual platform WorldWideScience.org, based on the observation that science – particularly “hard” science – is no longer done only in English.

The evolution of language presence on the internet is illuminating in this regard. English has diminished to well below the symbolic threshold of constituting 50 %, or even 30 %, of the total. Paraphrasing the famous words of Umberto Eco, for whom “*Translation is the language of Europe*”, we can say that the internet's *lingua franca* is multilingualism and, consequently, translation. We are seeing the rebabelisation of the world, a trend reconfirmed by the internet.

None of this is surprising. We now have enough experience to understand the factors behind this paradigm shift, a shift that demands to be understood in new terms, by implementing an approach that is not only linguistic, but resolutely multidisciplinary [OUSTINOFF, NOWICKI, MACHADO

DA SILVA 2010]. Behind the question of the *lingua franca* loom major economic, cultural and geopolitical questions; cyberspace is one of their most striking manifestations.

THE RESISTIBLE RISE OF “ENGLISH-ONLY POLICY” AND THE “PARADOX OF THE DOMINANT LANGUAGE”

A dramatic fall in the English portion of the internet is underway [PAO-LILLO, AL. 2005]. According to recent data on *Internet World Stats*¹, it stood at around 27.3 % in June 2010. Assuming these figures are reliable, this decrease is due mainly to four factors. First, the growth rate of the languages present: from 2000 to 2010, English presence increased by only 281.2 % (a considerable figure regardless), while Chinese, the second most represented language at 22.6 %, experienced a growth rate of 1277.4 %. But one language doesn't overtake another solely because of a higher growth rate: Arabic had the most impressive growth of the period at 2501.2 %, but still holds only a 3.3 % share, putting it at seventh place, behind Spanish (third, 7.8 %), Japanese (fourth, 5 %), Portuguese (fifth, 4.2 %) and German (sixth, 3.8 %) but before French (eighth, 3 %), Russian (ninth, also around 3 %), and Korean (tenth, 2 %). These ten languages make up around 82 % of the total, while the rest combine to make up the remaining 18 %.

The second factor is the rate of internet connection (*“Internet Penetration per Language”*), which varies according to the infrastructure and level of development in a given country. In Japan, the rate is 78.2 %, which represents over 99 million internet users and explains why Japanese comes in fourth even with its relatively modest growth rate (110.6 %), as opposed to Arabic, which has a much higher number of potential speakers (347 million, compared to Japan's 126 million, most of them concentrated in the territory of Japan), but is farther behind, due to its “penetration rate” of only 18.8 %. In this measurement, Arabic is similar to French, with its similar number of speakers (347 million worldwide) and a similar penetration rate (17.2 %), with each language closely following the other.

The third factor to be taken into account is a language's total number of speakers, which represents its potential breeding ground for increasing

1 Internet World Stats <http://www.internetworldstats.com/stats7.htm>

the number of internet users once the “digital divide” is narrowed. The current ranking of the top ten languages is not fixed, but dynamic, and changes with time. It doesn’t take a genius to observe that Arabic, with a growth rate of 2501.2 %, against French’s 398.2 %, cannot but outrun the latter so long as its penetration rate increases. Conversely, French, even at its current rate, could pass up Arabic if its penetration rate were to grow faster. Conversely, the higher the current penetration rate, the greater a language’s chance of being soon demoted in rank. This is the case of Japanese and German, for example, rated at 78.2% and 78.6% respectively, which receive high scores because of their status as major economic powers, and high wealth of Germanophone countries like Austria and Switzerland.

By adding to these three factors the fourth factor of economic and geopolitical power, all of them working together, it is clear that the once dominant English-only policy will soon have to compete against Chinese, Spanish, Arabic, and even Hindi and Indonesian – two languages that don’t yet appear on the top ten – for its spot on the top of the pyramid.

But there is also a fifth, essential factor to consider, one that internet World Stats has deliberately left aside: *“Indeed, many people are bilingual or multilingual, but here we assign only one language per person in order to have all the language totals add up to the total world population (zero-sum approach)”*². We must remember that since the existing 6000 to 7000 languages in the world are spread across only about two hundred countries, monolingualism is not the rule but the exception. To which I will add the uncommon but in no way negligible phenomenon of inter-comprehension: a Lusophone has direct access to 4.2 % of all internet, which may seem small, but s/he also has access to Hispanophone sites (7.8 %), because of the great ease of moving from one language to the other, especially in writing. That makes for 12 % of the internet that a Lusophone can access, a considerable number equivalent to the shares of Russian, French, Arabic, and German put together.

Now, suppose the Lusophone in question is a Brazilian who studied in France and who, moreover, speaks English, a relatively widespread situation in Brazil. This person’s access is not limited to 4.2 % of the total (Portuguese only), or even to 12 % (Portuguese + Spanish) or to 15 % (by adding the French) –but totals an impressive 42 % of the internet

2 <http://www.internetworldstats.com/stats7.htm>

(Portuguese + Spanish + French + English). Now, consider this person's Chinese counterpart, who can access nearly 50 % of the total (Chinese + English). But these are just as quantitative data. Imagine a French internet user, who speaks not only English, but also German. This person can access a quantitatively smaller share of the total (37 %), but the three languages accessed are three major languages of the European Union. In other words, the *qualitative* dimension must also be taken into account.

Conversely, Anglophones for whom English is the primary language tend to be overwhelmingly monolingual, precisely because English-only policy continues to be perceived as the universal panacea. They therefore have access to only 27 % of the total. The conclusion is inescapable: in a multipolar world, where globalisation is accompanied by the unprecedented development of information and communication technologies (ICT), to speak only the *lingua franca* is to be under-informed – a phenomenon that Louis-Jean Calvet has called the “*paradox of the dominant language*” [CALVET 2007]. And in today's world as in yesterday's, to be under-informed constitutes a handicap. In terms of the internet, when English constituted over 80 % of the total content, the lack of information could be seen as marginal. This is no longer the case.

UNITED STATES OR UNITED KINGDOM, SAME FIGHT: NO TO ENGLISH-ONLY POLICY

In the age of globalisation, being under-informed is a luxury we can no longer afford. Speaking the *lingua franca* is no longer sufficient. In this sense, the fact that countries like France continue to promote an English-only policy, while the English speaking world is beginning to fundamentally question it as a model, is rather amusing [MARTEL 2010].

I would like to examine three key moments of this questioning. First, let us revisit David Graddol's authoritative study undertaken on behalf of the British Council, entitled, *The Future of English? The Popularity of the English Language in the 21st Century* [GRADDOL 1997]. To my knowledge, this is the first substantive analysis conducted by an English speaker predicting that the future of English as an international, or “global”, language (*Global English*) is far from assured. Indeed, nothing prevents us from imagining that other languages will compete for influence, especially as de-westernisation carries on and emerging economies like the BRICS

(Brazil, Russia, India, and China) continue to rise. In 1999, this trend was only a hypothesis; in 2011, it is confirmed daily on the web. The bottom line is that English *monolingualism* is not risk-free: victim of its success, English has spread to the point where the number of “native speakers” worldwide has been overtaken by the number of people who speak English as second language. In a world gone multipolar, English is only one of many core competencies. What makes a difference (on the job market) is no longer mastery of two languages (English + 1), which has become the standard for non-English speakers – but mastery of *at least* two languages (English + 2 or more).

David Graddol’s second blow targets another misconception: English is not destined to be the *lingua franca* throughout the world. Why use English as the primary language of communication in Latin America when the interface of Spanish and Portuguese presents a far more practical solution because of the two languages’ high degree of intercomprehension? This point underscores the strategic value of a local *lingua franca*. But David Graddol goes even further, noting that it is always better to speak to others in their respective mother tongues, especially when it comes to business [BEL HABIB 2011]. If you wish to do business in Southeast Asia, for example, English is certainly useful, but Chinese is now the major language of communication in the region. Rather than passing through English, it is this “local” language that serves as the *lingua franca*. Moreover, these local *lingua franca* are going international, especially when it comes to transnational relations in southern countries: the Chinese are beginning to learn Portuguese and trade with Brazil; the Brazilians are reciprocating and learning Chinese. The growing number of Confucius Institutes for Chinese, Camoens Institutes for Portuguese, Cervantes Institutes for Spanish, and so on, is explained by the fact that soft power is no longer reserved for English, something China in particular has well understood. This explains, for example, the dramatic increase in Chinese language teaching in French public education, even though it’s reputedly much more difficult than German, while the language courses in the latter are on the wane.

In his third and final blow, Graddol does not exclude the possibility that English could be at the mercy of a “nightmare scenario” at a time when international public opinion increasingly sees the defence of linguistic and cultural diversity as necessary: “*These trends suggest a ‘nightmare*

scenario' in which the world turns against the English language, associating it with industrialisation, the destruction of cultures, infringement of basic human rights, global culture imperialism and widening social inequality" [GRADDOL 2000: 62]. This is why the following quote from Leonard Orban, former European Commissioner for Multilingualism, cannot be seen as unrealistic, but as news: "Employees should master, for the benefit of their employers, at least three languages: that of their country of origin, English of course, and a third from among the most widely spoken in the EU – German, French, Spanish or Italian. Without neglecting Russian, Arabic or Chinese" [RICARD 2007]. Put in economic terms, it is no longer English-only policies that constitute a hard cash boon (for which the primary beneficiaries have until now been English-speaking countries), but multilingualism that represents the real "competitive edge" under current and future globalising.

The second key moment was the publication of the *British Academy's* 2009 report *Language Matters*, which drove the point home, this time in the field of research, by sounding the alarm of British researchers' growing preoccupation with foreign languages (p. 3):

In the humanities, for example, fields such as history and philosophy need to draw on scholarship in other languages which is not translated into English, nor is likely to be. In the social sciences, comparative studies and cross-national work in subjects such as politics, sociology and development economics requires knowledge of other languages. And researchers in all disciplines (including the natural sciences) need skills in spoken as well as written languages in order to take up and make the most of opportunities to study and work overseas, or collaborate with overseas partners. With the increasing development in collaborative work, and the large sums of money attached to such work by national and international agencies, lack of language skills inflicts a real handicap on scholars in many parts of the British university system, and therefore weakens the competitive capacity of the system itself.

Who would have predicted, even thirty years ago, that such statements would one day come from a prestigious English institution like the *British Academy*? Let us remember that English-only policy was supposed to obviate the need for additional languages, like Koine in the Hellenic world, Latin in the Middle Ages, or French during the Enlightenment.

And even more so: while these *lingua francas* were reserved for the elites, English is accessible to a much wider audience in the era of mass education.

But there is a third and final key moment to keep in mind, which deserves to be fully developed separately: the launching, under the aegis of the Department of Energy, of the international multilingual platform WorldWideScience.org. This project calls into question the very foundation of English-only policy, and more generally, the use of a single *lingua franca* for global scientific communication.

LINGUA FRANCA, INFORMATION AND COMMUNICATION IN THE AGE OF GLOBALISATION

Before it was called into question, English-only policy was presented into the late 1990s as a model whose utility, if not outright necessity, seemed obvious. One famous English grammarian didn't shy away from putting it bluntly: "*English is the world's most important language*" [QUIRCK, AL. 1980: 2]. Apart from the fact that English was the international language with the greatest reach, four main advantages of this model were upheld:

- English is the most practical because it is an "easy language";
- English-only policy is the most economic solution;
- it is the most democratic and equitable solution;
- English as a *lingua franca* is "culturally neutral".

The latter is elaborated as follows (*ibid.* 6):

English [...] is pre-eminently the most international of languages. Though the mention of the language may at once remind us of England, on the one hand, or cause associations with the might of the United States on the other, it carries less implications of political or cultural specificity than any other living tongue (with French and Spanish also notable in this respect).

In other words, not only are we to presuppose that all languages are interchangeable, but also that English is somehow the most interchangeable of all because of its "neutrality". This is what ultimately legitimates it as a *lingua franca*. Nevertheless, this term is in itself ambivalent, if we return to its definition in the *Petit Robert 2011*, a French reference dictionary.

In the original sense, a *lingua franca* was a specific pidgin designated as a “language spoken until the nineteenth century in the Mediterranean ports; a composite language based on central Italian but also on French and Spanish, and also including Greek and Arab elements”, whose golden age came in the twelfth and thirteenth centuries. It was a purely utilitarian language. Conversely, today the term has enlarged to cover a different reality: “A common language used across a fairly large geographical area. Swahili, East Africa’s *lingua franca*”. (*ibid.*) The term “common language” means language “used for communication between different native language groups” as opposed to a “vernacular”. (*ibid.*) The gross asymmetry of English as a *lingua franca* is immediately apparent: in contrast to Latin, which was no one’s native language during the Middle Ages, English is both a common language and the vernacular of English native speakers.

Taking this fundamental asymmetry into account, one can explain why cyberspace trends toward spreading multilingualism and not the inverse, much less an English-only policy (English + 1) that leads to bilingualism or monolingualism whether or not English is one’s primary language. In other words, cyberspace is more likely to side with Wikipedia (which counts 280 languages) than with the portals of major international institutions, of which the most multilingual is the European Union, with its 23 official languages for 27 member countries.

At the polar opposite end from the “hypercentral” language of English (in the terminology of Louis-Jean Calvet) and other “central” languages with general circulation (Spanish, French and Arabic), we find, for example, articles in Navajo on Wikipedia. Although this language counts only 170,717 speakers [SHIN, KOMINSKI 2010], it remains the most widely spoken American Indian language according to the most recent US Census in 2007. We can safely assume that this language is unlikely to appear in the list of major *lingua franca* of the twenty-first century. However, this language holds 173rd position in the category of languages counting over 1,000 Wikipedia articles – 2,154 to be exact. And, of course, it is far from irrelevant when one uses Navajo rather than English on the internet in the United States, a point on which there is no need to dwell except to say that it brings further evidence to the idea that no mother tongue is truly “culturally neutral” – including, at the other end of spectrum, that of an English native speaker.

There is, however, one domain that seems to bypass all the differences that separate one language from another, and to transcend the “world views” (*Weltanschauungen*) that all languages contain, each one being irreducible to another, according to Wilhelm von Humboldt. It is that of science, as Descartes articulated at the beginning of his *Discours de la méthode*: “Those in whom the faculty of reason is predominant, and who most skilfully dispose their thoughts with a view to render them clear and intelligible, are always the best able to persuade others of the truth of what they lay down, though they should speak only in the language of Lower Brittany, and be wholly ignorant of the rules of rhetoric”³ [CASSIN 2004: 466]. Since languages are interchangeable in this regard, Descartes justifies writing *Discours* in French rather than Latin, the dominant language at the time, in order to reach the widest readership. Although he doesn’t express himself in Breton, it is not because the language is less capable of expressing just as complex thoughts as Latin and French, but simply because this would reduce his readership.

Transpose the same reasoning onto today’s world: if science is indifferent to its language of expression, so much the better to use the most internationally widespread language, English. Not that others, from the central (Spanish, Arabic) to the peripheral (Navajo), to those with the most speakers (Chinese, Hindi, Indonesian), are less worthy transmitting scientific texts, but none is more measurably able to reach the greatest possible readership. Certainly there exists an asymmetrical relationship according to whether the reader is a native speaker or not, but to paraphrase Winston Churchill on democracy, this is the worst form of communication, except for all the others.

Such reasoning seems flawless: the language of science – especially the “hard” sciences – isn’t it English? However, this is an optical illusion.

The USA are beginning to realise that their research output, once in first place by a landslide, is being caught up to with great speed by other countries, beginning with Europe and the BRICS. They have gotten the message that science doesn’t exist only in English, but in other languages as well.

This is why the US *Department of Energy* followed the lead of the *British Academy* by hopping onto the bandwagon of translation

3 Translation from <http://www.literature.org/authors/descartes-rene/reason-discourse/chapter-01.html>

via new technologies, in order to create the multilingual platform WorldWideScience.org at the end of June 2010, where you can search on 70 scientific databases from 66 countries in a diversity of languages. The whole thing is connected via a multilingual search engine and automatic translation program in ten languages: Arabic, German, English, Chinese, Korean, Japanese, Spanish, French, Portuguese and Russian, with other languages being progressively added to the list.

Take the example of China. Here's what comes up on the website: "*In 2008, while Chinese scholars published 110,000 papers on international journals recorded by SCI, they also published 470,000 papers on domestic Chinese journals. Without accessing these 470,000 papers, it is impossible to obtain a realistic feeling about the thrust of scientific and technological advancement in China. Therefore, the need for mutual translation between English and Chinese and for cross-language retrieval is increasingly urgent*".

It could not be clearer: to communicate is not simply a question of being informed [WOLTON 2003]; it is first and foremost, a question of having access to information at all. And in an increasingly multilingual world, with the West no longer the centre of gravity (including Japan in the famous economic Triad), a *lingua franca*, universal as it may be, does not, on its own, replace all others. This paradigm, inherited from the aftermath of World War II, which saw the emergence of The United States of America as a superpower, has run its course, and even the United States is beginning to take note.

CONCLUSION

We have understood that the question of the *lingua franca* must be re-framed within the broader context of geopolitics, just like translation, a similarly central issue, for which the etymology is illuminating. The term *translatio* (Latin for "transfer") was understood in the Middle Ages to mean "translation", a meaning that the English term "translation" retained. But it was also used [CASSIN 2004: 1312] to designate knowledge transfer (*translatio studii*) or power transfer (*translatio imperii*). The concept of knowledge as power was thus transmitted from Greece to Rome, and then to the West, a phenomenon that continues today in other parts of the world, particularly in Asia (India, China) and in the South (Latin America; Arab countries; perhaps, with time, Africa). Which means the

emergence of new powers, the redistribution of the card holders at the top of the pyramid. This explains why Chinese is becoming an international language, while in the past it wasn't spoken outside China, Taiwan, and their diasporas.

The case of Portuguese is revealing in this respect. Like English, Spanish, Arabic or Russian, it is a *lingua franca* across a linguistic area, in this case the Lusophone countries. But it is also one of the BRICs' major languages. In 1960, Brazil was a developing country with about 70 million people. Today it is the eighth world power and counts nearly 200 million inhabitants. From its original place on the periphery, Portuguese has come to play an increasingly central role. In a multipolar world, the continuation of English as the sole *lingua franca* appears less and less self-evident.

In 2005, Michael J. Barany, mathematician based in Princeton, published an article *Business Week* entitled "Science's Language Problem", which smartly summarises the above :

Tomorrow, when the number of researchers fluent in English will certainly dwindle in laboratories throughout the world, the English straightjacket will become increasingly uncomfortable – at a time when the volume of scientific information is about to explode.

In China and India, world-class scientific infrastructures are emerging, and more discoveries will be reported in their local languages. These could go unheeded or underappreciated elsewhere – just as work currently published in Japanese or French often fails to impress American scientists not fluent in those languages. Papers that undergo various translations and interpretations often emerge murky and hard to comprehend.

While automated translation is rapidly improving, it is unlikely that machines can ever attain the nuance and technical accuracy required for the ever-changing vocabulary of science.

The globalisation of science offers innumerable new opportunities for intellectual advancement. But unless we build better bridges between linguistic communities, countless ideas and innovations could be ignored and effectively lost.

What is true about the globalisation of scientific communication [LÉVY-LEBLOND 2007] is even more so in other spheres. As powerful as English may

be, cyberspace cannot content itself with a single *lingua franca*. It suffices to browse the internet to realise the immense benefit that multilingualism holds. To see the world only through English, by contrast, offers an increasingly weak vision as other languages grow in both absolute and relative terms. This is a paradigm shift of which we are only now beginning to measure the full extent. Communication is no longer conceivable on the scale of a single language; multilingualism and translation (including automatic or computer assisted) must be adjoined. These, however, are necessary but insufficient prerequisites.

“*War is too serious a matter to be entrusted to the military*”, said the french leader Clemenceau. The same applies to the language question, whether in cyberspace or in the broader context of rapid global rebabelisation. It cannot remain the preserve of linguists, translators, interpreters and translation scholars: it exists in the spirit of a truly multidisciplinary approach involving both the social sciences and the natural or “hard” sciences [OUSTINOFF 2011]. Even more broadly, it is now everyone’s business, since cyberspace is not addressed – far from it – only to the specialists.

BIBLIOGRAPHY

[BARANY 2005] Barany, M. J., *Science’s Language Problem*, Business Week, 16 mars 2005. http://www.businessweek.com/technology/content/mar2005/tc20050317_4179.htm

[BRITISH ACADEMY 2009] *The British Academy, Language Matters. Position Paper*, 2009. <http://www.britac.ac.uk>

[BEL HABIB 2011] Bel Habib, I., *Multilingual Skills Provide Export Benefits and Better Access to New Emerging Markets*, Sens Public, octobre 2011. http://www.sens-public.org/article.php3?id_article=869

[CALVET 2007] Calvet, L.-J., *La traduction au filtre de la mondialisation*, dans Oustinoff, M., Nowicki, J. (dir.), *Traduction et mondialisation*, Hermès, n°49, Paris, CNRS Éditions, 2007.

[CASSIN 2004] Cassin, B. (dir.), *Vocabulaire européen des philosophies. Dictionnaire des intraduisibles*, Paris, Le Robert / Le Seuil, 2004.

[CRYSTAL 1997] Crystal, D., *English as a Global Language*, Cambridge, Cambridge University Press, 1997.

[GRADDOL 2000] Graddol, D., *The Future of English ? A Guide to Forecasting the Popularity of the English Language in the 21st Century*, The British Council & The British Company (UK) Ltd, 1997 (nlle éd., 2000). <http://www.britishcouncil.org/learning-elt-future.pdf>

[MARTEL 2010] Martel, F., *Mainstream. Enquête sur cette culture qui plaît à tout le monde*, Paris, Flammarion, 2010.

Multilingual WorldWideScience.org. <http://worldwidescience.org/multi/index.html>

[LÉVY-LEBLOND 2007] Lévy-Leblond, J.-M., *Sciences « dures » et traduction*, dans Oustinoff, M., Nowicki, J. (dir.), *Traduction et mondialisation*, Hermès, n°49, Paris, CNRS Éditions, 2007.

[OUSTINOFF, NOWICKI 2007] Oustinoff, M., Nowicki, J. (dir.), *Traduction et mondialisation*, Hermès, n°49, Paris, CNRS Éditions, 2007.

[OUSTINOFF, NOWICKI, MACHADO DA SILVA 2010] Oustinoff, M., Nowicki, J., Machado da Silva, J. (dir.), *Traduction et mondialisation. Volume 2*, Hermès, n°56, Paris, CNRS Éditions, 2010.

[OUSTINOFF 2011] Oustinoff, M., *Traduire et communiquer à l'heure de la mondialisation*, Paris, CNRS Editions, 2011.

[PAOLILLO, PIMIENTA, PRADO 2005] Paolillo, J., Pimienta, D., Prado, D., et al., *Mesurer la diversité linguistique sur Internet*, révisé et accompagné d'une introduction de l'Institut de statistique de l'Unesco, Publications de l'Unesco pour le Sommet mondial sur la société de l'information, Paris, 2005. <http://unesdoc.unesco.org/images/0014/001421/142186f.pdf>

[QUIRK, AL. 1980] Quirk, R., Greenbaum, S., Leech, G., Svartvik, J., *A Grammar of Contemporary English*, Londres, Longman, 1980.

[RICARD 2007] Ricard, P. *Une étude britannique prône le multilinguisme en affaires*, Le Monde, 25 septembre 2007.

[SHIN, KOMINSKI 2010] Shin, H. B., Kominski, R. A., *Language Use in the United States: 2007. American Community Survey Reports*, Washington D.C., US Department of Commerce. Economics and Statistics Administration, US Census Bureau, 2010. <http://www.census.gov/prod/2010pubs/acs-12.pdf>

[WOLTON 2003] Wolton, D., *L'autre mondialisation*, Paris, Flammarion, 2003.

TECHNOLOGICAL INNOVATION AND LANGUAGE PRESERVATION

When industry is innovating new technologies, it is largely driven by a vision that is short to medium term. Languages on the other hand are evaluated in the long term. How does this meeting of time scales relate? Should we not think of ICT innovation in terms of linguistic diversity?

Original article in French.
Translated by Laura Kraftowitz.



ÉRIC PONCET founded Linguasoft to support the communities wishing to maintain their language. He has managed many projects using his Language Preservation Process and Tools.

ÉRIC PONCET

TECHNOLOGICAL
INNOVATION
AND LANGUAGE
PRESERVATION

This insert elucidates the influence of information communication and technology (ICT) on the evolution of world languages, and argues for ICT innovation to target multilingualism.

Languages, Technology, and Time

Writing may be considered the oldest language technology (not least because no other before it left a trace). However, five millennia are but a brief period in the existence of human language. Let's not forget that the skull of *Homo habilis* (2.5 million years old) contains Broca's area, the region of the brain that controls language. Even if this does not necessarily mean that human language is 2.5 million years old, it at least gives us an idea of just how young language technology is. Despite its youth, writing has had an influence as fast as it is profound on the evolution of human society. It is precisely this kind of lightning-fast technological revolution –but this time unfolding on a planetary scale– that we are now experiencing with the internet.

Researchers and language activists agree that half of the world's 6,900 languages will disappear within a century. It is likely that this view greatly underestimates the present and future impact of cyberspace on multilingualism.

Technological Linguicide

Since technological innovation is –at least within the industry– often guided by a business approach with short- to medium-range vision, ICT is for world languages a double-edged sword. It is easily conceivable that a software company would place its consumer products on the Chinese

market, thereby entering a market of almost one billion Mandarin speakers, to whom it can distribute millions of licenses. But what happens when that same company wants to localize that same product to serve a language with fewer than 10,000 speakers – and this statistic currently applies to half the world’s languages? Not only will the number of licenses sold count no more than a few dozen or at most a few hundred, but the localized aspect of the work will take longer (and therefore be more costly) than that of Mandarin. Indeed, many of these languages are not standardized, or do not even have a writing system. The company must thus accept to invest significantly more and earn much less. What company is willing or able to convince its shareholders and employees to adopt this as a business strategy? And what of the feasibility of localizing a piece of software for 6,900 languages? It comes as no surprise that not one of the world’s millions of software programs comes even remotely close to approaching this level of multilingualism.

Innovation and Multilingualism

So, given the technology factor, is there no salvation for multilingualism?

Current multilingual technology is simply too restricted, partly because of its limitation for the aforementioned reasons, to the world’s most powerful languages. It makes sense to extend multilingualism toward panlinguism. In other words, rather than *sprinkling* technologies with language, it becomes necessary to rethink them to the end of *integrating* language – as a strong characteristic of humanity.

Woe to the languages that do not have the critical mass, either in speakers or in financial resources. Their weak presence on the Web – if not their complete absence – means that their speakers are, by default, linguistically absorbed into languages that have an online presence. The astrophysical analogy of the black hole is no exaggeration: languages without sufficient inertia or mass to continue their trajectory will inevitably be absorbed, and the lower their mass the faster.

If the first aspect of critical mass (number of speakers) cannot be increased overnight by a simple magic wand, then second (financial resources) may be: language preservation programs can be launched with modest budgets, and all funding is in this sense a catalyst for productive energy into the language communities concerned.

Prospective

Let us return to the rate of language extinction within a century as mentioned above. The author, given current trends and the field work he has undertaken, predicts a language death rate of 80-95 %.

It took writing several centuries to conquer the world; it took a decade for the internet to revolutionize the way its users communicate, eat, work, and play; in short, how they live. It is important to note that language (thus, languages) is the vector of all these activities. Given that the Network of networks is a major language vector, what can we expect for the coming century – ten times the current age of the internet? Can we exclude the emergence of a technology more earth-shattering than writing or the internet? Such an event would leave the majority of world languages little chance of survival, and could lead in the short term to a cultural cataclysm.

“Science without conscience is but ruin of the soul”, wrote François Rabelais. This visionary quote, having brilliantly crossed the last five centuries, is all too easy to transpose onto the theme of this book: ICT without ethics is but the ruin of man.

PRESERVING THE HERITAGE OF EXTINCT OR ENDANGERED LANGUAGES

This article aims to show how cyberspace can support the preservation of extinct and endangered languages. Various case studies illustrate how cyberspace could fill this role and why the languages concerned are worth preserving. We do not favor a singular approach to the question, since multiple factors as diverse as the purpose of the preservation, attitudes toward the language in the user-communities, the existence of a diaspora and the accessibility of digital technologies all have a role to play.

Original article in English.



MAIK GIBSON is the head of the Department of Translation Studies at Africa International University, Nairobi, Kenya, and a Senior Consultant in Sociolinguistics for SIL International. He is also a teacher of sociolinguistics at the Institute for the Development of Languages and Translation in Africa (i-DELTA). His research interests include language shift, language policy and expanding the domains of lesser-used languages.

MAIK GIBSON

PRESERVING
THE HERITAGE
OF EXTINCT OR
ENDANGERED
LANGUAGES

The twenty-first century is witnessing two very different trends, both of which reflect the changing world in which we live. The first is the spread of the internet and communication technology in general; the second is a rapid decline both in the number of languages spoken, and of those which have a viable future. This chapter is concerned with how the internet may be used effectively as a tool for the prevention and mitigation of the effects of widespread language extinction and endangerment.

There are various estimates for the numbers of languages which are threatened by extinction, with up to 50% often being quoted as being at risk. For example *Unesco's Atlas of the World's Languages in Danger* [MOSELEY 2010] gives the estimate of "some 3,000" –but whatever the precise figure, it is clear that language extinction during the twenty-first century will most probably be a major trend, especially in the Americas and Australia. When languages die, local knowledge may also be lost; furthermore, the community may lose one of the main symbols of their ethnic identity, leading to possible social instability. It does seem certain that many languages will not serve as mother tongues for the next generation of children. However, documentation of these varieties can help the community with the maintenance of traditional knowledge, through at the very least the use of their language for heritage purposes, which can also assist with issues of identity. The availability of the internet can make this documentation a less arduous task, and also make such documentation more easily accessible –for both the community, where there is some internet access, and for interested others. It is the aim of this chapter to give the necessary background and consider various factors which may influence the usefulness of the internet in appropriate documentation of extinct and endangered languages.

EXTINCT AND ENDANGERED LANGUAGES

Extinct (sometimes referred to as *dead*) and endangered languages lie at the bottom of the range of language vitality, which can be evaluated by tools such as Unesco's methodology for assessing language vitality and endangerment [BRENZINGER 2003], Fishman's [FISHMAN 1991] Graded Intergenerational Disruption Scale (GIDS), or the Extended GIDS proposed by Lewis and Simons [LEWIS, SIMONS 2010]. Measures of linguistic vitality are principally concerned with the current or likely future of a language as the vehicle of children's primary socialisation. In the extreme case, that of extinction, we are faced with a situation where the language is not used by any people for matters of day-to-day communication, so that the opportunities for children to learn it as their first language are limited to cases of deliberate revitalisation. Often extinction implies complete loss of any knowledge of the language, but this is not always the case – it may be that the children and grandchildren of the last fluent speakers have some passive knowledge of the language from having overheard it when younger, or may even know how to say some words and phrases. And in some cases the language may have been lost as a vehicle of everyday communication, but still has some uses in ritual or religious practices. This is the case of some liturgical languages such as Ge'ez, Latin and Old Church Slavonic, where the extinct languages have been written down, but the scenario is also possible without a developed written form (eg for Lucumi in Cuba, according to Courlander [COURLANDER 1996: 20]). So the term *extinct* (like the other common term *dead*) does not necessarily mean that the language is totally lost, or that it has no functionality in the community, but refers to the lack of speakers for whom the language is one of their primary means of communication. Evidently an extinct language of which all knowledge has been lost cannot be preserved – some level of knowledge, whether from within the community, or consisting of documentation, is a prerequisite.

While the designation of a language as extinct should, in theory, be simply a matter of establishing the lack of first-language speakers, defining endangerment can be more complicated. Generally a language which is not being passed on to children will automatically be deemed as being endangered, as extinction is the most likely outcome after a couple of generations have passed on. There are however some exceptional cases where the community language is not the first one taught to children,

but it is still fully acquired at a later stage. For example, concerning Digo in Kenya, Nicolle [NICOLLE 2012: 4,5] writes: “*Most young children were addressed in Swahili by their parents and other adults... However, by the age of ten or eleven, most children had become proficient speakers of Digo and were habitually addressed in Digo by their elders*”. In addition, cases where only a certain percentage of children are learning the language may also be scenarios of endangerment –generally, the category can be applied if there is some risk of extinction. This risk can be quite difficult to quantify in individual cases, as language shift may sometimes be slowed, halted or even reversed. In addition, Moseley [MOSELEY 2010] classifies languages which are not used in all domains (eg Welsh) as vulnerable. This is however the norm in multilingual societies, where different languages tend to be assigned different functions, whether language shift is prevalent or not.

THE WIDESPREAD LOSS OF LINGUISTIC DIVERSITY

While we accept that language extinction has most probably accelerated due to rapid social changes accompanying globalisation over the last hundred years, it is by no means exclusively a modern phenomenon. Even languages committed to writing such as Sumerian, Old Prussian and Massachusett have died without leaving modern-day descendants. However countless other languages of which either very little or nothing was written down have disappeared, leaving only on occasion some now opaque place names. To accurately compare the twentieth-century loss of linguistic diversity with that of previous centuries is almost impossible because of the lack of documentation.

Despite this, it seems that the predictions of up to half the world’s 6,000+ languages being lost over the next 50-100 years would be an unprecedented reduction of linguistic diversity. The impact of this loss is multi-faceted. Often shift to another language goes along with a change of lifestyle, and thus much traditional knowledge and vocabulary may be lost, for example in the field of ethnobotany, where languages may distinguish between plants unknown to modern science. In cases where the lifestyle change is minimal, traditional vocabulary is sometimes passed on. Brenzinger [BREZINGER 1992] gives two Kenyan cases of this –Yaaku and Elmolo,

whose speakers shifted the communal language to dialects of Maa. Both however maintained aspects of distinct lifestyles for which Maa has a limited vocabulary (beekeeping and hunting in the first case, and fishing in the other), and kept the relevant vocabulary for these areas. Now the former Yaaku are assimilating their lifestyle to pastoralist Maasai culture, and hence losing beekeeping and hunting vocabulary. Even in cases where lifestyles are maintained, an understanding of the meaning and origin of specialist terms can disappear, as the language from which they stem is forgotten. For further coverage of the types of knowledge and world-views that can disappear, see Harrison's [HARRISON 2007] *When Languages Die*.

An additional consequence of extinction of an indigenous language is cultural disruption as evidenced in much higher suicide rates in communities where the language has been or is being lost [HALLETT 2007]. The authors find that in British Columbia, loss of the communal language was the most significant correlate with high suicide rates found among Aboriginal communities, while other "cultural continuity factors" also showed weaker correlation with lower suicide rates. Whether all cases of communal language loss lead to equivalent patterns of cultural disruption has not been established, but this study shows that the loss of a language does not occur in a social vacuum, and can have serious consequences for the community involved.

The wide diversity of languages which are spoken today demonstrate an amazing array of strategies of expressing thought and categorising the world around us. So from a scientific point of view, the extinction of one language means that there is less data concerning the workings of the human mind, not only concerning linguistic facts and theories, but also about more general capacities of the human mind, for example in cognition and perception. And each individual language is equally important for the advancement of knowledge in these areas; basing a universal theory of language or cognition on only a handful of languages is obviously flawed. Each extinguished language represents a lost opportunity to understand ourselves better.

EXAMPLES OF INTERNET-BASED INTERVENTIONS

The type of preservation activity, whether in cyberspace or not, will depend on its goals. Should the purpose be merely to preserve the language

for posterity, some texts, vocabulary and recordings published on the internet could be adequate. If it is decided that the language is to be learned as a second language by current and future generations, then having enough of a base to develop accurate teaching materials will be necessary, and an interactive online course could be developed for those interested in learning the language. If the language still has a desired role within the community, such as may be the case for certain rituals, developing relevant materials may be appropriate, if the society does not view the language of ritual as something secret whose knowledge is not to be shared outside a prescribed group of people.

Preservation activities with the goal of language revitalisation will require a wider range of efforts, but some quite simple online interventions, such as the setting up of websites and forums where people are encouraged to use the language, can be quite effective, depending on the social profile of the speakers of the language.

The usefulness of the internet as a tool for protecting endangered languages is nicely demonstrated in an online document by Mensching [MENSCHING 2000] for Sardinian, a Romance language with a high level of difference between dialects. In fact Moseley [MOSELEY 2010] lists four varieties separately as endangered languages, while the *Ethnologue* [LEWIS 2009] lists these four as varieties within the Sardinian macrolanguage, with a combined population of over one million. In some ways the criteria for potential success in this case are almost ideal, given the large number of speakers using the language today, and the context of a highly developed society with comparatively high levels of income, internet access and education. The main negative indicator is the high level of difference between dialects, but Mensching argues that the written medium helps minimise this, as people converge on a common Sardinian writing system. He mentions many advantages of the internet, and in particular the Sardinian language and culture website¹, in that it:

- “reinforces[s] the linguistic consciousness of speakers”, partially because Sardinian is not just the subject of the dedicated website, but also its medium, thus giving users practice in using the written form;

1 <http://www.lingrom.fu-berlin.de/SardischEngl.html>

- helps speakers by giving the users information about how to write the language, helping the written form become less divergent, and therefore more accessible, with practice, to the community;
- acts as a “*central node for accessing online information about the Sardinian language and the culture of the island*”;
- serves as a space for discussion, documentation and evaluation, as well as enabling all this to happen without the mediation of dominant languages (as is common in academic activities);
- takes advantage of the internet-specific “orality effect”, whereby “*E-mail communication is more similar to spoken language than to written language*”. This observation can be made concerning other minority languages and non-standard dialects, which have a significant presence in online writing, even more so in today’s Web 2.0 contexts such as social networking sites, e.g. Facebook². The normal categorisation of writing as a formal domain, where the dominant language predominates, does not seem to apply in some online writing, which bears some similarity to the language of text messages on mobile phones.

These factors seem to lie behind some of the success of minority language activism on the internet. One can hypothesise similar advantages for minority languages spoken in similar contexts (e.g. in Western Europe), where we may find the same social factors of high literacy and internet access, as well as the resemblance of both the writing system and grammar of the dominant and minority languages. Websites also enable members of the diaspora, not living in the language area, to use their language when it might otherwise be restricted to phone conversations, and so to maintain their competency and to enable them to participate in language activism³. However not all these factors will be as productive in contexts where the endangered language is spoken by a marginalised minority, or where there are lower rates of literacy, access to technology, or where there are significant differences in writing systems and language structure, particularly in the sound system – the phonology.

2 See in this book: Vassili Rivron, *The Use of Facebook by the Eton of Cameroon*.

3 See in this book: Viola Krebs & Vicent Climent-Ferrando, *Languages, Cyberspace, Migrations*.

Where this context is different, strategies for online language revitalisation may also look somewhat different. For example, in the case of some Native American communities, substantial language shift has taken place, but websites are being used as part of a broader strategy to pass the language on to younger generations (including, for example, summer language camps, where children are encouraged to use the language with each other). Examples are the Anishinaabemowin language⁴, and online lessons in Potawatomi⁵. Given that most members of these communities are dominant in English, the websites are also primarily presented in English, and are gateways into the languages they serve. This is in contrast with the case of Sardinian above, where the website does not need to use the dominant language. However the online strategy can still have a major impact, because of the high internet penetration within these communities. Galla [GALLA 2009] gives good coverage of issues related to technology for contexts such as these, particularly for the technologically advanced North American case.

Internet connectivity continues to increase around the world, though there are still significant issues of speed and ease of access especially in developing nations, and particularly in the rural contexts where the endangered languages tend to be spoken. An encouraging sign is the increased penetration of mobile phones into, for example, the Kenyan countryside, along with the reduction of the price of phones which are internet-capable. Because of this, internet access is no longer dependent on a constant electricity supply or broadband cables, as long as solutions are found to charge the phone, e.g. from a car battery, a bicycle, or a solar device. This means that language-revitalisation efforts can use this new technology, as has been demonstrated by K. David Harrison in his online dictionary of the Turkic language Tuvan⁶.

In Africa many of the most critically endangered languages are spoken by hunter-gatherers, a small minority of the overall population, who in many cases receive minimal respect from neighbouring peoples. For example in Kenya, according to Brenzinger [BREZZINGER 1992: 215] “*Hunters are looked upon as being ‘poor’, ‘primitive’, ‘living like animals’, etc. by the cattle herders*”. As such, when hunter-gatherer bands enter into symbiotic relationships

4 <http://anishinaabemdaa.com>

5 <http://www.potawatomilanguage.org/revitalisation.php>

6 <http://www.swarthmore.edu/SocSci/tuvan/dict> This dictionary is also available at no charge from the iTunes store as an application for the iPhone.

with neighbouring peoples, a pattern of abandoning their language and lifestyle can often ensue [DIMMENDAAL 1989]. However this lifestyle is one where levels of literacy in any language, and contact with the internet, are low. Hence internet-based assistance for language maintenance may be of little direct use to the community itself – in some ways the scenario is the opposite of that of Sardinian. However, cultural conditions permitting, posting video and audio material from such a language is still useful both to members of the communities as well as interested linguists and others, preserving aspects of the language for posterity, whatever use the community may wish to put it to in the future⁷.

We have presented three different types of scenario:

- That of Sardinian, a language with a large population of well-educated speakers, without significant difficulties in transferring writing skills from Italian, the dominant language, and good internet access. Here using Sardinian for a range of web-based activities presents relatively few challenges;
- The contexts of many Native American languages share some of these advantages, such as levels of education and access to the internet. But there is often quite a small pool of speakers, with the majority of the population having shifted to English, many of whom might like to learn the communal language. Also the level of linguistic difference between Native American languages and English is much greater than that between Italian and Sardinian. Hence the focus in these contexts is assistance with language learning, along with a strong element emphasising the communal culture;
- In the context of African hunter-gatherer groups, none of the above-mentioned factors favouring internet-based language activism pertain, and initiative may need to be taken by outside advocates in recording instances of language use, to the extent that this is deemed appropriate by the community itself. In such cases internet-based discussion groups and online pedagogical materials may be of limited value, until issues of both internet access and communal motivation to maintain the language are addressed. In some cases there may be no written form of the language, and while this in itself is not an indicator of language endangerment, developing a writing system alongside some written

7 See in this book: Tjeerd de Graaf, *How Oral Archives Benefit Endangered Languages*.

materials can help expand the domains of the language, increasing its prestige such that attitudinal motivations for shifting to another language are reduced⁸. The three scenarios presented are by no means exhaustive, but demonstrate that internet-based language interventions need to be sensitive to social and linguistic considerations if they are to stand a chance of battling language endangerment.

CHOOSING THE APPROPRIATE TYPE OF DOCUMENTATION

Given the threat of mass language extinction that faces us, preserving what will or might be lost is a high priority for linguists and those interested in preservation of the associated cultures. While it is evident that professional linguists have a role in this, there is much that other interested parties are able to contribute in this area. Preservation can take many forms including the linguist's grammar and dictionary, but less academic-oriented efforts such as audio and video recordings of the language being used, or collections of stories or accounts of local knowledge, are at least as valuable. In addition, a full dictionary or grammar may be a logistical challenge in some cases, but projects with more limited scope (e.g. a smaller lexicon or glossary covering particular aspects of the language, or a brief introduction to the grammar) may be easier to achieve, and have the added advantage of being accessible to a wider audience. Internet tools such as *WeSay*⁹ exist to help "*non-linguists build a dictionary in their own language*". Making such efforts available over the internet also reduces the costs and logistical challenges involved with book production and distribution. However in many contexts with limited access to modern technology, printed books may be the optimal solution – any intervention needs to consider the use of appropriate technology, as well as social values related to language, literacy and appropriateness of materials.

For example, in the case of *Munichi*, a now extinct language of Peru with a handful of semi-speakers remaining, the existing linguistic documentation [GIBSON 1996] has been placed on the web free of charge by SIL International, and so is accessible to the community. However such studies, destined for consumption by linguists rather than by members of

⁸ See in this book: Evgeny Kuzmin, *Linguistic Policies to Counter Languages Marginalization*.

⁹ <http://www.wesay.org>

the community itself, can be difficult for those without linguistic training to use; the vocabulary items and examples will be useful, but in the main, linguistic terminology is designed for precision rather than perspicacity to those not trained in linguistics. Fortunately for the people of Munichis, an additional research project has been initiated [MICHAEL 2009] with the goal of making material available for the community, including audio samples. Such materials destined for the community via the internet (or other media) should ideally be presented differently from those whose audience is primarily the academic community – though adaptation from one form to another is a possibility in most cases. Primary material (recording or transcriptions of stories, rituals, conversations, etc.) is useful for both members of the community, and for those who want to do further work in language analysis and development, and should not be supplanted by grammatical treatments. After all, recordings of spoken language are the linguist’s primary data.

In conclusion, it is certainly true that there are many opportunities for people to use the global *lingua franca* of English on the internet, with some even describing the internet as a force for linguistic uniformity rather than diversity. Nevertheless, the internet can and does facilitate the documentation and use of endangered languages, taking some of the space that may seem to be an aspect of globalisation for the continuance of linguistic diversity. In addition, it is a suitable repository for language materials in whatever form, whether the purpose is preservation, encouragement of use, or pedagogy etc. As we have seen, the manner in which the internet best helps with this task of maintaining diversity is variable, depending on various social, technological and linguistic factors, though these are dynamic, being subject to changing technological capacity and communal attitudes. Therefore we do not advocate a ‘one-size-fits-all’ approach for using the internet as a resource in combating language extinction and endangerment, but encourage interventions which best suit the existing language ecology, in the hope of maintaining diversity wherever it is found.

BIBLIOGRAPHY

[BRENZINGER 1992] Brenzinger, Matthias (1992). Lexical retention in language shift: Yaaku/Mukogodo-Maasai and Elmolo/Elmolo-Samburu. In: Brenzinger, Matthias (ed.) *Language Death: Factual and theoretical explorations with special reference to East Africa*. Berlin: Mouton de Gruyter. 213-254.

[BRENZINGER 2003] Brenzinger, Matthias, A. Yamamoto, N. Aikawa, D. Koundioubu, A. Minasyan, A. Dwyer, C. Grinevald, M. Krauss, O. Miyaoka, O. Sakiyama, R. Smeets, & O. Zepeda. (2003). *Language Vitality and Endangerment*. Paris: Unesco Ad Hoc Expert Group Meeting on Endangered Languages. <http://www.unesco.org/culture/en/endangeredlanguages>

[COURLANDER 1996] Courlander, Harold. (1996). *A Treasury of Afro-American Folklore: The Oral Literature, Traditions, Recollections, Legends, Tales, Songs, Religious Beliefs, Customs, and Humor of People of African Descent in the Americas*. New York: Marlowe & Company.

[DIMMENDAAL 1989] Dimmendaal, Gerrit. (1989). On Language Death in Eastern Africa, In: Dorian, Nancy C. (ed.). *Investigating obsolescence: Studies in language contraction and death* (Studies in the Social and Cultural Foundations of Language, 7). Cambridge: Cambridge University Press. 13-31.

[FISHMAN 1991] Fishman, Joshua. A. (1991). *Reversing language shift*. Clevedon: Multilingual Matters.

[GALLA 2009] Galla, Candace K. (2009). Indigenous Language Revitalisation and Technology: From Traditional to Contemporary Domains. In: Reyhner, John and Louise Lockard (eds.). *Indigenous Language Revitalisation: Encouragement, Guidance & Lessons Learned*. Flagstaff, AZ: Northern Arizona University. 167-182.

[GIBSON 1996] Gibson, Michael (1996). *El munichi, un idioma que se extingue*. Serie Lingüística Peruana N°42. Pucallpa: Instituto Lingüístico de Verano. [translated by Marlene Ballena Dávila]. <http://www.sil.org/americas/peru/pubs/slp42.pdf>

[HALLET 2007] Hallett, Darcy, Michael J. Chandler and Christopher E. Lalonde. (2007). *Aboriginal language knowledge and youth suicide*. *Cognitive Development*, 22: 3, July-September 2007, 392-399.

[HARRISON 2007] Harrison, K. David. (2007). *When Languages Die: The Extinction of the World's Languages and the Erosion of Human Knowledge*. New York: Oxford University Press.

[LEWIS 2009] Lewis, M. Paul (ed.) (2009). *Ethnologue: Languages of the World, 16th edition*. Dallas: SIL International. <http://www.ethnologue.com>

[LEWIS 2010] Lewis, M. Paul and Gary Simons. (2010). Assessing Endangerment: Expanding Fishman's GIDS. *Revue Roumaine de Linguistique LV:2, Special issue on Language Endangerment and Language Death*. 103-120. http://www.lingv.ro/resources/scm_images/RRL-02-2010-Lewis.pdf

[MENSCHING 2000] Mensching, Guido. (2000). *The internet as a rescue tool of endangered languages: Sardinian*. <http://www.gaia.es/multilinguae/pdf/Guido.PDF>

[MICHAEL 2009] Michael, Lev (2009) *National Science Foundation Award Abstract #0941205RAPID: Munique Rapid Documentation Project*. <http://www.nsf.gov/award-search/showAward.do?AwardNumber=0941205>

[MOSELEY 2010] Moseley, Christopher (ed.). (2010). *Atlas of the World's Languages in Danger*, 3rd edn. Paris: Unesco Publishing. <http://www.unesco.org/culture/en/endangeredlanguages/atlas>

[NICOLLE 2012] Nicolle, Steve. *A Grammar of Digo: A Bantu language of Kenya and Tanzania*. Dallas: SIL International and The University of Texas at Arlington.

CYBERSPACE AND MOTHER TONGUE EDUCATION

After having redefined what is a mother tongue, we will focus in a second phase on evaluating the use of the mother tongue as a medium of instruction, and to describe such teaching. Then we will focus in a second phase on evaluating the use of the mother tongue as a medium of instruction, and on describing such teaching.

Original article in French.

Translated by Laura Kraftowitz.



MARCEL DIKI-KIDIRI, Central African Republic, is now Consultant in Applied Linguistics. Before he retired in 2010, he was senior researcher at the CNRS in the unit Language, Languages and Black African Cultures (LLACAN : CNRS, INALCO)

MARCEL DIKI-KIDIRI

EDUCATION
MOTHER TONGUE
CYBERSPACE AND

The question of Cyberspace and Mother Tongue Education can only be posed if the very sense of the term “mother tongue education” is clarified. In most countries, with the exception of those in a state of post-colonial dependency, or for minority and under served communities, education takes place in the language of the majority. In monolingual and weak multilingual communities, the majority language is the mother tongue of the majority of its speakers. In strong multilingual societies, the majority language, if it exists, is often a second, third or even fourth language for most of its speakers. Nevertheless, the fact remains that in all cases, a child learns best, that is, more content more rapidly, by receiving instruction in the language s/he uses most frequently, which is often considered the mother tongue.

WHAT IS A “MOTHER TONGUE” ?

The term itself initially came from the belief that all children learned their first words in the lap of their mothers. This simplification is often far from reality. In many patrilineal societies¹, it is the father’s language that is taught to the child, whether or not the mother does the teaching. In this case, we could speak of a “father tongue”, to better illustrate this socio-cultural reality. In strong multilingual societies, it is not uncommon for a child to learn several languages simultaneously beginning at infancy. We use the term “individual multilingualism” to describe a person who speaks several languages, and the term “multilingual society” when multiple languages are spoken by a community living in the same area. A multilingual society is in turn characterized as either diglot or polyglot,

1 Which is based solely on the paternal ancestry in terms of parentage, family organization and social group, of a clan (Trésor de la Langue Française <http://atilf.atilf.fr>)

depending on whether two or more languages are spoken by at least a majority, if not the entirety, of the population, with specialized functions for each language. But a multilingual society can also be composed of several local and essentially monolingual populations, each with its own mother tongue. Finally, where a common language is in the majority, it quickly becomes the first language of younger generations. Thus, the definition of mother tongue as “the first language a child learns” conflicts with its qualifier, “mother”. Furthermore, education specialists prefer to speak of a “first language” as a more accurate, precise and neutral term in relation to the conditions of language learning. Throughout the remainder of this article, the term “mother tongue” should thus be understood as the equivalent of “first language”.

WHAT IS “MOTHER TONGUE EDUCATION” ?

We use the word education to mean training and instruction given to children, adolescents and young adults, to prepare them to fully assume their place and responsibilities in society as adults and citizens. Educational systems vary between countries and human communities, based on core values, as well as essential knowledge and skills that society chooses to transmit to the younger generation. In the case of adults who have already exited the primary and secondary educational system, but who find themselves in a situation of learning, we speak of training rather than education. The central question that interests us here is what language(s) should be used to teach the various disciplines that are programmed into an educational or vocational training system?

IN A MULTILINGUAL CONTEXT, WHAT LANGUAGE SHOULD BE USED FOR EDUCATION ?

One would expect that any teaching or training would be conducted in the language best understood by the learner, assumedly that person’s first language and mother tongue. While this is the case in independent countries where institutions function in the local majority language, in many strong multilingual countries, the language of institutions doesn’t always correspond to the mother tongue of the local majority. Child learners are taught in a language of instruction they don’t know or haven’t mastered, which demands considerable and disproportionate effort to reach each

level of knowledge in the various disciplines covered. Does this mean that strong multilingual countries must give up the dream of mother tongue education and training? Thanks to extensive research in the educational sciences, along with experiments carried out in several strong multilingual countries, it has been established that it is possible, even in such a country, to develop diverse educational systems that judiciously and positively use multilingualism to provide populations with access to knowledge in each of their languages, while at the same time improving the ability of learners to speak several languages, by teaching those languages. In many cases, this requires profound educational reform and long-term effort, but is ultimately less expensive and more profitable than maintaining the *status quo* of the existing inadequate systems.

Educational reform targeting the integration of mother tongues as the languages of instruction and as subjects taught not only changes curricula, programmatic organization, term lengths, and so on, but also teaching and learning methods that can integrate new pedagogical tools, such as ICTs, and lead to new ways of doing things, or even new behaviours. In addition, the use of a given language as the language of instruction assumes that it is adequately equipped with the technical terminology of the discipline in question to fully convey the specialized knowledge of that discipline. The systematic development of specialized vocabularies belongs to the field of terminology as a specialized branch of linguistics and requires an organized and methodical implementation by the necessary public institutions (academies, offices, high commissioners, delegations, institutes, etc.). Mother tongues that are thus equipped must then be taught in formal curricula and written into the national qualification exams in order to become fully attractive. As with all modern life's fields of activity, ICTs are ubiquitous within the field of education, notably in language teaching, for which ICTs have totally revolutionized teaching methods, particularly in the field of distance education. However, the installation of ICTs requires heavy infrastructure that is not always available in certain countries, particularly in rural and remote areas.

WHAT INFRASTRUCTURE IS REQUIRED FOR ICT USE IN MOTHER TONGUE EDUCATION ?

Two types of infrastructure should be considered prerequisites: that of the educational system itself, and that of communication networks. After that, special equipment is needed to use ICTs in mother tongue education.

Educational Infrastructure

The density and distribution of educational and training institutions (all schools from kindergarten to higher education, vocational training centres, centres for training, etc.) ; and the number of learners (pupils, students, apprentices, trainees, etc.) by class and institution, are essential parameters that should guide the choice of technological solutions promoting ICT access to the greatest number possible. Since in general, cities have a higher population concentration than rural areas, they also have better equipped educational facilities. On the other hand, rural areas, even when highly populated, are much more spread out, which leads to lower infrastructure density. As a result, rural youth have a much less easy time than their urban counterparts to even get to school, let alone get to a computer. The divide observed in rich countries is even more significant in poor countries, where the number of students in a class is frequently extremely bloated.

Communication Infrastructure

The oldest communication network still in use is that of the landline, which is giving way to the radio waves of wireless networks. These require the erection of many antennae just to cover a city, let alone countryside or an entire continent. Aerial solutions (geostationary, ultralight gliders) aren't any cheaper, but are less suited to applications requiring fast action in real-time (video games, remote surgery). However, they can cover large areas with no on-ground apparatus aside from wireless receiver-transmitters. Finally, fiber-optic technology allowing the transfer of large amounts of data at a very high speed is fast approaching: as usual, first in large cities in developed countries before later reaching the rural areas; and still later, cities in developing countries.

Special Equipment

Just as a classroom should be equipped with desks or the equivalent, it is necessary to equip any institution wishing to integrate ICTs into its teaching methods and school management with the appropriate tools [BASQUE 1998]. Technological products are extremely varied and can match any budget. The choice of equipment also depends on the desired objectives and modes of operation. The computer is an essential basic tool, regardless of format (desktop, laptop, tablet). No institution, however, can afford to provide each learner with their own computer as a work tool. The development of a computer room, where a set number of computers are locally connected, is a much more manageable solution. In a university, where the number of students is usually much greater than in a primary or secondary school, it may be useful to have several computer rooms together on a digital campus. Specialized software permits the management of all institutional activities, including course management, grades, exams, and the flow of information between teachers, students, and administrators [PELGRUM 2004]. This last point requires the establishment of an intranet email system, in which each member of the institution has a personal account. Finally, document management is one of the ICT's main areas of expertise. Not only does it facilitate the complete management of a local library, but also all the relations between a library and other national and international holdings, allowing to locate and access their available resources.

USING ICTS TO TEACH A MOTHER TONGUE

When the above infrastructure and equipment conditions are met, the use of ICTs for language teaching, especially of the learner's mother tongue, varies greatly depending on the pedagogical approach taken; and the educational sciences have developed many, each with its own advantages and shortcomings [BASQUE 2002]. This is also one of the main reasons they are constantly replaced. Whatever the theories, none of them considers ICTs to be more than a learning aid [DEMAIZIÈRE 2007]. The whole question is what is meant by an "aid". The idea itself of what constitutes a learning process conditions what we consider an aid. When we consider, for example, that to attain an acceptable level of proficiency in a given language, one must successfully pass a number of tests at several calibrated levels of difficulty,

it's not hard to imagine that the use of these indications would structure the learner's advancement. It would help her or him arrive rapidly at the correct answer for each test. The computer is in that context used as a tutor and assessor [TAYLOR 1880]. This is consistent with a conception of learning that sees the learner as an empty brain to fill with new information, which, once assimilated, constitutes the knowledge gained. From this perspective, we create tutorials to lead a "failure" step by step to success. But today, in light of ICT advances, together with those of the educational sciences, pedagogy emphasizes the learner-centred approach [DE VRIES 2001], [ANGRIST 2002]. The learner is no longer a passive recipient, but an actor on a learning path, whose choices and actions lead and build according to her or his progression. Also, the computer becomes a resource-rich tool of production [FORCIER 1999]. The student still has goals to achieve, but they are achieved intuitively, by trying out various tools and taking different paths according to temperament, prior knowledge, psychological state, relationship to work and to others, and so on. The others are simply more present, because the computer no longer being a resource provider, the bulk of the "aid" becomes human again. Indeed, peer support via collaborative work is made possible by ICT networks and distance learning [WILEY 2002] and also through the "human resources", such as coaches and experts [CAZADE 1999], guest lecturers, and so on. The teaching profession is fragmenting into several specific activities that each requires the intervention of specific specialized persons. The learner is invited to "play" with this panoply of technological tools to create and produce using the imagination [JONASSEN 2000], [LEBRUN 2002], and moves through the appropriation of the original focus of studies, that of the mother tongue, within this act of production and creation.

SOME ILLUSTRATIVE EXAMPLES

In Canada, as in most industrialized countries, governments have invested significant resources into networking and equipping primary and secondary schools, as well as institutions of higher learning. The initial teacher training has been adjusted to optimize the capacity of young teachers to take full advantage of these new technologies. However, many studies show that teachers make only marginal use of ICTs, restricting them to certain types of classroom activities. In many cases, activities are experimental and the results are watered down and overly general. According

to [LAROSE 2010], the United States has the best documented and most convincing experience:

A single methodologically rigorous study with a large sample finds a stable effect of ICT use on building writing skills in the mother tongue. This evaluative study was undertaken about the Maine Middle School Laptop Program², where the performance of high school students on standardized tests after two years in the program, both for the organization of ideas and for grammar and syntax, were significantly higher compared to their baseline levels.

However, caution [LAROSE 2010], the conditions for obtaining such results are particularly demanding. In all cases where there was a significant improvement in the construction of competence in language skills, both oral and written, with the result being attributable to the use of ICTs, the following four conditions were observed:

- Students had stable individual access and used networked computers, both at school for all periods of language instruction, and at home;
- Their teachers received a high level of support, both with technology and with managing an approach to the integration of ICTs;
- The school curriculum was also adapted to the availability of personal computers and special educational devices;
- Exhaustive information databases of teaching-learning situations (TLS), adapted to widespread computer use, were developed and updated regularly.

An example of these databases is the digitized *Trésor de la Langue française* (“*Treasury of the French Language*” or TLF), which can be used to teach French:

The digitized Trésor de la Langue française allows students to discover the historical facets of our language. The digital version has 100,000 words with their history, 270,000 words with their definitions, 430,000 examples and 500,000 citations. The search engine offers three levels of search: simple, advanced and complex. All the words displayed in a TLF article can lead to a hypernavigation, providing access to diverse resources: The digital Trésor de la Langue française, the Dictionnaires de l’Académie (‘Dictionaries of the Academy’)

2 Silvenrail and Gritter, 2007 quoted by Larose et al., 2010.

(8th and 9th editions), the *Altif Lexical Knowledge Base*, the *Frantext Base*, and the *Historical Database of French Vocabulary*)³.

To conclude, [LAROSE 2010] state the following :

More generally, our research, together with the available literature, suggests that the adoption of ICTs to teach mother tongues or second languages, just as their integration into the teaching practice of all scholastic disciplines, depends on the strategies of educational intervention that teachers adopt. In a relatively traditional and frontal view of education that leaves little room for student initiative in terms of research, analysis and integration of information, ICT use remains marginal. On the contrary, in the perspective of an education that places a greater emphasis on research and integration of information, the role of ICTs as tools to support the students' construction of knowledge is more evident.

SOME INDICATORS FOR DECISION AND ACTION

In conclusion, a few basic principles should be considered in the decisions and choices made to ensure optimal use of ICTs in mother tongue education and teaching :

- The mother tongue is the first language of the child, and it must serve as the first language of knowledge acquisition. When controlled, it can facilitate the learning of other languages, such as a majority language when it differs ;
- The use of a mother tongue as language of instruction requires, firstly, the development of specialized vocabularies in that language for each discipline, and secondly, the teaching of that language ;
- When education is not undertaken in the mother tongue, the integration of mother tongue(s) in education requires a (rather deep) reform of the system at several levels : teacher training, the valuing of mother tongues by teachers, and their registration in official exams ;
- ICTs provide many forms of assistance to learning. Their use requires making carefully considered choices between different pedagogical

3 Retrieved from http://www.cafepedagogique.net/lesdossiers/Pages/2010/indis2011_francais.aspx accessed on 8 Apr 2011. My translation.

approaches, because they are decisions on which an educational system's selection and development of methods and tools depend;

- For optimal use of ICTs in education, especially mother tongue education, students must have access to a personal computer that is networked throughout the school year; teacher training and curricula should be adjusted to optimize the use of technological resources; and various databases in and on the mother tongue should be developed, or at least made available to students via intranet or internet;
- One should always keep in mind that as computer hardware and software evolves and diversifies, it increasingly opens up new pedagogical possibilities for teaching professionals. Distance learning, already quite popular with businesses, is now on the rise in some schools, which have begun to use podcasting (via digital media players and mobile phones) and video conferencing (via webcam). Teachers should strive to integrate these tools, by now commonplace for their students, into their classes.

BIBLIOGRAPHY

[ANGRIST 2002] Angrist, J. ; Lavy, V. 2002 « New evidence on classroom computers and pupils learning ». *Economic Journal*, vol. 112, n°482, p. 735-765.

[BASQUE 1998] Basque, J., Rocheleau, J., Winer, L., Michaud, P, Bergeron, G., Paquette, G., Paquin, C. *Un modèle adaptable d'une école informatisée*, Montréal, École informatisée clés en main du Québec inc., 1998.

[BASQUE 2002] Basque, J., Lundgren-Cayrol K. 2002 « Une typologie des typologies des applications des TIC en éducation » *Sciences et techniques éducatives*, n°9 (3-4), p. 263-289.

[BAUMGARTNER 1998] Baumgartner, P, Payr, S. « Learning with the Internet: A typology of application », *Proceedings of ED-MEDIA/ED-TELECOM 98 (World Conference on Educational Multimedia and Hypermedia & World Conference on Educational Telecommunications)*, Charlottesville, AACE, 1998, p. 124-129.

[BONNET 2009] Bonnet, Annick (Coord.) 2009. Elisabeth Brodin, Micheline Maurice, Chirine Anvar, Pernelle Benoit, Séverine Blache, Fiorella Casciato, Concetta Cirocco, Catherine Clément, Stéphanie Favre, Olivier Gisselbrecht, Haydée Maga, Marianne Mavel, Olivier Steffen, Dominique Satgé, Nicole Thierry-Chastel. *Guide SAEL, guide pratique pour la conception, l'animation et l'amélioration des sites d'accompagnement pour les enseignants de langues*. http://www.eurosael.eu/sites/default/files/3/guide_sael_2009_0.pdf

[CAZADE 1999] Cazade, A. « De l'usage des courbes sonores et autres supports graphiques pour aider l'apprenant en langues ». *Apprentissage des langues et systèmes*

d'information et de communication (Alsic). Vol. 2, n°2, décembre 1999, pp 3 - 32.
<http://alsic.revues.org>

[CRABÈRE 2010] Crabère, Béatrice. *Interview. Comment enseigner une langue difficile et minoritaire?* http://www.cafepedagogique.net/lexpresso/Pages/2010/02/Fourgous_BCrabere.aspx

[DEMAIZIÈRE 2007] Demaizière, Françoise. « Didactique des langues et TIC : les aides à l'apprentissage », *Alsic*, Vol. 10, n°1 | 2007. <http://alsic.revues.org/index220.html>

[DE VRIES 2001] de Vries, E., « Les logiciels d'apprentissage : panoplie ou éventail? », *Revue Française de Pédagogie*, n°137, octobre-décembre 2001, p. 105-116.

[FORCIER 1999] Forcier, R. C., *The computer as an educational tool: Productivity and problem solving*, Prentice-Hall, 1999, 2^e édition, 1999.

[JONASSEN 2000] Jonassen, D. H., *Computers as mindtools for schools: Engaging critical thinking*, Prentice Hall, 2^e édition, 2000.

[LAROSE 2010] LAROSE François, GRENON Vincent, CARIGNAN Isabelle et HAMMAMI Abdelhakim, 2010 « Les TIC en enseignement des langues au Québec : objet obscur d'un désir prescrit? » *Québec français*, n°159, 2010, p. 71-72. <http://id.erudit.org/iderudit/61597ac>

[LEBRUN 2002] Lebrun, M., *Des technologies pour enseigner et apprendre*, De Bœck, 2^e édition, 2002.

[MUNN 2011] Munn, Yves 2011. *Outils TIC en langues (ESL)* <http://www.reptic.qc.ca/bibliotheque/enquetes-inventaires-compilations/outils-tic-langue-esl.html>

[PELGRUM 2004] PELGRUM, W.J., LAW, N. 2004 *Les TIC et l'éducation dans le monde : tendances, enjeux et perspectives*. Unesco.

[SILVENRAIL 2007] Silvenrail David L. et Gritter Aaron K. 2007 *Maine's middle school laptop program: creating better writers*, Gorham, ME, University of Southern Maine, Maine Education Policy Research Institute.

[WILEY 2001] Wiley, D. A. « Connecting learning objects to instructional design theory: A definition, a metaphor, and a taxonomy », In D.A. Wiley (éd.), *The instructional use of learning objects*, Bloomington, Indiana, Association for Educational Communications of Technology, 2001.

[WILEY 2002] Wiley, D. A., Edwards, E. K. « Online self-organizing social systems: The decentralized future of online learning », *Quarterly Review of Distance Education*, <http://www.opencontent.org/docs/ososs.pdf>



DIGITAL SPACES

PART 2

Cyberspace is like so many invisible continents where billions of people develop or extend conversations, relationships, networks, creations, translations, knowledge, social links. What linguistic bridges exist to facilitate these human links within and beyond the technical constraints and cultural barriers?

MULTILINGUALISM AND THE INTERNET'S STANDARDISATION

The Internet exists in compliance with standards, protocols, formats and other collective rules necessary to interconnect and exchange data. Can these standards become a constraint limiting the range of possibilities? What are the standards that manage languages on the internet? Who defines them, what organizations elaborate them? Are they well adapted to multilingualism or conversely are they lagging behind? Are they a positive factor for languages?

Original article in French.
Translated by John Rosbottom.



STÉPHANE BORTZMEYER is a computer engineer, particularly specializing in TCP/IP networks. He works for AFNIC and maintains a blog where he talks from his own personal view mainly about technology but sometimes also of culture or politics. <http://www.bortzmeier.org>

STÉPHANE BORTZMEYER

MULTILINGUALISM
AND THE
INTERNET'S
STANDARDISATION

I f I write in French on the internet, for example the word “café”, will my readers still see the correct word, or “cafi”, or “caf=E9”, or “cafÃ©”? This phenomenon, known in Japanese as *mojibake*, was common fifteen years ago, but it has been considerably reduced by the progress of standardisation.

This article discusses the various practical standards that define the protocols, formats and other rules to be followed by software found on the internet. As in other sectors, the standards are both a benefit (without them, no internet, because no communication: imagine a Web where websites are visible by Firefox or Internet Explorer but not by both), and a constraint because they can limit what is possible. What are the standards that apply to multilingualism today? Who defines them, and what *SDO* (*Standards Development Organization*) elaborates them? Are they well adapted to multilingualism or conversely are they lagging behind? Are they a positive factor for languages or rather one of the reasons for the problems of multilingualism?

Note that this article is about *standards*, and not their implementation issues. A standard isn't everything; if it isn't applied in a programme, or if nobody is using it, it has no utility; it becomes a “forgotten standard”. We will see that, once a standard is complete, there is still work for programmers (to implement it), for system and network administrators (to deploy it) and for users (to use it).

Let us briefly review the treatment of multilingualism in computers. The internet is based traditionally on the written word, and we must distinguish between language and alphabets. Scripts¹ can serve many

1 As defined by the IETF, RFC 6365, *Terminology Used in Internationalization in the IETF*, a script is a set of graphic characters used for the written form of one or more languages. <ISO/IEC10646> Examples of scripts are Latin, Cyrillic, Greek, Arabic, and Han

languages (for example the case of the Latin alphabet) and a language can use multiple alphabets (for example in Azerbaijani² or Tamashek³). It is therefore important to distinguish when we need to handle multiple languages, and when to handle multiple alphabets.

On the other hand, human languages have not been scientifically designed: they are weird, full of incomprehensible rules and many idiosyncrasies. If we could redo languages from scratch, with the sole aim of facilitating their computerisation, we could greatly simplify the problems of multilingualism. But this is evidently not the policy of standards bodies, who all consider the current state of languages and alphabets as a given: it is impossible to envisage changing these. So when some people say that the Unicode standard, for example, is complex, they are missing the point. It is the alphabets, these human creations, that are complex and Unicode only reflects the real world⁴.

Another obstacle to multilingualism: the lack of scientific knowledge about certain languages. The digital divide is evident here, some languages are not represented on the internet because they have not yet undergone rigorous modelling.

Finally, an additional difficulty comes from the sensitivity associated with language and alphabets. Any technical differences are quickly perceived as an affront to the sensibilities of a nation or a people, which often makes discussions rather less calm.

Another point that may be necessary to take into consideration: the functioning of the internet itself. It is important to understand that the internet has no central leadership that could give instructions such as “On January 15, 2011, all email software must accept email addresses in Unicode”. On the contrary, the deployment of any technology depends on the decisions of a number of players (sometimes a great number) and it is necessary therefore to persuade them to agree. As a result, and because of the huge investments that were made during the last thirty years, the weight, or significance, of history is high: you cannot throw

(the characters, often called ideographs after a subset of them, used in writing Chinese, Japanese, and Korean). RFC 2277 discusses scripts in detail.

2 Written today with the Latin alphabet or Arabic, and has long been Cyrillic.

3 Written in the Latin alphabet but especially Tifinagh.

4 The complexity –very real– of Unicode is due also to technical constraints, such as the need for compatibility with pre-existing encodings.

away everything to try to remake the internet “better”, any more than you can erase a town to improve urban planning.

STANDARDS THAT MAY INFLUENCE MULTILINGUALISM

In what ways can a standard, a document in general rather technical and dry, help or hinder multilingualism? Consider a few examples of the importance of standards. One of the best known cases is that of *character sets*. To write a human language in a computer we need to represent each character in the text by a number, the only objects that programs can handle. We must therefore establish a list of characters (this is already a formalisation step that is not obvious, and is not yet done for all alphabets) and we assign each a number. If two people who write in Tamil use two different character sets⁵, their messages will be incomprehensible to one another. A standard character set is needed. But even then it has to cover all the alphabets of the world! One of the first such standards, ASCII (*American Standard Code for Information Interchange*), designed in 1969 for English usage, does not include characters using diacritical marks (like é or ç) or, *a fortiori*, the characters of the Arabic, Devanagari or Hangul alphabets. After ASCII, several incompatible standards⁶ were designed, none of them covering all alphabets of the world, until the advent of Unicode, presented later.

Once there is a standard for characters, it is still necessary for different formats to permit its use. Thus, in its infancy, e-mail on the Internet only allowed the use of ASCII. The users of another alphabet could either renounce certain characters, this was the case for those using the Latin alphabet⁷, or develop a system of transliteration of their alphabet into ASCII⁸. It took considerable discussion within IETF (*Internet Engineering Task Force*), the agency that manages standards for the Internet, to establish a standards document (an RFC: *Request For Comments*) to expand the

5 Supposing that one is written in Unicode and the other in ISCII.

6 And covering only the “large” alphabets. The “small” ones are those that benefit most from Unicode.

7 For example, in French, you can substitute composed characters by their ASCII equivalent and the text remains relatively readable in most cases.

8 Such as Japanese romaji, which in practice has hardly been successful.

range of character sets. Other character sets were accepted into e-mail in 1992 with the release of RFC 1341.

Another case where standards and multilingualism have stirred a hornet's nest: *identifiers*. All internet protocols define legal identifiers as well as non-legal ones. As these identifiers are very visible (for example on advertisements, business cards, etc.), embarrassment can be considerable if they are misused. Thus, in mail addresses, such as `stephane@bortz-meyer.org` only ASCII was permitted until the release of RFC 4952. Now (but beware, this RFC is still considered experimental; only its successor, under development, will be a real standard), one can have addresses like `st ephane@bortzmeyer.org`.

Another case that has generated much discussion⁹ are the Unicode domain names, IDN (*International Domain Name*). For various reasons¹⁰, the names were traditionally restricted to US-ASCII only. The IDN standard of 2003, in RFC 3490, marked the beginning of names in Unicode, which have become commonplace today. Unlike other standards for internationalisation, they have been implemented very quickly in software and deployed in several registries of domain names.

A FIRST EXAMPLE OF A STANDARD : UNICODE

Historically, all standard character sets were limited to one writing system, an alphabet, or a set of similar alphabets¹¹. One of the consequences of this Babel of character sets was that it was very difficult to write a text with multiple alphabets (a course in Hindi written in Spanish, for example). Some character sets included a way to include ASCII but not, in general, characters of the Latin alphabet outside the ASCII set, not to mention other (non-Latin) alphabets. On the other hand, managing a collection of texts written in different alphabets (although each text has only one alphabet) was difficult. For example, for a web server, it would not be possible without Unicode, to configure a global parameter `Charset`, to indicate the character encoding for the entire site, even multilingual.

⁹ And not necessarily in a justified way.

¹⁰ Among which does not feature a "DNS limit". This last has been accepting all characters from the beginning. But it is used in many contexts, some presenting a problem with names in Unicode.

¹¹ For example, ISCII covers all the alphabets that have official status in India and that are derived from the Br hm  script.

Unicode has changed all that: Unicode¹² is a character set that includes all the alphabets of the world¹³. What is the content of the Unicode standard? Firstly, Unicode is a list of characters. Compiling such a list doesn't seem like much, but in fact is a difficult task. Where some alphabets are highly standard it is sufficient to reuse this standard. In other cases, there is no official list. The authors of the standard must therefore ask themselves questions such as “Does the German capital letter ß exist?”¹⁴ or “Are the Japanese and Chinese characters the same?”¹⁵. Once we have answered these questions, we may publish the list¹⁶. It currently consists of 109,449 characters, from the most mundane such as the Latin “a” to the most astonishing such as the “*sunrise over the mountains*”¹⁷.

Once the list is established, Unicode gives each character a unique number, which facilitates communication: when two characters resemble each other visually, or when necessary fonts are not installed, this number enables an exchange without ambiguity. To give an example, the two characters cited above, “a” and “*sunrise over the mountains*” are respectively numbers U+0061 and U+1F304¹⁸.

These numbers also serve as the basis for the subsequent encoding of these characters. In effect, we have to represent these characters as a sequence of bits in files or on a network. There are several methods for doing this, known by names such as UTF-8, UTF-32, etc., which all start from the number used, to represent it in a manner appropriate to certain uses. In practice, this last point only concerns computer programmers.

Just as technical, but perhaps more necessary to understand, are the concepts of canonicalisation: there are several ways to represent the same visual character in Unicode. For example, the é in my name can be represented by U+00E9 (e with acute accent) or with U+0065 U+0301 (e


12 <http://www.unicode.org>

13 In November 2010, the current version of Unicode is 6.0; some alphabets are still missing but these are almost all dead alphabets, the lack of which affects only researchers.

14 No up to Unicode 5.2, yes afterwards.

15 The answer was yes, this is called the “Han unification” and was certainly the most controversial decision of Unicode.

16 One of the important points in Unicode is that not only the text of the standard, but also the data – such as lists of characters – are publically distributed.

17 If you have the right configuration, you'll see it here: 
<http://www.fileformat.info/info/unicode/char/1f304/index.htm>

18 The numbers are conventionally preceded by U+ and written in hexadecimal.

followed by a combining acute accent). Current operations in computing such as comparison (imagine a user whose login name was the first name Stéphane ...) would fail if they were applied to Unicode characters naively. It is therefore necessary to canonicalise the strings of characters, reducing them to a canonical form. The most common standard for canonicalisation, on the internet (see RFC 5198) is known as NFC, and in the case presented above, would reduce all the é to the form U+00E9.

So, who writes and maintains this standard? The Unicode Consortium is a coalition of several organizations, including major companies in computing (Google, Apple, IBM, Microsoft, etc.). Recently, some nonprofit organizations have begun to address this problem and have joined the consortium. There is a very interesting list for public discussion, `Unicode@Unicode.org`, but most of the work is done in private, only the results are public.

For those who want to deepen their understanding of Unicode, I recommend *Unicode explained* by Jukka Korpela (O'Reilly editor) for the authors of documents and *Unicode demystified* by Richard Gillam (Addison-Wesley editor) for the programmers [ANDRIES 2008].

AN EXAMPLE OF SDO: IETF

Let's take a detour through a SDO (*Standards Developing Organization*), a particularly important one, the IETF (*The Internet Engineering Task Force*). This organization is, among others, responsible for e-mail standards, for the instant messaging protocol XMPP (eXtensible Messaging and Presence Protocol), the HTTP protocol, the DNS protocol, etc. One of the peculiarities of the IETF is its extreme openness: there is no formal membership, so no fee; anyone, whether individual or company, can participate. If a member cannot travel to physical meetings (which are, themselves, quite expensive), it is not a big deal; some IETF working groups have never met face to face¹⁹. Even if the group meets physically, most of the work is done online, via public mailing lists (and publicly archived), and working papers also public. Wikileaks would not have much to do to ensure the transparency of the IETF :-)

19 Such as the working group LTRU, who created the language tags described below.

What is the IETF policy on multilingualism? This is explained in RFC 2277. In two steps, the IETF separates the protocol elements, internal to the operation of the protocol, from the text that is shown to users. The former are visible only to programmers. Thus, a web browser requesting the resource `/faq.html` to a server sends the command `GET /faq.html`. The verb `GET` is indeed derived from the English but it is not really an English word, rather an element of the HTTP dialog. The ordinary user never sees it and so there is no reason to translate it. On the other hand, the text of the web page will be retrieved to be viewed by a user. Here, RFC 2277 establishes how in principle it must be able to be encoded in any character set and should certainly not be restricted to ASCII.

These are excellent principles, but obviously the reality is more complex. Two cases are not directly addressed by this RFC, one, very sensitive, is the case of *identifiers* (such as domain names or email addresses listed above) which are both protocol elements and text read by users. Much of the controversy around the IDN system (*Internationalised Domain Names*) for example comes from the clash of two points of view, those who see a domain name as a formal identifier, devoid of any semantics (and therefore can be written in a foreign alphabet to a user) and those who consider it an identity marker, which must be user-readable.

Note that the W3C, the organization responsible for the standardisation of web technologies, operates relatively closely to the IETF, and has a similar policy²⁰.

A SECOND EXAMPLE OF A STANDARD: EMAIL

Email is one of the less visible and yet most used applications on the internet. Despite some predictions concerning what might be diverted to instant messaging, or to communication tools controlled by closed services such as Facebook, millions of messages continue to be exchanged every day. How does email manage internationalisation? There are two separate problems, the content of messages and their addresses.

Previously, the only content that was accepted was plain text, exclusively in US-ASCII. That changed in 1992 with the publication of the MIME standard (*Multipurpose Internet Mail Extensions*). This allowed a message

²⁰ <http://www.w3.org/International/getting-started>

to contain instructions for formatting text, and also sound, images, etc. Another aspect of MIME was less noticed at the time: the character set of the text was no longer obliged to be US-ASCII; any character set was accepted, provided it was properly defined... and that the receiver software could use it. Since then we can say that email standards enable writing messages in any language, but it has taken long enough for all software authors to adapt to them. Here we see that setting standards is only the first pillar of a language policy in cyberspace. Incentives for users, or programmers, to apply them are also part of an informed decision.

Until very recently, there was a lack: email addresses themselves were not internationalised. There was no question of putting on a business card `stéphane@coopération.com`. This limitation shrank in 2008 with the experimental RFCs modeled on RFC 4952. Scarcely deployed at the moment, the possibility (of people using their own name in their own language and alphabet as mailbox names) should become more widespread with its forthcoming access as a standard. One can easily imagine the interest for writers using non-Latin alphabets, for which the transliteration of names would no longer be necessary.

A THIRD EXAMPLE OF A STANDARD: LANGUAGE TAGS

Much less well known, because less visible, than MIME or IDNS, language tags are short identifiers used to indicate the language of a document. They are essential for librarians and archivists, linguists who exchange their documents, but also for authors of web sites when they want to indicate the language of a document, for presentation purposes (typographical rules are not the same for all languages) or for research (to facilitate the work of a search engine when asked “only documents in Portuguese”). A format such as XML allows the language of a document to be specified, avoiding editing software having to guess it, a complex and not always safe operation. While, today, language tags are unfortunately little used on the Web (both by the authors of pages, and by search engines), they are significantly present in large documentary catalogues.

Standardised in RFC 5646, language tags²¹ can indicate not only the language but also the form of writing used, the national or regional

21 <http://www.langtag.net>

variations, etc. Thus, if the label `el` is simply modern Greek, without more explanation, the more complex label `yue-Latn-HK` indicates Cantonese used in Hong Kong, and written in the Latin alphabet.

The standardisation of such identifiers was not a path of roses. Everything related to languages is extremely sensitive, and, for instance, considering an idiom as a language or as a dialect is not neutral, and can lead to anger and misunderstanding. It is partly to limit this risk that language tags relate to, wherever possible, other standards such as ISO 639 for languages. RFC 5646 provides, in relation to these standards, the free availability of the standard, the possibility of combination (as in the example above) and stability (unlike ISO identifiers, a tag is still valid, even if the identifier is removed or reassigned by the ISO).

HISTORY

If the status of multilingualism on the internet today is quite good (almost perfect in terms of standardisation, less so for implementation and deployment), it was not always the case. All users of a certain age can remember when simply to send a message with a composite character required good computer skills and reading lots of documentation. The author remembers reading about the first MIME software²² and has fond memories of the long struggle of the 1990s enabling two French people to be able to send messages in correct French. Let us congratulate, a posteriori, therefore the GERET²³ group members, who have done such a great job of consciousness-raising and training.

This long struggle has left a legacy, particularly the persistent urban legend that “the internet does not support accents” which has led some French speakers to self-limit themselves, even today, into using only ASCII. Of course, their attitude is justified by the fact that we cannot, today, guarantee 100% success, but, unfortunately, this blocks progress: in 2011, users should no longer tolerate a system that prohibits the use of all characters in their language!

22 It was named metainmail and was a computer program that even the most dinosaur computer scientist of today would not like.

23 Groupe d'Exploitation des Réseaux Ethernet TCP/IP (Ethernet and TCP/IP Network Operations Group). It was a working group whose aim was to provide a forum for the exchange of experiences for engineers operating predominantly Ethernet and TCP/IP networks.

And what does the future hold? In early 2011, the standardisation work is 95 % finished²⁴ and the problem henceforth concerns above all programming, deployment and content. The concrete work awaits those who really want help multilingualism on the internet!

GLOSSARY

American Standard Code for Information Interchange (ASCII)

An old (but still widely used) character set standardised in the U.S.A. having only the characters needed for English. As it was one of the first, and as its birth was in a founding country of computing, it has long been used as the basis for many network protocols.

Domain Name System (DNS)

This term refers to both the system of domain names, the tree structure for creating identifiers such as `cooptel.qc.ca` or `vélipianchiste.com`, and the protocol enabling the retrieval of information such as IP address, the name of the mail server, etc. from such a name.

Internationalized Domain Names (IDN)

The term IDN designates domain names expressed in Unicode, such as for example, سنوت.تيرح²⁵. It sometimes uses the acronym IDNA (*Internationalised Domain Names in Applications*) for the specific technique, in current use, that goes through a local conversion to ASCII before sending to the DNS.

Internet Engineering Task Force (IETF)

The main standards organization for the Internet, notably charged with layers 3 (routing) to 7 (applications). It is distinguished by its great openness, its debates and its standards (the famous RFC) being public. <http://www.ietf.org>

Indian Script Code for Information Interchange (ISCII)

An old (but still widely used) character set standardised in India that covers much of official paperwork in India (a very rare case in the world, India, like the European Union, has not only several official languages, but also several alphabets.)

Multipurpose Internet Mail Extensions (MIME)

An IETF standard giving structure to the content of an email message. This opens the possibility of using in a mail message sound, images, files of any format, and also text in any character set.

Requests for comments (RFC)

A numbered series of official documents describing the technical aspects of the Internet, or of different associated hardware (routers, DHCP servers). Note that not

24 The two major gaps in the standard are in the Unicode FTP –File Transfer Protocol– and the passing of mail addresses in Unicode.

25 Using the national top-level domain domain of Tunisia.

all RFCs are official standards, some are qualified as “for information only”, and others as “experimental”.

Standards Development Organization (SDO)

An organization, usually not-for-profit, that develops and maintains standards. The term is generally reserved for those relatively open organizations (like IETF, ITU or W3C) rather than those representing a cartel of businesses.

World Wide Web Consortium (W3C)

The standards organization for Web-related formats such as HTML (format for web pages), XML (format for structured data) or CSS (layout of web pages). <http://www.w3.org>

BIBLIOGRAPHY

[ANDRIES 2008] Patrick Andries. *Unicode en pratique*. 2008. Dunod.

[UNICODE STANDARD] The Unicode Consortium. *The Unicode Standard, Version 6.0.0*. 2010. The Unicode Consortium.

[GILLAM 2002] Richard Gillam, *Unicode Demystified: A Practical Programmer's Guide to the Encoding Standard*, Addison-Wesley, 2002.

[KORPELA 2006] Jukka K. Korpela, *Unicode Explained*, O'Reilly, 2006.

RFC

[RFC 1341] N. Borenstein. N. Freed. *MIME (Multipurpose Internet Mail Extensions): Mechanisms for Specifying and Describing the Format of Internet Message Bodies*. 1992.

[RFC 2277] H.T. Alvestrand. *IETF Policy on Character Sets and Languages*. 1998.

[RFC 3490] P. Faltstrom. P. Hoffman. A. Costello. *Internationalizing Domain Names in Applications (IDNA)*. 2003.

[RFC 4952] J. Klensin. Y. Ko. *Overview and Framework for Internationalized Email*. 2007.

[RFC 5198] J. Klensin. M. Padlipsky. *Unicode Format for Network Interchange*. 2008.

[RFC 5646] A. Phillips. M. Davis. *Tags for Identifying Languages*. 2009.

**MIKAMI YOSHIKI
& SHIGEAKI KODAMA**

MEASURING LINGUISTIC DIVERSITY ON THE WEB

The issue of localization in the information society has aroused curiosity, but also a great concern among many researchers. An interest that prompted the authors of this paper to create in 2003 the Language Observatory Project with the intention of measuring the extent of utilisation of each language. If everyone agrees on the need for such an assessment, the methodology of the observatory and its findings deserve attention for anyone who wants to understand the state of linguistic diversity in the digital world.



MIKAMI YOSHIKI is the director of the Language Observatory Project at Nagaoka University of Technology (Japan). The project was initiated in 2003.

SHIGEAKI KODAMA joined as a researcher in 2006. This project studies linguistic diversity in the Cyberspace and has done a periodical research on the status quo of the linguistic diversity in the Cyberspace.

With the collaboration of **CHEW YEW CHOONG, PANN YU MON, OHNMAR HTUN, TIN HTAY HLAING, KATSUKO T. NAKAHIRA, YOKO MITSUNAGA**

MIKAMI YOSHIKI
& SHIGEKI KODAMA

MEASURING
LINGUISTIC
DIVERSITY
ON THE WEB

Rapid development in information technology is drastically changing communication around the world, extending its reach and enriching its mode. However, new technologies do not evenly benefit all language communities, thus creating the possibility for a “digital language divide”. Let us consider an episode from the era of the printing revolution. In 1608, while stationed on the southwestern coast of India, Thomas Stephens, a Jesuit friar wrote to Rome:

“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 alphabets 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”. [PRIOLKAR 1958]

At the time that the friar wrote this letter, more than one hundred and fifty years had passed since Gutenberg’s innovation, but the new printing technology would not reach his parish until the XIXth century. As he mentioned, the main obstacle was the difficulty of introducing printing technology in the regional languages. Current terminology interprets this as a “localization problem”. The difficulty of casting a large number of metal typesets would take on a different form in the age of computers and the Internet.

The question of the localization problem in an information society has aroused interest and concern for a number of researchers, including the authors. In 2003, we launched the *Language Observatory Project*, intending to measure the use of each language in cyberspace.

The first section of this article describes why such measurements are necessary. The second section introduces the *Language Observatory Project*, and the third section provides recent results obtained from our observations.

WHY MEASURE ?

Localization Still Matters

The typing mould for printing technology was the equivalent of today's computer technology's character code. We now have an international standard on character code for information interchange, the ISO/IEC 10646 Universal Coded Character Set, abbreviated as UCS, or Unicode¹. As the name implies, it covers an entire universe of character codes, from ancient writing systems such as Egyptian hieroglyphs and cuneiform, to minority scripts like those used in the deep mountainous regions of Southeast Asia by the speakers of Thai-Kadai languages.

But many problems in language processing remain. The most fundamental of them is that the UCS, contrary to its name, does not include the entirety of character sets used by humankind; according to our study, many language users still face the same obstacles encountered by the Jesuit friar in XVIth century India.

The Mongolian language, for example, is written either in Cyrillic script or in its own historical and traditional script, for which at least eight different codes and fonts have been identified². No standardisation of typed font exists, causing inconsistency, even textual mistranslation, from one computer to another. As a result, some Mongolian web pages are made up of image files, which take much longer to load.

Indian web pages face the same challenge. On Indian newspaper sites proprietary fonts for Hindi scripts are often used and some sites provide their news with image files. These technological limitations prevent

1 Unicode is a standard created by the Unicode Consortium Inc. But its development and revisions are completely synchronised with the *de jure* standard ISO/IEC 10646. These two standards can be treated as one single standard.

See in this book: Stéphane Bortzmeyer, *Multilingualism and the Internet's Standardisation*

2 In addition to UCS/Unicode, BeiDaFangZheng, GB18030, GB8045, Menksoft, Sayinbillig, Boljoo and SUDAR. Most of them are proprietary, local codes used only by a limited group of users.

information from being interchangeable, and lead to a digital language divide. Our research shows that use of UCS Hindi fonts is spreading, but that many web pages still depend on image files or proprietary fonts.

Such technical challenges maintain gaps not only between languages but between scripts. The authors' initial motivation stems from this issue. When the *Language Observatory Project* was launched, one of its founders wrote the following statement:

“My recent study based on statistical data provided by ITU and Unesco gives a rough sketch of global digital-divide “among scripts”. Latin alphabet users, 39% of global population, consume 72% of world total writing/printing paper and enjoy 84% of access to the internet. Hanzi – Chinese ideograph, users in China/Japan/Korea, 22% in global population, consumes 23% of paper and have 13% of internet access. Arabic users, 9% in population, consume 0.5% of paper and have 1.2% of Internet access. Cyrillic script users 5% in population consume 1.1% of paper and have 1.6% of Internet access. Then how about Indic script users? If all Brahmi-origin scripts widely used in Southeast Asia – Myanmar, Thai, Lao, Khmer, etc. included, Indic scripts users occupy 22% of world population, consume 2.2% of paper and have just 0.3% of internet access” [MIKAMI 2002].

The Language Observatory was launched to address and close these divides to ensure equality and diversity online.

English Dominance on the Web

The second reason for the digital language divide is the dominance of the English language on the Web, which may also reflect the economic aspect of the Web's evolution. This topic was first referred to in 1995 at the Francophone Summit in Cotonou. At that summit, the presence of English was publicly quoted as being above 90%. Funredes (*Fundación Redes y Desarrollo*), reacting to the figure, attempted to obtain accurate measurements of several languages including English [PIMIENTA ET AL. 2010].

Our research shows that in Asian and African ccTLD domains, English continues to dominate a full ten years after the summit. From 2006 to 2009, we conducted annual surveys of language presence on the web. A direct comparison is impossible because the numbers of pages collected

differs between studies (we are currently attempting to identify a methodology to normalise size disparity in sample collection using analysis of variance), but we can nonetheless make the observation that in all surveys English was the language most widely used in Asian and African domains.

The 2010 study found that English was used in 82.7 % of pages collected from the African ccTLD domains; French came in second with 5.5 %. For Asian domains, English also placed at the top, but with a smaller proportion, about 39 %. This is because with the Asian domains, some regional languages, including Hebrew, Thai and Turkish, are strongly dominant in the ccTLD of each language.

In 2010, we extended the survey to include Caribbean domains, and found that Spanish was the most frequently used language with a ratio of about 55 %. English came in second with a ratio of about 33 %.

Multilingual Knowledge

Our study indicates an interesting result in terms of the gap between the availability of practical and professional knowledge among languages. We analyzed the availability of 100 science and engineering terms in the online encyclopedia Wikipedia. One term, nitrogen, was found in 171 different languages. All 100 terms were found in English; among the European languages, the average number of terms found was about 30. In Asian languages, only 18 terms were available, while in African languages, the average number of available terms was just 7. It is often said that the internet revolution will universalise accessibility to the knowledge society; in reality, however, we find significant gaps between languages in terms of accessing professional knowledge. More details on this study can be found in [HTUN *ET AL.* 2010].

Table 1. Availability of 100 science and engineering terms on Wikipedia, by language

Number of available terms N	European*	Asian	African	Others
0-9	0	2	0	0
10-19	31	30	6	8
20-29	10	8	1	0
30-39	4	4	1	1
40-49	3	2	0	0
50-59	11	1	0	0
60-69	7	4	0	0
70-79	5	4	0	0
80-89	6	1	0	0
90-99	2	0	0	0
100	1	0	0	0
Average number of available terms	30	18	7	15
Number of languages checked	80	56	8	17

*European languages include English.

Unesco Initiatives

Since language is a key purveyor of culture, language diversity cannot be avoided in the wider discussion of cultural diversity. Unesco has been engaged in the preservation of cultural diversity since its founding on the grounds that “*intercultural dialogue and respect for cultural diversity and tolerance are essential to building lasting peace*” [UNESCO 2003]. Unesco has since repeatedly published declarations and recommendations relating to cultural diversity.

In October 2003, the Unesco Member States adopted the *Cyberspace Recommendation*, affirming the importance of cultural diversity emphasizing Unesco’s responsibility in maintaining it. Unesco is now responsible for developing multilingual content and systems as well as public domain content, and facilitating access to networks and services. Under this recommendation, various activities including ws1s (World Summit on

the Information Society), IGF (Internet Governance Forum), and IMLD (International Mother Language Day), were planned and carried out.

THE LANGUAGE OBSERVATORY

The Language Observatory Project was founded in 2003 after Unesco's *Cyberspace Recommendation*. The project's main objective is to observe and provide data on the real state of language use on the web to the end of examining language diversity on the web.

How it Works

The Language Observatory is designed to measure use of each language on the World Wide Web. Measurement is effected by counting the number of pages on the Web written in each language.

The project consists of two major components. The first is a data collecting instrument from the Web using crawler robots, which, together with high-performance parallel crawler software developed at the University of Milan, [MIKAMI ET AL. 2005] can collect millions of Web pages per day.

The second component is a language identification instrument. As the Language Observatory has developed software to identify language, script and encoding properties of Web pages with high accuracy and maximum coverage. The first version of the identification algorithm LIM (Language Identification Module) was developed by Suzuki et al. in 2002 [SUZUKI ET AL. 2002] and implemented by Chubachi et al. in 2004. It was later improved by Chew in 2008 for a second version and the one currently in use, G2LI.

G2LI is capable of identifying 184 languages in ISO Language Code (ISO 639-1) with an average accuracy of 94 % according to a recent verification examination. In addition to a wide coverage of languages, it can identify various types of legacy encodings³, which are still extensively used by many non-Latin-script user communities, as mentioned in the first part of this article. The second version employs improved preprocessing

3 Legacy encodings are non-standardised, and often proprietary encodings.

procedures and is capable of properly handling HTML entity encoding⁴, which is also extensively used in many non-Latin scripts. Due to these special features, the authors believe that G2LI is the most suitable language identification instrument for the measurement of language on the Web.

A Hidden Component: The Universal Declaration of Human Rights

Hidden inside the language identification instrument is a set of training texts for the software. The technical details are provided in [SUZUKI *ET AL.* 2002], but it should be mentioned that the richness and the quality of training texts is the most critical in language identification task. A set of texts translated from the *Universal Declaration of Human Rights* (UDHR) provided by the UN Higher Commission for Human Rights (UNHCHR) were used for this purpose because of their wide coverage of the world's languages.

Of note is that not all translated UDHR texts are provided with encoding; some are available only as image files. Image files can be read by humans but not directly by computers, necessitating that we transform images into text data. Table 2 illustrates how many transformed texts are given in image format (322 languages were available at the date of the first search, in early 2004). More than two hundred languages use Latin script, with or without diacritics, and only three of them were given in PDF or GIF file format. In contrast to this, among languages using Abugida script⁵, not a single language was presented in the form of encoded text.

4 HTML entities represent characters using only ASCII letters (e.g. `&a1pha`; entity represent greek character α).

5 Abugida scripts are syllabic scripts, most of which are generated from Indian Brahmi scripts and currently used in South and Southeast Asian regions. Another important Abugida script is Amharic.

Table 2. Number of available UDHR texts from UNHCHR website by format

	Latin	Cyrillic	Other alphabet	Abjad	Abugida	Hanzi	All others	Total
En-coded	253	10	1	1	0	3	0	268
PDF	2	4	2	3	10	0	4	25
GIF	1	3	0	9	15	0	1	29
Total	256	17	3	13	29	3	5	322

NOTE: **Other alphabets:** Greek, Armenian and Georgian; **Abjad:** Arabic and Hebrew; **Abugida:** Amharic and all Brahmi origin scripts used in south and southeast Asia; **Hanzi:** Chinese, Japanese and Korean; **All others:** Assyrian, Canadian syllabics, Ojibwa, Cree, Mongolian and Yi.

This fact might itself point to the existence of a digital language divide, or in this particular case, a “digital script divide”. Upon first encountering this problem, one of the authors elaborated in an essay for the Indian journal *Vishbha Bharat*:

“Recently I visited a website of the United Nations Higher Commission for Human Rights⁶ which introduces more than three hundred different language versions –from Abkhaz to Zulu of the Universal Declaration of Human Rights. The site claims that this text is the most widely translated text in the world, and has been awarded the Guinness World Record for having done this great job. Thus the Universal Declaration is “the most universal text” in the world.

Try now! And you can find all eighteen Indian official language versions of the 1,778 words text, with only two exceptions –Konkani and Manipuri. But really disappointing for you would be the fact that all Indian language versions are just posted as “gif” files, not in the form of encoded texts. And actually many other non Latin scripts users in the world have to feel the same kind of sadness after visiting”. [MIKAMI 2002]

Since then, many collaborators have voluntarily helped us to create a text version of these image files⁷. For certain languages, we are still seeking

⁶ <http://www.unhchr.ch/udhr>

⁷ Sinhala, Vietnamese, Bahasa Melayu, Lao, Persian (Farsi), Mongolian, Tamil, Uyghur, Nepali, Malayalam, Hindi, Magahi, Marathi, Sanskrit, Bengali, Saraiki, Punjabi, Gujarati, Kannada, Myanmar, Vietnamese in TCVN5712, VIQR, VPS, Assamese, Azeri, Dari, Kyrgyz,

appropriate collaborators and have had to renounce the inclusion of training texts in those languages.

Around the same time as we launched the Language Observatory Project, Eric Miller launched UDHR-in-Unicode. The objective of this project was to demonstrate the use of Unicode for a wide variety of languages, using the Universal Declaration of Human Rights (UDHR) as a representative text. Currently, UDHR-in-Unicode is housed on the Unicode Consortium website and the texts are used in the study of natural language processing⁸.

Sponsors and Collaborators

The *Language Observatory project* was initiated by the authors in 2003 and received funding from Japan Science and Technology Agency (JST) through its RISTEX program from 2003 to 2007. The kick-off event, held at Nagaoka University of Technology on February 21, 2004, included guest Paul Hector, then director of Unesco's Communication and Information (CI) section.

The project interacted and collaborated with many partners from various parts of the world, and joined with the African Academy of Languages (ACALAN) at WSIS in Tunis in November 2005 at a session on African languages. Among the attendees were the President of ACALAN, Adama Samassekou, Daniel Pimienta of Funredes, and Daniel Prado of Union Latina.

We agreed to organise a joint African web language survey project. The project's initial target was initially the African ccTLD domain. In 2006, we held a workshop in Bamako, Mali, with the cooperation of ACALAN and the support of JST. Many African researchers interested in language diversity and the digital divide on the web attended the workshop.

After a fruitful workshop in Bamako, we planned a workshop to publicise our project and the digital language divide. We also held workshops at

Marwari, Sindhi, Tajiki, Tamang, Telugu, Turkmen, Urdu, Uzbek. Unless otherwise noted, texts were prepared in UTF-8 encoding. UTF-8 text is not enough for our purpose in some languages which use non-standard, legacy encodings.

For more details and contributors' names, visit our site: <http://gii2.nagaokaut.ac.jp/gii/lopdiary.php?itemid=480>

⁸ NLTK (Natural Language Tool Kit) by Steven Bird *et al.* is one example.

Unesco headquarters in Paris for International Mother Language Day in 2007 and 2008.

The first complete observation report, published in 2008 [NANDASARA *ET AL.* 2008], was the first article addressing language distribution on the Asian web. The report confirmed a significant digital language divide. English used in more than 60 % of web pages in south Asian and southeast Asian countries. In west Asia, English dominance was less outstanding, and in some countries, Arabic was most widely used. In central Asia, Russian was the dominant language, except in Turkmenistan where English was used in 90 % of web pages. A minority of indigenous languages, including Turkish, Hebrew, Thai, Indonesian, Vietnamese and Mongolian, were the most used languages in their country domains. The study signified a breakthrough in understanding online language disparity, and provided a basis for future work.

LINGUISTIC DIVERSITY ON THE WEB

In this section, some results of the Language Observatory's language surveys will be introduced.

Lieberson's Diversity Index

Lieberson's Diversity Index (LDI) [LIEBERSON 1981] is a widely used index of linguistic diversity that is defined by the following formula, where P_i represents the share of i -th language speakers in a community:

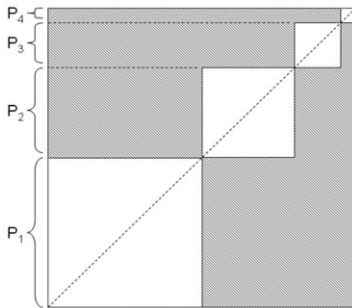
$$LDI = 1 - \sum P_i^2$$

If anyone in a community speaks the same language, then $P_i = 1$ and for the speakers of other languages, $P_i = 0$. Thus the LDI of a completely monolingual community is zero. If four languages are spoken by an equal number of people, then $P_1 = P_2 = P_3 = P_4 = 0.25$ and the LDI of this multilingual community can be calculated as $LDI = 1 - (0.25)^2 \times 4 = 0.75$. Thus a higher LDI means larger linguistic diversity and a lower LDI means lower diversity.

Lieberson also took into account the fact that bilingual or multilingual speakers would render the formula a bit more complicated. But the basic idea of LDI can be explained by the illustration in Figure 1. A square of P_i

means the probability that the i -th language speaker meets with a speaker of the same language. And the sum of P_i squares represents the combined probability of any speaker meeting with a speaker of the same language in the community on average. Finally the sum of P_i squares is subtracted from 1, indicating the probability that any speaker will encounter different language speakers in a society. The dark-colored areas of the square in Figure 1 correspond to this probability.

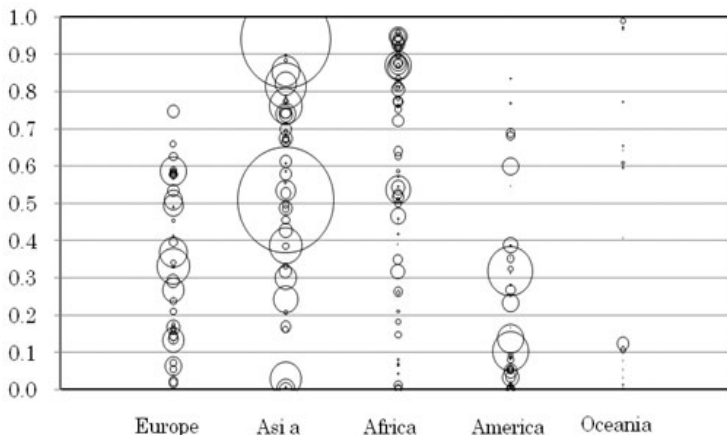
Figure 1. A graphic interpretation of LDI



Ethnologue provides a complete list of LDI data for each country or region, together with population size and the number of indigenous and immigrant languages. Based on this data⁹, Figure 2 was prepared by the authors to show how LDI changes across countries and across continents. Each circle represents a country in this chart. The circle's size corresponds to the country's population, and its vertical axis represents the country's LDI. The two large circles on the axis of Asia correspond to India (LDI = 0.94) and China (LDI = 0.51).

⁹ Based on the web version, an equivalent of the 16th edition of *Ethnologue*.

Figure 2. Lieberson's Diversity Index of countries by continent



(based on data from *Ethnologue*)

As the chart illustrates, countries in the African continent have the highest language diversity among the continents, followed by Asia, Europe, America (North and South America included) and Oceania.

The highest LDI in Africa is of the Central African Republic (LDI = 0.96); nine other countries have an LDI over 0.90 (the Democratic Republic of Congo, Tanzania, Cameroon, Chad, Mozambique, Uganda, Benin, the Ivory Coast and Liberia). Thirteen countries with an LDI above 0.80 (Togo, Zambia, Kenya, South Africa, Mali, Guinea-Bissau, Nigeria, Ethiopia, Congo, Sierra Leone, Angola, Namibia and Ghana), and seventeen have an LDI of over 0.5. The lowest LDI countries on the African continent are Rwanda and Burundi, with 0.004.

In Asia, the highest diversity is observed in Papua New Guinea (0.99). This country is known for its abundant language resources, and its LDI is the highest of all countries on the earth; in Asia it is followed by India, East Timor, Bhutan, the Philippines, Iran and Indonesia. These seven countries have an LDI of over 0.8, and another 22 Asian countries have an LDI of over 0.5. On the opposite end, Korea (0.003), the Maldives (0.01) and Japan (0.03) appear to be quite monolingual societies.

In Europe, the highest LDI belongs to Belgium (0.75). It is followed by Bosnia (0.66), Serbia (0.63), Moldova (0.59), Italy (0.59), Latvia (0.58), Georgia (0.58), Macedonia (0.58), Switzerland (0.58), Albania (0.57), Andorra (0.57), Austria (0.54), Monaco (0.52), and Spain (0.51). These

fifteen countries have an LDI over 0.5. Countries with a dominant mother language, such as Germany (0.37), Russia (0.33), the Netherlands (0.29), and France (0.27), generally have lower LDIs. The lowest in Europe is Hungary (0.02).

In the American continent, only three countries have an LDI of over 0.5: Belize (0.77), Trinidad and Tobago (0.70), and Canada (0.60). Spanish dominant countries generally have a low LDI.

In Oceania, the separation of peoples on small isolated islands has meant that islands tend to develop unique languages. Countries composed of multiple islands accordingly tend to display a higher LDI. The LDI of Vanuatu is 0.97 and the highest among the Oceania countries; over 100 languages are spoken in its islands. Other archipelago countries also show a high LDI: the Solomon Islands (0.97), New Caledonia (0.83), Micronesia (0.77), Fiji (0.61), and Nauru (0.60).

Local Language Ratio

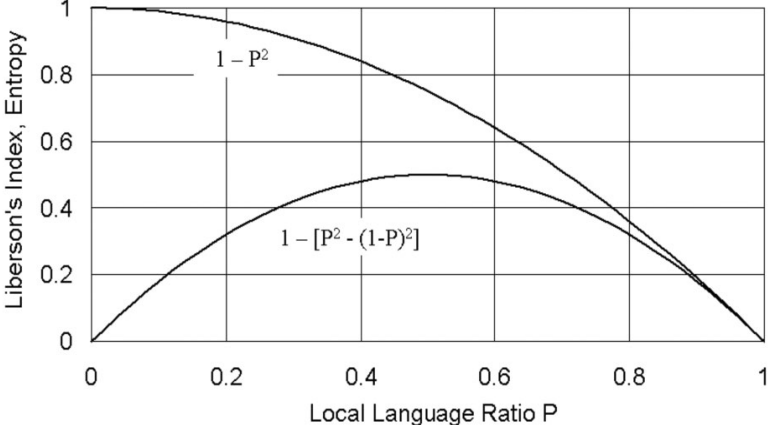
In the previous section, we reviewed the overall condition of linguistic diversity of the world based on data provided by *Ethnologue*, data that reflects the situation *in the real world*. Now we would like to move onto the main theme of this article, language diversity *in the cyber world*.

Since being launched, the Language Observatory has focused its attention on two continents, Asia and Africa. As mentioned above, the first observation results were reported during a workshop organised at Unesco headquarters in February 2005; they are fully documented in an article published in 2008. Recently, the project has completed another round of surveys of Asia, Africa and the Caribbean region based on 2009 data. The following sections will introduce an overview of this most recent study.

Here we propose a two-dimensional chart, which is tentatively named the LL-chart, because the chart has the Local Language Ratio on the horizontal axis and the LDI on the vertical axis. The purpose of this chart is to solve a problem we encountered when preparing an LDI chart based on data from cyberspace. It often happens that languages used on the Web are completely different from languages spoken in the real world. In many cases, the latter consists of local languages while the former mainly consists of global languages like English, French or Russian. And in those cases,

the high LDI of languages in cyberspace and that in the real world are not considered to be the same. We have to take into account some measurements about the presence of local languages, as presented in Figure 3.

Figure 3. Schematic diagram of the LL-chart



Notice that all countries with a local language ratio P fall within the area between the two curves $1 - [P^2 + (1-P)^2]$, Liberson's index in the case of two languages, and $1 - P^2$, which gives the maximum value of Lieberson's index¹⁰. When P becomes larger than 0.5, the LDI becomes smaller and the plotted point will move towards the bottom-right corner. When P is small, there are two possibilities: either the vacancy of local language is filled by a dominant foreign language, in which case the LDI shrinks and the point moves down and to the left; or the vacancy of local language is filled with multiple foreign languages, in which case the LDI grows and the point moves up and to the left.

10 Two curves provide the upper and lower limits. The upper curve indicates the LDI of a two-language community. As the addition of a third-language speaker to this community increases the average probability to encounter different language speakers, this value is the minimum LDI of more than two language communities. The lower curve indicates the LDI of a very special case, where each member, in addition to the local language, speaks an additional language, or the maximum LDI.

Comparison by Region : Asia, Africa, Europe and the Caribbean

Based on data collected in November 2009, the LDI and local language ratio were calculated for all country domains in Asia and Africa. As we do not have data for European countries, we used Google's page count by language. Figures 4, 5 and 6 show the Local Language Ratio – LDI chart for these three regions.

Asian LDIs are plotted in Figure 4. China, Japan and Korea and some Arabic-speaking countries (Iraq, Saudi Arabia, and Jordan) are found in the bottom-right corner, while Vietnam, Thailand and Indonesia, Israel, Turkey, Georgia and Mongolia show a relatively high local language presence.

Of note here is the context of central Asian countries. Their web spaces are composed of local languages, with major components of English and Russian, although the emphasis changes by country. Kazakhstan, Kyrgyzstan, Tajikistan, and Uzbekistan have a major emphasis on Russian, while only Turkmenistan has an emphasis on English.

On the other hand, web contents in the Indian subcontinent have a nearly negligible local language presence. More than 70 % of these web contents are written in English.

The case of Laos is particular and deserves mention here. According to *Ethnologue*, the country's LDI is only 0.674. Why then does it have such a high LDI on the Web? The major reason for this is that the “.1a” domain is actively marketed to foreigners, including customers connected to Los Angeles. As the domain is sold mainly to foreign industries and peoples, in the “.1a” domain, just 8 % of web pages are in Lao.

LDIs of African domains are plotted in Figure 5. The presence of local languages in African domains is far rarer than in Asian domains. For Arabic-speaking countries, the local language claims the majority only in Sudan and Libya; Egypt, Mauritania, Tunisia and Tanzania, along with the rest of Africa, show very little local language presence on the Web. However, several countries nevertheless show high Web LDIs.

The LDIs of European and some Anglophone domains are plotted in Figure 6. Local language presence is above 50 % with the exception of Slovenia and Denmark, whose countries' web spaces are dominated by

English, resulting in a lower LDI. At the opposite extreme is the United Kingdom, which joins other Anglophone countries (USA, Australia and New Zealand) in displaying a characteristically low LDI.

Table 3. Language composition of the Asian and African web domains

African Domains			Asian Domains		
Language	# of pages	%	Language	# of pages	%
English	30,327,396	78.40%	Chinese	7,832,521	20.46%
French	2,737,455	7.08%	Japanese	5,287,655	13.82%
Afrikaans	660,510	1.71%	English	4,867,355	12.72%
Arabic	592,746	1.53%	Russian	1,611,339	4.21%
Chinese	391,745	1.01%	Korean	1,100,232	2.87%
Portuguese	348,131	0.90%	Vietnamese	710,048	1.86%
Russian	307,178	0.79%	Thai	544,561	1.42%
Spanish	276,126	0.71%	Indonesian	308,894	0.81%
Japanese	158,992	0.41%	Hebrew	89,076	0.23%
Others	879,605	2.27%	Others	14,055,334	36.72%
Not identified	2,005,311	5.18%	Not identified	1,867,355	4.88%
Total	38,685,195	100.00%	Total	38,274,370	100.00%

NOTE: Web data was obtained from the country-code domains of Asia and Africa in November 2009. For a list of domains, see the ANNEX.

Figure 4. Local Language Ratio and LDI of language composition for Asian domains

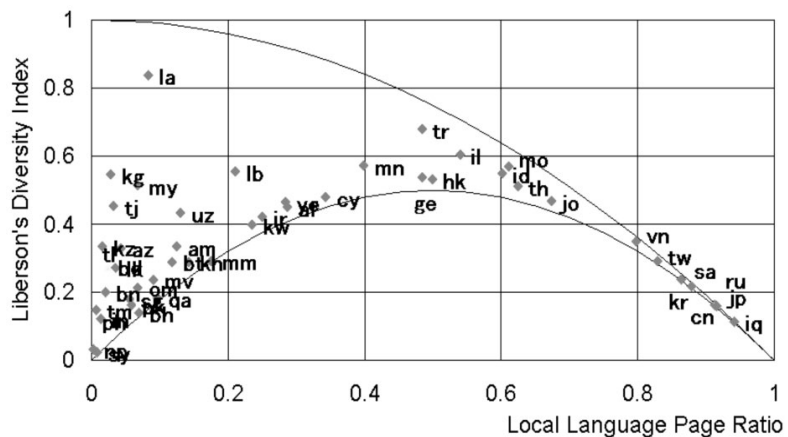


Figure 5. Local Language Ratio and LDI of language composition for African domains

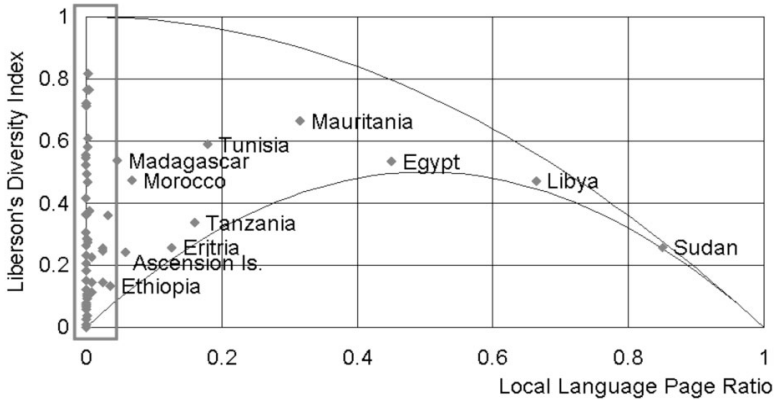
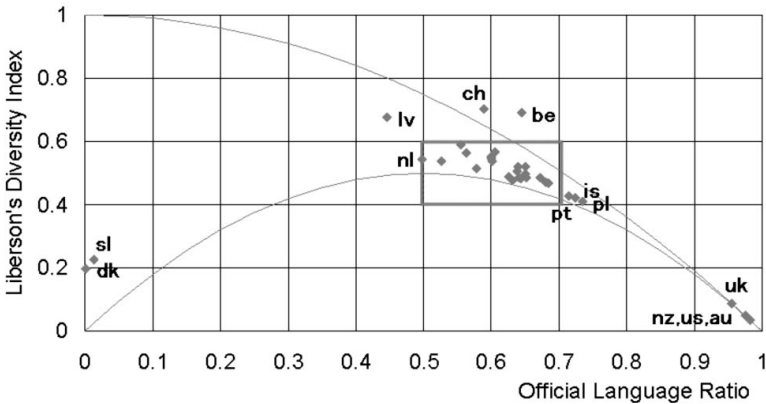


Figure 6. Local Language Ratio and LDI of Language Composition on the European Web for selected Anglophone countries



Challenges and Directions

The most serious challenge to measurement efforts comes from the sheer size of the growing Web. Nobody knows exactly how many web pages exist on the entire Web. In 1997, the number was estimated at only 320 million; by 2002, it had grown to 8 billion [MILLER 2007]. In 2008, Google

announced 1 trillion URLs on the Web, but has since stopped providing data. Nor do other search engines provide such data, which leads us to conclude that it is not currently possible to count all the existing pages on the Web.

Another strategy is needed to create a sampling method of pages that can reflect the entire Web. We are currently developing what we believe to be a promising method using ANOVA (Analysis of Variance).

Another advantage of sampling is extending the research target to other ccTLDs that have not yet been targeted because of their huge size.

We have provided yearly reports on the statistics of language use on the Web at events held by Unesco or IGF, and will continue to provide them in the future, with the following improvements:

- Extension of target;
- Extension of identifiable languages;
- Diversification of analysis method.

The first improvement was mentioned above. Our research target areas currently include Asia, Africa, the Caribbean, and Europe. Many ccTLDs are still lacking in our research because of the storage capacity of our system.

The second improvement will help draw a more accurate image of language use on the Web. Our identification engine can identify more than 300 languages, but by *Ethnologue's* estimation, over 7,000 languages exist on Earth. As many do not have a written form and are only spoken, as shown in Table 3, our identifier could not identify about 5 % of collected pages, leading us to conclude that we are overlooking many languages. As mentioned in Section 1.1, we need to collect local encodings to investigate problems with legacy encoding.

A prototype of the third improvement was displayed in Section 3. The most basic data we can provide is a list of the number of pages in each language on each ccTLD. But these data do not tell us much about language use on the Web. We need to increase the sophistication of our interpretation to enable deeper reflection on digital language use.

With those improvements, we hope to increase the usefulness of statistics as fundamental data for considering language usage and diversity on the Web.

BIBLIOGRAPHY

[PRIOLKAR 1958] A. K. Priolkar. 1958. *The Printing Press in India*. Bombay: Marathi Samsodhana Mandala.

[MIKAMI 2002] Yoshiki Mikami. 2002. Global digital-divide among scripts. VishwaBharat. October 2002 Issue, p.1.

[PIMIENTA ET AL. 2010] Daniel Pimienta, Daniel Prado and Álvaro Blanco. 2010. *Twelve Years of Measuring Linguistic Diversity in the Internet*. Paris: Unesco.

[HTUN ET AL. 2010] Ohnmar Htun, Shigeaki Kodama and Yoshiki Mikami. 2010. Analysis of Terminology Terms in Multilingual Terminology Dictionary. *Proceedings of the 8th International Conference on Computer Applications* 2010, pp. 122-128.

[UNESCO 2003] *Recommendation Concerning the Promotion and Use of Multilingualism and Universal Access to Cyberspace*. Unesco: Paris.

[ISC 2010] Internet Software Consortium. 2010. *Internet Domain Host Count* <http://www.isc.org/solutions/survey>

[UNESCO 2001] *Universal Declaration on Cultural Diversity*. Unesco: Paris.

[MIKAMI ET AL. 2005] Yoshiki Mikami, Zavorsky Pavol, Mohd Zaidi abd Rozan, Izumi Suzuki, Masayuki Takahashi, Tomohide Maki, Irwan Nizam, Massimo Santini, Paolo Boldi, and Sebastiano Vigna. The Language Observatory Project. 2005. *Proceedings of the 14th International World Wide Web Conference*, p. 990.

[SUZUKI ET AL. 2002] Izumi Suzuki, Yoshiki Mikami, Ario Ohsato. 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.

[NANDASARA ET AL. 2008] S. T. Nandasara, Shigeaki Kodama, Chew Yew Choong, Rizza Caminero, Ahmed Tarcan, Hammam Riza, Robin Lee Nagano and Yoshiki Mikami. 2008. An Analysis of Asian Language Web Pages. *The International Journal on Advances in ICT for Emerging Regions (ICTer)*, Vol.1 No.1. pp. 12-23.

[LIEBERSON 1981] Stanley Lieberman and Anwar S. Dil. 1981. *Language Diversity and Language Contact: essays*. California: Stanford University Press.

[MILLER 2007] Miller, Colleen. 2007. Web Sites: Number of Pages. *NEC Research*. IDC. 6 June 2007.

HOW LANGUAGE TECHNOLOGIES SUPPORT MULTILINGUALISM

The issues of multilingualism are many, and the need for it is important, both in Europe and internationally. Language Technologies can help us respond, but it is necessary to develop infrastructures and generate the resources needed to conduct research on the different languages. Some programs support this domain, but suffer from a lack of scale, continuity and cohesion. This effort deserves to be coordinated among nations and international agencies to facilitate multilingualism in Europe and globally.

Original article in French.
Translated by John Rosbottom.



JOSEPH MARIANI is currently director of the French-German Institute for Multilingual and Multimedia Information (IMMI). He was Director of LIMSI-CNRS and Head of its Human-Machine Communication department, then Director of the Information and Communication Technology Department at the french Ministry of Research.

JOSEPH MARIANI

HOW LANGUAGE
TECHNOLOGIES
SUPPORT
MULTILINGUALISM

Since the divine punishment of Babel, mankind must live with the wealth of a multitude of languages and cultures. The difficulty and costs of sharing information and communicating, despite the language barriers, while preserving these languages, could benefit from the support of automatic language processing systems (that we will call language technologies), which are the object of a major research effort, although still insufficient and insufficiently coordinated.

THE ISSUES OF MULTILINGUALISM

The issues of multilingualism are twofold :

First, to take care of preserving cultures and languages, i.e. to allow citizens to express themselves in their first language. This question takes on a particular depth in the context of the construction of Europe, given the strong linguistic diversity within a single political entity. Thus, 75 % of Germans citizens questioned prefer to find websites in their own language rather than in a foreign language. One can also note that currently it is estimated that less than 30 % of the web is in English, a proportion that has declined sharply from a rough estimate of 50 % in 2000¹. 50 % of European citizens speak only one language and when they speak a second one, it is not necessarily English. Only 3 % of Japanese speak a foreign language. In India, less than 5 % of people fluently speak English. Preserving languages and, through them, their corresponding culture responds to a strong demand from citizens.

The second challenge is to enable communication among humans, usually in the framework of common democratic structures. We are facing

1 See in this book : Michael Oustinov, *English Won't Be the Internet's Lingua Franca*.

it in the European Union, where there are now 27 member countries and 23 official languages, representing 506 language pairs. If one considers all the European languages, one can count over 60, which represents almost 4,000 pairs of languages to translate! The European Commission employs more than 2,500 translators who in 2007 translated over a million and a half pages. This covers only a fraction of the needs. To cover the totality would require 8500 translators to process 6.8 million pages annually. Taking into account the EU linguistic diversity represents 30% of the budget of the European Parliament, or about 300 million euros per year, with the use of 500 translators and interpreters. The estimated total cost of multilingualism for the European Union is a little over one billion euros per year; but considering the number of Europeans, that represents only 2.2 euros per citizen per year, which ultimately is not prohibitive. A similar situation exists within some nations, like India, but also internationally, with about 6,000 major languages that are spoken, or 36 million pairs of languages to translate... And a simple statistic: at present YouTube, every minute, uploads thirty two hours of new videos in all languages.

NEEDS RELATED TO MULTILINGUALISM

At the European level, the needs related to multilingualism are very numerous: needs for the establishment of the European Digital Library (Europeana, which included, in January 2011, 14.6 million documents in 26 languages), for which it is necessary to provide crosslingual and multilingual tools to enable access for all; for the realisation of a multilingual platform for alert and information exchange planned by the European Security Agency (ENISA) for the Member States; for the European Patent Office – *The London Protocol* has reduced the number of official languages to three (English, German and French) for reasons of cost, whereas, with more automated tools, more languages could be handled; for meetings of the European Commission, of the European Parliament or of the European Court of Justice, where English tends increasingly to become the only working language...

Such needs respond to a real democratic necessity, to be met more generally at the international level. If we take the example of Internet governance within the UN Internet Governance Forum (IGF), only English is accepted as a working language, and a lively debate concerns the possibility of using

different spellings and different accents in the domain names. The World Digital Library in Unesco had 1,500 documents filed in 7 languages at its inception in April 2009². Dubbing and subtitling of audiovisual works; writing technical manuals, in the aerospace or automotive industries, or instruction manuals for the consumers; live super-titling of works of performing art; translation of texts, videos, and radio or television programmes that are innumerable, and in all languages; simultaneous interpreting at multiple meetings, conferences, workshops, courses, which take place throughout the world: there are many applications where language technologies can offer opportunities. Think also of the urgent needs related to scientific articles written in a mother tongue, which are diminished markedly due to the overvaluation of English by bibliometrics, risking the loss of specialised terminology in other languages.

Add to this picture the many needs related to the accessibility of information by the visually or hearing impaired, requiring the translation of information from one medium to another: written to oral, oral to written, oral to gesture (sign language), and more generally to the accessibility of information by people who do not speak fluently the language in which it was encoded, including, notably, migrants³.

FINDINGS

The extent of these needs shows very well that they cannot all be covered by existing or even future human resources of professions dealing with language processing.

Taking into account multilingualism is not a top priority in any economic sector. If we ask the boss of a big company what is his/her priority, none will say it is multilingualism. But if we add up the priorities in each area where it is necessary to take it into account, then we reach a very large sum. This therefore requires, in our opinion, thought and political action to bring out this awareness and provide appropriate responses.

Even when multilingualism is seen as a necessity, its cost is still very important. It is this gap that calls for the development of language technologies and their utilisation when their performance is up to the needs of target applications.

² <http://www.wdl.org/fr>

³ See in this book: Viola Krebs & Vicent Climent-Ferrando, *Languages, Cyberspace, Migrations*.

It should be noted that currently, language technologies have not yet reached maturity for all languages, with strong imbalances among languages. And they do not provide for human intervention. Thus, automated translation is not good enough to translate literary works or, in general, texts which require high quality translation. This must be said clearly. But on the other hand, it can help a human translator in his or her work and has a sufficient quality to give an approximate translation, of web pages for example, thus meeting the needs of the general public. Language technologies can more fully participate in solving the issue of multilingualism, which justifies drawing attention to their merits, especially in the funding of research programmes.

LANGUAGE TECHNOLOGIES

Language technologies are said to be *monolingual* when they handle a single language, *multilingual* when the same technology processes several (individual) languages, or *crosslingual* when they allow for switching and transferring from one language to another.

Language technologies cover the processing of written language, whether monolingual (morphosyntactic and syntactic analysis; text understanding; text generation; automatic summarisation; terminology extraction; information retrieval; Question & Answer systems, etc.) or crosslingual (automatic or computer-aided translation; crosslingual information retrieval, etc.).

For the processing of spoken language, there are also monolingual technologies (speech recognition and understanding; speech-to-text transcription (textual transcription of what has been said); speech synthesis; spoken dialogue; speaker recognition, etc.) and crosslingual (identification of a spoken language, speech translation, real-time interpretation, etc.).

Finally, we must not forget gestural communication, particularly for processing Sign Languages (recognition, synthesis and translation)⁴.

These technologies can be intermedia, i.e. translating from one medium to another, with numerous applications to enable accessibility for

⁴ See in this book: Annelies Braffort & Patrice Dalle, *Accessibility in Cyberspace: Sign Languages*.

the disabled (Text-To-Speech synthesis for the visually impaired, automatic transcription (subtitles or supertitles), aids to lip reading, Sign Language processing... for the hearing impaired, voice commands for the motor-impaired...).

In language science and technology, research initially covered two areas under two different scientific communities:

- The processing of written language (also called automatic language processing, or natural language processing (NLP)), coming from linguistics and artificial intelligence;
- The processing of spoken language (called “speech communication”), coming from acoustics, signal processing and pattern recognition.

These two communities have gradually come together, due to a political will and to the use of complementary methods based on machine learning with statistical modelling.

Research in these two major areas has made great progress on the lower levels of language processing: regarding written language processing, in text segmentation, lexical analysis, morpho-syntactic and syntactic analysis; and regarding spoken language processing, in speech recognition, Text-To-Speech synthesis, or speaker recognition.

Numerous resulting applications are now in everyday use, such as, regarding written language processing, spelling and grammar checkers, monolingual and crosslingual search engines, online machine translation... and, regarding spoken language processing, talking GPS systems, dictation systems, transcription and automatic indexing of audiovisual content... This list shows that many of these existing applications are related to linking spoken and written language (transcription of speech into text, speech synthesis from text). Spoken dialogue systems, including voice recognition and synthesis, are also growing, but in very specific applications: Voice command on mobile phones, Call centres, tourist or public transportation information, etc.

*Basic architecture of a natural language processing system*⁵



⁵ META-NET White Paper Series, 2011.

Research in the field of automatic machine translation illustrates particularly well the meeting of these two communities. This area has traditionally been studied by researchers in NLP, using a rule-based approach including a combination of rules and linguistic knowledge (bilingual dictionaries, grammars, etc.). Researchers working in the field of spoken communication have for their part experimented in machine translation the machine learning methods that they have successfully used in speech recognition: matching the same text in two languages (parallel corpora), with the same approach used to match a speech signal and its written transcription. This statistical approach has resulted in significant progress leading to the recent development of hybrid translation systems, mixing statistical approaches and linguistic knowledge.

The challenge now is to process information related to meaning, at the semantic and pragmatic levels, in order to establish a natural dialogue between human and machine, or to give the machine the ability to participate in communication between humans. To do this, we need to take into account other communication modalities (multimodal communication, processing of multimedia documents), as well as the processing of paralinguistic information (prosody, expressions of emotion, analysis of opinion and feelings).

LANGUAGE RESOURCES AND EVALUATION

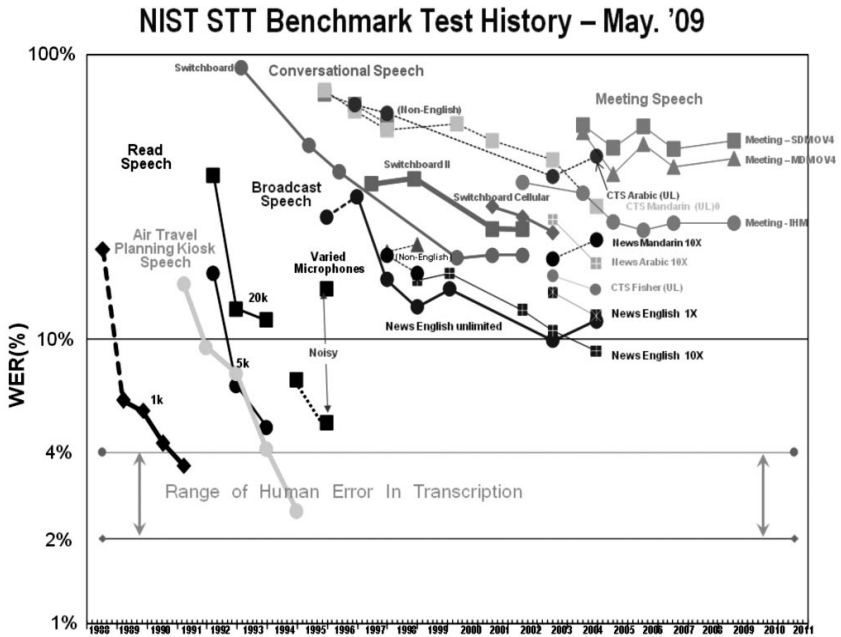
It is crucial for conducting research aimed at developing language technologies to provide a base that includes both language resources and evaluation methods for the technologies that are developed.

With regard to language resources, the data (corpus, lexicons, dictionaries, terminology databases, etc.) are both necessary for conducting research investigations in linguistics and for training automatic language processing systems that are based in most cases on statistical methods. The greater the amount of data, the better the statistical model and therefore the better the system performances. The interoperability of language resources also invites us to think more deeply on the standards to be put in place in order to organise, browse, and transmit data.

It is also necessary to have a means for evaluating these technologies in order to compare the performance of systems, using a common protocol with common test data, in the context of evaluation campaigns. This allows for comparing different approaches and using an indicator of the quality of

the research and of the advances of technology. We now speak of “coopetition”—a mix of international competition and cooperation—and this has become a way to carry out technological research. The Defense Advanced Research Projects Agency (DARPA) of the Department of Defense in the United States, was the initiator of this approach in the mid 80s, through the National Institute of Standards and Technology (NIST) [MARIANI 1995].

History of speech recognition since 1987 according to the NIST⁶ evaluation campaigns



This table shows the progress of Automatic Speech Recognition over the years, through the international evaluation campaigns conducted by NIST. Shown on the chart are the best performances obtained that year, in terms of Word Error Rate (WER) in a logarithmic scale. The effort to go from 100 % error (where the system does not recognise any word) to 10 % is comparable to that required to go from 10 % to 1 % error rate. The tasks became increasingly difficult over the years (first with voice command, using an artificial language of 1,000 words, then voice dictation

⁶ <http://itl.nist.gov/iad/mig/publications/ASRhistory/index.html>

(20,000 words), radio/TV Broadcast News transcription (in English, Arabic and Mandarin Chinese) telephone conversations transcription (also in English, Arabic and Mandarin), meeting transcriptions...), with variable conditions (real time or not, different qualities of sound recording). We see that for some tasks, the performance of systems is similar to those of a human listener, making these systems operational and marketable (such as for command languages). On the other hand, it is clear that for more complex tasks, performance improves more slowly, justifying the continuation of the research effort. Knowledge of these performances helps us to determine the feasibility of an application based on the quality level it requires. Thus, contrary to voice dialogue systems, an information retrieval system for audiovisual data does not require error-free performances in the transcription of speech.

A similar approach was used to monitor progress in machine translation (MT), using the BLEU metrics, proposed in 2000 [PAPINENI ET AL. 2001], whereas the research had been conducted in MT for about fifty years without systematically measuring the quality of results to guide future research. This measure is based on a rudimentary comparison between the results of the systems and the translations of human translators.

*Performance of machine translation systems
in 22 official languages of the EU (Ph. Koehn et al. 2009)*

Translating between all EU-27 languages ³



	Target Language																					
	en	bg	de	cs	da	el	es	et	fi	fr	hu	it	lv	mt	nl	pl	pt	ro	sk	sl	sv	
en	-	40.5	46.8	52.6	56.0	41.0	55.2	34.8	38.6	50.1	37.2	50.4	39.6	43.4	39.8	52.3	49.2	55.0	49.0	44.7	50.7	52.0
bg	61.3	-	38.7	39.4	39.6	34.5	46.9	25.5	26.7	42.4	22.0	43.5	29.3	29.1	25.9	44.9	35.1	45.9	36.8	34.1	34.1	39.9
de	53.6	26.3	-	35.4	43.1	32.8	47.1	26.7	29.5	39.4	27.6	42.7	27.6	30.3	19.8	50.2	30.2	44.1	30.7	29.4	31.4	41.2
cs	58.4	32.0	42.6	-	43.6	34.6	48.9	30.7	30.5	41.6	27.4	44.3	34.5	35.8	26.3	46.5	39.2	45.7	36.5	43.6	41.3	42.9
da	57.6	28.7	44.1	35.7	-	34.3	47.5	27.8	31.6	41.3	24.2	43.8	29.7	32.9	21.1	48.5	34.3	45.4	33.9	33.0	36.2	47.2
el	59.5	32.4	43.1	37.7	44.5	-	54.0	26.5	29.0	48.3	23.7	49.6	29.0	32.6	23.8	48.9	34.2	52.5	37.2	33.1	36.3	43.3
es	60.0	31.1	42.7	37.5	44.4	39.4	-	25.4	28.5	51.3	24.0	51.7	26.8	30.5	24.6	48.8	33.9	57.3	38.1	31.7	33.9	43.7
et	52.0	24.6	37.3	35.2	37.8	28.2	40.4	-	37.7	33.4	30.9	37.0	35.0	36.9	20.5	41.3	32.0	37.8	28.0	30.6	32.9	37.3
fi	49.3	23.2	36.0	32.0	37.9	27.2	39.7	34.9	-	29.5	27.2	36.6	30.5	32.5	19.4	40.6	28.8	37.5	26.5	27.3	28.2	37.6
fr	64.0	34.5	45.1	39.5	47.4	42.8	60.9	26.7	30.0	-	25.5	56.1	28.3	31.9	25.3	51.6	35.7	61.0	43.8	33.1	35.6	45.8
hu	48.0	24.7	34.3	30.0	33.0	25.5	34.1	29.6	29.4	30.7	-	33.5	29.6	31.9	18.1	36.1	29.8	34.2	25.7	25.6	28.2	30.5
it	61.0	32.1	44.3	38.9	45.8	40.6	26.9	25.0	29.7	52.7	24.2	-	29.4	32.6	24.6	50.5	35.2	56.5	39.3	32.5	34.7	44.3
lt	51.8	27.6	33.9	37.0	36.8	26.5	21.1	34.2	32.0	34.4	28.5	36.8	-	40.1	22.2	38.1	31.6	31.6	29.3	31.8	35.3	35.3
lv	54.0	29.1	35.0	37.8	38.5	29.7	8.0	34.2	32.4	35.6	29.3	38.9	38.4	-	23.3	41.5	34.4	39.6	31.0	33.3	37.1	38.0
mt	72.1	32.2	37.2	37.9	38.9	33.7	48.7	26.9	25.8	42.4	22.4	43.7	30.2	33.2	-	44.0	37.1	45.9	38.9	35.8	40.0	41.6
nl	56.9	29.3	46.9	37.0	45.4	35.3	49.7	27.5	29.8	43.4	25.3	44.5	28.6	31.7	22.0	-	32.0	47.7	33.0	30.1	34.6	43.6
pl	60.8	31.5	40.2	44.2	42.1	34.2	46.2	29.2	29.0	40.0	24.5	43.2	33.2	35.6	27.9	44.8	-	44.1	38.2	38.2	39.8	42.1
pt	60.7	31.4	42.9	38.4	42.8	40.2	60.7	26.4	29.2	53.2	23.8	52.8	28.0	31.5	24.8	49.3	34.5	-	39.4	32.1	34.4	43.9
ro	60.8	33.1	38.5	37.8	40.3	35.6	50.4	24.6	26.2	46.5	25.0	44.8	28.4	29.9	28.7	43.0	35.8	48.5	-	31.5	35.1	39.4
sk	60.8	32.6	39.4	48.1	41.0	33.3	46.2	29.8	28.4	39.4	27.4	41.8	33.8	36.7	28.5	44.4	39.0	43.3	35.3	-	42.6	41.8
sl	61.0	33.1	37.9	43.5	42.6	34.0	47.0	31.1	28.8	38.2	25.7	42.3	34.6	37.3	30.0	45.9	38.2	44.1	35.8	38.9	-	42.7
sv	58.5	26.9	41.0	35.6	46.6	33.3	46.6	27.4	30.9	38.9	22.7	42.0	28.2	31.0	23.7	45.6	32.2	44.2	32.7	31.3	33.5	-

(using the Acquis corpus)

[from Koehn et al., 2009]

This table gives the best performance obtained for 462 pairs of official languages of the European Union (lacking Irish Gaelic), in terms of their BLEU score (the higher the score, the better the translation, a human translator scoring around 80). The best results correspond to the languages that benefit from research efforts in coordinated programmes, and from the availability of many parallel corpora (English, French, Dutch, Spanish, German,...), the worst are languages that have not seen similar efforts, or that are very different from other languages (Hungarian, Maltese, Finnish ...).

Referring to the initial issues, we can pick up the two key elements that are necessary for a language technology policy: the availability of monolingual resources and technologies in each language, in order to ensure the preservation of culture (and therefore of languages) and, at the same time, the availability of crosslingual resources (such as parallel corpora) and technologies for each pair of languages to be processed, in order to enable communication between humans.

But there is also an interest in developing concurrently monolingual technologies for each language in order to better address crosslingual technologies. This facilitates the coordination of efforts: standards for data exchange and tools, feedback of experiences, collections of Best Practices, and it is a necessity for applications such as speech translation (speech recognition in the source language, translation, then speech synthesis in the target language) or retrieval of crosslingual information (in order to produce a summary of information that has been found, regardless of the source language), or, more generally, for the localization of documents, which requires both crosslingual technology (translation...) and monolingual (such as spelling and grammar checkers...). And it also facilitates a shared effort between various laboratories around the world, working too often mainly on their national language, or on English only.

THE DIGITAL DIVIDE AND LANGUAGE COVERAGE

There is currently a two-speed situation and a “digital divide” between languages for which technologies exist, and others. This is related to the “weight of languages”⁷ [GASQUET-CYRUS, PETITJEAN 2009] It should be noted that 95% of languages are spoken by only 6% of world population. Some

7 See in this book: Daniel Prado, *Language Presence in the Real World and Cyberspace*.

linguists believe that 90 % of languages will have disappeared within a century. We can therefore classify languages according to the data and automatic processing systems that exist for these languages: whether they are well, less or not at all “resourced”, or indeed if they have only an oral tradition and no writing system at all. The availability of data is crucial for the development of usable systems, often based on statistical approaches. Machine translation therefore requires parallel corpora, whose number is reduced. Therefore we try to overcome this gap by developing methods using noisy parallel corpora, comparable corpora (texts dealing with the same topic in different languages) or quasi-comparable corpora, which are more readily available, thanks especially to the extension of the Web.

In order to resolve this digital divide, how can we take into account “minority” languages, regional languages, languages spoken by migrants, foreign or regional accents? Who bears the cost when these languages are of no economic or political interest, or are unrelated to armed conflicts or natural disasters that justify addressing them? How to ensure that citizens in a community of states are able to communicate among themselves? How to reduce the risk of conflicts and crises by allowing exchanges between people? This is now a major social and political issue, which is the subject of much debate. Thus, the International Forum of Bamako, organised in January 2009 in pursuit of the outcomes of the World Summits for the Information Society in Geneva (2003) and Tunis (2005), concluded on a commitment to promote an ethical use of information in its linguistic dimension, allowing mother tongue education and ensuring the existence of a multilingual cyberspace, both in terms of content availability on the Web and of technologies to access it.

RESEARCH EFFORTS IN THE DOMAIN

To produce the language resources and technologies that are needed to address multilingualism, different initiatives can be identified:

- those of big companies like Google or Microsoft;
- national programmes in some countries, with different objectives: to process an internal multilingualism (tdil in India, nhn in South Africa); to understand foreign languages for geopolitical reasons (gale or ears in the United States, funded by the Department of Defense – darpa); to ensure the use and promotion of a national or transnational language

- (TechnoLangue for French, stevin for Dutch/Flemish); or to maintain a place in an economic and cultural competition (Quaero in France);
- efforts to support R&D programmes of the European Commission;
 - international efforts to network the actors of the field, to better coordinate activities and promote greater sharing of resources (Oriental Cocosda, Clarin, FL&ReNet, META-NET...) and the establishment of distribution agencies for linguistic resources, such as ldc in the United States or elra in Europe.

These various initiatives to address multilingualism have their advantages and drawbacks: sustainability, links with the scientific community, links with existing applications, quality control...

Producers of Information Technology

First, it must be underlined that large U.S. companies in the information technology sector make a major effort in multilingualism and crosslingualism. Thus, the Google search engines work in 145 languages (national and regional), and Google has made available “free” tools for machine translation and crosslingual information retrieval online: in April 2011, 52 languages (including Catalan and Galician) and 2,652 language pairs were available on the internet, and 58 languages and 3,306 language pairs were available on smartphones (including 16 languages with voice input, and 24 languages with voice output). The Google Book Search Library contained 7 million documents in 44 languages and in December 2010 Google provided statistics on the evolution of human language from a corpus of 500 billion words (including 361 billion words in English and 45 billion words in French and Spanish). Also Microsoft provides the MS Word spell checker in 126 languages (233 if we consider regional variants) and a grammar checker in 6 languages (61 if we consider regional variants).

National programmes addressing the issue of language technologies to help multilingualism: TDIL in India, NHN in South Africa

Major programmes were launched as part of public policy. The TDIL⁸ programme (Technology Development for Indian Languages) is an important programme, which is one of ten priorities of the Indian national programme on the information society. The target is to process (Indian) English and eighteen “recognized” Indian languages⁹, with several language technologies: machine translation, Text-To-Speech synthesis, speech recognition, search engines, optical character recognition (OCR), spelling checkers, language resource production; all this for the group of nineteen languages. A comparable programme (NHN¹⁰: National Human Language Network) is taking place in South Africa for the automatic processing of the eleven national languages¹¹.

TechnoLangué: a programme for processing the French language

In France, TechnoLangué [CHAUDIRON, MARIANI, 2006]¹², conducted from 2002 to 2006, was a national programme aimed at producing language resources (monolingual, specialised and bilingual dictionaries, lexicons, corpora, databases of terminology, and language processing tools, etc.) and at conducting evaluation campaigns for written and spoken language processing. Different campaigns have been conducted for the processing of French, on parsing, on automatic extraction of terminology, on search engines that provide answers to questions (Q&A), on text-to-speech synthesis, on spoken dialogue, and on the transcription of speech (for the automatic indexation of radio or television broadcast). In this framework, an important corpus was produced with 1,600 hours of speech, including 100 hours of transcriptions, that represents a million words and 350 registered speakers. A corpus of this size had not previously existed

8 <http://tdil.mit.gov.in>

9 Assamese, Bengali, Gujrati, Hindi, Kannada, Kashmiri, Konkani, Malayalam, Manipuri, Marathi, Napali, Oriya, Punjabi, Sanskrit, Sindhi, Tamil, Telegu, Urdu.

10 <http://www.meraka.org.za/nhn>

11 Afrikaans, (South African) English, isiNdebele, isiXhosa, isiZulu, Sepedi, Sesotho, Setswana, SiSwati, Tshivenda, Xitsonga.

12 <http://www.technolangué.net>

for languages other than American English. It was therefore important to establish one for the French language, and it remains important to do the same for most languages in the world if we are to develop systems that process those languages automatically with a sufficient quality. TechnoLangue also conducted two evaluation studies on crosslingual technology. One on the agreement of parallel texts, first between French and English, German, Italian and Spanish, and secondly, between languages having different alphabets : French and Arabic, Mandarin, Greek, Japanese, Persian and Russian. Finally, an evaluation study was conducted on the automatic translation between English and French and between Arabic and French, including a study of the evaluation metrics employed in machine translation.

QUAERO : a French programme for processing multilingual and multimedia documents

The QUAERO¹³ programme was launched in France in May 2008. It covers the processing of multilingual and multimedia documents. The programme is structured around the development of about thirty technologies involving different media (text, speech, image, video, music...) which meets the needs of a group of five different applications (digitisation platform; media monitoring and social impact; personalised video; search engines; communication portals). It is based on the use of corpora and of systematic performance assessment. It is expected to handle more than twenty languages. This programme, consisting of 26 public and private partners, has a budget of 200 million euros, with a 100 million euros public funding provided through the OSEO agency, over five years (2008-2013). Initial results have succeeded in the audiovisual area¹⁴ (radio, television, online video...), in the Voxlead search engine, working in six languages (English, French, Spanish, Arabic, Mandarin and Russian) and developed by Exalead; in an aggregator of plurimedia news (text, radio, television) developed by Orange; or in a system to read e-books developed by Jouve.

13 <http://www.quaero.org>

14 <http://voxaleadnews.labs.exalead.com>

Actions of the European Union

From 2007 to 2010 the European Union benefited from having a commissioner specifically for multilingualism¹⁵, who established a High Level Group on Multilingualism that produced a report¹⁶, and who made a presentation to the Parliament and the European Council in September 2008¹⁷. As President of the European Union, France in September 2008 organised the *États-Généraux du Multilinguisme* (Multilingualism Summit) at La Sorbonne (Paris), that was followed in November 2008 by a resolution of the European Council of Ministers on multilingualism, taken up by the European Parliament in March 2009¹⁸. The idea of a “Single European Information Space” was highlighted.

The European Commission has supported several important projects on multilingual technologies under the 6th Framework Programme for Research and Development (CLEF, TC-Star, CHIL, etc.). In particular, the TC-Star¹⁹ Integrated Project covered speech translation in three languages: English, Spanish and Chinese, through an application performing automatic translation of the speeches at the European Parliament. Working in this context is very interesting because all the necessary resources exist at the European Parliament: members speeches in their own language, their (speech) interpretation in the different languages of the Parliament, their transcription into written form, and the translation of the transcripts in the different official languages. Thus, these data allow for training the automatic interpretation systems, including recognition in the source language, translation from the source language to the target language, and speech synthesis in the target language, thus utilising both monolingual and crosslingual technologies. A demonstration of a system for the English-Spanish language pair is available on line²⁰. TC-Star has also produced and distributed a report in five languages on Language Technology in Europe [LAZZARI, STEINBISS, 2006]²¹.

15 http://ec.europa.eu/commission_2004-2009/orban/index_en.htm

16 http://ec.europa.eu/education/policies/lang/doc/multireport_en.pdf

17 <http://europa.eu/rapid/pressReleasesAction.do?reference=IP/08/1340&format=HTML&aged=0&language=EN&guiLanguage=en>

18 <http://www.europarl.europa.eu/sides/getDoc.do?type=TA&language=EN&reference=P6-TA-2009-0162>

19 <http://www.tc-star.org>

20 See “Demo JM.asf” on the website http://audiosurf.org/demo_video

21 http://www.tc-star.org/pubblicazioni/D17_HLT_ENG.pdf

In the seventh European Framework Programme, FP7 (2007-2013), this area is mainly conducted by the “Language Technology, Machine Translation” Unit. In addition to R&D projects, an infrastructure and two networks have been established: CLARIN (Common Language Resources and Technology Infrastructure)²², FLareNet (Fostering Language Resources Network)²³, and META-NET (Multilingual Europe Technology Alliance)²⁴.

CLARIN is an infrastructure supported by the programme ESFRI (European Strategy Forum on Research Infrastructure) of the European Commission. Its objective is the distribution of language resources and tools for Humanities and Social Sciences.

FLareNet is a Thematic Network supported under the e-Content European Programme, with a budget of €0.9 million over 3 years (2008-2011). Its purpose is to serve as a think tank for the promotion of language resources in European programmes.

The META-NET Network of Excellence was established within the T4ME (Technologies for a Multilingual Europe) project. This project has a budget of €6 million over a period of 3 years (2010-2013) and is structured in three parts:

- pushing the research frontiers in machine translation;
- establishing an Open Resources Infrastructure (META-SHARE), including the production, annotation, standardisation, validation and distribution of language resources, and the evaluation of language technologies;
- to conduct a reflection on the place of multilingual technologies in the context of drafting a Strategic Research Agenda for the next Framework Program (2014-2020).

EUROPEAN AND INTERNATIONAL PERSPECTIVE

The resolutions of the European authorities demand a major effort to process all European languages, national and regional. However, if one considers the number of languages or language pairs that are to be

²² <http://www.clarin.eu>

²³ <http://www.flarenet.eu>

²⁴ <http://www.meta-net.eu>

addressed, and multiply it by the number of technologies, we see that the size of the effort is probably too large for the European Commission alone. It would therefore be interesting to share this effort among Member States, or Regions, and the European Commission, in perfect harmony with the “principle of subsidiarity”.

Language technologies are well suited for a joint effort. The European Commission would have the primary responsibility for overseeing and ensuring coordination of the programme (management, provision of standards, technology evaluation, communication...) and of developing core technologies around language processing. Each Member State would have as a priority to ensure the coverage of its language(s): to produce the language resources essential for the development of systems (corpora, lexicons, dictionaries), and to develop or adapt technologies to the specificities of its language(s). This model would be easily adaptable to an international effort, combining the efforts of the participating countries and of international organizations.

Unfortunately, until now the topic of Language Technologies has been regrettably considered just as one research area among many others in Europe, not as an essential element of the the European construction, requiring a high priority effort to handle the corresponding issues. This weakness is all the more dangerous given the liveliness of the European Union and its needs to increase economic, informational and cultural exchanges between countries, and to address the citizens of each State and help them in their communication. Let's hope that the political awareness of the issues attached to multilingualism will see research in language technologies receive adequate attention in future Framework Programmes.

CONCLUSION

Language technologies are a major tool to facilitate multilingualism in Europe as well as in the rest of the world. To achieve this, we need to agree to coordinate the efforts of States, even regions, and international organizations (European Commission, United Nations, Unesco, the African Union, etc.), involving industry and public research laboratories. Care should be taken to produce for each language the language resources

needed, and organise the research effort in an open way, based on the interoperability and objective benchmarking of technologies.

We could then add a nod to the famous phrase of Umberto Eco, saying: “Translation is the language of Europe... with the support of technology”, and extend this assumption to the global village.

BIBLIOGRAPHY

[CENCIONI, ROSSI 2008] R. Cencioni, K. Rossi. Language based Interaction, *EC-ICT Conference*, Lyon, 26 Novembre 2008.

[CHAUDIRON, MARIANI 2006] S. Chaudiron, J. Mariani. Techno-langue: The French National Initiative for Human Language Technologies (HLT), *Proceedings LREC'06*, Genoa, Italy, May 2006.

[ECO 1993] U. Eco. *La langue de l'Europe, c'est la traduction*, Assises de la traduction littéraire, Arles, 1993.

[GASQUET-CYRUS, PETITJEAN EDS 2006] M. Gasquet-Cyrus, C. Petitjean eds. *Le poids des langues*, L'Harmattan, 2009.

[KOEHN, BIRCH, STEINBERGER 2009] Ph. Koehn, A. Birch and R. Steinberger. 462 Machine Translation Systems for Europe, *Machine Translation Summit XII*, p. 65-72, 2009.

[LAZZARI, STEINBISS 2006] G. Lazzari, V. Steinbiss. *Human Language Technologies for Europe*, TC-Star Report, April 2006.

[MARIANI 1995] J. Mariani, ed. *Evaluation chapter in Survey of the State of the Art in Human Language Technology*, R. A. Cole, J. Mariani, H. Uszkoreit, N. Varile, A. Zaenen, A. Zampolli, V. Zue eds., Cambridge University Press, 1995.

[PAPINENI, ROUKOS, WARD, ZHU] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In: *Proceedings of the 40th Annual Meeting of ACL*, Philadelphia, PA.

THE USE OF FACEBOOK BY THE ETON OF CAMEROON

Opened in 2006, Facebook currently has hundreds of millions of users worldwide, who use several dozen languages. Taking the example of Facebook groups run by and for members of the ethnic group Éton (Central Cameroon, approximately 250 000 speakers), we can observe how the use of the latest communication technologies enables new ways to affirm traditional culture.

Original article in French.

Translated by Laura Kraftowitz.



VASSILI RIVRON is an anthropologist, lecturer at the University of Caen Basse-Normandie. He is attached to CERReV (UCBN) and associate researcher at the École des Hautes Études en Sciences Sociales and at CSU (CNRS)

VASSILI RIVRON

THE USE OF
FACEBOOK BY
THE ETON OF
CAMEROUN

Facebook, a “social network” launched in 2006, has gained hundreds of millions of users worldwide. The site interface is available in dozens of languages (including regional languages like Basque and artificial languages like Esperanto), and uses multiple writing systems, alphabetical or other. It allows for diverse uses, among which one finds projections onto the network web of pre-existing cultural practices and references, but also of cultural innovation.

By considering the example of several “Facebook groups” that are run by and for members of the ethnic group the Eton (located in Central Cameroon, counting approximately 250,000 speakers), we can observe how the use of recent communication technologies has enabled new forms of affirmation of traditional cultures. One observation is the extension of a mother tongue outside its habitual context and uses, and the correlated development of its graphic system.

In a country with two official languages (English and French) and over two hundred local languages, the Eton¹ language appeared as though it would remain, as many others, confined to its specific geographical and social context. These specific contexts are the inhabitants of the Lékié department, and domestic space, family circles, and traditional hierarchies for the urban migrants and emigrants; and all this principally within the context of a interpersonal communication (including the telephone), or at least in the presence of interlocutors (i.e. speeches).

Some Cameroonian languages are or have been written (including by specific graphic systems). Some were taught during the period of the German protectorate (1884-1922), before this was abolished under French colonization. Proposals to teach the local languages in public schools have

1 A tonal Bantu language, located to the North of the Beti/Bulu/Fang linguistic complex.

periodically made their way onto the national agenda since independence (1960), but have been blocked by a republican imaginary that fears tribal schisms. Eton thus continues to enjoy a primarily oral existence, its graphic system remaining uncodified. Its alphabet appears mainly in scholarly works: ethnographic transcriptions, linguistic studies, bilingual folklore collections and oral literature compilations. It also serves as a mnemonic resource to its users for personal annotations (journals, sayings, notes) or in passages of collusion or scholarship embedded within correspondence that is written principally in another language². The lack of education and the republican principles inherited from France explain why regular readership in Eton is impossible (in print media, books, government, and politics).

From the very beginning of our study (2004), we were able to observe an extension of Eton written on the web, mainly in two contexts. One was the creation of a number of cultural heritage sites (on folklore, culture, language, ethnic group or regional history), which were often individual initiatives. On these sites, Eton might appear in fragmentary lexicons, or collections of sayings, for example. These sites rarely addressed themselves directly to the public in written Eton, but rather transcribed or taught the language's oral form. For obvious reasons, written dialogue in Eton appeared mainly in blogs and forums, and were especially prominent in the comment sections of YouTube videos by local artists and Cameroonian online news sites. But in both cases, the writing rarely exceeded two sentences without being translated and often boils down to illustrations, posing questions, witticisms or talk of complicity (hence excluding non-speakers).

The emergence of social networks (i.e. Facebook) has prompted the development of a wide variety of communities ("groups"), which generally overlap individual entities (photo and personal data) with cooptation procedures ("friend requests"), and which provide access to "profile" or "group" contents (photos, texts, videos, games, etc.) that can be shared and customized. For our purposes, we highlight several interesting Facebook groups: *The Etons* (ethnic referent, 53 members), *Sons and Daughters of Lékié* (territorial referent, 220 members), and *Ongola –Fang-Bulu-Beti*

2 For more on techniques and languages of correspondence with the diaspora, see Sayad, Abdelmalek, "Du message oral au message sur cassette: la communication avec l'absent", *Actes de la recherche en sciences sociales*, n°59, 1985, pp. 61-72.

Culture (enlarged cultural referent, 1,445 members). Apart from transposing various aspects of the groups' social structure and cultural practice onto a digital resource, we also find forms of writing that are evolving in comparison to our initial observations. Among the many topics covered in these Facebook groups, a significant portion is devoted to generic issues relating to "traditional" culture, marriage, parenthood, initiations and sayings. The question of language is rarely specifically addressed.

Sustainable exchanges are established in the social networks, using different registers. French and English are clearly dominant for generic substance (group description and instructions), or for comments and dialogues themselves. But at the heart of personal messages, profile "walls", and discussion forums, Eton may be used for section titles, opening and continuing a debate (other languages may also be included).

There seems to be less hesitation to write in Eton in the "among friends" atmosphere of Facebook groups, where mutual understanding is postulated despite the absence of codification and official teaching of customary spelling. The inequality of language skills is evident in these exchanges, along with the diversity of resources mobilized to design graphic solutions (in Francophone, Anglophone, and Ewondophone environments³).

Efforts to make themselves understood in writing in this tonal language, for example to decrypt messages by sounding them out, are visible in exchanges and during our direct observations of users. The pleasure of this activity is also clear; it is an expression of solidarity and cultural pride.

Of interest to the researcher is that this textual corpus is supplemented by "profiles" that provide valuable information on active group members. Far from being a populist trend, the reinvestment and movement of resources maintained for "traditionals" through the internet are often the doing of cosmopolitan literati, if not expatriates. In the aforementioned cases of Facebook groups, we found that the founders and moderators/facilitators were, respectively: a white Cameroonian, a Cameroonian expatriate to the United States, and a Cameroonian intellectual. It could not be otherwise, because of the simple fact that Facebook cannot be used on many of the outdated computer hubs in Cameroon. But beyond that, and as we can see in the codification of European popular culture

3 Ewondo is a language spoken in the southern part of Cameroon, especially in the capital, Yaounde.

by the folklorists of the nineteenth and twentieth centuries, the desire and ability to acquire technological resources and to transpose linguistic resources from one system to another, as from oral to writing, are socially determined and enshrine the decisive role of the “cosmopolitan elites” and especially those who “join the other side” (e.g. the diaspora). Such a detour is likely necessary to ensure the vitality of an oral language within the written context of new technologies.

BIBLIOGRAPHY

- [ABÉLÈS 2008] Abélès, Marc, *Anthropologie de la globalisation*, Paris, Payot, 2008.
- [AMSELLE 1999] Amselle, Jean-Loup et Mbokolo, Elikia (dir.), *Au cœur de l'ethnie*, Paris, La Découverte, 1999.
- [BOURDIEU 1994] Bourdieu, Pierre: « Esprits d'État – Genèse et structure du champ bureaucratique », In: *Raisons Pratiques*, Seuil, Paris, 1994, pp 99-135.
- [GOODY 1994] Goody, Jack P., *Entre l'oralité et l'écriture*, PUF, Paris, 1994.
- [GUICHARD 2003] Guichard, Eric, « Does the 'Digital Divide' Exist? », In: *Globalization and its new divides: malcontents, recipes, and reform* (dir. Paul van Seters, Bas de Gaay Fortman & Arie de Ruijter), Dutch University Press, Amsterdam, 2003.
- [GUYER 2000] Guyer, Jane I., « La tradition de l'invention en Afrique équatoriale », *Politique africaine*, n°79, octobre 2000, pp.101-139.
- [SAYAD 1985] Sayad, Abdelmalek, « Du message oral au message sur cassette : la communication avec l'absent », *Actes de la recherche en sciences sociales*, n°59, 1985, pp. 61-72.
- [THIESSE 1999] Thiesse, Anne-Marie, *La création des identités nationales (Europe XVIII^e-XX^e siècle)*, Seuil (coll. Univers Historique), Paris, 1999.
- [VAN VELDE 2006] Van Velde, Mark, *A description of Eton: phonology, morphology, basic syntax and lexicon*, thèse de doctorat, 2006.

PANN YU MON
& MADHUKARA PHATAK

SEARCH ENGINES AND ASIAN LANGUAGES

Although many search engines are available in the languages most used in the digital world, they do not work when dealing in less computerized languages. In recent years, the number of non-English resources on the Web has grown rapidly, especially in Asian languages. This article raises the difficulties faced by search engines in this situation.

Original article in English.



PANN YU MON holds a PhD from the Department of Management and Information Systems Engineering of Nagaoka University of Technology, Japan. Her research interests are indexing, archiving and Web requests.



MADHUKARA PHATAK holds a Bachelor of Engineering in computer science from JSSATE, India. His research interests are cloud computing and distributed systems.

PANN YU MON
& MADHUKARA PHATAK

SEARCH ENGINES AND ASIAN LANGUAGES

In today's world, search engines play a critical role in retrieving information from the borderless Web. Although many search engines are available in major languages, they are not functional when it comes to less computerised languages. Over recent years, the number of non-English resources on the Web has been growing rapidly and it has been estimated that English is not the native language for more than 60 % of Web users. Even if the above numbers are not exact, it is clear that non-English language pages and users cannot be ignored. Although current popular search engines work on non-English queries, it is just pattern matching, the sequence of symbols entered by users appears somewhere in the web document, but more sophisticated method, based on natural language analysis of the language¹, are not used (e.g. dealing with stemming, word breaking, stop words retrieval, etc.).

For the convenience of users speaking different languages, Google has developed more than 136 interface languages, and proposes around 180 local search engines. Among these only 20 % are dedicated to Asian languages. More than half of all web pages use Asian languages. Some articles were written about the difficulties of search engine queries in western languages; but only a few articles were focused on Asian language queries. Our main aim was to discuss the additional problems faced in non-English web queries and to suggest techniques to improve the response of searching systems. In this paper, we study the difficulties met by search engines when they handle queries for Asian languages. We give examples using five different languages or language families: Indian, Malaysia, Myanmar, Indonesian and Thai.

1 See in this book: Joseph Mariani, *How Language Technologies Support Multilingualism*.

INTRODUCTION

Search engines have to crawl billions of web pages in order to index the constantly changing hypertext, which contains information in a variety of languages and all sorts of formats. The size of the Web is growing exponentially and the number of indexable pages² on the Web is considered to be near one hundred billions pages. It has become more and more difficult for search engines to keep an up-to-date and comprehensive search index, resulting in low precision and low recall rates. Users often find it difficult to search for useful and high quality information on the Web using general-purpose search engines, especially when searching for information on a specific topic or in a non-English language. Search engines should also support users who globally have different computer handling abilities, cultural backgrounds and most importantly, who speak different languages. The majority of current popular search tools supports only English and ignores the diacritics and special features of non-English languages. One search tool alone cannot be perfect for all languages. For that reason, search engines need to be localised in a local language. Many domain-specific or language-specific search engines have been built to facilitate more efficient searching in different areas. Thus, the main aim of this article is to figure out the different kinds of difficulties that search engines encounter when handling Asian languages queries. Although comprehensive software tools enabling the creation of search engines exist, most of them cannot function with non-English languages such as various European, Asian and Middle Eastern languages.

The major modules of a Web search engines are

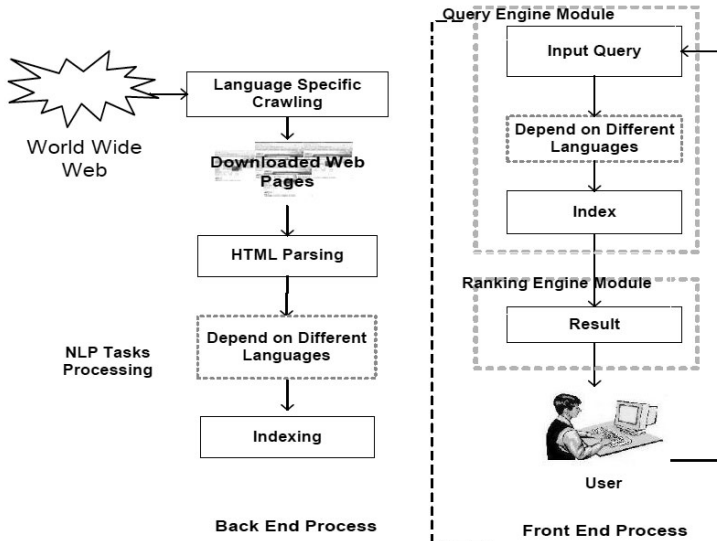
- Crawler;
- Natural Language Processing Module (NLP);
- Indexer;
- Query Engine Module;
- Ranking Engine Module.

A crawler is small program that browses the World Wide Web and downloads web pages. These programs are given a starting set of seed URLs to visit, from which they copy the pages and identify other hyperlinked

2 The surface web (also known as the visible web or indexable web) is that portion of the World Wide Web that is indexed by conventional search engines.
(wikipedia - http://en.wikipedia.org/wiki/Surface_Web).

URLs to be visited. In order to implement the Language Specific Search Engine, we need a small part of the World Wide Web which is related to our interest.

Figure 1. Architecture of General Web Search Engine



In order to download only the interesting parts of the World Wide Web, specific crawling criteria are needed. The next step is HTML parsing. It is relatively easy. The work after parsing is the NLP processing tasks. Non-English web pages are complicated in this step. The scope of this module varies depending on the language. It includes transliteration, word tokenization, stemming, preprocessing on input compound words, stop words removal and so on.

The next module is an indexer module. This module extracts all the words from each page and records the URL where each word occurs. The result is a generally very large mapping of URLs to pages where a given word occurs. It includes the tasks such as transcoding, word breaking, stemming and stop words removal.

The query engine module is responsible for receiving and answering users search requests. The task given to the ranking engine is to sort results so that highly ranked results are presented at the top of the list. All search

engine modules have the same function for different languages except the NLP module, which depends on the specific features of the language.

PROBLEM IN EACH LANGUAGE

In this section, we will explain the tasks of the NLP processing module. These tasks may vary depending on each different language. Here, we will figure out the different kinds of task by giving examples based on different kinds of language families.

Encoding handling

One issue that should be taken into account during indexing is the existence of different encodings for the Web documents. This is especially relevant in the case of Asian languages. Here, we give examples for Indian and Myanmar languages. Indian languages have different encodings for different languages. But for the Myanmar language, there are different kinds of encodings for just one language. Following are the detail explanation of each language.

Indian Language

More than 95 % of Indian language content on the web is not searchable due to multiple encodings of web pages. Most of these encodings are incompatible and hence need some kind of standardisation for making the content accessible via a search engine³.

Indic scripts are phonetic in nature. There are vowels and consonant symbols. The consonants become a syllable after the addition of a vowel sound to it. Further to compound the problem there are “compound syllables” also referred as ligatures. For instance, if we consider “tri” in “triangle”, there are three letters corresponding to three sounds “ta”, “ra”, “yi”. But in the case of Indic Scripts the three are built together to make a single compound consonant having a non-linear structure unlike Latin based languages⁴.

³ See in this book: Stéphane Bortzmeyer, *Multilingualism and the Internet's Standardisation*.

⁴ Prasad Pingali, Jagadeesh Jagarlamudi, Vasudeva Varma, “WebKhoj: Indian language IR from Multiple Character Encodings”, In: *WWW '06 Proceedings of the 15th international conference on World Wide Web*, 2006. <http://dl.acm.org/citation.cfm?doid=1135777.1135898>

In India, many languages use the same script called Devanagari script. So language detection becomes more complex. For example the query “मधुकर” means “honey pot” in Hindi but means “rifle” in Oriya.

Webkhoj is a search engine, which gives users the choice to search in ten different Indian languages. The engine supports Hindi, Telugu, Tamil, Malayalam, Marathi, Kannada, Bengali, Punjabi, Gujarati, and Oriya. In order to search Indian language websites, Webkhoj transliterates all the encodings into one standard encoding (Unicode/ucs) and accepts the user’s queries in the same encoding and builds the search results.

Myanmar Language

Myanmar language uses various encodings. The problem is that when a user puts the keywords in one specific encoding, the search engine searches for pages using the same encoding as the user query. So some pages that are written in different encodings may be missed. The problem being that some pages, which are relevant, may be excluded in the result due to different encodings.

Several alternative ucs/Unicode encodings have also been implemented to encode Myanmar web pages by different groups of people. These can be divided into three groups.

Graphic encodings: Actually it has been pretending to be English (technically Latin 1 or Windows Code Page 1252) fonts and is substituting Myanmar glyphs to English Latin glyphs. This means that they are using the code point allocated for the Latin alphabet to represent Myanmar characters.

Partially followed ucs/Unicode encodings: These kinds of encoding have different mappings and none of these follows the Universal Coded-character Set (ucs) or Unicode standard. They partially follow the ucs/Unicode standard but they are not yet supported by Microsoft and other major software vendors.

ucs/Unicode encodings: These fonts contain not only Unicode points and glyphs but also the Open Type Layout (OTL) logic and rules.

Some Myanmar Web Pages are made by using the so-called Mixture Encoding Style format. It is the mixture of ucs/Unicode code points and HTML-entity, like `ଠံ`; `ଠြ`; `ଠ္`; `ଠƞ`; `ଠ္`;

မက္ (သံလွင်ဒဿိမ်မက်). These are coded in decimal value. For these kinds of web pages, the HTML-entity should be converted to UCS/Unicode point by converting the decimal values to hexa-decimal values. Some of the web page publishing software automatically encodes Myanmar words in Mixture Encoding Style format. For this current popular search engines cannot search Myanmar words properly.

Word segmentation issues on input keyword

Each language has its own characteristics for word segmentation, so special attention needs to be paid to the segmentation method in the indexing process. Appropriate segmentation is still a problem for search engines.

Myanmar Language

Word segmentation is even harder in Asian Languages such as Chinese or Myanmar, since words are not segmented by spaces. Foo and Li (2004) conducted experiments to study the impact of Chinese word segmentation on Information Retrieval (IR) effectiveness. Accuracy varied from 0.34 to 0.47 (on a scale going from 0 to 1) depending on the segmentation method.

Similarly we compare English queries with one of the Asian languages (Myanmar). In the case of English language search, it is the common practice for search engines to return those pages which contain a sub-set of the words used in the user's query. For instance, when a user enters the words "chocolate ice-cream", the search engine returns not only the web pages that includes the exact phrase "chocolate ice-cream" but also returns those pages with the words "chocolate" and "ice-cream" alone. It is made possible by tokenization of the input query. In contrast to this, when searching for Myanmar words in a general search engine, it behaves like a "phrase search" in English language search. That is equivalent of putting double quotes (" ") around the query, telling search engines to consider the exact words in the exact order without any changes. For example; document A contains a Myanmar compound word XYZ. And another document B contains every components of the word XYZ in non-consecutive manner, like "X...Y...Z". If a user searches a query XYZ, the search engine retrieves document A, but does not retrieve document

B, because word segmentation is not done by the search engine. That's why special treatment is needed for the Myanmar language in a search engine. It requires the word segmentation process both at the indexing stage and in the input keyword processing stage.

For the majority of languages such as Chinese or Japanese, it performs the work breaking process properly. In the following, the authors want to give a comparison between the operation of search engines on majority and minority languages. We use Japanese as an example for majority languages and Myanmar as a minority language.

Figure 2. Manner of Search Engine on Majority Languages

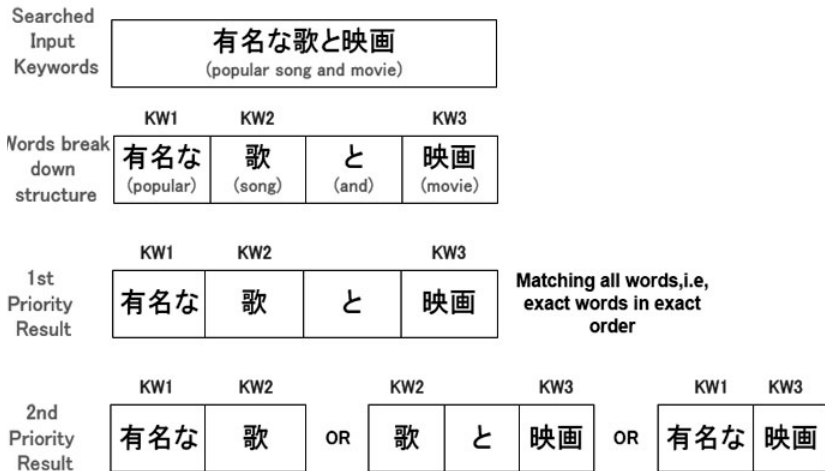


Figure 3. Operation of Search Engines on Minority Languages

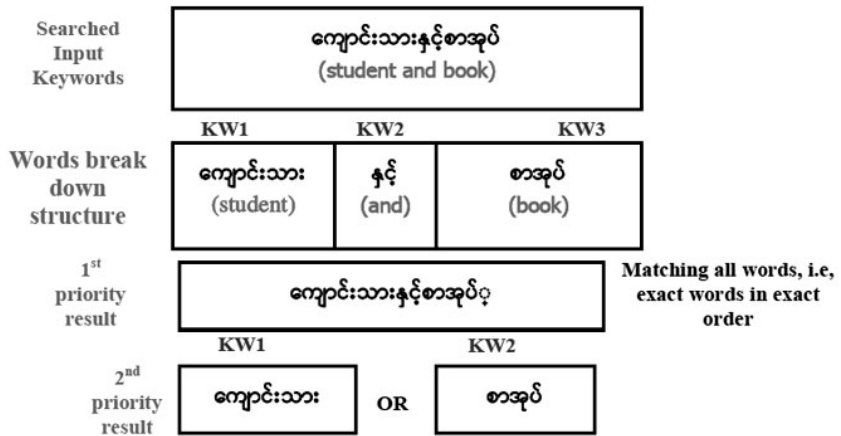


Figure 4. Google Result for keyword “ကျောင်းသားနှင့်စာအုပ်”



As shown in figure 2, for the majority languages, search engines return the exact phrase as first priority. For the second priority, the web pages that included segmented words are retrieved. It is shown that search engines work properly for the majority of languages.

Figure 3 shows the operation of a search engine on a minority language. When a user searches for “ကျောင်းသားနှင့်စာအုပ်”, the result should be the pages that include “ကျောင်းသား” or “စာအုပ်”. Instead, it searched for the words that are exactly the same as the input query. If there are no pages that included the exact words in the exact order, they simply say that “Your search “ကျောင်းသားနှင့်စာအုပ်” did not match any documents” as shown in Figure 4. It is obvious that most general search engines do not implement the natural language processing task for minority languages.

Thai Language

A recent study on the Web Search Engines on Thai queries shows that word segmentation is still challenging. According to “Evaluation of Web Search Engine with Thai Queries” by Virach Sornlertlamvanich *et.al.*⁵, when the word “ข้าว” is submitted to several search engines, most of them returned results that do not contain that word. This shows that most search engine do not handle Thai language word segmentation properly.

Compound words

Compound words are another major issue for the indexing of web pages in major search engines. Some Asian languages make extensive use of compound words. Below we give the example of compound words in the Myanmar language as well as the Indian Kannada, and Thai languages.

Myanmar Language

In Myanmar; two simple words, ခဲ (lead) and တံ (rod, stick etc.) combine together to form a compound word ခဲတံ (pencil). Similarly, ကျန်းမာ (healthy) combines with ချော်ရွှင် (happy) to form ကျန်းမာချော်ရွှင် (healthy and happy). Although compound words are widely seen in every language and are not a specific feature of the Myanmar language, they differ greatly from one language to another.

⁵ Virach Sornlertlamvanich, Shisanu Tongchim and Hitoshi Isahara “Evaluation of Web Search Engines with Thai Queries” *Proceedings of Workshop on NTCIR-6 and EVIA-1, NII*, National Center of Sciences, Tokyo, Japan, May 15-18, 2007. <http://research.nii.ac.jp/ntcir/ntcir-ws6/OnlineProceedings/EVIA/15.pdf>

Indian Language

Similarly with the Indian Kannada language, two simple words “ಸು” (good), and “ಆಲೋಚನೆ” (thinking) combine to become a compound word “ಸುಲೋಚನೆ” (good thinker).

Thai Language

According to the study made by Virach Sornlertlamvanich *et. al.* (*op. cit.*) some Thai queries are indivisible units although each query can be considered as a set of words. For example, a query found in the query log is “กรมอุตุนิยมวิทยา” which is the “Thai Meteorological Department”. This word can be considered as two words: “กรม” (Department) and “อุตุนิยมวิทยา” (Meteorology). Since this word represents a unique entity, it may be recognized as an indivisible unit. There are also some queries that resemble to this word, but they are ill-written. For example, at least three queries that can be considered to refer to this word: “กรมอุต”, “กรมอุตุนิยม” and “กรมอุตวิทยา”. The use of these keywords usually leads to websites that have improper forms of the word “กรมอุตุนิยมวิทยา”, rather than the website of the “Thai Meteorological Department”.

Pre-processing of the different writing system

Different writing systems for the same word may occur. Here, we would like to give an example in the Myanmar language.

Myanmar Language

The Myanmar writing system has been strongly influenced by Pali and Sanskrit. In ancient times, words were written on a piece of stone in subscripted-form because of the limitations of space availability. Later, some of the subscripted-words had been changed to expanded-form, but they can still be written in subscripted form. Everybody can write those words freely, as convenient.

More complex, some Myanmar words can be written in different forms without omitting any character in both forms even though they do not belong to the same consonant group. The two forms have exactly the same meaning, the same pronunciation but different representations. An

example is “ထွင်း” in subscripted form. It can be expanded as “ထမင်း” (rice). Similarly, the word “သို့” is sometimes an abbreviation of “သမီး” (daughter). These words are not found as native Myanmar words, except for the purpose of abbreviation.

Hence, if those kinds of words are given as a query in a search engine, the expanded form should be treated as a phrase.

Stemming

To help search engines retrieval processes become more effective, one of the practices used is word stemming. In this case, morphological variants are different from one language to other.

Indonesian Language

The Indonesian language is morphologically rich. There are around thirty five standard affixes (prefixes, suffixes, circumfixes, and some infixes inherited from Javanese)⁶. In Indonesian language, affixes can be attached to virtually any word and they can be iteratively combined. The wide use of affixes seems to have created a trend among Indonesian speakers to invent new affixes and affixation rules⁷.

Malay affixes consist of four different types, which are the prefix, suffix, prefix-suffix pair, and infix. Unlike an English stemmer which works quite well just by removing suffixes alone to obtain the stems, an effective and powerful Malay stemmer must not only be able to remove suffixes, but also prefixes, prefix-suffix pairs, and infixes as well⁸. Without removing all these affixes, stems cannot be efficiently used to index Malay documents.

Myanmar Language

In the case of the Myanmar language, stemming focuses on the removal of inflectional suffixes, derivational suffixes, inflectional prefixes and derivational prefixes for a given Myanmar word. According to Myanmar

⁶ Kridalaksana, Harimurti., *Pembentukan Kata Dalam Bahasa Indonesia*. PT. Gramedia, Jakarta 1989.

⁷ Tim Penyusun Kamus, *Kamus Besar Bahasa Indonesia*. 2ed. Balai Pustaka, 1999.

⁸ F. Ahmad, *A Malay Language Document Retrieval System: An Experimental Approach and Analysis*, Universiti Kebangsaan Malaysia, Bangi, 1995.

grammar book, it seems there are ninety one different affixes for the main four word classes: verb, noun, adjective and adverb. A Myanmar language stemmer involves the straightforward removal of the affixes to get the correct stem. More details on a Myanmar language stemmer are given in one research thesis⁹.

The stemming algorithm is different for each language. The search engine should pay attention to each individual language.

CONCLUSION

Our conclusions are that it would be more effective if search engines took more account of the properties of individual languages, and that there is a need for more studies of real user behaviour in practical situations.

As the Web continues to become more multilingual, and as languages other than English continue to gain ground on the Web, the need to develop search engines to handle all these languages has become more apparent. It might be unrealistic at this point to suggest that search engines should be equipped with an entire linguistic toolkit capable of handling all languages, but a gradual progress towards this goal is not far-fetched. The developers of search engines should start looking at implementing the basic requirements of truly multilingual engines. Web pages published in languages that do not share the linguistic characteristics of English are more likely to be missed or improperly indexed by major search engines than English web pages.

Overall, it can be argued that the processing and searching of non-English text poses additional difficulties that are not faced in English texts. Search engines need to be localised, in a local language.

⁹ San Ko Oo, Yoshiki Mikami, *Development of Myanmar Language Stemmer*, Master thesis of Management of Information system engineering department, Nagaoka University of Technology, Japan, 2010.

DIGITAL LIBRARIES

How can we preserve cultures and traces of different languages in digital libraries? How can we add value through many translations of the same work so that users, especially the young, can understand the diversity and wealth of human thought? How to participate locally, with one's own language and culture, in the construction of a huge interconnected library, offering everyone access to the works of the entire world?

Original article in French.
Translated by Laura Kraftowitz.



HERVÉ LE CROSNIER is a senior lecturer at the University of Caen Basse-Normandie, where he teaches Internet technologies and digital culture. He is currently working with ISCC, the Institute for Communication Sciences of the CNRS. His research focuses on the impact of the Internet on social and cultural organization, and extending knowledge in the public domain. He is one of the founders of C&F éditions.

HERVÉ LE CROSNIER

DIGITAL LIBRARIES

From the Egyptian papyrus collections, to the clay tablets of Mesopotamia covered with cuneiform writing, whenever and wherever knowledge and culture have been able to be transcribed onto a carrier, documents have been found. The famous Library of Alexandria's loss to flames remains a founding moment for those invested in the transmission of knowledge so that future generations can benefit from the advances of the preceding ones. The desire to collect documents, organise them and make them available is one of the main preoccupations of scholars. When Europe witnessed the birth of movable type printing, the number of documents available to the public grew rapidly, which catalysed book exchange, and the need was perceived for legal deposit libraries, as a way to store and accumulate knowledge constituted in this way. The first sound recordings, reels, and later disks from the beginning of the twentieth century, were deposited into audio archives.

For oral languages, use and conservation go hand in hand: they are transmitted directly by their speakers. The entry of oral-only languages into libraries is recent, and owes its existence to the spread of audio and video recording¹. Within such spaces, multimedia technology encounters writing cultures whose existence and history have endured by being engraved onto a medium, and can inherit the organizational *savoir-faire* that has been constructed around the book.

Storing, organising and making available all records are the founding pillars of the institution of the library, which, by adhering to these fundamentals, has been able to incorporate each new process of recording knowledge and emotions.

1 See in this book: Tunde Adegbola, *Multimedia and Signed, Written or Oral Languages*.

Today, documents are witnessing an essentially “digital moment”, which is changing our approach towards their durability and transmission. Word processing is eclipsing the manuscript. Web pages are often regarded as services offering news or commerce, feedback and comments, too often in a continuous flow that moves contrarily to the accumulation logic of libraries. On the other hand, performances (concerts, local and global events, readings), and even quotidian life (through the proliferation of digital photography and home video) are recorded. The range of media is growing, allowing the traces left by cultural and scientific activities to be made permanent and transformed into audio, video or multimedia recordings.

Digital libraries are at this new crossroads, between the proliferation of new documents linked to digital media’s ease of production and distribution, and traditional libraries. In the present discussion, I will attempt to define digital libraries as distinguished from other forms of document access, and evaluate approaches and needs as they relate to multilingualism. Finally, I will expand upon the legal and technical constraints, as well as new cultural practices, that frame digital library activity.

LIBRARIES AND ARCHIVES

Traditionally, we have distinguished between three types of organizations :

- Libraries preserve and make available “duplicates”, that is, existing works in multiple copies that have been publicly released, usually by a publisher, and sometimes via reprography for reports or scientific papers (dissertations, “grey literature”). Before finding their place in the library, a selection for content is made by the editorial circuit; the materials entering the library are thus broadly homogenous in their approach to the writer-reader relationship, and by the industrial circuit facilitating the transmission of the work of the former towards the latter;
- Archives preserve the internal documents of a structure (company or institution) or individual (personal records). In general, archives deal with “unicates”, that is, documents existing in only one copy, managing them into groups that are organised by the archivist: files, archival boxes, or sets of bound documents. These groupings are often the only ones described in the catalogue, which is why archives

always contain as yet unknown documents, treasures for the curious researcher to discover. By extension, some have spoken of archiving the media, a term that considers an audio or audiovisual stream as consisting of a single copy, even if its vehicle for distribution (movies, series and music recordings) consists of multiple documents. Before the innovation of cloud computing, this stream was impossible to maintain in its entirety. Instead, archives use “sampling”, keeping only randomly selected examples to inform future historians of the archived period’s social practices;

- Museums also preserve unique objects, but accompanied by a tracking file that tells the object’s known history, including its various owners and any restorations made.

But these distinctions, which describe different approaches to documents, are becoming increasingly tenuous in the age of digitisation and the internet.

The digital document’s trademark feature is its duplication at a marginal cost close to nil. The main costs relate to its creation (the cost of the prototype, whether of an original work, or of an already existing work that is digitized), and infrastructure (from datacenters to user terminals, and the communication network that connects them). Yet many of these documents are never duplicated: blog pages, newspaper websites, online shopping catalogues, photographs uploaded to services such as Flickr or Picasa, and so on. The existence of a single, centralised access point allows website publishers to profitise production and infrastructure, either by selling advertising space, or through subscriptions. Libraries, with their collective and multi-centennial experience, distribute document storage to render their collections permanent, accessible, and in close proximity to readers. This makes the web as a whole operate more like an archive, consisting primarily of unicats despite the low duplication cost. One even speaks of “web archiving” to describe the process of copying onto high-capacity external hard drives samples extracted from websites at regular intervals – or at least copying their appearance at a given point in time, so that future readers can access not only textual content, but also the “materiality” of reading as it existed at the time of a document’s production.

So it is more than just an enigmatic recasting that generates the name “Internet Archive” for the main online copy registration service since 1995. The service does more than archive the Web itself; it also attempts to create an accessible online audiovisual “archive” (film collections, musical recordings and digitised books). In France, the web archiving division between two institutions emphasises its contradictory nature: the National Library of France (Bibliothèque nationale de France) conserves websites with the .fr² extension, dominated by “text” or “published” documents, while the National Audiovisual Institute (l’Institut national de l’Audiovisuel) conserves “broadcast” streams, whose numbers are increasing (online radio, television, and music websites, among others). The entire stream flowing through the web is considered a document for archiving. In the United States, the Library of Congress has an agreement with Twitter to “archive” all that is exchanged on that particular social network, that is, messages of up to 140 characters and the conditions surrounding their release (who publishes, to which “followers”, what is “re-tweeted”; in short, the author’s social graph image)³.

This allows us to measure the work required to save a trace of the present for future generations. The size of the Web always exceeds expectations. Documents deposited into this vast global network haven’t been organised by anyone, save the producers of information themselves. One now speaks of *autoritativité*⁴, literally “authorativity”, to describe this new phenomenon that puts authors in the position of being the sole decision-makers of a publication, for example by clicking their blog’s “publish” button. The editorial process, which transforms a document into a book or other publicly-accessible medium, has been replaced by the immediacy of publication/broadcast. And services whose goal is to save memory are put in the position of filling a bottomless pit: who should “judge” the “value” of a document and save it for the future? This has never been the library’s prerogative, or its mission. Should we leave this role to an algorithm that measures and evaluates based on usage (number of clicks, number of links, etc.) at the risk of sentencing to oblivion major

2 Which doesn’t include the entirety of websites published in France, but only about a third.

3 Olivier Ertzscheid, Twitter, un patrimoine superflu(x)? *Affordance*, 9 mai 2010, <http://affordance.typepad.com/mon-weblog/2010/05/twitter-le-patrimoine-du-superflux-.html>

4 Autoritativité, In: *Dictionnaire des concepts info-documentaires*, Savoir-CDI. <http://www.cndp.fr/savoirscdi/index.php?id=593>

works that only time and the patience of readers can measure? This has happened so often in the past⁵ that we know popularity is not necessarily synonymous with quality. Finally, in this universe of algorithmic performance, how can documents that are written or spoken in minority languages fight back? The media model cannot be used as a guide for the actions of librarians/archivists.

THE LIBRARY FUNDAMENTALS

As we have seen, “digital libraries” must be distinguished from “web archives”. The function of the former is not to record the Web, but to extract and organise its document content for future readership. It is also a way to gauge the world’s digitisation, by relating it to the documents themselves. Building digital libraries, which store, organise and make available these types of documents is becoming an urgent task to ensure knowledge development and sharing. When we speak of storage, we must also mean preservation, long-term if possible, meaning document encoding formats are regularly refreshed (for videos, images, and even digital books).

To understand the functions of digital libraries, it is first important to recall the library fundamentals. A library is a public organism whose functions include building document collections, describing them (through cataloguing and indexing), conserving them, and offering them up for reading by a specific readership. All the terms of this definition are important.

First, the library is not governed by commercial imperatives. The whole of society, through public and therefore collective financing, feels the need to create spaces that guarantee long-term access, available to all, of the products of human knowledge. Like any collective solution, this has its drawbacks, notably a “reaction time” that differs from the interest and attention that the media can engender. But this ensures that the documents forming bibliographic stores are selected so that each idea, theory, perspective and language is fairly represented, while time and criticism allow the most significant documents to emerge.

⁵ Samuel Beckett, for example, who was later awarded the Nobel Prize for literature, sold only 150 copies in France of his play *Waiting for Godot* the year it was published.

Second, a library is a collection of documents that is organised, that is coherent, and that represents a collective will. Digital production, and even more so the Web, tends to be judged by collection size. One speaks of millions of digitised books, millions of photographs, the quantity of video posted each minute, and so on. Is this race, this desire to “up the numbers” effective? Is it not a tangent deriving from the simple fact that it has become technically possible? A collection, on the other hand, pursues a goal. It seeks to ensure either the completeness of a restricted domain (as with research libraries), or public satisfaction (the diversity necessary to meet the needs of a public library situated in the heart of an informal neighbourhood). In the case of digital production, one must distinguish the time of the collection’s creation from the time of document access. This can be done by the digital library’s own catalogue, but is usually done through collective catalogues, a role often performed by external actors who form the libraries’ collections. Search engines can index documents in several collections. Protocols are in place to promote the creation of global indexes, despite the specificity of each collection. Thus, the OAI-PMH⁶ protocol allows external search engines to create indexes from metadata for each document in a collection that is open to such “harvesting”.

Finally, library collections are audience-specific. The mission of a university library, which caters to students, is not the same as that of a specialised laboratory, or in the opposite direction, actions promoting reading and literacy conducted by “street” or neighbourhood libraries, or sites of exclusion (prisons, hospitals, and so on). By defining a specific audience, one obliges the acquisition of staff and services tailored to that audience; a library is above all a set of services, from the reception desk to its guiding mission – helping readers find documents they would not otherwise be able to discover.

By this logic, a library is defined more by its audience, activities, positioning and project, than by documents themselves, which are, let’s recall, “duplicates” existing in multiple copies in many places worldwide. Should we not maintain this approach in the digital world? Rather than focusing on documents, number of pages scanned, and catalogue size, shouldn’t

6 François Nawrocki, “The OAI Protocol and Its Uses in Libraries” (Le protocole OAI et ses usages en bibliothèque), Ministry of Culture, France, February 2005.
<http://www.culture.gouv.fr/culture/dll/OAI-PMH.htm>

we return to the relationship between the collection and the public, which is guaranteed apart from any commercial pressures on the librarian? And in any case, isn't this the only approach that not only emphasises the conservation of documents in minority languages and cultures, but also offers for these languages' speakers access to documents from other linguistic universes that are suited to their needs?

METADATA

The traditional library is not just a set of books, or even a collection tailored to its readership. It also performs two tasks: inventory and description on the one hand, and organising the knowledge included in its collection on the other. It is through metadata management that we encounter the performance of these tasks in digital libraries.

Metadata denotes all the information that we can obtain about a document. It is thus a primarily descriptive cataloguing of information: author, illustrator and prefacer references; edition indicators (dates of publication and, if applicable, re-release; collections, publisher references, date of scanning, etc.), the work's collation (number of pages, book specifics, especially if it contains marks indicating the *ex-libris* history, scanning method, file format, duration of sound or audiovisual recordings, etc.) and finally, references regarding the book's filiations to other works in the same "family" (including the indication "original edition" if it is an original with translations, a multi-volume work, and so on). Descriptive cataloguing provides a work with a material and editorial context, whether it was printed onto media first (CD, DVD) or published directly online. This question of context is even more important when it comes to digital documents. Too often, the ease of using text itself to retrieve documents makes us forget the need to place a given document in its broader context (date and terms of publication, document type and *genre*, etc.). Let me add that digital production with its unique ability to link information, allows us to go even further than we are usually able in material libraries. Thus, links to author biographies, photographs, the reproduced covers of all linguistic versions of the same book, or a list of available critical readings, can increase a book's range of perspective, placing it in the context of an entire publishing venture.

Metadata also includes information describing or summarising the knowledge contained in a book, thereby ensuring the consolidation of books or data covering the same topics. First, there is classification, the ability to place a given work in the same semantic field as others. In this way, we can scan a subset of knowledge to measure complexity, and to discover within it works with a unique viewpoint regarding classics in a given field. All scientific sectors have classifications adapted to the production of documents in their area of knowledge. Besides the ability to group classifications to promote inadvertent or serendipitous discovery, keywords and tags offer precise indicators that a document addresses a specific subject. For this, there are two different approaches that can be complementary in the case of digital libraries:

- Follow established descriptor rules by selecting from a pre-existing and closed list that is shared by several libraries (the concept of the “authority list”), or in accordance with a pre-established framework (for example if a document addresses a particular place, person, event, historical period, and so on);
- Let each person decide on tags, possibly adding them to a list. This model is called a “folksonomy”. It creates independent information from each tag or descriptor, provided by the readers themselves, which raises questions of validity and disperses research, but by the same token considerably enriches description through the action of anonymous amateurs⁷.

Here as well, the digital network allows for all these approaches to cohabit. Professional metadata workers can transform their professions to create and validate the experience and knowledge of readers, using independent tags to build structured descriptors. They thus offer a framework to allow amateurs and fans to exercise their collaborative will to share knowledge.

7 The fruitfulness of this approach has been demonstrated with heritage photographs, when communities of anonymous individuals add information to the photographs, enabling librarians to improve the accuracy of the information. The Library of Congress’s experience of making numerous photographs available on Flickr is significant, as is PhotosNormandie, established by a group of historians, about the Battle of Normandy:

Patrick Peccatte, “PhotosNormandie at five years – a record in the form of FAQ” (PhotosNormandie a cinq ans – un bilan en forme de FAQ), *Culture Visuelle*, 27 January 2012. <http://culturevisuelle.org/dejavu/1097>

For the Common Good: The Library of Congress Flickr Pilot Project, 30 October 2008. http://www.loc.gov/rr/print/flickr_report_final.pdf

Finally, metadata is akin to a fact sheet that accompanies a book. One may then create a catalogue by grouping those sheets and making them searchable. Or one can facilitate discovery by allowing a reader to browse between abstracts/summary sheets before diving into an entire book or document. Finally, the metadata allow to track a document's history, both physically (for example, the history of a found work that has been scanned, such as the Timbuktu Manuscripts⁸) and intellectually (by linking to translated versions, audio versions or films, access to preparatory documents, or even a manuscript's digital archives).

As a separate sheet of paper, the presentation of the metadata itself can also promote multilingualism and shared understanding. Regardless of a document's language, the description can preliminarily provide a multilingual approach, allowing for the discovery of a book written in a foreign language. This is especially useful for publicising scientific research written in local languages. If the metadata record is translated into the scientific community's several major languages, a document intended for local users (students, young researchers, civil society and policy makers) can place itself within the global field of knowledge. It can also be translated later, depending on the interest it generates.

Metadata should be exchanged between multiple systems, and should be deployed in different ways to meet specific needs. They are the basis of the semantic web, and as such, they are written primarily in a computer-readable format. The most widespread model is currently the RDF format (Resource Description Framework)⁹, which is standardised and extensible, and can handle all the languages of the world, each piece of information being preceded by a language code in multilingual cases. Even though RDF is a computer-readable metadata format, it still lacks software that can easily capture the contextual information. The format's flexibility on the one hand, and the growing quantity of information that we maintain in the descriptor flyer on the other, means that metadata management systems are not yet entirely user-friendly.

There is thus a contradiction between the results obtained by software only, such as optical character recognition (OCR), which indexes an entire text using search engines; and the results from working directly with the metadata. The first method is of the "industrial" variety, with a digitisation

8 Tombouctou Manuscripts Project, <http://www.tombouctoumanuscripts.org/>

9 *RDF primer*, 2004, <http://www.w3.org/TR/rdf-primer>

process chain. It is economically more efficient, and works well as long as one accepts to view “the book divided into pages”, in the words of Jean-Noël Jeanneney¹⁰. Conversely, the use of metadata is close to artisanal or skilled labour, with every professional being able to add information about the card, including translations of terminology, titles, and summaries. This type of work can also be opened to users themselves in a process known as crowdsourcing.

Documents are often translated, re-edited, or subtitled. Printed works tend to disperse the various translated versions of a text, just as CDs or DVDs do with video and sound. But computerised catalogues, based on extensible metadata records, can compensate for this dispersion, offering the reader a list of the various versions. This concept was developed by libraries in the late 90s under the acronym FRBR (Functional Requirements for Bibliographic Records¹¹). It offers a mechanism for discovering the translated versions of a work best suited to each reader, and can be extended to digital documents, articles from scientific journals, blog pages, or videos with overdub and sign language animation¹². In the opposite direction, adopting the FRBR model to access documents will strengthen the desire to translate. We often hear of a document through a title or other criterion linked to a specific language, usually that of the original document, and everyone searches for that version. However, a version may exist in the reader’s own language, which can be discovered using FRBR. This process of discovery reinforces interest in translation, quite simply because there will be readers to take up the translated version.

DIGITISATION

Digitising consists of reproducing pre-existing analogue documents in digital form. One may digitise books, films, videos, photographs, or sound recordings. Digitisation means controlling two elements: the digital file formats, and rendering the digitised document searchable.

10 Jean-Noël Jeanneney, *When Google Defies Europe: The Case for a Jump Start (Quand Google défie l'Europe: plaidoyer pour un sursaut)*, Fayard/Mille et une nuits, 2005.

11 *Modèles FRBR, FRAD et FRASAD*, Bibliothèque nationale de France.

http://www.bnf.fr/fr/professionnels/modelisation_ontologies/a.modele_FRBR.html

Barbara Tillett, *What is FRBR? A Conceptual Model for the Bibliographic Universe*, Library of Congress, 2004. <http://www.loc.gov/cds/downloads/FRBR.PDF>

12 See in this book: Annelies Brafort & Patrice Dalle, *Accessibility in Cyberspace: Sign Languages*.

Digitising transforms an original into a digital copy. One obtains an image whose quality corresponds to the technical means available at the time of scanning—which is advancing rapidly. Scanned images from only twenty years ago are a far cry from the definition and quality of those scanned today. In addition, so-called “bitmap” files, which retain colour information for every pixel (a point, in computer code), are quite memory heavy and difficult to send. More manageable systems and compression formats, such as .jpeg or .png, have been developed. The same phenomenon exists for audio (.mp3, .flac, .vorbis, etc.) and video (.mpeg4, .ogg-theora, etc.). A format is always a balance between quality and manageability. One of the functions of digital libraries is then to maintain the original digital file in the highest possible definition, and to manage format changes to make the best use of the available technological environment (network quality, screen resolution, and so on).

Making files searchable and findable is more complex. For example, it means recognising characters, words, sentences, and above that the sense of a scanned text¹³. It may mean transcribing an audio stream into text, identifying a video’s sequences or an image’s regions. We can then retrieve documents based on words (text search), or even by entering images to retrieve similar images¹⁴.

Even more complex are the operations for adding contextual information and metadata to digitised files, compressing them to make them manageable, so that they can be retrieved via automatic indexing, which can then help to reorganise pages into a book, a collection of photos into a catalogue, videos into their various translations in overdub and subtitling. These actions require human intervention, with global coordination, to benefit from the leverage of the library network. But these are the operations that ensure that scanned documents will be treated with the same respect as their earthly originals: that they will be kept in an orderly form, are fit for public consumption, and can thus constitute the digital library collection.

It is also in this third, often neglected phase, that multilingualism intervenes, for its indexing utility (for example, to have the names of cities in a country’s language and script in addition to names in other languages),

13 This process differs following languages. See in this book: Pann Yu Mon & Madhukara Phatak, *Search Engines and Asian Languages*.

14 See for example <http://tineye.com>

its ability to link documents with a common original source (translation), or documents dealing with similar topics (classification).

THE E-BOOK

Besides web pages and their archives, and in addition to documents that are scanned for online distribution, we are now seeing the emergence of the e-book. This new digitised book is a documentary object that has inherited from web technology the ability to be read on machines or computers. Most often they are read on reading-specific electronic devices called e-readers (Nook, Kobo, Kindle); tablets (iPad, Kindle Fire, Samsung); and smartphones that function both as a phone and a pocket computer (iPhone, Android). But as “books”, they have two qualities: portability—being easy to carry and use without any network connection; and, similarly to printed books, all of their content “resides between two covers”¹⁵. The e-book’s techniques may come from the online experience, but its core concept – organising content and shaping reflection – comes from the world of publishing.

E-books today are often designed as complementary versions of printed books. The book you are currently reading, which is available in both printed and digital formats, is no exception. However, we will increasingly see books come out that have a book’s organization and ease of exchange and citation, but are only published in a digital version. In particular, this will be used for confidential professional exchange, essays, text collections, rapidly-evolving documents like guides or tutorials, and even textbooks.

Libraries will be faced with the question of how to store and maintain the availability of these sorts of public documents. This challenge will expand their roles and responsibilities, including obligating them to develop a system of distance lending. As far as that goes, we are witnessing the birth of “chronodegradable” books, which cannot be read after a specified loan period. Libraries might also offer remote e-readers (usually within web browsers).

But language’s big challenge is whether published e-books are mono- or multilingual. These documents’ ease of creation, the ability to improve their content over the course of re-publishing and to provide the books

15 Michel Melot, Nicolas Taffin (ill.), *Books (Livres)*, L’œil neuf éditions, 2006, p. 27.

with tools to facilitate reading, just as dictionaries, glossaries and notes, facilitate reading books in minority languages. These e-books can also integrate text with videos or pictures. Finally, some e-readers include a Text-To-Speech feature that facilitates content access for the visually impaired.

The proliferation of creative tools¹⁶ and the potential for professionals to truly agree on e-book formats are two issues at stake to help these materials spread widely and push along the development of digital books, especially in countries without printing infrastructure or access to paper or bookstores. Libraries thus have a major role to play. Some online bookstores like Amazon are trying to pass off their book selling as akin to the function of a “library”, reducing the latter’s work to that of a commercial transaction. This reinforces the need for actual libraries that handle digital documents to remain independent from the book market (and more widely the document market), as an organization that services all of society, and guarantees the balance of ideas and the rights of readers.

AUDIO AND VIDEO RECORDINGS

Recording tools are advancing. This will propel the formation of audio and video collections, which in their turn will enter the digital libraries. Which means that familiarity and use of the three steps discussed above will become vital: format compression management, maintenance and updating, retaining documents in the best possible condition for later reissuing, and finally, adding metadata to provide context and summarise ideas and knowledge in order to facilitate reading overview and selection.

The proliferation of these recording tools provides an opportunity for libraries to enrich and diversify their collections from a linguistic viewpoint. The standardisation of grammar in writing is far from having an oral equivalent. French, for example, changes in pronunciation, intonation, inflection and even flavour between the inhabitants of Paris, Lille, Marseille, Quebec, Senegal and the Antilles. Even within the major languages, there exist strong differences that grammar books and academic works cannot always harness, insofar as these different ways of speaking spring from a linguistic community’s history. Additionally, dialects evolve

16 See for example the online tool Polifile (<http://polifile.com>) for creating digital books in ePub format.

rapidly, especially under the uniformising force of the media that construct recognised and distinguished forms of pronunciation.

The capacity to build an oral archive is also a means for transmitting the knowledge stored in endangered languages, or to rebuild collections of extinct or endangered languages and dialects to our best ability, by using original recordings¹⁷.

This oral information gathering process¹⁸, which aims to collect popular phrases from stories, voices, songs and music, has long been part of the mission of libraries and ethnographic museums. It is “language preservation” in the sense that it maintains a record of historical developments and changes in linguistic phenomena. Technology proliferation, democratisation (lower costs), and the ability to build and make available these collections will strengthen this work. Libraries, whose mission and culture projects them into the temporal dimension of documents, have a special role to play in this process.

FREELY ACCESSING DOCUMENTS

Digital documents have that rare quality of being easily duplicated and transmitted across the net. To defend the document’s traditional mode of economic management, which is generated from an era of high reproduction costs, some feel the need to reduce this exchange capacity by installing digital document locks. To read a digital document that has one of these Document Rights Management (DRM) systems, one must have a specific encrypted key.

Unfortunately, this system ignores the reality of social practices. Copyright laws worldwide have always envisaged numerous specific cases, grouped in international negotiations at WIPO¹⁹ under the heading “Limitations and Exceptions”. Therein, copy for private use, expanded rights for schools and libraries, as well as the right of citation, satire, free availability of speeches

17 See in this book: Tjeerd de Graaf, *How Oral Archives Benefit Endangered Languages*; and Maik Gibson, *Preserving the Heritage of Extinct or Endangered Languages*.

18 Larry Rother, Folklorist’s Global Jukebox Goes Digital, *The New York Times*, 30 January 2012. <http://www.nytimes.com/2012/01/31/arts/music/the-alan-lomax-collection-from-the-american-folklife-center.html>

19 WIPO: World Intellectual Property Organization.

The official document on limitations and exceptions can be found at: <http://www.wipo.int/copyright/en/limitations/index.html>

by influential people, and so on, are registered as legitimate rights that can be mobilised by readers or by cultural and educational institutions. Locking systems are quite rigid, and cannot take into account the type of practice or institutions wishing to use the locked document. There are, therefore, two trends: to do away with DRM and find other ways of financing creation and publishing, as when the music industry's major labels decided to drop DRM in 2006; or else profoundly modify copyright law, turning it into a "natural" law like that of property law on material goods. Unfortunately, it seems the second approach currently dominates, if we take a look at the intellectual property laws and projects being filed around the world.

When libraries are involved in document creation, as with the aforementioned examples of oral information gathering and digitisation, they must ask themselves what rights will be awarded to readers. Permit uses, including Creative Commons²⁰ licenses, are a good way to expand the rights of readers and public institutions. Negotiations with publishers to obtain free unlocked versions for library use may also be envisaged.

The example of scientific publications, too often published in extremely expensive journals, and therefore inaccessible to many libraries and universities worldwide, is significant to this concept. The "open access" movement, initiated by researchers, proposes to make versions of their articles and work available to all²¹. Many research institutions, such as the CNRS in France and the NIH in the U.S., already require members to submit their articles to "open archives", or to publish in journals with open access.

This question of the legal and financial means that allow people to access and share information contained in documents is central to the future of digital libraries. Libraries everywhere are mobilising to enforce their collective approach and public service. The IFLA (International Federation of Library Associations), for example, has committed itself to the global movement known as A2K – *Access to Knowledge*²².

20 See in this book: Melanie Dulong de Rosnay, *Translation and Localization of Creative Commons Licenses*.

21 Jean-Claude Guédon, Knowledge, Networks and Citizenship: Why Open Access? (Connaissances, réseaux et citoyenneté: pourquoi le libre accès?), In: *Open CourseWare, The Commons of Knowledge (Libres Savoirs, les biens communs de la connaissance)*, C & F Éditions, 2010, p. 67-75.

22 For an overview of this issue that, beyond the library, also affects access to knowledge recorded in property, such as drugs or food:

CONCLUSION

Digital libraries, web archiving, and the creation and preservation of new documents (e-books, audio and video recordings) are complex issues that are nevertheless essential for the future of languages worldwide.

Every new medium brings opportunities to better understand, preserve, disseminate, and at times even revive knowledge and cultures in the very languages that created them. But technical and legal regulations may obliterate the opportunity.

Documents without metadata also lose their strength when they don't offer the historical, cultural, and publishing context of their creation. Digitisation, which cuts knowledge into parts, does not account for a book in a library, which is much more than the sum of its parts.

That's why libraries cannot truly meet their missions unless they are at the service of all society, by being public services open to all, defending readers' rights to free access of information, and providing the necessary supplemental indexing, classification, and contextualisation. It is libraries that must show how knowledge sharing is also a precondition for the expansion of the cultural property market, and that library-managed open access is the best way to accustom readers to cultural practices, which then nourish the publishing market.

Libraries as institutions, and librarians as actors, have responsibilities, competencies and goals that policy makers should encourage and listen to, to strengthen the presence of all languages in the global mental universe. Culture is a tool for building world peace, which gives us the essential means for addressing the challenges and plagues²³ that primarily affect poor countries – those whose languages are less well-endowed with publishing, media and archival resources. Cultural empowerment, involving schools and libraries, deserves our full attention if we are to consider it a common infrastructure for sustainable and peaceful development.

Gaëlle Krikorian and Amy Kapcsynski, *Access to Knowledge in the Age of Intellectual Property, Sone Book* (MIT Press), 2009, 246 p.

23 See in this book: Adama Samassékou, *Multilingualism, the Millennium Development Goals, and Cyberspace*.

SOFTWARE LOCALIZATION: OPEN SOURCE AS A MAJOR TOOL FOR DIGITAL MULTILINGUALISM

What can the ordinary citizen do to promote his language in the digital age? We show in this article that the open source and open content movements provide an ideal platform for individuals to contribute to the continuing existence of their language. We will see how these movements change the situation for those wishing to promote their language, change perceptions and influence the political space.

Original article in English.



DWAYNE BAILEY is the director of Translate, a project advancing the localization of software into the languages of South Africa and assisting other in the world to do the same for their languages. He is the research director of a large network of African experts on localization (ANLoc).

DWAYNE BAILEY

MULTILINGUALISM
FOR DIGITAL
A MAJOR TOOL
OPEN SOURCE AS
LOCALIZATION:
SOFTWARE

How **Open resources** can create opportunities to develop and promote languages and influence the language landscape.

SOFTWARE, CONTENT AND APPLICATIONS

But firstly, what is open source and open content and how do these differ from freeware?

Open Source and its allied *Free Software* movement are both philosophies concerning how software should be developed. The Free Software movement preceded the Open Source movement. Free Software advocates believe that the ability to change the inner workings of software is a fundamental right of owners of software products. They also allow software to be copied and shared. Obviously this is a very different view from proprietary software vendors who use terms like “pirate”. Open Source advocates believe that wide public access to the inner workings of the software leads to better and higher quality software. The philosophies may be debated and their claims disputed but one thing is clear, this software now plays a critical role in today’s economy.

One commonality between Open Source and Free Software is access to the inner workings of the software and the right to change the software. And these are critical for its influence on local languages. With the right to change the software you now have the right to change the text of the software. The industry calls this *localization*, yet it is enough to know that this is the translation of the software interfaces into local languages.

Taking two open source software products as examples we can see the influence that this power has for local languages. Firefox is an open source

web-browser. It commands a market share of 30 % of all internet users worldwide¹. At time of writing, the current version of Firefox is available in 68 languages and a further 14 languages are in development.

OpenOffice.org and LibreOffice are open source office suites compatible with Microsoft Office. Market share is very hard to estimate for such products for the simple reason that anyone can distribute the software and no one needs to register. However, Webmaster Pro² attempted to determine OpenOffice.org market share in 2010. Their data shows values from 0.2 % in China to 22 % in Poland. Usage figures seem to average 10 % across the markets surveyed, which translates into millions of users. When it comes to languages those software are translated in 85 languages by communities of volunteers.

Both OpenOffice.org, LibreOffice and Firefox present opportunities to provide software in local languages at little cost to communities.

It is worth noting that, using traditional approaches to software localization, industry will translate into at most 35 languages. But industry players such as Microsoft have adapted to this pressure from Open Source through programs like their Local Interface Pack or LIP to help raise those numbers.

Open Content is often best thought of as Wikipedia, but includes text and content released under Creative Commons³. Wikipedia is an online encyclopedia available in multiple languages, the English version of which contains over 3 million articles. Other large European languages also have large numbers of articles in the 1,000,000 range.

Although the two largest African language Wikipedias, Afrikaans and Swahili, are much smaller with 17,000 and 21,000 articles respectively, they still represent a large body of local language knowledge. These two Wikipedias contribute 6,4 and 3,8 million words in each language to the Internet (based on data of May 2010). They represent a free resource that can be used by teachers teaching students in their mother tongue.

1 <http://gs.statcounter.com/>

2 <http://www.webmasterpro.de/portal/news/2010/02/05/international-openoffice-market-shares.html>

3 See in this book: Mélanie Dulong de Rosnay, *Translation and Localization of Creative Commons Licenses*.

It is worth spending some time on *Free Applications*. In this we can categorise, Google applications e.g. Search, Gmail, Maps, as well as Facebook and other social media applications. While these applications are free to use they do not provide access to the source code. However, almost all of these large applications have a combination of internally developed and community driven localization programs.

Communities choose to translate these for free mostly because it increases the utility of the program for the users themselves. While some see these large, and often rich companies' behaviour as exploitative, the users clearly regard this as a fair trade as these applications are translated into many languages. Facebook, after releasing their community translation platform quickly had over 100 active translation communities.

Clearly Open Source and Free Applications are delivering more languages than traditional approaches: 70+ and 100+ languages compared to 35 languages. What makes this possible and what is the response from commercial software product providers?

The reason why it is possible is simple. When the barriers that prevent you from providing a translation are lower many more people end up translating. In Open Source the process is clear and open and you are unlikely to ever be prevented from contributing.

Contrast this with the traditional software localization process. The company decides which language it wishes to translate based on market share and other pressures. It then contracts localization companies and delivers the translations; an often long and tedious process that follows the release schedule of the product, which is usually a 3 year cycle. The process is not designed for community contribution, in contrast to Open Source products that see release cycles of 6 months with new languages added when ready, this speed places commercial translations at a disadvantage.

This disadvantage is one of the reasons why Microsoft started their Local Language Program (LLP). This allows Microsoft to localise applications outside of the normal release cycle of a product, allowing them to release new translations after the release of the product. It would seem unlikely that they would translate into a new language close to the release of a new version of the same product. In the early days of LLP local universities were contracted to perform the work but certainly in South Africa this has reverted to commercial localization companies once again.

Microsoft's LLP programme was a very good answer to the power of Open Source localization and has grown the number of languages that Microsoft Windows and Office are available in worldwide.

Now that we understand some of the approaches and impacts of Open Source, Free Applications and Desktop software providers it is good to understand the motivations of communities before we then explore the motivations of commercial companies.

WHY WOULD YOU LOCALISE?

Localization is a costly exercise. Even when people volunteer we are talking of cost in terms of time. Depending on the applications we are talking about months and months of time required to localise some of the applications we have mentioned. So what would motivate someone to volunteer or organise such volunteers?

The first motivation is that people want software in their own language. This is the simplest motivation, it's not about the value or the effectiveness of software in local languages. It stems simply from the fact that the user feels more comfortable in their language.

This concept of comfort is expressed well in the story of an Afrikaans speaking computer programmer in South Africa. He refused to use software in Afrikaans. Since he had always used computers in English he found English easier and more natural to use on computers, Afrikaans to him seemed foreign in this context. However, he used bank ATMs in Afrikaans. Why? Because he has always done banking in Afrikaans and struggled with words like account, statement, withdrawal and deposit while using an ATM. Afrikaans in this context was so much easier for him to understand. Comfort was the motivator and is in many cases where people translate computer software.

The second motivation is that local language software actually makes a difference to people's lives. Unesco and other researchers promote mother tongue education as it has a number of quantifiable benefits that lead to better education of children. This includes participation in the learning process and the learning of fundamental concepts⁴.

4 See in this book: Marcel Diki-Kidiri, *Cyberspace and Mother Tongue Education*.

The same findings can easily be extrapolated to software. In the same way that local language education leads to better marks, better retention of concepts and student participation we could expect local language computing to lead to more active and involved users. Users who know the position of menu items needed to execute a task are disempowered compared to a user who can read the entry and make an association based on their understanding of the language of the interface. Anecdotal stories of training conducted in the Western Cape province of South Africa using English software but providing all training in Xhosa showed a marked improvement in participation and results from the students.

The previous two reasons related to the computer and the users, but there are a number of other valid reasons for localising software that have more to do with the development of the languages. Foremost is the development of skills in the language.

It's no secret that the first localizations into any language are poor. Localisers will deny this, especially commercial localisers since they have been paid to produce good translation. The low quality has little to do with the lack of translation skills of those employed to do the localization. They are poor for the following reasons. Firstly, you are using translators who have no experience in software translations. Second, there is usually no developed terminology in the domain. Third, general style guides are not available in the language.

But the act of localising helps to develop the skills of translators so that they can translate computer software. The act of localization develops the language. We develop the language when we coin terms for words like *wordprocessor* and *spreadsheet*.

Without style guides and terminology we get inconsistent translations. Many people approach the issue of poor quality of localization by developing these first. Unfortunately anyone who hasn't localised and who develops a style guide or terminology will get much of it wrong. Which goes back to our thesis that the first localization will be bad. But using knowledge and developing style guides based on using the language leads to better and better translations.

This is one place in which open source comes to the fore. With slow release cycles there is the very real concern that poor terminology choice and style is accepted as correct. Many languages have examples where the translators

followed English style closely without much thought to the style of the language itself. The easiest of these to spot is that of Title Capitals used in software. While English does this regularly most languages use sentence case. Yet software is often translated, even in these languages, using Title Case. Where open source wins is that these mistakes can be corrected very quickly as the style and terminology emerges for the languages.

Localization helps to promote the status of a language and change the perception of the language. In many minority languages, the topic of translation often creates heated debate about whether software should be translated. The debate has two perspectives: “*We don’t need translation, we should use it in English, French, etc.*” or “*Our language cannot be used for computer programs*”. Interestingly both arguments against localization prove to be false. While most computer users in a given language will use software in a dominant language, it is precisely because they speak these languages that they don’t have any trouble. Their reflection of a lack of need is expressed based on their own experience not based on the experience of the majority of speakers in their language.

The fact that a language does not have a word for a computer concept is easily addressed by creating these missing words. The reality is that many words needed in translation already exist. Words like *proxy*, *password* and *authorise* will already exist in a language. Thus many words perceived as being about computers have much wider application and have been used for centuries.

Usually the debate about whether a language can handle software is rendered moot with the release of the first translation. Whether it is usable or useful will be determined by the longterm uptake of the software. But the idea that the language cannot be used for software is proved wrong.

Why is open source a better environment than commercial software for the promotion and development of languages ?

Localization requires a given set of skills which can be learnt in training but can really only be honed by practice. While efforts like Microsoft’s LLP program are creating localised software they are not safe platforms on which to create skills. When commercial companies hire localisers they expect their software to be localised well. For a new language these

localization skills and resources do not exist. A professional translator is a good substitute for an experienced localiser, but they are not the best localisers.

Why do we say that open source is a safe environment? Does it have lower standards? It is a safe environment because in Open Source mistakes are tolerated and quickly fixed. Quality is enhanced by acknowledging that mistakes will happen and allowing them to be corrected easily.

Both open source and commercial translators try to ensure quality through these additional steps in their translation process. Firstly, translations are reviewed by a second translator. Second, an in-product review is performed by translators using the actual product. Performing a review assumes that the translators know the idiosyncrasies in their language when translating software. Our argument is that for a language that has never been localised, nobody knows these issues. These have to be learnt. A review by someone with no knowledge of these issues is a superficial review. An in-product review by a person who does not use localised software and is often not even an expert in the translated product, is of questionable value.

Open Source has two benefits for the development of the language. It has a lively community that can easily give feedback about terminology, style and other translation issues. Release cycles in open source vary from 6-12 months, while commercial software is more like 3 years. This allows new translations to appear quickly. In contrast the LLP translation of Windows XP into Zulu, once it was reviewed and released, was not updated. The next Zulu version was Windows Vista. Mistakes will happen and open source creates a proper environment for testing where users can give feedback that can result in changes in the product in a short space of time. This prevents bad mistakes from becoming entrenched in the language.

Open source benefits languages through the creation of open resources. A number of commercial companies have added terminology and style guide development as a prelude to translating into new language. This is a good step as these two resources are very important for localization.

However, these resources are easily created for well established languages but fit badly for emerging languages. For established languages the resources exist in some form or another and simply need to be repurposed for the job at hand. Spelling and grammar rules and bilingual dictionaries

can be used to create the resources needed for localization. Now, imagine creating a style guide when there is no general guide in a language and no experience in localization. It is a useful starting point but it is definitely a work in progress and needs to be thought of as that. Terminology creation has the same issue. Gathering a number of language experts together to coin 6000 terms without really understanding their use in software will always be problematic.

But probably of most concern for an emerging language is the availability of these resources once created. In a case in South Africa, Microsoft created 6000 computer terms in various languages. These were reviewed by the Pan South African Language Board (PanSALB) and approved. However, the list of words were not generally available to the public or to other localisers. This was shortsighted on the part of the language board, as it would have been easy to require that these resources be made generally available since the approval made use of state resources. To Microsoft's credit these terminology lists (and style guides) have been made more generally available, but it's taken a number of years.

In contrast open source by its very nature must create open resources. A terminology list or style guide created for an open source translation will be released under an open license allowing others to reuse and improve the guide as needed.

By translating open source it is possible for translators to hone their skills, develop new skills and develop new markets. In an environment where dominant languages are accepted there is no need for translation skills. In an environment in which users make use of localised software there is a growing need for more localization. By localising open source software it is possible to stimulate a demand from users for local language software, which leads to a demand from software producers and thus creates new work for the translation industry.

While this market is growing, open source allows professional translators to build their skills. They get to actively translate without the business risk of poor localization. They get to work with localization software, follow formal localization processes and receive feedback from localization engineers. All of these are skills that they can reuse in commercial localization.

In a university or training environment, open source is a valuable tool with which to teach localization principles to students. It presents a real localization environment, as opposed to a sterile simulation. And students are very aware that their work will be used by real people in real software.

Open source and open content can grow a language community that can achieve much more than individual people and companies. Open source is predominantly a volunteer movement. This is important when it comes to local languages. When people are aware that their time and effort can impact their language then it ensures that their language has a future.

If five people are paid to translate Wikipedia at one article a day for a year then we'd have one thousand new articles at the end of the year. If fifty volunteers translate one Wikipedia article a month it would result in six hundred articles in that year. The speed is not as great, but the effort is more sustainable. Each of the fifty volunteer translators has the potential to bring in new people to translate, while the five paid translators won't be doing any recruiting. Once the payments stop the five translators disappear while the fifty volunteers are more likely to continue. A combined approach would be very powerful as it would allow professional translators to lay the groundwork for volunteer translators. After all it is much more exciting to add your one translated article to a corpus of one thousand articles than it would be to add your one to the existing twenty.

In the same way a large group of software localisers can localise much more software than one or two individual translators.

ANLOC: AFRICAN NETWORK FOR LOCALIZATION

The African Network for Localization (ANLoc) has been able to apply these principles of open source localization. Within ANLoc we've seen open source provide an amazing platform in our 'localise software' and 'training' programme. The end result has been a number of pieces of software translated into local language. These included Firefox in Northern Sotho, Luganda, Akan and Songhay. Tuxpaint, a children's drawing program. Abiword, a word processor and VLC, a media player.

During the ANLoc training we used the Pidgin instant messenger client and Virtaal Computer Aided Translation tool to teach localization principles. Both Pidgin and Virtaal are open source application softwares.

Using Pidgin allowed us to easily demonstrate the results of the translation to the participants. One of the participants went on to complete the translation of Pidgin into Swahili, on the way becoming a better localiser while stimulating demand for more localization.



This is Firefox operating in Swahili with the preference dialog open. The user is looking at the Afrikaans Firefox download page. Both languages are supported by our efforts in Anloc.

By being able to demonstrate localization issues in Pidgin we had feedback from a commercial localiser who'd localised Microsoft products, that for the first time he understood some of the issues of localization that he hadn't before. While you don't require open source to demonstrate these principles, it is worth realising that none of the trainers developed Pidgin, yet the software was easy to modify to add the translations. The students could see these issues in a real piece of software that they knew would be used by real people.

In the *ANLoc localise software programme* we awarded small grants to teams across Africa who were tasked to translate a number of pieces of open source software. What was different in this programme is that these teams followed a strict translation programme. This programme was created by experienced localisers, allowing the teams to focus on the work rather than having to choose which software to translate. In addition, a team of technical experts guided the translators and handled all the behind the scenes technical issues that would hamper inexperienced localisers.

In ANLoc we showed that anyone can localise. The quality of these localizations, of course, needs to be evaluated by the respective language communities. Using technical experts able to hide the software complexities, or the hard parts of localization, from the localisers, coupled with an easy to use web-based localization platform, that required no installation and was easy to learn, we showed that small teams working in minority languages could be assisted to make a dramatic impact for their language.

The future of open source and community translation is web-based, distributed translation platforms. Within ANLoc we have continued to develop Pootle, a web-based translation platform created by Translate.org.za. This is critical piece of software for volunteer translation environments because it dramatically lowers the barrier to entry.

It also ensures a behaviour that can benefit localization of minority languages in the long run. This behaviour is the sharing of resources. For minority languages it is important that terminology lists and translation expertise are shared. For dominant languages this matters less. But in a language like Xhosa if you create an ICT terms list you need to share it, so that it's used but also so that people are aware of it.

What motivates companies to localise? Companies localise based around market share, market advantage and policy. If a company can sell more products by localising then they will localise. If a company can gain market advantage then they will localise. Lastly, if a policy says that something must be localised then a company will comply.

While companies will lose no time in portraying this to be about their concern for communities and language, below the surface it rarely is more than the above motivators. The most intriguing motivator though is market advantage and its twin, market disadvantage. With an eleven languages policy in South Africa and no software translation policy, there is no need to translate. With the market mostly not needing localised software and buying capacity of the other languages being low there is no push to drive market share through localization. There is also no market advantage to having software localised.

When OpenOffice.org was translated in South Africa it created a market disadvantage to others participating in the field of office software. To eliminate that disadvantage Microsoft translated their software. Remove the names of the players and you have a general principle that can be

used by language activists to advance local languages. When you see an opportunity to create a change in market conditions you can often use open source to execute that opportunity.

Open source allows local languages to create the language environment that they desire. Commercial translation programs were driven by market value until open source pressurised them to change and include other motivators such as the number of speakers, local language policies, etc. Open source localizations remove the excuse not to have a dominant commercial product localised.

Countries like Iceland can ensure that Microsoft is translated into Icelandic and Canada ensured that Windows 95 was simultaneously released in English and French through their relevant language policies. While most minority languages do not have that policy power, through open source they can garner that power by changing the rules of the market. Where no local language software exists it is easy to suggest that none should be created as there clearly is no market. Once open source translations are available it easily becomes a checklist requirement for commercial software to eliminate their market disadvantage.

If we examine the Microsoft LLP program we can see that in many cases a LLP language follows open source translations and initiatives. This is true in South Africa, Nepal, Nigeria, Tanzania and other countries. While LLP gave Microsoft the ability to respond to this pressure it was open source translation initiatives that created the real pressure for change.

In South Africa, Afrikaans language bodies made many requests for Microsoft localization, including letter writing and meetings with Bill Gates' father. While the constitution of 1994, which introduced eleven official languages, had made the need clear, it would take more than a decade before LLP allowed Microsoft to respond to this need. This period included five releases of Microsoft's Windows product, and to date not all eleven languages have been covered by LLPs.

It is interesting to note that three weeks after the release of OpenOffice.org in three of South Africa's official languages in 2004 a press release from Microsoft stated that they would be creating local language version in six months. It would take Microsoft two years to deliver on this promise, they could be accused of a vapourware announcement, but they did deliver in the end. Thanks to open source.

The lesson from this is not that Microsoft doesn't care about language. It is that, Open Source creates a motivator that forces companies to localise software because of a market disadvantage. A language activist translating open source is much more powerful than a letter writer making public appeals to commercial companies. Open source is a powerful tool in the hands of language activists.

OPEN SOURCE DOES NOT REQUIRE THAT LOCALIZATION MAKES ECONOMIC SENSE, JUST THAT IT MAKES SOCIETAL SENSE.

It's interesting to note that commercial interests are not always aligned with societal interests. Commercial interests are not concerned about languages of ten thousand speakers. While in open source it simply takes one person to make their language a personal priority to enable that language. Open source allows non-commercial motivations to be prioritised and to take root.

These non-commercial motivations can include the desire to promote and advance the language. The desire to make the language a primary language of education. They could include efforts at language recovery to bring a language back from the brink. They could include political motivations, both good and bad, to advance language. Ultimately these are driven by people who care about language. Open source releases them from the obligation of justifying their efforts in terms of economics.

While we have not dealt with open content in much detail, almost all of these principles can be applied to open content. The difference being that while open source provides the tools to create, open content provides the content that stimulates the creation of more and more content to the point where local language content is large enough to make the creation self-sustaining.

The message is clear. If you want to empower local languages in the digital world then you have to localise open source software. By doing that you create the environment, demand, incentive and skills that will lead to more localised software and ultimately a digital ecosystem in the local language.

TRANSLATION AND LOCALIZATION OF CREATIVE COMMONS LICENSES

Creative Commons is a set of legal rules and licenses for smoothing the flow of creative work. Ease of legal adoption in local jurisdictions goes far beyond mere translation from one language to another. It also calls for building a community of experts throughout the world.

Original article in English.



MÉLANIE DULONG DE ROSNAY is a researcher at the ISCC, the Institute for Communication Sciences of the CNRS and Creative Commons France legal lead at CERSA CNRS University Paris 2. She co-founded Communia international association on the digital public domain. After receiving a PhD in law, she was a research fellow at the Berkman Center for Internet & Society of Harvard Law School and at the Institute for Information Law of the University of Amsterdam.

MÉLANIE DULONG

LICENSES
OF COMMONS
AND LOCALIZATION
AND LOCALIZATION
AND LOCALIZATION

With the extension of copyright law duration and the expansion of its scope, possibilities to access and reuse works are being reduced, while digital technologies can (and should) be used to facilitate their usage, instead of locking them even more. Copyright law grants automatically to authors an exclusive right to control the copying, distribution and modification of their works, leaving few rights available to the public without authorisation, such as parody, private or educational use.







However, creators can choose to let others copy and reuse their works for free. By deciding to be more generous, they will get more exposure, maybe even foster citizen participation, creative remix or translation by volunteers. Creative Commons (cc), a non-profit organization, is offering a set of open content licenses to the public, in order to remove barriers to access and creativity by facilitating sharing of works⁴. When distributing their work under a cc license, authors authorise the public to copy their work given that some conditions are respected, such as providing appropriate credit, reserving commercial rights or requiring modified versions to be made available under the same freedoms.

cc licenses are applicable to works which are covered by copyright law: text, blog posts, articles, books, images, websites, audiovisual creations, photographs, music, etc. They are used by individual artists and institutions such as Wikipedia, Al Jazeera for footage, MIT for educational material or Hindawi Open Access for scientific journals. Other tools are available for data or public domain works which are not covered by copyright: the White House, the Dutch and the Piedmont governments use a cco (o for zero) instrument in order to indicate they waive their rights on public data to facilitate citizen access to information and innovation based on public sector data reuse.


1 *Net.lang* is itself distributed under a CC-by-sa license to facilitate translation and publication in every country and language.

CC licenses are made available to the public through an online user interface² asking authors to specify which rights they wish to grant to the public and to choose optional elements. Licensers may (or may not) request their work be used for non-commercial purposes only or in a non-derivative way only, or request the derivatives (such as translations) be licensed under the same conditions. Based on the answers to these questions, the user will be delivered one of six licenses to be displayed on their website or on the physical copy of their works in order to indicate to the public which freedoms are granted in advance and which rights are still reserved.

The six licenses are the following:

License	Logo	License	Logo
Attribution (BY)		Non Commercial (BY NC)	
Non Commercial - No Derivative Works (BY NC ND)		No Derivative Works (BY ND)	
Share Alike (BY SA)		Non Commercial - Share Alike (BY NC SA)	

Licenses are made available in various formats when clicking from one to the other:

- A button with the CC logo, containing a link to the license’s human-readable summary. 
- Embedded machine-readable code containing metadata to be processed by search engines.
- A human-readable summary of the license’s core freedoms and optional restrictions: <http://creativecommons.org/licenses/by-sa/3.0>
- The legal code, e.g. the full license: <http://creativecommons.org/licenses/by-sa/3.0/legalcode>

Originally, the legal code was drafted according to US copyright law, as the organization is based in this jurisdiction. It has later been drafted in reference to international conventions. Licenses are being translated in over fifty languages and seventy countries. This process called international porting goes beyond a mere translation. For instance, the definitions are expected to be extracted from copyright legislations in each jurisdiction.

² <http://creativecommons.org/choose>

The purpose of having local licenses is to provide a linguistic and legal translation, as well as to increase access, acceptability and understanding by users and judges who need to interpret the licenses in their jurisdiction. The internationalisation process also provides local teams of affiliates who are expert in copyright and open content licensing. Beyond ensuring the translation and porting of the legal code, jurisdictions project leads work with local user communities and governments to explain the licenses and facilitate their adoption³. Jurisdiction teams also collaborate with cc headquarters staff to perform research, provide suggestions to improve the licensing system, report on users' questions, use cases and issues arising in their jurisdiction. They translate and create educational material and constitute a network advising on questions affecting user communities around the world.

However, the legal porting process comes with a caveat due to the lack of harmonisation among copyright legislations. As copyright law varies among countries, licenses do not exactly cover the same scope of rights. As cc licenses are declared compatible among themselves⁴, an author is expected to consent that future adaptations of her work be licensed under unidentified terms, which can be a problem in contract law. The legal porting process has been a useful constitutional event for the development of an international network, and ported versions facilitate understanding and adaptation in diverse legal cultures and systems. But it is a time-consuming task in a complex international law environment. In any case, linguistic translations improve access, acceptability and understanding by non-native English speakers. The license's human-readable translations, summarising the legal text in a few sentences written in plain, non-legalese language, are making it clear to all creators that works can be reused.

Translation is not only a matter of local language, it is also a question of making concepts accessible to non-specialised audiences, and Creative Commons licenses are providing a means of accessibility to legal knowledge, towards access to knowledge and creativity in general.

3 Hala Essalmawi, *Partage de la création et de la culture : les licences Creative Commons dans le monde arabe*, in: *Libres Savoirs, les biens communs de la connaissance*, C&F éditions 2012, p. 145-155.

4 Works licensed under a Share Alike license can be remixed with works licensed under a Share Alike license from another jurisdiction, and the resulting derivative work may be relicensed under the Share Alike license of another jurisdiction, all versions having slightly different conditions.



DIGITAL MULTI LINGUALISM: BUILDING INCLUSIVE SOCIETIES

PART 3

Nearly two billion people use the internet worldwide. However, can we say that cyberspace is accessible to all, regardless of language, culture or abilities? How can diasporas, children, disabled people rebuild their world with the help of cyberspace? How can acceptance of diversity of cultures progress the internet, and how can this network in turn contribute to enhanced social and global inclusion?

VIOLA KREBS
& VICENT CLIMENT-FERRANDO

LANGUAGES, CYBERSPACE, MIGRATIONS

How have migrants taken possession of cyberspace? Do they speak in their mother language(s) or do they use a *lingua franca* such as French, English or Castilian? How does their participation contribute to linguistic diversity in cyberspace and what are the needs that may be partially or completely covered to help migrants in their new environment? What languages are then used in this context?



VIOLA KREBS is a sociolinguistic and communication specialist. She is the Founder and Executive Director of ICVolunteers (<http://www.icvolunteers.org>), a non-profit organization focusing on communications (communication technologies, culture & languages and conference support).

ICVolunteers works with a network of 13,000 volunteers worldwide. In the area of research, Viola's expertise is in communications, volunteerism, language and migration and education bilingual. She has written a number of scientific articles and reports and has contributed to several books.



VICENT CLIMENT-FERRANDO [College of Europe, Bruges and Phd candidate at UPF Barcelona] is associate professor at Universitat Pompeu Fabra (UPF) of Barcelona and policy advisor on language, immigration and international relations at the Language Policy Directorate of the Government of Catalonia. He is also the Executive Secretary of the European Network for Linguistic Diversity (NPLD's Think Tank) and a research fellow at the Interdisciplinary Research Group on Immigration (GRITIM) at UPF.



YOANNA RACCIMOLO contribute this article during her time in ICVolunteers. She was a researcher in sociology a University of Lausanne, where she participate an interdisciplinary study on the interpretation of migrants stories. She also conduct research in Lebanon and Russia about human rights and internal or transborder migrations.

& VICENT CLIMENT-FERRANDO
VIOLA KRIBS

MIGRATIONS
CYBERSPACE
LANGUAGES

Migration is not a new phenomenon, and statistics show that population mobility worldwide is on the rise. The latest edition of the Atlas of Migration¹ reported that there are an estimated 200 million migrants and displaced people around the world, representing about 3 % of the world's population. People may move within their own country from rural contexts to cities or leave the place where they were born to become international migrants, often seeking a better livelihood and conditions that correspond better to their daily needs. Many migrants are able to leave their country because they have the necessary psychological and educational resources to do it. Leaving one's land is never an easy decision and during recent months, migration waves linked to insurrections in Tunisia, Egypt and Libya have shown again that large-scale migration can initiate irrational fears and extremist reactions in European countries². The human factor then becomes entrenched in political and economic debates, which often forget the origins of the population living in modern western countries.

“In the 1600s and 1700s, by forced exile, by lures, promises, and lies, by kidnapping, by their urgent need to escape the living conditions of the home country, poor people wanting to go to America became commodities of profit for merchants, traders, ship captains and eventually their masters in America”³ This description could fit what happened to the people of Africa turned into slaves and thrown in boats to build the riches of

1 Wihtol de Wenden, Catherine. *Atlas des migrations dans le monde, Réfugiés ou migrants volontaires, Alternatives Economiques*, éd. Autrement, Paris, 2009.

2 Virginie Guiraudon, directrice de recherche au CNRS et au centre d'étude européen de Sciences-po Paris.

http://www.humanite.fr/01_03_2011-linvasion-de-leurope-par-bateaux-est-un-fantasme-466348

3 Howard Zinn, *A People's History of the United States*, Harper Perennial Modern Classic –Persons of Mean and Vile Condition– P. 43.

a nation. But the “rogues and vagabonds” precisely described above are those the Elizabethan society was trying to expel from its cities in England. Hundred of thousands of them, English, Irish or Welsh, came to America during that period, becoming servants and slaves to other rich European immigrants.

Though Howard Zinn, in his People’s *History of the United States*, studies the origins of the United States population, his example can be extended to many nations nowadays.

LANGUAGES IN MOVEMENT

It is estimated that close to one half of the world’s population is bilingual⁴. Many factors contribute to these high statistics, including political, economic and religious migration, as well as language policies in individual countries.

Many migrants may find themselves in a country of residence where the official language or languages (L2) are different from their mother tongue (L1) and the language(s) they spoke back home. Typically, the children of migrant families adapt very quickly to the new context and assimilate L2 easily. However, this is a much slower process for the parents who may have great difficulties learning and mastering the host country’s language(s).

Beyond the exchange of information, languages are closely related to individual, collective and national identity. In this sense, they represent power, often a struggle for control, beyond questions of merely functional communication. This is nothing new and looking back in history, it has existed for many centuries.

Immigration is undoubtedly one of the main factors leading to social, economic and political transformation. And the issue of migration *management* has been included in a much broader debate, that of multiculturalism and the rights of minorities. By reinforcing fundamental rights already granted to all individuals in a democratic state, multiculturalism tends to extend public recognition, and encourage support for

⁴ Comrie Bernard et al., *The Atlas of Languages, Facts On File, Inc.*, New York, USA, 1996.

ethnocultural minorities to maintain and express their distinct identities and cultural practices⁵.

This does not happen without creating questions of language acquisition, linguistic presence and influence, and legal restrictions established around languages, as well as a debate on linguistic diversity. Thus, in this increasingly interconnected world, where people are more and more mobile, the acquisition and mastering of multiple languages is of growing importance⁶.

This complex context is paired with the fact that globalisation and the introduction of new information and communication technologies represent both an opportunity and a threat for the approximately 7,000 languages spoken in the world today⁷. Indeed, they have made the world a much more interconnected place, where information from any part of the world is relayed on screens within minutes, and where people can create, display and exchange content through the internet.

SCALE OF MIGRATION

Migration has gone from 75 million in 1965 to over 200 million people in 2008 following growth of the world's population, which has more than doubled during the same period, from 3.2 billion to close to 7 billion⁸ people. One third of them are family-related migrants, one third are refugees and one third move for work reasons. In addition to the publications by the International Organization for Migration (IOM)⁹, the *CIA World Factbook* provides regular updates by country on net migration rates¹⁰.

For the last twenty years, South-South migration has been on the rise. Indeed, Asia accounts for the largest number of migrants with 40 to 50 million Chinese and 20 million Indian migrants.

5 Banting, K. & Kymlicka, W (eds.). *Multiculturalism and the welfare state: recognition and redistribution in contemporary democracies* Oxford University Press, 2006: 1.

6 Anna Lietti, *Pour une éducation bilingue*, Payot, 1994.

7 Malherbe Michel, *Les langues de l'humanité*, Robert Laffont, coll. Bouquins, 1993.
http://portal.unesco.org/es/ev.php-URL_ID=1864&URL_DO=DO_TOPIC&URL_SECTION=201.html

8 *World migration 2003: managing migration challenges and responses for people on the Move*, IOM International Organization for Migration, Geneva, 2002.

9 <http://www.iom.int>

10 <https://http://www.cia.gov/library/publications/the-world-factbook/fields/2112.html>

In addition, migration movements have reached a complexity and a scale unprecedented in history. To traditional immigrant-receiving countries such as the United States and Canada, we must now add countries which, until very recently, were emigrant countries such as Spain, Italy and Portugal, among others. There are therefore immigrant groups in practically all western democracies today

In the European Union, the countries that received more than half of all immigrants in 2008 are Spain, Germany and the United Kingdom¹¹.

A very representative example is given by the Catalonia migrant phenomenon. Out of the 6,147,610 inhabitants living in Catalonia in 1998, only 1.97 % were foreign immigrants. In 2009, the population had increased by more than one million people, with almost 17 % of all citizens being immigrants, hence Catalonia represents the highest receiving territory in Spain. In addition, the arrival of people of foreign origin represents 77 % of the average growth of the population in this territory during this period

In Geneva, Switzerland, 38.3 % of all citizens are foreigners, with 184 nationalities represented, who speak some 150 languages¹². For 25 % of the population residing in Geneva, French is not the primary language and some of them are not able to understand it and/or speak it¹³. In London alone, three hundred different languages have been inventoried. The latest census in Canada listed 6,293,110 allophones, which means that 20,1 % of the global population of the country speak a language other than English or French. The United States also faces the same phenomenon: the census of the year 2000 reports that more than three hundred languages were spoken in the country.

While cities like Barcelona, London, New York or Geneva are particularly international, the trend of multiculturalism is very broadly applicable. If one assumes that every citizen has the right to health care and education, we should examine which languages are used to communicate in these

11 Source: Eurostat, European Statistics available online at http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Migration_statistics, figures from 7 November 2010.

12 Portrait statistique des étrangers vivant à Genève. Résultats du recensement fédéral de la population et autres sources. Office cantonal de la Statistique, *Études et documents* n° 37, Genève, septembre 2005.

13 *La politique cantonale de préformation des non-francophones à risque d'exclusion: Évaluation des mesures de soutien*, Commission externe d'évaluation des politiques publiques, Genève, Septembre 2005, page 15.

contexts. There is a risk of exclusion of the allophone population both socially and professionally.

IMMIGRATION POLICIES AND APPROACHES

For a long time, the most important immigrant receiving countries – Australia, Canada, and the United States – adopted assimilationist approaches to immigrant groups where immigrants were expected to fully assimilate into mainstream society, as it was hoped that over time they would become indistinguishable in their way of life from native born citizens.

The late 1960s and beginning of the 1970s saw two major changes. First, the adoption of race-neutral admissions criteria, so that immigrants to these countries from non-European (and often non-Christian) societies were not excluded; and second, the adoption of a more “multicultural” conception of integration, one which expects that migrants will express their ethnic identity, and which accepts an obligation on the part of public institutions to accommodate these ethnic identities¹⁴. It must be stated, however, that this trend has not taken place in all western societies and the degree of public recognition varies from country to country.

Along with new traditions, religions and cultures, immigrants have brought a whole array of languages. As an example, Tamazigh – a language spoken in scattered areas of Morocco, Algeria, Tunisia, Egypt and Mali – is the third most widely spoken language in Catalonia, after Catalan and Spanish, the two official languages.

This unprecedented change in the linguistic kaleidoscope of all western societies does not seem to have led to a profound scholarly debate on how to apply multicultural policies to the languages of migrants and only scattered references on this issue are found in literature.

When dealing with the languages of migrants within the framework of multiculturalism, one of the most commonly referenced issues is mother-tongue education for migrants¹⁵, that is, the establishment of the language of specific migrant groups as the vehicular language in some

14 Banting, K. & Kymlicka, W (eds.). *Multiculturalism and the welfare state: recognition and redistribution in contemporary democracies*. Oxford University Press, 2006: 54.

15 See in this book: Marcel Diki-Kidiri, *Cyberspace and Mother Tongue Education*.

extra-curricular schooling. Scientific studies on bi- and multilingualism show that bilingual education is a good way of transmitting multiple languages and that it is positive for the cognitive development of a child¹⁶.

Bilingual education or mother-tongue teaching were two forms of instruction introduced in the 1970s in a number of countries (Germany, the Netherlands, etc.) when the idea prevailed that migrants would return home. It was not therefore originally conceived of as an element of multiculturalism policy. Furthermore, in more recent years, bilingual education in the Netherlands and Germany has decreased, as it has sometimes been considered as an impediment to assimilation. More generally speaking, the recognition of the language of migrants is now being questioned as it is argued that it goes against effective integration of migrants¹⁷. Indeed, it is argued that the recognition and support for the languages of migrants could foster ghettos and linguistic enclaves, which would be detrimental both to the host society – which regards the learning of its language as a source of patriotism and loyalty – and to immigrant groups – as their lack of fluency in the host country’s language(s) may eventually lead to marginalization and economic disadvantage. One key term in this debate is parallel society, that is, the fear that promoting immigrant particularities – cultures, religions and languages – could create self-secluding ethnic communities, leading to what has been referred to in the literature as “Balkanization”.

Emphasis therefore seems to be placed on the host societies’ languages. The number of proposals to strengthen language tests for naturalisation appears to be on the increase. Proposals of this sort have surfaced throughout Western nations, where difficulties in immigrant integration are often blamed on the inability or unwillingness of immigrants to learn the state languages and point out that in some European countries, there is even talk about legally requiring immigrants to attend language classes

16 Cummins Jim. *Bilingualism, multiculturalism, and second language acquisition: the McGill conference in honor of Wallace E. Lambert*. L. Erlbaum, London, 1976, 1981.

CDIP, Conférence suisse des directeurs cantonaux de l’instruction publique. *Rapport sur la question des langues, Quelles langues apprendre en Suisse pendant la scolarité obligatoire ?* Berne, 1998.

Crawford James. *Bilingual Education: History, Politics, Theory and Practice*. Bilingual Educational Service, Inc., Los Angeles, 1999.

17 Kymlicka, W & Patten, A. (eds). *Language Rights and Political Theory*. Oxford University Press, 2003 : 8-9.

as a precondition for access to social benefits. A recent study¹⁸, which analyses the programmes applied by Germany, France, Canada and the Netherlands, clearly points in this direction. This growing requirement for immigrants to learn the language of the residency country could be interpreted as a “return to assimilation”.

Another example of how policies impact migration is the ‘Open Door’ policy put in place by the Chinese government in the last few decades, which has greatly influenced migration statistics. In a recent article, Phoebe H. Li describes these changes¹⁹. In 1997, the total number of Chinese citizens going overseas was 5.32 million. A decade later, in 2007, there were almost 80 million national border crossings by mainland Chinese. These Chinese nationals departed for most parts of the world and comprised a wide range of permanent and temporary migrants such as international students, contract workers and tourists, of whom many were potential permanent migrants. Phoebe H. Li describes the immigration criteria applied in New Zealand, which include employability and English language skills, among other things.

In recent years, many Chinese have also massively moved to Africa. Many of them are workers in the construction sector; others have established businesses in Africa, creating an unwelcome competition for local businesses in textiles, road building, gastronomy, etc. In some cases, they also organise business activities in the informal sector. From a linguistic point of view, many of these new immigrants do not speak the local languages. Also, they tend to stay among themselves and generally do not integrate as easily as migrants from other parts of the world²⁰.

There is a category of migrants who remain cut off from the local society, depending only on one or two people to stay in contact with the language of the host society. Immigrants from Sri Lanka in Lebanon perfectly exemplify this scenario. Most of them are employed as housekeepers. The majority are not entitled to leave the house without the owner who usually confiscates their passport and forces them to live isolated lives, sometimes

18 Biles, J et al. *Policies and Models of Incorporation. A Transatlantic Perspective: Canada, Germany, France and the Netherlands*. Documentos Cidob. Migraciones, num. 12, June 2007.

19 Phoebe H. Li, University of Auckland, New Zealand, *New Chinese Immigrants to New Zealand, a PRC Dimension*, 2010, http://international.metropolis.net/pdf/fow_newzealand_imm.pdf

20 *Speak to me, speak here. Linguistic Situation in Barcelona*. Éd. ICVoluntaris.org. 2007.

close to slavery. To avoid their isolation, NGOs, such as *Caritas Migrant*, provide these populations with legal and administrative support²¹.

All of the above examples show the complexity of the situation in relation to immigration policies and approaches. The resolution of linguistic issues depends heavily on the way each government addresses immigration. Nevertheless, they are also related to two other main criteria.

The first one is the type of migration, which could be seasonal, transitional or circulatory. In this case, migrants are spending a few years in another country, then returning to their home-base or moving on to another place. The linguistic integration is typically not considered a high priority by the migrants, as their main focus is around earning money or building a career rather than integrating into the new country and society. Consequently, many of these migrants are not interested in learning the language of their temporary host country.

The second one is more related to the psychological aspects of the migrants that may present an obstacle for learning the native language of a country. Many of migrants intentionally decide to not make efforts to learn the language. This refusal symbolises their own intention not to establish themselves for long in the foreign country. They have a deep desire to return home and do not see the need to acquire a language for a supposedly impermanent stay. Actually, many of them stay more than twenty years without gaining a significant understanding of the language of the host country, because they do not want to renounce their hope of returning home.

CYBERSPACE AS AN OPPORTUNITY FOR MIGRANTS: A FEW EXAMPLES OF EXISTING TOOLS AND NETWORKS

The internet and the web present important resources and tools for migrants, social care givers, translators, interpreters and policy makers working in the field of migration issues.

Firstly, migrants might appear on the web as the topic of articles. Many journalists write about migrants and the impact they have on the countries that have to deal with sometimes massive immigration influxes.

21 Caritas Lebanon Migrant's Center : <http://www.caritas.org.lb>

Secondly, online resources are supposedly created in order to help migrants in their new environment. Some of them are in the language of the host country; others are translated into a series of languages spoken by the migrant populations living in a given country.

Thirdly, migrants use the Web to stay in touch with their families left behind, as well as their country and culture of origin. This may be through online video and phone services, or by reading contents on web sites published in the local language of their country of origin.

Sites presenting research studies related to migration issues

MPI²², an independent non-profit think tank, is dedicated to the study of the movement of people worldwide. Its online database includes many scientific articles and is a source of in-depth showcases with the latest data on migration trends and patterns in the United States and around the world. Research tools include US State Data on the Foreign Born, Maps of the Foreign Born, the World Migration Map, Comparative Charts and Tables, the Global Remittances Guide, and asylum data.

Another interesting research group is the “*Programme d’études sur l’usage des TIC dans les migrations*” (TIC-Migrations, “the Program for the Study of the Use of Information and Communication Technologies [ICTs] in Migrations”). It defines itself as a research program “*exploring the impact of new technologies on the world of migrants (paths, personal connections, relations with origin and host countries, etc.)*”. Their objectives “*are to open a new field of research, to bring together two previously separate disciplines (diaspora theory and web exploration), and to develop generic tools to be used in the social sciences and humanities*”. They also present the concept of the “connected migrants”; trying to figure out where migrants stand in a world made of discontinuous continuity.

The International Metropolis Project²³ is a forum for bridging research, policy and practice on migration and diversity. The Project aims to enhance the academic research capacity, encourage policy-relevant research on migration and diversity issues, and facilitate the use of that research

22 <http://www.migrationinformation.org>

23 <http://international.metropolis.net>

by governments and non-governmental organizations. Research articles are available online and the conferences of Metropolis bring together several hundred researchers every year who are involved in migration-related academic research.

The last project we can list here is called Bridge-IT Network²⁴ which aims to raise the question of the potential of ICT for promoting the integration of migrants and cultural diversity in Europe.

Information sources to help migrants in the host country

Sites aimed at migrants typically provide practical information and useful links about everyday life in the country of residence. One example of such sites is the Swiss Migraweb²⁵, which presents practical information gathered by independent organizations from civil society, as well as relevant government offices and specialists in each field. This information is then translated into languages spoken by migrants by a team of volunteers who are all settled migrants from various communities. These volunteers are fluent in both their own tongue and in French or German. They serve as a bridge between different cultures.

In Spain, MIGRAweb.es²⁶ provides access to a team of professionals giving legal advice, specifically related to immigration issues, asylum and nationality.

MigraLingua.org²⁷ aims to provide practical information about the community interpreting services coordinated by ICVolunteers in order to accompany non-French speaking migrants in Geneva in their daily lives. The aim is to increase mutual understanding between migrants and teachers, to encourage respectful dialogue, facilitate access to institutions, and help migrants break out of isolation and thus promote the desire to learn French in order to achieve greater autonomy. With the support of a network of volunteers, this service is available to any person whatever their origin and social status.

24 <http://bridge-it.ning.com>

25 <http://www.migraweb.ch>

26 <http://www.MIGRAweb.es>

27 <http://www.MigraLingua.org>

There are also different software tools developed to bridge the gap between patients and doctors in the health system. Worth mentioning in this context is the Universal Doctor Speaker Project – a tool developed by a group of doctors – a project used in many European hospitals aimed at facilitating communication between health professionals and patients from different backgrounds and in ten languages, including Urdu. Applications of this project exist for iPods²⁸.

Mobile phones can also provide support to migrants. Mobile Voices (VozMob) represents one of the best illustrations for this, where cell phones are used to transmit testimonials, to provide support to members of the same community, and to permit them to keep in touch with their relatives or compatriots. VozMob describes itself as being “*a platform for immigrant and/or low-wage workers in the region of Los Angeles to create stories about their lives and communities directly from cell phones*”²⁹.

Online language courses

Learning languages now is possible online. Paid language lessons, forums and education platforms are easily available to learn and improve a range of different languages spoken in host countries. An increasing number of websites are offering a wide range of paid language lessons in up to 40 languages, including English, German, Spanish, Italian, Russian, Arabic, Chinese (Mandarin), Farsi, Hindi, etc.³⁰ However, the most frequently available languages are English, French, Spanish, German and Italian. These kinds of classes can be an excellent alternative, when it is difficult for migrants to follow regular language courses due to work schedule constraints.

WHEN TECHNOLOGY BECOMES A CHALLENGE RATHER THAN A SOLUTION

Overall, Information and Communication Technologies (ICT) connect people who are geographically in very different locations. However, they can also be a true challenge for the poorest, who may be excluded

²⁸ <http://itunes.apple.com/us/app/universal-doctor-speaker-for/id364812043?mt=8>

²⁹ <http://vozmob.net/en/about>

³⁰ <http://www.livemocha.com>

from cyberspace. Indeed, technology has a cost: equipment, internet subscriptions and cybercafés are still not affordable for everyone. This applies to both immigrants and those left behind. In terms of budgets, not all migrants living in an industrialised country are easily able to use computers and the internet.

Several research studies confirm the positive impact of ICT on the migrants' interactions with their family or the rest of the community. The "culture of the link"³¹ is usually promoted, and quite often internet services, such as Skype or MSN messenger with the possibility to use webcams – are so positively promoted that only the positive aspects are put forward. Even though migrants can effectively keep in touch with their family, and see them very often with the webcam, these practices can also provoke the opposite effect. Instead of having the impression of being near one's relatives or friends back in the native country, the migrant may feel nostalgic not to be physically present for a birthday or a special event. In any case, the communication has to be interrupted and he or she is "condemned" to remain apart. Even though ICT may provide an improvement, it cannot replace the direct inter-personal relationships. The idealisation of ICT might underrate the real difficulties met by the migrant geographically cut off from his or her loved ones.

Another downside of such changes consists in the fact that more intense obligations are placed on family members in other parts of the world, and the process of "keeping in touch", which can result in intense emotional pressures. Indeed, oftentimes, relatives in the country of origin idealise the living conditions of their family members who have left and are convinced that their relatives earn a great income, which may or may not be the case. Pressure to send money back home may be important in this case, and has probably increased in recent years with the many means of communication available nowadays.

Today, many on-line services, administrative or private, are replacing what was once provided by a public utility or customer services agency with a physical person at the other end. In France, as an example, a web service site of the national agency for employment³², requires online pre-registration in order to benefit from unemployment services. These sites are often only provided in the local language (French in this case), which

31 <http://ticmigrations.fr/fr/whoweare/manifeste>

32 <http://www.pole-emploi.fr>

is not really helpful for allophone speakers. In such cases, administrative services are in fact reducing their costs and responsibilities by interposing a machine between the applicant to seeking services and the employees dealing with their requests.

The electronic administration has also a large impact on identification. Biometrical systems of identification are increasingly more commonly required and accepted. They reinforce a society of control, and strongly regulate the influx of immigrants³³. Migrants are submitted to intensive controls motivated by the fear of illegal migrations.

Moreover, migrants may be illiterate, or may not have any practical experience with the Web. As a result, they become completely dependent on their friends or, if they have other family members present, on their children sent to school who know how to use technologies and who also speak the language of the host country. The two previous situations create asymmetrical relations which can negatively affect each member of the relationship. The solicited friends are required to deal in private matters, even if they would prefer not deal with those kinds of problems. But friends may not have the courage to refuse to help their friends in genuine need. On the migrant's side, a dependency can develop on someone who does not belong to the family circle and with whom he or she shares private and sometimes even intimate information.

Concerning the children, they often find themselves in situations where they are the ones having to deal with administrative procedures, whether physical or online. This point is one of the challenges presented by the recent study conducted in Geneva about the situation of language services provided to migrants³⁴. Many testimonials show that these children, being the only ones who have some level of understanding of the local language, suffer greatly, often in silence, because they bear the full weight of migration and acculturation issues. Several studies show that children of migrants are already subject to psychological stresses provoking voluntary autism caused by the feeling of disloyalty toward their parents. They are torn apart between two spaces; one is the space of their home, their mother tongue and their parents, and the other one is the universe of hostility in the unfamiliar society and the foreign language ignored

33 <http://ticmigrations.fr/fr/whoweare/manifeste>

34 *Voix au Chapitre, Rapport sur l'accompagnement linguistique des migrants non-francophones à Genève*, éd. ICVolontaires.org, 2008.

by their own parents. The only solution they find is silence as a barrier between two antagonistic worlds³⁵.

The two attitudes of silence, depicted above, inevitably influence each other and make more difficult the children's situation, trapped in a vicious circle.

RECOMMENDATIONS / OPPORTUNITIES

While an article like the present one can never be exhaustive, it shows that migration and language issues are a complex phenomenon relevant to many millions of people around the globe. Technology can be an asset for migrants, including situations where they live in an environment with a dominant language(s) other than the one(s) they understand and speak. Online translators and other resources can be very helpful. However, the fact that many administrations do increasingly require all citizens to register for online services is a challenge especially for allophone and illiterate speakers.

The role of children in these contexts is particularly delicate and must be taken into consideration. Children should not replace social workers and should not be placed in situations where they have to play adult roles.

Efforts, such as the one for a World Charter of Migrants³⁶ can put the debate into perspective and provide tools that can be used to develop decent, practical migration policies. In the same line, a comprehensive list of conventions and declarations related to migration policies is available on the website of the International Migration Organization.

The role of online tools and information sources is to help share information, provide practical guidance and promote the use of languages beyond the ones often used in international debates. This, however, requires efforts when it comes to the development of technical tools and scripts, on one hand, and translation carried out by individuals, on the other hand.

35 Dahoun Zerdalia *Les couleurs du silence. Le mutisme des enfants migrants*, Calmann-Lévy, 1995.

36 <http://www.cmmigrants.org>

BIBLIOGRAPHY

[ANAYA JAMES 1996] Anaya James. *Indigenous Peoples in International Law*. New York : Oxford University Press, 1996.

[BAUBÖCK] Bauböck R. "Cultural citizenship, minority rights and self-government". In : T. Aleinikoff and D. Klusmeyer (eds) *Citizenship Today: Global Perspectives and Practices*. Washington : Carnegie Endowment for International Peace.

[BRUBAKER ROGERS 2003] Brubaker Rogers. "The Return of Assimilation? Changing Perspectives on Immigration and its Sequels in France, Germany and the United states", In : Joppke, Christian/Morawska, Ewa (eds.), *Toward Assimilation and Citizenship. Immigrants in Liberal Nation-States*, New York, 2003.

[CLIMENT-FERRANDO 2011] Climent-Ferrando, Vicent « La recherche sur l'immigration en Catalogne : bilan 2000-2010 » [Research on Migration in Catalonia: An Overview 2000-2010] In : *Migrations et Société. Centre d'études sur les migrations internationales*, vol XXIII, 134-135, Paris, Mai-Juin 2011, 251-268.

[EUROSTAT] Eurostat, European Statistics available online at http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Migration_statistics, figures

[JOPPKE 2004] Joppke Christian. "The retreat of multiculturalism in the liberal state: theory and policy" *The British Journal of Sociology*, Volume 55/2 ; 237-257, 2004.

[JUNYENT 2005] Junyent Carme (ed.). *Les llengües a Catalunya. Quantes llengües s'hi parlen?* Barcelona, Editorial Octaedro, 2005.

[KREBS 2007] Krebs Viola. Bilinguisme, inter culturalité et communication politique, dans *L'anglais et les cultures : carrefour ou frontière ? Droit et Cultures*, éd. Harmattan, Paris, 54/2007, <http://droitcultures.revues.org/79>

[KYMICKA, NORMAN 2000] Kymlicka, W. & Norman, W. *Citizenship in diverse societies*. Oxford, Oxford University Press, 2000.

[KYMICKA 2001] Kymlicka, W. *Politics in the vernacular: nationalism, multiculturalism, and citizenship*. Oxford. Oxford University Press, 2001.

[KYMICKA 1995] Kymlicka, W. *Multicultural citizenship: a liberal theory of minority rights*. Oxford, Clarendon, 1995.

La politique cantonale de préformation des non-francophones à risque d'exclusion : Évaluation des mesures de soutien, Commission externe d'évaluation des politiques publiques, Genève, Septembre 2005, p. 15.

[MCROBERTS 2001] McRoberts, Kenneth. "Canada and the Multinational State", *Canadian Journal of Political Science*, 683-714, 2001.

[ZAPATA-BARRERO 2007] Zapata-Barrero Ricard. "Immigration, Self-Government and Management of Identity, The Catalan Case". In : Korniman, M. ; Lauglanf, J. *The Long March to the West: 21st Century Migration in Europe and the Greater Mediterranean Area*. Vallentine-Mitchell, 2007.

[ZAPATA-BARRERO 2007] Zapata-Barrero Ricard. "Setting a research agenda on the interaction between cultural demands of immigrants and minority nations" *Journal of Immigration and Refugee Studies* vol.5, n°4; 1-25, 2007.

Sites web

BABELFISH <http://babelfish.yahoo.com>

CATALAN TRANSLATOR http://traductor.gencat.net/index_en.jsp

EUROSTAT. STATISTIQUES DE LA COMMISSION EUROPÉENNE http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Migration_statistics

GOOGLE AUTOMATIC TRANSLATOR <http://translate.google.com>

INSTITUT STATISTIQUE CATALAN

<http://www.idescat.cat/dequavi/Dequavi.?TC=444&V0=1&V1=8>

INSTITUTO NACIONAL DE ESTADÍSTICA. GOBIERNO DE ESPAÑA

http://www.ine.es/inebmenu/mnu_cifraspob.htm

MAIRIE DE BARCELONE, CATALOGNE <http://www.bcn.cat>

METROPOLIS <http://international.metropolis.net>

MIGRALINGUA <http://www.MigraLingua.org>

MIGRATION INFORMATION <http://www.migrationinformation.org>

MIGRATION ONLINE <http://www.migrationonline.cz/themes/eu>

MIGRAWEB <http://www.migraweb.ch>

MIGRAWEB.ES <http://www.MIGRAweb.es>

TRANSLATION CAFÉ translatorscafe.com

ANNELIES BRAFFORT
& PATRICE DALLE

ACCESSIBILITY IN CYBERSPACE: SIGN LANGUAGES

Sign languages are the natural languages of deaf communities. Their lexicons and grammars are very different from spoken languages. In many countries, spoken language, even in written form, is usually a second language often poorly mastered by deaf speakers. The automatic processing of sign languages is a very new field. In this paper, we make an inventory of existing resources, both in terms of a corpus and systems and tools based on new technologies and associated uses.

Original article in French.
Translated by John Rosbottom.



ANNELIES BRAFFORT is Research Director at LIMSI-CNRS, where she coordinates research on Natural Sign Language Processing (NSLP). This research concerns the study of the body, French Sign Language (FSL) modelling, and text-to-FSL translation. Current applications focus on automatic generation via the animation of virtual signers, which are 3D characters speaking in FSL.



PATRICE DALLE is a professor at Toulouse University 3. He leads the team for IRIT image processing and comprehension. His research focuses specifically on Sign Language modelling, processing and comprehension. Applications include communication in Sign Language and the development of educational tools.

ANNELIES BRAFFORT
& PATRICE DALLE

ACCESSIBILITY IN CYBERSPACE: SIGN LANGUAGES

Sign languages (sL)¹ are natural languages used by the deaf, and some hearing people alongside the deaf. These languages are classified as visual-gestural (issued by the body and received through vision). They are expressed in space, in front of the speaker, by means of gestural units composed of hand and arm gestures, chest, shoulders or head movements, facial expressions, gaze directions, etc., that are carried out simultaneously.

Since Spoken languages (spL) cannot be perceived, sL are the only type of languages available to the profoundly deaf to communicate with their environment. As deafness is hereditary in only 4% of cases, signed languages are native only to a minority of deaf people. For the other 96% of deaf sL speakers, then, the transmission of language is *a priori* not parental². sL provide all the functions performed by non-gestural natural languages and, for the deaf, are really the only appropriate linguistic mode, the only languages that enable psychological and cognitive development in a manner equivalent to the way spoken languages function for hearing people [DALLE 2005].

As with spL, there is no one universal sL. Rather, there exist as many variations as there are different deaf communities, each with its own sL history, signifying units and lexicon³. However, unlike speakers of two different spL, two deaf people operating in two different sL may come to understand each other and communicate in a very short time. This is due to the close proximity of linguistic structures between sL: certain very iconic are characterized by the absence of so-called “standard”⁴ signs

1 Throughout this article, we will use SL for Sign Languages, and SpL for Spoken Languages.

2 <http://corpusdelaparole.in2p3.fr/spip.php?article117>

3 When we speak of LSF and French, these words can be generalized to the SL in a given country and the VL in this country.

4 A lexical unit institutionalised for a given SL, possibly listed in a dictionary.

(each of them different for each language) [CUXAC 2000]. The provenance of these shared structures is probably the very nature of the channel and its propinquity to the mental representations of deaf speakers. What we often refer to as “visual thinking” challenges conventional perceptions of what belongs to the field of linguistics.

Only a few among the hundreds of sign languages in the world have attained legal recognition; the remainder hold no official status. In France, *Law N°2005-102*⁵ for the equal rights, opportunities, participation and citizenship of disabled persons has officially recognised French Sign Language (LSF) since 2005.

SL, because they have no writing system, are eminently oral. This all the more strongly differentiates them from spoken languages: SL are gestural rather than vocal, and with no writing system are uniquely oral. They are thus spoken languages without recourse to forms of transmission or teaching other than the immediate face-to-face, or with a delay through video.

One of the main reasons for the difficulty of creating or borrowing a written form for use in SL lies in their modality: SL exist within space and time, through gestures, postures, mimicry and gaze, all meaningful and potentially simultaneous. This languages' mode of expression is therefore multilinear and multidimensional. By contrast, the ensemble of human writing follows from the (mono)linearity of spoken languages. Currently, no graphics technology makes it possible to provide SL with the primary and daily functions of writing (e.g. recording, note-taking on the fly, linear reading), even if video is sometimes used to fill some of these roles [BRUGEILLE 2006]. The only form of writing available to the deaf is generally that of the sPL of the country where they live. But the majority of the profoundly deaf do not read and write well enough to access a high level of education and training, to access a writing-based means of communication, or to even assume their role as citizens. These facts hinder their professional and personal development.

The fact that SL are unwritten languages, and that access to writing is difficult for the vast majority of the profoundly deaf, results in very incomplete access to information in cyberspace. A few rare sites offer a translation of written text with video in SL, but this remains a very limited phenomenon, with few updates.

5 <http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT00000809647>

This need has generated an interest among researchers in Sign Language Processing (SLP), which would include elements of recognition, generation and machine translation. This research relies on corpus analysis, principally video.

The following section takes inventory of the available SL corpora. The two sections thereafter cover recognition and generation, and the last section addresses existing or future applications [DALLE 2007].

CORPUS : ON LESS-RESOURCED LANGUAGES

SL are less-resourced languages. Indeed, they have access to too few of the resources, and in some cases none at all, that are commonly available to other languages. Specifically, this includes a writing system, reference books describing a language's operation (grammars, dictionaries), mass publishing and distribution (books, press, cultural works), technical and learning books (technical publications, scientific, educational), communication media of everyday life (letters, instruction manuals), as well as computer applications in that language. Similarly, the corpora, which are the only way to establish and maintain a permanent patrimonial record in sign language, are few and small-scale.

Some sign languages have available reference works, cultural works on media like DVDs, and a few TV shows that are interpreted in SL as a locket embedded in the screen, but it is still very limited, and most shows lack even these limited resources. Research suffers in this situation; works detailing the operation of SL are rare and quite limited, corpora are rare and too small, existing SLP systems are only prototypes in laboratories, and are rarely generalisable or even reusable.

Technological advances in video capture, storage and handling have recently (in the XXIst century) enabled the initiation of several projects to create SL corpora, even multilingual corpora, in different countries (Figure 1). Workshops associated with conferences, or funded by national projects, have enabled the scientific community to share experiences on the creation, annotation, analysis and archiving of video corpora⁶. Some recommendations provided by the most experienced researchers are

⁶ <http://www.ru.nl/slcn>; <http://www.sign-lang.uni-hamburg.de/lrec2008/programme.html>

beginning to be included in different projects [JOHNSTON 2008]. However, it is too early to define standards and norms.

Figure 1: Extract from the LSF corpus of the European project Dicta-Sign⁷



The following table lists some SL corpora that already exist or are under development. There are many others, but they rarely exceed a few dozen hours or speakers. This illustrates how recently most countries have begun to finance the development and production of large corpora. The funds themselves are short-term, which often requires large projects to be conducted as a sequence of smaller projects, with the notable exception of Germany, which has funding for fifteen consecutive years. Its corpus, which will be the largest in the world, will comprise only 400 hours of video, far from even approaching the size of the written or audio corpora currently available.

Country	SL	Name of corpus	Size	Status
Australia	Auslan	<i>Auslan Corpus</i> Variously financed since 1990 http://www.auslan.org.au	300 hours 100 speakers	Completed
Great Britain	BSL	<i>BSL Corpus</i> Financed for 3 years : 2008-2010 http://www.bsllcorpusproject.org	249 speakers	Completed
France	LSF	<i>Corpus Creagest</i> Financed for 5 years : 2007-2011 http://www.umr7023.cnrs.fr/-Realisation-de-corpus-de-donnees-.html	130 hours 125 speakers	In process
Allemagne	DGS	<i>DGS Corpus</i> Financed for 15 years : 2009-2023 http://www.sign-lang.uni-hamburg.de/dgs-korpus	400 hours 300 speakers	In process

⁷ <http://www.dictasign.eu>

The formation of significantly large and varied corpora (lexicon, monologue, dialogue, group discussion), that are both archived and accessible, is one of the most important mechanisms both for archiving the heritage of these less-resourced languages, and also for enabling further research on SLP that requires, as with all languages, the analysis of large corpora.

ANALYSIS AND RECOGNITION

The ultimate goal of video analysis in SL [ONG 2005] is to understand automatically the meaning of an utterance, to translate this utterance into another language, or to perform an action, such as a request to a database, or a search for information in a document in SL. In general, this task remains out of reach for computer programs. This leads to limited areas of processing (reduced lexical and semantic fields), imposed constraints on expression and its context, or targeting only semi-automatic programs that assist a human operator. However, more accessible intermediate stages of processing already enable the production of applications.

Where do these difficulties come from, and what can explain the performance gap with NLP or speech processing?

First, the context of this research. As mentioned in the introduction, SL have only recently achieved the status of a language, and in keeping research on SL is also recent. Furthermore, the linguistic models, which could reinforce the computer models, are not yet stable. There exist persistently few researchers studying image analysis with SL as the research topic and not a single applicable framework. In addition, almost none of these researchers understand sign language, the very object of their study. Finally, we know that the tools of speech recognition (vocal language) progressed strongly when they were able to integrate statistical approaches built on large corpora. Due to the lack of data, it is not possible at present to follow the same approach in the case of SL [COOPER 2009].

Another difficulty lies in the nature of the video signal, which is extremely complex to analyse. Different body elements are brought into play to perform in conjunction; their analysis must occur in very different spatial and temporal scales (e.g. simultaneously estimating a fleeting change in gaze and a repetitive swaying of the body), all resulting from a projection onto a plane of 3D postures and movements, which engenders a substantial loss of three-dimensional information and introduces numerous occlusions.

This spatial and multi-component character makes the use of analysis or speech recognition tools, developed for linear and mono-source spoken languages, unpredictable when applied to SL [DALLE 2006].

The processing chain generally presents in two main steps:

- **analysis**, that is, detecting (by identifying and tracking in every video frame) the characteristics of the relevant bodily elements and estimating their parameters; and
- **recognition**, that is, performing a temporal segmentation into units and identifying them by assigning them to a class. These linguistic units have different levels of granularity.

The recognition step is preceded by a learning phase, based on examples and prior knowledge (grammatical rules), which enables an estimation of the parameters of these classes in the framework of a given SL linguistic model.

Early work focused on the recognition of the dactylographical alphabet, that is, using manual configuration to realise the alphabet of the written language. So, not really SL, but rather recognising hand configurations, either isolated (single letters), or chained (to spell a word).

The next step aimed at recognising isolated signs, most often based off of four manual parameters: hand shape, orientation, movement and location of the hands in relation to the body; the given sign involving either one or both hands. Some works report recognising hundreds of signs, with a recognition rate of over 90 %. In reality, performance varies greatly, according to the nature of the data and whether they were acquired using 2D and conventional cameras, 3D featuring stereovision devices, or multi-camera motion capture systems with face and body markers. Generally, these latter systems are used only for learning models (body geometry, movement dynamics, sign signatures), while recognition is realised on 2D videos.

The transition to a continuous signed production analysis is indirect; a statement in SL is not simply a sequence of isolated signs concatenated together. It involves not only the hand signals, but also the postures and movements of the trunk, head and facial expression. The direction of the gaze also plays an important grammatical role. Yet all these elements are

difficult to detect and characterise. Lastly, the signer⁸ uses the space in front of her (the signing space) to support and structure her discourse. Signs, then, are located in this space, and many pointing and referencing operations are observed. Here also, the loss of 3D information and the great variability in possible ways to use this space make it difficult to characterise and model its exploitation by the signer [LENSEIGNE 2005]. Furthermore, processing a continuous production reveals the problems of transitions between signs and coarticulations [SEGOUAT 2010].

The analysis of statements in SL can target varying objectives:

- **Corpus annotation:** This assists the annotator by enriching the signal (3D information reconstruction), by automatically performing certain measurements (gesture dynamics, characterization of facial expressions, etc.), or by detecting specific events (hand contact, particular areas of the face, etc.).
- **Identification and demonstration of syntactical structures:** In particular the exploitation of the signing space (to identify instances of pointing and locate their target.), and to structure the utterance.
- **Sign recognition:** The search for a sign within a continuous stream of SL imply using
 - standardisation methods, to overcome the variations in aspect and scale between signers;
 - time alignment methods, because signing can be made at different speeds;
 - characterization methods, to extract a sign's intrinsic properties rather than the variability introduced by each speaker;
 - comparison methods [ALON 2009].

The most common methods are based on Hidden Markov Models and their variants (coupled, paralleled), attempting to take into account the parallelism of signs and their synchronisation and spatialisation. However, for these methods to be successful, they must be based on decomposing a sign into smaller units of a phonetic nature, whose relevance, definition and detection are still problematic [THEODORAKIS 2010].

- **Sentence comprehension:** The results in this area remain modest [JUNG-BAE 2002]: finding and explaining the grammatical characteristics

⁸ The one who use Sign Language.

of a sentence, or attempting to translate it into an sPL are not simple tasks. The order of signs is not the same as the order of words in a spoken language, and there is not always a systematic sign-word correspondence. On the other hand, the signer has the option to choose between two forms of expression: an illustrative form (“showing while saying”, utilising structures that exploit iconicity) and a non-illustrative form (using standard signs). The illustrative form calls on perceptual-practical experience, and the extent to which a machine can interpret is debatable.

GENERATION TO ANIMATION

sL generation software, coupled with the analysis one presented before, is conceived to enable a complete and bidirectional access to information, whether for the expression and understanding in sL. Moreover this is a response to the difficulty the vast majority deaf adults have in mastering writing. Its potential applications are manifold: web accessibility, sL sub-titling, educational software using sL, and so on.

The process consists of two steps: the generation of the utterance from a linguistic point of view, followed by the generation of the signal received, in the form of an animated virtual character called “virtual signer”. No such sL-generating software using these two steps yet exists, but research in this area is undergoing very active development.

To generate an utterance, there are three main approaches:

- **Generating utterance by concatenation.** This method is used if the utterances are known in advance, are of a finite number, and contain variable parts, typically, information messages or alerts in public places. It is at this point the most developed approach and is implemented into actual use in France (Figure 2) for announcements in railway stations [BRAFFORT 2011]. The system can generate announcements like “Train N°. xx from xx will arrive at platform xx and will be xx minutes late”, the ‘xx’ representing the variable elements of the announcement.

Figure 2: Transmission of Messages in French SL (LSF) at Gare de l'Est in Paris.



- **Generating utterance from scripts.** This approach is based either on video editing tools, or linguistic mechanisms [ELLIOTT 2008], [FILHOL 2010]. The latter approach requires a database of lexical descriptions and a set of grammar rules describing the SL's functioning. These elements are still being investigated in foundational, long-term research, and the SL are not yet fully described.
- **Generating utterance from text.** These methods of machine translation, often inspired by those used for sPL, are statistical mechanisms necessitating very large corpora, which are not yet available for SL. Their other drawback is that they don't model the SL characteristics of iconicity, spatiality, and multilinearity. Another, more recent approach [VENDRAME 2010] is to go through a semantic representation that better resembles SL discourse organization [GUITTENY 2007].

For animation generation, three approaches can also be identified:

- **Generating animations via rotoscoping.** This approach uses video to create a model for a specialised graphic designer, who then uses specialized software to transpose the signer's filmed postures and movements onto an avatar. This method, which is widely-used in 3D animated film production, enables the high quality of animations that is indispensable for a satisfactory content comprehensibility and SL user acceptance. This approach's disadvantages are that the quality of the result is heavily dependent on the graphic designer's experience and talent, and that it entails a certain period of production. However, this approach makes it possible to construct an animation database that can then be concatenated and adapted. This is also the method that is currently used in SNCF software [BRAFFORT 2011].

- **Generating animations via motion capture.** This approach, like the previous one, consists of concatenating and adapting predefined animations, but in this case using motion capture databases [LU 2011]. It similarly requires prior construction of an SL corpus, but in this methodology the corpus is captured by infrared cameras. It has the advantage of providing data directly in 3D, but nevertheless requires a certain production period for a data “cleaning” process that is required prior to use. Moreover, it requires the use of a motion capture system.
- **Automatic animation generation.** This refers to generating animation from symbolic description. It stems from the research problematic of the fields of computer graphics and biology (physiology, movement modelling, etc.). Some websites, especially those dedicated to bilingual dictionaries, incorporate software built using this approach⁹. Automatically generated animations are still rather robotic and lacking in realism, particularly when it comes to the naturalness of movement and the animation of facial expressions. The latter difficulty arises from the current lack of knowledge about non-manual elements function in SL. Some studies are now beginning to address this issue [CHÉTELAT 2011].

The advances in SL generation are indisputable, and internationally increasing numbers of research teams are gaining interest in this field. To produce significant advances, upstream studies on SL function analysis must be developed, by creating large corpora and supporting linguistic and multidisciplinary studies that incorporate the results of this research in their models and software.

CYBERSPACE : TOWARDS NEW APPLICATIONS

Until now, devices for the deaf enabling them to live bilingually, that is, to practice their own sign language in a predominantly hearing world using the country’s spoken language, consisted mainly of human assistance, essentially, the model of face-to-face interpreters, a model more or less unrelated to cyberspace.

New technologies and the internet have enabled these devices to be expanded using video conferencing systems: remote communication

⁹ www2.cmp.uea.ac.uk/~jrwg/Dictionary094

between signing and non-signing individuals via interpreters at a relay centre, or directly in SL between signing individuals¹⁰, is now possible. However, these mechanisms may only be used for communication between people.

New systems [DALLE 2011], focusing on content access or on the SL itself [LEFEBVRE 2010B], are now beginning to appear in several fields.

In terms of content access, the first applications targeted accessibility for existing websites. They aimed to integrate online assistance with SL video format, either to translate parts of a document or to add explanations in SL (to alleviate the difficulty many deaf people experience when reading). However, this complementary form of SL can only be used for pages that rarely change; and so, despite initial enthusiasm, these devices haven't continued to develop. The WebSourd company website (Figure 3), which provides daily translations of AFP (Agence France Presse) bulletins in SL video, remains an exception.

Advances in SL generation and analysis have not yet translated into new products in this area, but we can see in experimental products some hints of future developments:

- **Website enrichment:** ECA (Embodied Conversational Agents) are used to support navigation; virtual signers, viewable on demand, can be used to provide complementary SL¹¹; for example, by translating a site's FAQ page. Compared to a video of a real signer, they offer the advantage of anonymity, they conform to the look of the site, and they can be modified, especially if the generation module is powered by a translation module from text to SL;
- **Wiki in SL:** An even more ambitious prototype application, which has been explored during the European project DictaSign, is the realisation of a *wiki* in SL. The wiki consists of partial recognition of information in SL that the user provides via webcam, along with messages generated and performed by a virtual signer;
- **Bilingual sites:** SL is now well established on the web, through high quality bilingual information websites¹². Beyond the useful and cultural value for those involved, they also serve to exhibit SL and make

10 <http://www.afils.fr/index.php> conseils "Du bon usage des centres relais".

11 <http://www.limsi.fr/Individu/jps/online/diva/geste/geste.main.htm>

12 <http://www.websourd.org>

it accessible to all. Furthermore, these sites have an impact on the language itself: as they employ a more academic level of language than that used in everyday signing, these sites have a diffusive effect on the local lexicon and the spread of neologisms;

Figure 3: Intensive video use on the website WebSourd.



SL dictionaries: Cyberspace is the perfect venue for this type of document. Although they remain rare, several encyclopaedia projects are underway (Elix¹³, Ocelles [MOREAU 2010]), together with some thematic glossaries (UVED¹⁴);

The use of this SL content consists of several components [DALLE 2011]:

- Content realisation : This refers to SL and bilingual documents. Easy-to-use tools [LEFEBVRE 2010B] for including a comment in SL using video (SL video, video of a text or a presentation) now exist;
- Acceptance of signs used : In a rapidly evolving but little standardised language, this issue is particularly acute, and ongoing projects are providing mechanisms to identify the geographical origin of signs and their level of acceptance within the deaf community. This is especially helpful with neologisms;
- Presentations in SL : Before working with a document in a SL, you need to know of its existence and be able to assess its relevance. As SL do not have a writing system, apart from text usage, special techniques¹⁵ for presentation in image form, photosigns (Figure 4), and mini-video enable tables of contents and indexes to be provided in SL.

¹³ <http://www.signesdesens.org/-e-learning-.html>

¹⁴ <http://www.irit.fr/GlossaireDD-LSF>

¹⁵ http://www.usherbrooke.ca/liaison_vol41/n08/a_avaglyphe.html

- Content navigation in SL: Finally, the routes through an SL document should be able to be consumed non-linearly, and experiments are underway¹⁶ to replicate in video the equivalent of a clickable link in text. For this, it is necessary to know how to display the link, and subsequently how to use it. On the other hand, research in pattern recognition should help in the short term to perform queries directly in SL, at least in the form of isolated signs.
- **Distance education in LSF:** Beyond creating online support for LSF, distance education in LSF is beginning to develop (In France: DAEU¹⁷ Nancy¹⁸, DU LSF Rouen¹⁹, DU IELS Toulouse²⁰). While the courses are most often offered in the form of downloaded SL videos, student progress monitoring can occur directly in LSF using online resources (videoconferencing, LSF forums, etc.) [TANAKA 2010];
- **Learning LSF:** The first websites in LSF were language presentation sites, essentially French lexicons providing each entry in French with its signed equivalent; terms were listed either alphabetically or by theme. In LSF, an order of presentation is defined, usually parametrically, with signs being classified at a first level according to hand shape (choosing from among around fifty configurations). It will soon become possible to make a request in LSF via webcam [LEFEBVRE 2010A]. However, these sites are of limited value for learning LSF. LSF teaching websites²¹ are often online registered courses repositories where students can send their exercises in video form. Again, we can expect some interesting developments for self-learning, thanks to tools comparing signs and generation by virtual signers [ARAN 2009];
- **Teacher resource platform of/in LSF:** The recognition of SL and its introduction in education (bilingual education, an LSF option for the French Baccalaureate) has led to the creation of teacher resource sites (CNDP, e-LSF) and extensive use of SL video, but less so the most advanced LSF techniques.

16 <http://www.signlinkstudio.com/en/index.php>

17 Diplôme d'accès aux études universitaires – diplôme d'université.

18 <http://erudi.free.fr/index.php?page=daeulsf>

19 http://formations.univ-rouen.fr/LSA31_864/0/fiche___formation

20 <http://www.irit.fr/iels>

21 <http://www.signingsavvy.com>

Figure 4: Examples of photosigns
DEVELOPMENT and *SUSTAINABLE* in LSF



CYBERSPACE, AN OPPORTUNITY FOR SL

Because deafness is not seen, it isn't taken into account except when the deaf express themselves in SL. Cyberspace presents an important opportunity for SL:

- **For minority languages that are not geographically localised:** Their practice is doubly penalised; it is therefore crucial to increase sites of SL expression, that is, to recreate social spaces;
- **Multimedia languages, for which video is the “written” form:** For these languages, the web and its ability to disseminate image, animation and video, as well as its potential for interaction, constitutes a privileged space;
- **Languages for which the grammatical component is fairly universal:** These languages are facilitating international exchanges.

Cyberspace should enable mother-tongue education for deaf children. Communication in sign language beginning at birth can restore a normal environment in which deaf children can build a real language, provided the SL used is correct. For these children, only SL can play the role of native language. To do this, it must be or must also become the parents' language. Most children who are born deaf have hearing parents who have not mastered (or don't know) sign language, and they must be trained early and at home. These families are dispersed throughout the country (there is no community of families of deaf children), so it is difficult to establish regular group training.

Distance learning in SL, and the dissemination of language resources and sites via internet provide one answer to this problem. Self-teaching systems for SL, including corrective mechanisms via recognition or generation, would facilitate SL learning for families and its early acquisition by deaf children, who would then enter a life with language.

Deaf people often lack quality training and mastery of the written form of their country's spoken language(s). SL supports, the web resources to distribute them, SL encyclopaedias, and online training can help fill these gaps. While production techniques for these supports exist, there are currently few operational solutions for SL content access (queries in SL, indexing, navigation in SL) and generation (content production by virtual signers). However, these systems would be very useful for the deaf, for their cultural development, and thus their social integration.

The development of these systems entails more complete and realistic computer modelling of SL, incorporating models of variations, either to recognise spontaneous utterances or to generate realistic productions by a virtual signer. Campaigns for the constitution of large high-quality corpora should be undertaken to accelerate the development of these models, support research into SL recognition and generation, and thus enable SLP to approach the performance level of NLP.

Network performance, web server storage capacity, and the power of personal computers now make possible the exchange of video in fluid sign language. Remote communication between signers is operational, as along with access to pre-recorded content in sign language. On the other hand, enabling SL cyberspace interactions between user and software resources remains to be accomplished. The absence of writing in sign language must be compensated for by researching methods and tools that facilitate the chain-editing of SL documents and thereby feed SL content into cyberspace. Access to such content must be SL-interactive, which requires the development of research into the recognition of requests in SL so that users can initiate, as well as the generation by the system of contextual responses produced by virtual signers. Under these conditions, cyberspace will become a real resource for the deaf, allowing them access information in their own language, and a real opportunity for sign languages by multiplying and enriching their fields of expression.

BIBLIOGRAPHY

- [ALON 2009] Alon J., Athitsos V., Yuan Q. & Sclaroff S. (2009). "A Unified Framework for Gesture Recognition and Spatiotemporal Gesture Segmentation", *IEEE Transactions of Pattern Analysis and Machine Intelligence (PAMI)* vol. 31 no 9 p. 1685-1699
- [ARAN 2009] Aran O., Ari I., Akarun L., Sankur B., Benoit A., Caplier A., Camp R, Carrillo AH., Fanard FX. (2009). "SignTutor: An Interactive System for Sign Language Tutoring", *IEEE Multimedia*, vol. 16 p.81-93.
- [BRAFFORT 2011] Braffort A., Bolot L. & Segouat J. (2011). "Virtual signer coarticulation in Octopus, a Sign Language generation platform". *International Gesture Workshop*, Athènes.
- [BRUGEILLE 2006] Brugeille JL., Dalle J. & Kellerhals MP (2006). "Une expérience d'utilisation de formes graphique dans la scolarité des enfants sourds: méthode de travail et premières observations", *Colloque Syntaxe, interprétation, lexicque des langues signées*.
- [CHÉTELAT 2011] Chételat E. & Braffort A. (2011). "Investigation and analysis of non manual gestures involved in LSF: blinking". *International Gesture Workshop*, Athènes.
- [COOPER 2009] Cooper H., Bowden R. (2009) "Sign Language Recognition : Working with Limited Corpora", *Universal Access in HCI, Part III*, HCII 2009, LNCS 5616, pp. 472–481.
- [CUXAC 2000] Cuxac C. (2000). "La langue des signes française – les voies de l'iconicité". *Faits de langues* n°15-16 Ophrys.
- [DALLE 2005] Dalle P. (2005). "Histoire et philosophie du projet bilingue", *Nouvelle Revue de l' AIS*, Hors série "Enseigner et apprendre en LSF".
- [DALLE 2006] Dalle, P. (2006). "High level models for sign language analysis by a vision system", *2th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, 5th edition of *Language Resources and Evaluation (LREC)*.
- [DALLE 2007] Dalle P., Braffort A., Collet C. (2007). "Accessibilité et langue des signes: modélisations, méthodes, application", *Conférence internationale sur l'accessibilité et les systèmes de suppléance aux personnes en situations de handicaps (ASSISTH 2007)*, Cépaduès, p.209-217.
- [DALLE 2011] Dalle P. (2011). "TIC au service de la LSF", colloque GERS Grandir et apprendre en LSF, *Revue Contact sourds entendants*, l'Harmattan.
- [ELLIOTT 2008] Elliott R., Glauer J.R.W., Kennaway J.R., Marshall I. & Safar E. (2008). "Linguistic modeling and language-processing technologies for Avatar-based sign language presentation". *Universal Access in the Information Society*, 6/4, Springer.
- [FILHOL 2010] Filhol M., Delorme M. & Braffort A. (2010). "Combining constraint-based models for Sign Language synthesis". *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, 7th edition of *Language Resources and Evaluation Conference (LREC)*.
- [GUITTENY 2007] Guitteny P. (2007). "Langue des signes et schémas", revue *Traitement Automatique des Langues (TAL)* Vol 48 2007. 3. "Modélisation et traitement des langues des signes". <http://www.atala.org/-Modelisation-et-traitement-des->

- [JOHNSTON 2008] Johnston T. (2008). “Corpus linguistics and signed languages: no lemmata, no corpus”. *3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, 6th edition of the *Language Resources and Evaluation Conference (LREC)*.
- [JUNG-BAE 2002] Jung-Bae K., Kwang-Hyun P, Won-Chul B., Zenn Bien Z. (2002). “Continuous Korean sign language recognition using gesture segmentation and Hidden Markov Model”, *FUZZ-IEEE'02. IEEE Int Conf on Fuzzy Systems*, p. 1574-1579.
- [LEFEBVRE 2010A] Lefebvre-Albaret F., Dalle P (2010). “Requête vidéo dans une vidéo en langue des signes: Modélisation et comparaison de signes”, RFA.
- [LEFEBVRE 2010B] Lefebvre-Albaret F., Dalle J., Piquet J., Dalle-Nazébi S., Gache P, Bacci A., Dalle P (2010). “Analyse des langues des signes. Démarche de conception pluridisciplinaire d’outils d’analyse de discours en langues des signes”, *Technique et science informatiques (TSI) n° spécial L’informatique à l’interface de l’activité humaine et sociale Vol. 29 N° 8-9 pp.959-989*
- [LENSEIGNE 2005] Lenseigne B., Dalle P (2005). “Using Signing Space as a Representation for Sign Language Processing”, *6th International Gesture Workshop*, Springer-Verlag, p. 25-3.
- [LU 2011] Lu P, Huenerfauth M. 2011. “Collecting an American Sign Language Corpus through the Participation of Native Signers”. *International Conference on Universal Access in Human-Computer Interaction (UAHCI)*.
- [MOREAU 2010] Moreau C., Masclet B. (2010). “Organizing data in a multilingual observatory with written and signed languages”. *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, 7th edition of *Language Resources and Evaluation Conference (LREC)*.
- [ONG 2005] Ong S.C., Ranganath S. (2005). “Automatic Sign Language Analysis: A Survey and the Future beyond Lexical Meaning”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 6, pp. 873–891.
- [SEGOUAT 2010] Segouat J. (2010). “Modélisation de la coarticulation en Langue des Signes Française pour la diffusion automatique d’informations en gare ferroviaire à l’aide d’un signeur virtuel”, Thèse doctorat de l’université Paris-Sud 11, Orsay.
- [SCHNEPP 2010] Schnepf J., Wolfe R. & McDonald J. C. (2010). “Synthetic Corpora: A Synergy of Linguistics and Computer Animation”. *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, 7th edition of *Language Resources and Evaluation (LREC)*.
- [TANAKA 2010] Tanaka S., Matsusaka Y., Nakazono K. (2010). “Development of E-Learning Service of Computer Assisted Sign Language Learning: Online Version of CASLL” *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, 7th edition of *Language Resources and Evaluation (LREC)*.
- [THEODORAKIS 2010] Theodorakis S., Pitsikalis V., Maragos P (2010). “Model-Level data-driven sub-units for signs in vidéos of continuous Sign Language”, *IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, p. 2262, 2265.
- [VENDRAME 2010] Vendrame M. & Tiotto G. (2010). “ATLAS Project: Forecast in Italian Sign Language and Annotation of Corpora”. *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, 7th edition of *Language Resources and Evaluation (LREC)*.

HOW ORAL ARCHIVES BENEFIT ENDANGERED LANGUAGES

The work of the Fryske Academy (Frisian Academy) and the Mercator European Research Centre on Multilingualism and Language Learning is dedicated to the study of minority languages in Europe. The list of languages being at risk of extinction has increased dramatically in Central and Eastern Europe, and more in Russia and eastern Siberia. This article presents the scientific projects aimed using linguistic materials from archives or harvested from fieldwork data.

Original article in English.



TJEERD DE GRAAF holds doctoral degrees in arts, in linguistics, as well as in theoretical physics. He was an associate professor of phonetics in the language department of the University of Groningen from 1975 to 2003. He now conducts his research in the Frisian Academy and the European Mercator Research Centre on Multilingualism and Language Learning.

LEERD DE GRAAF

HOW ORAL
ARCHIVES
BENEFIT
ENDANGERED
LANGUAGES

The **Fryske Akademy** (Frisian Academy) and the Mercator European Research Centre on Multilingualism and Language Learning are devoted to the study of minority languages in Europe. The Academy's primary involvement lies in the history, literature and culture related to the West-Frisian language. Speakers of its nearest relatives, the East- and North-Frisian languages in Germany, are less numerous and these languages are included into the list of endangered languages of Europe. This list increased significantly after the extension of the European Union with new member states in Central and Eastern Europe. Further eastwards, in the Russian Federation and Eastern Asia, a large number of endangered languages can also be found. This report presents existing and potential projects related to some of the endangered languages in the Russian Federation, in particular those based on the use of material from sound archives and fieldwork data.

HISTORICAL DATA IN SOUND ARCHIVES

In the last half of the XIXth century, Thomas Edison drastically changed the possibility of doing linguistic research [DE GRAAF 1997, 2002c] with his 1880s invention of the phonograph, which could record sounds. For the first time, people were able to store and rehear acoustic data, in particular speech, and to reproduce it for other sound carriers. Not long after this invention, ethnologists, folklorists, linguists, composers, and amateurs began to use the new machine to collect information on the oral and musical data of cultural groups at home and abroad.

Prior to 1890, linguists in the field had to take notes by hand, which required many repetitions of spoken utterances, a laborious process for both investigator and informant. The phonograph changed all this; linguists

could now obtain an accurate, objective and instantaneous record of a single performance. It was possible to capture the nuances and subtleties of the spoken word, duplicates could be played repeatedly for transcription and analysis, and the original recordings preserved for future use.

For best results in the reproduction of sound from the old wax cylinders, several modern cylinder players have been built which employ lightweight pick-up cartridges for mechanical extraction of the signal. In order to minimise degradation of cylinders by replay, and to make contents retrievable from broken cylinders, several optical methods have been developed for contactless, non-destructive replay. The first was introduced by a Japanese research group [ASAKURA *ET AL.* 1986]. In 1988 I was invited to work a few months with this group in Sapporo (Japan) where I could apply this method to some wax cylinders and learn from the experience of my Japanese colleagues.

Using the phonograph over the years from 1902 to 1905, the Polish anthropologist Bronisław Piłsidski recorded the speech and songs of the Ainu people on Sakhalin and Hokkaido on wax cylinders to study their culture. These wax cylinders were discovered in Poland and taken to Japan, where Prof. Asakura's research group contributed to the reconstruction of this valuable material. During my stay in Japan Prof. Kyoko Murasaki introduced me to the last speakers of Sakhalin Ainu, who were living on Hokkaido [MURASAKI 2001] and suggested that we might go together to Sakhalin to conduct fieldwork. Until 1988 Sakhalin was completely isolated from the outside world; Gorbachov's perestrojka made it possible to organise the first international ethnolinguistic expedition to the island, which I joined in 1990 [DE GRAAF 1992]. We found no remnants of the Ainu population, but visited various parts of Sakhalin where the Nivkh people are living. The following sections of this article will report on the projects related to the use of sound archives for the study of minority languages.

SOME PROJECTS RELATED TO ENDANGERED LANGUAGES AND SOUND ARCHIVES

Our research group on Phonetics and Ethnolinguistics has investigated various aspects of the languages spoken in the Russian Federation. In this report we shall describe a few projects that have been undertaken by the research group and elsewhere for the study of the minority peoples of

Russia and for the description of endangered languages. For this purpose, data from archives have been used and combined with results of modern fieldwork in several parts of the Russian North, Siberia, the Russian Far East and the border areas of Russia and Japan. Since 1992, these projects have been financially supported by the Netherlands Organization for Scientific Research (NWO), the Organization INTAS of the European Union, and the Sakhalin Energy Investment Company LTD. We have collaborated with colleagues in Russia and Japan and part of our work is simultaneously related to Japanese research projects.

When recordings were made, it became obvious that a central facility was needed to preserve the valuable data that had been collected. At the beginning of the XXth century, this led to the establishment of sound archives, the earliest of which in Europe were located in Vienna, Berlin and St. Petersburg. The sound archives of the Russian Academy of Sciences in the Museum of Russian Literature (*Pushkinsky Dom*) in St. Petersburg contain about 7,000 wax cylinders of the Edison phonograph and more than 500 old wax discs. In addition, an extensive fund of gramophone records and one of the largest collections of tape-recordings of Russian folklore represent the history of Russian ethnography and contain a wide range of materials [DE GRAAF 2001, 2002A]. Many of these recordings form one of the basic collections used in our joint projects with St. Petersburg.

The first of these projects, on the Use of Acoustic Data Bases and the Study of Language Change (1995-1998), was financially supported by the organization INTAS of the European Union in Brussels. We reconstructed some of the many recordings in the *Pushkinsky Dom* and to made them available for further research, which is not only important for historical and cultural reasons, but also for language description and for the study of possible direct evidence of language change. In a second INTAS project, *St. Petersburg Sound Archives on the World Wide Web* (1998-2001), some of the sound recordings were placed on the internet and are now available at a special website for further study [DE GRAAF 2004]. In both projects, the *Phonogrammarchiv* of the Austrian Academy of Sciences was both partner and technical supervisor.

For these projects, we first completed the reconstruction of the sound archive material of the Zhirmunsky collection. Zhirmunsky was a famous linguist who worked in St. Petersburg/Leningrad in the early XXth century. One of his main interests was the study of German dialects spoken in

Russia. Between 1927 and 1930, he recorded many utterances, in particular songs by German settlers, on waxed cardboard discs, which were transferred to the Vienna *Phonogrammarchiv*. Within the framework of the INTAS project, this collection has been copied onto tape and part of the material is now stored in a special database. A special study covered the language of the Siberian Mennonites [DE GRAAF 2005].

For our third INTAS Project, on The construction of a full-text database on Balto-Finnic languages and Russian dialects in Northwest-Russia (2000-2003), we inventoried the Finno-Ugric minority languages in the vicinity of St. Petersburg and the southern and middle parts of Karelia. They represent a specific linguistic picture of an area where endangered languages such as Vepsian, Ingrian, Votic, Ingrian-Finnish and Karelian and various types of Russian archaic dialects are spoken in close proximity to this day.

The St. Petersburg sound archives also contain important data on Yiddish, the language of the Jews in Eastern Europe, which at the beginning of this century was spoken by millions of speakers in the Russian empire. In the archives we found an unpublished manuscript, *The Ballad in Jewish Folklore*, which corresponded to Yiddish material on wax cylinders. Together with specialists in St. Petersburg, we further explored the acoustic data in the sound archives and prepared the edition of the book. This took place as part of the project *Voices from the Shtetl, the Past and Present of the Yiddish Language in Russia* (1998-2001), for which we obtained financial support from the Netherlands Foundation for Scientific Research NWO [DE GRAAF, KLEINER AND SVETOZAROVA 2004].

Modern fieldwork and reconstructed data from sound archives provide important information for the preparation of language descriptions, grammars, dictionaries and edited collections of oral and written literature. During fieldwork expeditions to Northern Yakutia, the Altai Region and Sakhalin, we studied the processes of language shift and language death of the aboriginal populations of Russia, collecting much interesting data.

THE LANGUAGES OF SAKHALIN

As mentioned above, our first international expedition in 1990 took place to the island of Sakhalin, where we were looking for the Ainu, Nivkh and Uilta people and making recordings of their languages.

The island of Sakhalin belongs to the Sakhalin area (*Sakhalinskaya Oblast'*), one of the most eastern territorial units of the Russian Federation, 87,100 square kilometres large and only 980 km long from North to South. The Kurile Islands, a chain of 1200 km length with 36 islands, are also part of this territory. The original population of Sakhalin consisted of some Paleo-Siberian and Tungusic tribes, in particular the Nivkh (Gilyak) and Uilta (Orok) in the North and Center, and the Ainu in the South. Their numbers were rather small and during the colonization process by the Russians from the North and by the Japanese from the South, they were quickly numerically dominated by these stronger nationalities. Due to their isolated life far from the political centre, they were able to maintain their native language and culture for a long time, but since the beginning of the xxth century the assimilation process has gradually become stronger.

In the summer of 1990, I took part in the first international field work expedition to Sakhalin, aiming to investigate the linguistic and ethnographic situation of the smaller nationalities on the island. The idea was to look for the remnants of the Ainu population and for the other small minority groups, in particular Nivkh (Gilyak) and Uilta (Orok). Unfortunately, during our expedition no Ainu people were found, and the only person representing the Sakhalin Ainu language and culture was probably the informant we met on Hokkaido, Asai Take san [DE GRAAF 1992], [MURASAKI 2001].

Ainu is the only small endangered indigenous language of Japan, whereas Nivkh is a representative of the many minor languages of Russia. From the available demographic data, we concluded that in 1989 the aboriginal peoples of the North formed a very small minority within the total population of Sakhalin: for the Nivkh ethnic group, which is the largest group, the percentage was only 0.3 % [DE GRAAF 1992].

Among the small nationalities in the Russian Federation, the minority peoples of the North play a special role. There are nearly thirty different groups, all living in the northern parts of the country bordering the Arctic Ocean from Scandinavia to the Bering Sea and the Pacific. The Peoples of

the North were the last to be put under effective Soviet rule. In the early thirties the Soviet regime tried to extend its grip on these peoples and to encourage Russian culture and literacy among them. A “Committee for the Assistance and Protection of the Small Peoples of the North” was founded in 1923 and a writing system developed for many of the minority languages. Initially the Latin alphabet was used, but in the later thirties this was changed to Cyrillic.

The Nivkh language is classified as Paleo-Siberian and spoken by tribes inhabiting the lower reaches of the Amur River in the Far East of the Asian continent and the northern and central parts of Sakhalin [GRUZDEVA 1998]. One of its linguistic complications is the fact that the language has at least two dialects: Amur and Sakhalin. Both groups are rather small: all together about 4400 people have the Nivkh nationality, and less than 15 % of them speak Nivkh. A very small group speaks the southern Poronaisk dialect, and for this dialect it is very difficult to find speakers. After the war, several of them emigrated from their homeland in Southern-Sakhalin to Japan, where Japanese and other non-Soviet linguists studied their language.

The first all-Russian census was organised during the czarist regime in 1897. The total number of people on Sakhalin, belonging to the Nivkh ethnic group, was given as 1969. All identified Nivkh as their mother tongue and most were probably monolingual. In the census in 1926, the first organised by the Soviet Union, the total number of Nivkh people was lower, due to the fact that the inhabitants of the Japanese southern part of Sakhalin were not counted. Practically all still identified Nivkh as their mother tongue. Since then, however, a decrease in the percentage of Nivkh speakers has been observed even if the number of Nivkh on Sakhalin has remained stable (about 2000). In 1989, over 80 % of Nivkh people no longer spoke Nivkh, and identified Russian as their first language.

The transition of the Sakhalin Nivkh to Russian can be explained in a number of ways. One of the most important factors was the growing contact of the Nivkh population with the other inhabitants on the island, many of them Russian speakers from the motherland who came to the island to exploit its natural resources (oil, coal, wood, fish, caviar). Before then, the Nivkh people lived as fishermen and hunters in their isolated villages, but they increasingly came into contact with the immigrants,

who began an active policy of educating and influencing the aboriginal inhabitants of the eastern parts of the Russian Federation.

Recently, a development is taking place in favour of the native languages and cultures of the small minorities in the Russian Federation, in particular the Nivkh [DE GRAAF, SHIRAISHI 2004]. Attempts are being made to revive the Nivkh language, for example by introducing language classes in several schools in Nivkh. In 1980, the Ministry of Education of the Russian Federation initiated a program for primary and secondary schools, for which textbooks and dictionaries were printed in Nivkh. Special instruction was given to teachers of Nivkh descent about the education of Nivkh children in their own language. This teaching program was introduced in the special boarding schools for children from the ethnic minorities in Nogliki, Chir-Unvd and in Nekrasovka. We were able to visit these schools and to learn about the teaching methods for Nivkh used in the primary education.

During our fieldwork expeditions on Sakhalin, important linguistic material was collected on the languages of the minority groups. Most subjects were elderly people with a strong motivation to use their language, for example as members of a folkloric group. Practically all young people we met had no active knowledge of the language, and communicated only in Russian with their parents. During interviews with Nivkh informants, they were very positive about the value of keeping and cultivating their own culture and expressed a desire to participate as members of the Russian Federation's group of nations. They agreed that Russian language and culture play a very important role in their lives, but expressed wanting to see the survival of their native language and culture stimulated by all possible means.

VOICES FROM TUNDRA AND TAIGA

Important activities related to linguistic databases in St. Petersburg concern the recordings of Russian dialects and minority languages in the Russian Federation, such as Nivkh, Tungus, Yakut and others [DE GRAAF 2004]. One of our aims is to use these recordings to construct a phonetic database of the languages of Russia, which will have many scientific, cultural and technical applications. Within the framework of the research program *Voices from Tundra and Taiga*, which began in 2002, we combined

the data from old sound recordings with the results of modern fieldwork to give a full description of the languages and cultures of ethnic groups in Russia. The endangered Arctic languages and cultures of the Russian Federation must be described rapidly before they become extinct. Our earlier work on reconstruction technology for old sound recordings found in archives in St. Petersburg has made it possible to compare languages still spoken in the proposed research area with the same languages as they were spoken more than half a century ago, which provides a fortunate start for these projects. The sound recordings in the St. Petersburg archives consist of spoken language, folksongs, fairy tales, and so forth, in Siberian languages among others [BURYKIN ET AL. 2005], [DE GRAAF 2004A].

*Map of the languages of Siberia*¹



In these projects, the techniques developed earlier have been applied to some of the disappearing minority languages and cultures of Russia, such as Nivkh and Uilta on Sakhalin and Yukagir and Tungusic languages in Yakutia. Our goal is to set up a phono- and video-library of recorded stories, folklore, singing and oral traditions of the peoples of Sakhalin and Yakutia. Thus the existing sound recordings in the archives of Sakhalin

1 *Map of the languages of Siberia*, by courtesy of the Max Planck Institute, Leipzig. © MPI for Evolutionary Anthropology.

and Yakutia will be complemented by new fieldwork results. The data obtained will be added to the existing archive material in St. Petersburg and made partly available on the internet and CD-ROM.

This research project and the related documentation are carried out in close cooperation with scholars in local centres such as Yuzhno-Sakhalinsk who participate in archiving sound recordings and fieldwork expeditions. Specialists from St. Petersburg and the Netherlands visit them to set up new centres for the study and teaching of local languages and related subjects. For this purpose, we organised a special seminar for Nivkh teachers in Yuzhno-Sakhalinsk in October 2003.

Spontaneous speech and the reading of prepared texts is collected for (ethno)linguistic as well as for anthropological, folkloristic and ethno-musicological analysis. These data are videorecorded and analysed and can illustrate the art of storytelling and language use. The above-described texts will be published in scientific journals and books with audiovisual illustrations on CD-ROM and/or on the internet. The materials will thus become available for further analysis to scholars working in the field of phonetics, linguistics, anthropology, history, ethno-musicology and folklore.

Using a phrase book for school children of Nivkh [TAKSAMI *ET AL.* 1982], we recorded a native speaker during our fieldwork trip in 1990. The texts with the illustrations of the book are now shown on the Internet together with acoustic data. The separate phonemes are supplied on a special table; by selecting one of them the student can listen to various speech sounds. This has as the advantage that students will be able to learn the distinction between various separate phonemes (e.g. four k-sounds) of Nivkh, which are variants (allophones) of one phoneme in Russian. One of our research students and his Nivkh colleague published a series of books with Nivkh stories, songs and conversation in which for the first time the corresponding texts are recorded on a CD. The series, *Sound Materials of the Nivkh Language I - III* [SHIRAISHI, LOK 2002, 2003, 2004] appeared as part of the Japanese program on Endangered Languages of the Pacific Rim (ELPR) and the research program *Voices from Tundra and Taiga*. This unique material is used not only by linguists, but also by the language community itself, where it can be used for teaching purposes. In 2006, Hidetoshi Shiraishi finished a dissertation on this topic entitled

Aspects of Nivkh Phonology, which he defended in September 2006 at Groningen University [SHIRAISHI 2006].

ENDANGERED ARCHIVES

In summer 2005, we reported on the NWO research project *Voices from Tundra and Taiga*, and published a catalogue of existing recordings of stories, folklore, singing and oral traditions of the peoples of Siberia [BURYKIN ET AL. 2005]. This material was thus made available for further analysis by researchers working in the field of phonetics, linguistics, anthropology, history, ethno-musicology and folklore. The information is also highly important for the development of teaching methods for representatives of the related ethnic groups and for the conservation and revitalisation of their languages and cultures.

At present, many old recordings remain hidden in private archives and places where the quality of preservation is not guaranteed. In a project, which from September 2006 until September 2008 was financially supported by a special Programme on Endangered Languages at the British Library, we made part of these recordings available and added them to the database developed in St. Petersburg.

Our partner in this new project on Endangered Archives has again been the *Phonogrammarchiv* of the Austrian Academy of Sciences. The aim was to re-record the material using updated technology [SCHÜLLER 2005], and to store them in a safe place together with the metadata, which is obtained from the related secondary data. The storage facility provided by the project could modernise the possible archiving activities in the Russian Federation and bring them to updated technological standards.

The original open reel and cassette tapes were copied onto hard discs. In the total collection of more than 111 hours (218 GB) of digitised sound material, the following languages are represented: Azerbaijani, Balochi, Chagatay, Chatror, Dari (Farsi-Kabuli), Enets, Kati, Kerek, Mendzon, Nenets, Nganasan, Parachi, Pashai, Pashto, Russian, Shugni, Tajik, Udeghe, Vaygali and Wakhi (Vakhan). The data in this digital sound archive provide information about the historical development of these languages and can be used for language description, the study of folklore, and ethnomusicology of many of Russia's endangered minority languages.

In other parts of Russia, similar important collections can be found, not only in established institutions, but in private hands where they often lack sufficient protection, for example the private collections on Nivkh, available in Yuzhno-Sakhalinsk, in Vladivostok, in London and elsewhere. For most of these, it can be said that the preservation quality is below standard. Following our long-standing collaboration with scholars from Sakhalin, we plan to create facilities in Yuzhno-Sakhalinsk to store sound material related to the aboriginal languages of the island. Most important are the above-mentioned Nivkh collections, but we should also like to add material on Sakhalin Ainu and Uilta. For some of these private collections, the size is approximately known, but in others a preliminary estimation is required. Within the framework of our project and future new projects, we hope to obtain access to these collections, copy them on modern sound carriers, make a catalogue available and publish part of the material together with the related recordings in St. Petersburg. On Sakhalin and in other parts of Russia, local scholars will be involved in the preparation of these projects with the support of colleagues in St. Petersburg, Austria, the Netherlands and Japan.

CONCLUSION

A joint effort by researchers from Russia and the Netherlands to analyse data from audio archives and at the same time apply modern fieldwork techniques in studying endangered languages such as Nivkh, Nenets and Yukagir, is described above. The results are language descriptions, grammars, dictionaries and edited collections of oral and written literature on and in these languages. In seminars, the use of these learning and teaching materials within the modern facilities of information technology can be passed on to local teachers. Formal language teaching of former mother tongues can be directed to those younger members of the communities who have not learned their native language informally at home. In Russia, special methods for teaching the former mother tongue as a foreign language must be applied. Selected parts of the acoustic databases used for specific projects are available online and provide an opportunity for the exchange of information on these languages with institutions in other parts of the world.

At the local community level and over the past several decades, many people have been working to develop language education programmes,

usually with extremely limited technical resources. Unlike teachers of major languages of the world, they lack not only formal training in language teaching, now often required by local governments, but also language curricula and, even more crucially, usable basic language descriptions. The Mercator European Research Centre intends to be instrumental in coordinating these activities. It will be important to exchange ideas with similar institutes in other parts of the world. Together we shall be able to develop an effective and viable strategy for sustaining the world's endangered languages.

BIBLIOGRAPHY

[ASAKURA ET AL. 1986] Reproduction of sound from old wax phonograph cylinders using the laser-beam reflection method. *Applied Optics*, vol.25, no. 5, pp.597 – 604.

[AUSTIN, PETER K. (ED.) 2002 AND 2004] Language Documentation and Description. *The Hans Rausing Endangered Languages Project*, SOAS, London. (Vols. 1 and 2).

[BURYKIN ET AL. 2005] Burykin, A., A. Girfanova, A. Kastrov, I. Marchenko and N. Svetozarova 2005. *Kolleksii narodov severa v fonogrammarxive pushkinskogo doma*. [Collections on the peoples of the North in the phonogram archive of the Pushkinski Dom]. Faculty of Philology, University of St.Petersburg.

[DENISOV VICTOR. 2008] Zapisi udmurtskogo jazyka i folkloru v Fonogrammarkhive Istituta russskoi literatury (Pushkinskij Dom). In: *Rossija i Udmurtija: istorija I sovremennost'* [Recordings of the Udmurt language and folklore in the Phonogram Archive of the Institute of the Russian literature (Pushkinskij Dom)]. In: *Russia and Udmurtia: History and Present*. Izhevsk, 879-884.

[DE GRAAF 1992] The languages of Sakhalin. Small languages and small language communities: news, notes, and comments. *International Journal of the Sociology of Languages* 94, pp. 185-200.

[DE GRAAF 1997] The reconstruction of acoustic data and the study of language minorities in Russia. In: *Language Minorities and Minority Languages*. Gdansk: Wydawnicstwo Uniwersytetu Gdanskiego, pp 131-143.

[DE GRAAF 2001] Data on the languages of Russia from historical documents, sound archives and fieldwork expeditions In: Murasaki, K. (red.) *Recording and Restoration of Minority Languages, Sakhalin Ainu and Nivkh*, ELPR report, Japan, pp. 13 – 37.

[DE GRAAF 2002] The Use of Acoustic Databases and Fieldwork for the Study of the Endangered Languages of Russia. Conference Handbook on Endangered Languages, Kyoto. *Proceedings of the Kyoto ELPR Conference*, pp. 57-79.

[DE GRAAF 2002] Phonetic Aspects of the Frisian Language and the Use of Sound Archives. In: *Problemy i metody eksperimental'no-foneticheskikh issledovanij*. St.Petersburg, pp. 52-57.

[DE GRAAF 2002] The Use of Sound Archives in the Study of Endangered Languages. In: *Music Archiving in the World*, Papers Presented at the Conference on the Occasion of the 100th Anniversary of the Berlin Phonogramm-Archiv, Berlin, pp. 101-107.

[DE GRAAF 2004] Voices from Tundra and Taiga: Endangered Languages of Russia on the Internet. In: Sakiyama, O and Endo, F (eds.) *Lectures on Endangered Languages: 5 - Endangered Languages of the Pacific Rim* C005, Suita, Osaka, pp. 143-169.

[DE GRAAF 2004] The Status of Endangered Languages in the Border Areas of Japan and Russia. In: Argenter, J A and McKenna Brown, R (eds.) *On the Margins of Nations: Endangered Languages and Linguistic Rights*. Proceedings of the Eighth Conference of the Foundation for Endangered Languages, Barcelona, pp. 153-159.

[DE GRAAF 2004] With Kleiner, Yu. And Svetozarova, N. Yiddish in St.Petersburg: *The Last Sounds of a Language*. Proceedings of the Conference "Klezmer, Klassik, jüdisches Lied. Jüdische Musik-Kultur in Osteuropa". Wiesbaden, Harrassowitz Verlag, pp. 205 - 221.

[DE GRAAF 2005] Dutch in the Steppe? The Plautdiitsch Language of the Siberian Mennonites and their Relation with the Netherlands, Germany and Russia. In: Crawhall, N. and Ostler, N. (eds.): *Creating Outsiders. Endangered Languages, Migration and Marginalization*. Proceedings of the IXth Conference of the Foundation for Endangered Languages, Stellenbosch, 18-20 November 2005, pp. 32-31.

[DE GRAAF 2004] With H. Shiraiishi. Capacity Building for some Endangered Languages of Russia: Voices from Tundra and Taiga. In: *Language Documentation and Description, Volume 2*, The Hans Rausing Endangered Languages Project, School of Oriental and African Studies, London, pp. 15-26.

[DE GRAAF 2008] With V. Denisov. Sokhraneniye zvukovogo nasledija narodov Udmurtskoi Respubliki: opyt veduschikh zvukovykh arkhivov mira. *Rossija i Udmurtija: istorija i sovremennost'* [Preservation of the sound heritage of the peoples of the Udmurt Republic: experience of the World's leading archives]. In: *Russia and Udmurtia: History and Present*. Izhevsk, 866-878.]

[GRUZDEVA 1998] *Nivkh*. München: Lincom Europa.

[IASA-TC 03. 2004] *The Safeguarding of the Audio Heritage: Ethics, Principles and Preservation Strategy*. International Association of Sound and Audiovisual Archives (IASA).

[IASA-TC 04. 2007] *Guidelines on the Production and Preservation of Digital Objects*. International Association of Sound and Audiovisual Archives (IASA)

[MAGID, S. 1936] Spisok Sobranii Fonogramarkhiva Folklornoi sektsii IAE Akademii Nauk SSSR [List of the Collections in the Phonogram Archive of the Folklore Section of the Institute for Anthropology and Ethnographics, Academy of Sciences of the USSR]. *Sovetskii Folklor*, n°4-5, pp. 415-428.

[HINTON, L. 1994] *Flutes of Fire. Essays on California Indian languages*. Berkeley, California: Heyday Books.

[MATSUMURA, KAZUTO (ED) 1998] *Studies in Endangered Languages*. Papers from the international symposium on endangered languages, Tokyo, November 18-20, 1995. Tokyo: Hituzi Syobo.

[MURASAKI, K. 2001] *Tuytah: Asai Take kojutsu, Karafuto Ainu no mukashi banashi* (Old stories of the Sakhalin Ainu). Tokyo: Sofukan.

[SCHÜLLER, D. (ED.) 2005] *The Safeguarding of the Audio Heritage: Ethics, Principles and Preservation Strategy*. IASA Technical Committee – Standards, Recommended Practices and Strategies, IASA-TC 03.

[SHIRAIISHI, H. AND G. LOK. 2002, 2003] *Sound Materials of the Nivkh Language 1 and 2*. Kyoto, ELPR Publications A2-15 and 36.

———. 2004. With G. Lok. *Sound Materials of the Nivkh Language 3*. Publication of the International NWO project “Voices from Tundra and Taiga”, University of Groningen.

———. 2006. *Topics in Nivkh Phonology*. Dissertation at Groningen University, September 2006.

[TAKSAMI, CH.M., PUKHTA, M.N., VINGUN, A.M. 1982] *Nivkhgu bukvar'*. Leningrad, Prosveshchenie.

[TAMURA, S. 2000] *The Ainu Language*. Tokyo: Sanseido.

[WITSEN, N. 1705] *Noord en Oost Tartarye, ofte bondig ontwerp van eenige dier landen en volken, welke voormaels bekend zijn geweest*. Amsterdam, Halma.

LINGUISTIC POLICIES TO COUNTER LANGUAGES MARGINALIZATION

Urbanization and globalisation promote unification of ethnic cultures and strongly reject the vast majority of cultures in the margins. Knowledge, as well as historical and cultural experience of these cultures and their languages are dwindling. A singular culture disappears when its language dies. Also the development of communication technologies brings hope and opportunities. What actions can be taken to stop or even just slow marginalization of languages, enhancing the vitality, the representation and use of endangered languages? Who can do it, and especially who has the responsibility?

Original article in English.



EVGENY KUZMIN is Chairman of the Intergovernmental Council and Russian Committee of the Unesco Information for All Programme. He headed the Department of Libraries of the Russian Ministry of Culture in 1992-2005 and participated in the elaboration of Russia's national cultural policy and strategies of information society building. He initiated a number of projects on linguistic diversity in cyberspace, including two international conferences.

EVGENY KUZMIN

MARGINALIZATION
LANGUAGES
TO COUNTER
POLICIES
LINGUISTIC

Many nations in the contemporary world enjoy neither statehood nor sovereignty. Their languages are not state languages because a majority of countries are multiethnic and multilingual. Even in the best possible scenarios, when governments and dominant ethnic groups are rigorously protective of ethnic and linguistic minorities, most languages are still marginalised to varying extents. They develop or decline in the shadow of the country's dominant language, which is used in all spheres of influence – political, economic, educational, cultural, and scientific.

Globalisation, migration, and the rapid pace of urbanization have made many ethnic minorities undervalue their native language. Meanwhile, state and international languages garner a wealth of attention and research.

No language develops outside the context of its corresponding ethnos. At the same time, urbanization and globalisation encourage smaller cultures to merge with the majority, and marginalize themselves. The knowledge and historical and cultural experience stored within these cultures gradually vanish, as well as the culture's/language's potential. Cultural and linguistic marginalization is thus interrelated and multifaceted process; with the death of a language, its unique carrier culture vanishes¹.

These issues are salient nearly every country where two or more languages cohabit. But the situation is not entirely hopeless: the development of Information and Communications Technology (ICT) provides some optimism.

1 In the context of this article, the term “culture” is used in the broadest sense to denote the entirety of salient material, intellectual and emotional features of a given community or social group, comprising the arts and literature, as well as lifestyle, the status of human rights, value systems, education, customs, traditions and philosophy.

What can we do to hinder the process of language marginalization, and to enhance the fitness of endangered languages? Who can do it, and whose duty is it?

Let us examine the case of Russia, one of the most multiethnic, multi-lingual and multi-religious countries of the world; how it tackles these issues, and to what an extent it solves them.

RUSSIA'S LINGUISTIC LANDSCAPE

As of 2009, the Russian Federation is home to 141.9 million inhabitants. Ethnic Russians account for close to 80% (114 million), while the remaining 20% speak 180 languages, over a hundred of which belong to indigenous ethnic entities historically formed within the present-day Russian borders or living there for centuries. However, more than 127 million people regard the official language of the country – Russian – as their native language. Many representatives of indigenous peoples know Russian better than their mother tongue; some use it more fluently than many ethnic Russians.

The most widely used minority languages include Tatar (5.35 million speakers), Bashkir (1.38 million speakers), Chechen and Chuvash (1.33 million speakers each). A further nine languages have between four hundred thousand and a million speakers². A further fifteen are spoken by between fifty thousand to four hundred thousand people.

All languages except Russian are thus minority languages, and all are marginalized to varying extents. More than a third are endangered or extinguishing, with a full 39 of them less than fifty thousand speakers strong, mainly those of the indigenous populations of the Far North, Siberia and the Far East. Despite official efforts at every level of Russian bureaucracy to nurture these languages and their corresponding cultures, the risk of extinction remains high.

The Constitution of the Russian Federation declares all languages of Russia to be common cultural assets. Almost all languages use graphic systems, even if some have acquired them somewhat recently.

2 Avar, Kabardian-Circassian, Dargin, Osset, Udmurt, Kumyk, Yakut, Mari and Ingush.

Unlike many other major multilingual countries, Russia offers primary education, television and radio broadcasting, internet resources (many of them catalogued), books and newspapers in nearly all of its languages. Russia is unique in another respect as well: close to forty of its indigenous languages enjoy official status. All languages can find support at federal, regional and municipal levels.

The Russian Federation possesses a sophisticated administrative-territorial structure that includes 83 constituent entities: fifty regions, eight territories, twenty republics, four autonomous areas and one autonomous region.

A region, or *oblast*, is an administrative territorial entity highly dominated by ethnic Russians, where other ethnic entities account for less than 1% of the population.

A territory, or *krai*, is a major administrative territorial entity that includes autonomous areas with dense ethnic minority populations.

Republics are constituent entities populated by numerically comparable communities of Russians and representatives of another ethnic group. Republics take the names of these ethnies. They have their own constitutions and enjoy greater autonomy from the federal centre than territories, regions and autonomous areas. Both Russian and the language of the titular ethnic group are recognised as the state languages, even if the non-Russian ethnic group is a minority within its republic. In certain republics, more than two languages enjoy official status. In this sense, considerable attention is paid to the equal rights of the languages, by way of their constitutional protection, special laws, and other modes of federal intervention.

Republics have the right to establish their own official languages, to publish federal and republican laws in those languages, and to place them on equal standing with Russian during elections, referendums, and industrial, official and judicial activities.

Establishing some languages as official does not obviate the protection of others. The republic of Tatarstan, for example, is undertaking efforts to preserve the culture and language of its local Bashkirs, Udmurts, Chuvashes; the republic of Chuvashia – of the Tatar and Bashkir living there, etc.

These protections are not the products of Russia's government policy towards languages, but its very foundation. They are explicated and promulgated in the *Law on the Languages of the Peoples of the Russian Federation* (RSFSR), which was ratified in the Soviet era but retains its legal force, the *Law on the State Language of the Russian Federation* along with a number of other laws protecting culture and education. This whole body of legislation stipulates state-sponsored protection, promotion, and development of all minority ethnic languages, as well as bilingualism and multilingualism. Russian citizens without a working command of Russian are thus given the tools to participate in meetings and conferences at government agencies, offices, industrial companies, and the court of law in the language they speak fluently, and an interpreter is provided when necessary.

The *Law on the Languages of the Peoples of the Russian Federation* envisages opportunities to organise teaching in a speaker's native language, irrespective of that language's numerical strength and in conformity with its demands.

The *Federal Law on Education* stipulates that the state vouches assistance in training experts for teaching in the languages of the people of Russia who have no statehood in the form of a republic or autonomous area within the Russian Federation.

Russian regional authorities are the biggest contributors to official ethno-linguistic policy formation and implementation, as they are the ones most directly encountering the issues of preserving multilingualism on a daily basis.

As an example, let us more closely examine the case of language policy and status in the Republic of Sakha (Yakutia), specifically as it pertains to the Yakut language.

LINGUISTIC STATUS, ETHNO-CULTURAL AND LANGUAGE POLICY IN THE REPUBLIC OF SAKHA (YAKUTIA)

Ethnic makeup and expression

The Republic of Sakha (Yakutia)³, with an area of 3,103,200 km², is Russia's largest constituent entity. As home to a population of slightly over one million, 450,000 of whom speak Yakut, it is also one of the most striking examples of Russian multiethnic areas.

Yakut is the eponymous language of its corresponding ethnies (historically "Sakha") that gave name to its home republic. Though the Yakut are genetically related to Mongolians, their language belongs to the Turkic group.

In terms of local population, the Yakut rank first at 45.6 %, ethnic Russians compose an additional 41.1 %, and the other 126 ethnic groups represent a combined 13.3 %. This 13.3 % consist of aboriginal Northern minorities – the Evenki, Even, Yukagir and Chukchi, who densely populate 69 settlements, mostly in the Far North – and 238 registered nomadic clans.

In Yakutia, 93.3 % of the population has a fluent command of Russian; on the other hand, 87.4 % of the Yakut, 37.7 % of the Chukchi, 20.7 % of the Even, 19.5 % of the Yukagir and 6.5 % of the Evenki regard their native language as their mother tongue. While Yakut is a minority language in terms of Russia as a whole, in Yakutia it is the majority language.

The Yakutian constitution grants official language status not only to Yakut and Russian, but also to Even, Evenki, Yukagir, Dolgan and Chukchi, guaranteeing all of them unlimited development and protection. The republic's Presidential Council on Language Policy plays an extensive role, through which the government launches multifarious targeted language development programmes. Cultural events are held to promote intercultural dialogue; public holidays are dedicated to the republic's assortment of languages.

3 Hitherto referred to as Yakutia.

Of the 1,059 books published there in 2009⁴, 318 were written in Yakut, two in Even, three in Evenki and five in Yukagir. There are 30 Yakut-language newspapers (13 republican and 17 district) and 12 magazines. *Tatkachiruk* is published in Evenki; *Ilken* in Russian, Yakut and minority languages. There is also the multi-language almanac *Khalarkhat*. As for republic-wide TV broadcasting, Russian language accounts for 62 %, Yakut 38 %, and Northern minority languages 1 %. Increasingly, however, broadcasts about Yakutian cultures and history come out in Russian.

Mechanisms of preservation

As stated above, no language can survive, let alone develop, outside the confines of its ethnic culture. Language preservation necessitates above all the preservation of ethnic identity: culture, customs, traditions, folklore, ethnic sports, cuisine, economic know-how and environment. Let us now turn to what is being done to preserve and develop the authentic Yakut culture and language while actively supporting culture and arts in general.

The republic has 528 libraries, 79 museums, 12 theatres, 565 cultural centres and 90 children's art schools where music, painting, dancing and other arts, including ethnic arts and crafts, are taught. Yakutia holds ethnographic festivals and organizes travelling art exhibitions and tours of the republic's leading performers at home, across Russia, and abroad.

Its capital of Yakutsk, with a population of 240,000, is home to the National Drama Theatre, which stages classic Russian, foreign, and contemporary locally written plays all in the Yakut language. Among the city's other cultural institutions are also the Russian Drama Theatre, the Puppet Theatre and the Youth Theatre; the National Opera, where operas and ballets by Yakut composers are staged alongside world classics; the Yakutia Symphony Orchestra alongside the Virtuosi of Yakutia Violin Ensemble; the National Museum, which exhibits historical artifacts of all the republic's peoples alongside contemporary paintings and sculptures; the open-air Ethnography Museum, the Mammoth Museum; the National Library of the Republic of Sakha (Yakutia)⁵ – the largest depository of books

⁴ 120,000 book titles were published in Russia in 2009 alone.

⁵ Available online <http://nlib.sakha.ru>

and other printed matter in Yakut and indigenous northern languages⁶ and 18 public libraries; the National Academy of Music with a boarding school, where the most gifted children spend twelve years studying all instruments of the symphony orchestra plus ethnic instruments; five art schools; four cinema theatres; and two exhibition halls and art galleries.

As of 2009, the republic possessed 654 educational institutions, 415 (67%) of them with Yakut-language teaching of all disciplines in primary school. The languages indigenous to the North are taught as special disciplines in 38 schools (some of them nomadic)⁷.

Northern languages are also studied at the Institute of Humanitarian Studies and Problems of Northern Ethnic Minorities, part of the Siberian Branch of the Russian Academy of Sciences and the Research Institute of Ethnic Schools.

In the republic's 6 higher educational establishments as well, minority languages are given attention. Northeastern Federal University offers higher education in history and philology taught in Yakut (all the other disciplines are taught in Russian only). In 2008 and 2011 the Russian Committee of the Unesco Information for All Programme and the Interregional Library Cooperation Centre together with the University and the Government of Yakutia held two international conferences "*Linguistic and Cultural Diversity in Cyberspace*". In 2010, following the 1st Conference's recommendations, the Centre for the Promotion of Multilingualism in Cyberspace was inaugurated under the Northeastern Federal University.

Memory of Yakutia, a programme launched in 2000, is tasked with creating a database for the Memory of Yakutia website⁸, which aims to make available to the public rare books in Yakut, archival documents pertaining to crucial elements of Yakut history and culture, and rare culturally and historically germane audio recordings of Yakut performers.

When Unesco entered the Yakut heroic epic *Olonkho* on its list of *Masterpieces of the Oral and Intangible Heritage of Humanity* in 2005, it impelled the Yakutian government to launch a state targeted programme

6 To visit the online website of the Library of Northern Ethnic Minorities.
<http://nlib.sakha.ru/knigakan>

7 Even (22 schools), Evenki (14 schools), Yukagir (3 schools), Chukchi (2 schools), Dolgan (1 school).

8 See <http://www.sakhamemory.ru>

from 2007 through 2015 to preserve, develop and circulate the epic. The programme's aims include searching and collecting epics, promoting folk narrators, including *Olonkho* in curricula, and establishing a pedagogy of folk recitation. The Northeastern Federal University is the base for the preservation, study and popularisation of Yakut folk heritage via research and education, and the *Olonkho* online portal⁹ provides access to local folklore and epic texts in many languages.

Non-textual internet services are also developing. Sound dictionaries of the Yakut language are being made. Educational establishments organise regular online Yakut language and literature conferences. Relevant internet content and resources are multiplying, and the Yakut-language Wikipedia¹⁰ is frequently replenished.

But in the face of all these measures, recent studies indicate a surprising trend: a steady decrease in Yakut as a first language among ethnic Yakut speakers, while Yakut proficiency is conversely growing among ethnic Russians.

RECOMMENDATIONS FOR THE PROMOTION OF MULTILINGUALISM

As we analyse the experience of the Russian Federation and one of its entities, the Republic of Sakha (Yakutia), we can attempt to draw out some more general conjectures regarding the question of how to guarantee the continued functioning of minority languages in the shadow of a dominant language in a national context.

Potential contributors to the promotion and development of a language are manifold and diverse, and include national authorities, local authorities, educative systems, research establishments, memory institutions, artistic establishments working in close contact with local painters, sculptors and architects; film studios; cultural centres, principally in remote settlements, which unite the functions of memory institutions and art and educational centres; book publishers and traders, media outlets, the ICT industry, public organizations and private persons and businesses.

⁹ See <http://www.olonkho.info>

¹⁰ See <http://sah.wikipedia.org>

As ICTs rapidly penetrate all spheres of contemporary life, opening ever-increasing possibilities to preserve languages and cultures, the development of cultural and linguistic diversity in cyberspace acquires newfound importance.

The *World Summit on the Information Society's Declaration of Principles* stresses that the information society should be founded on respect for cultural identity, cultural, linguistic and religious diversity, and foster dialogue between cultures and civilisations. Special attention in building an inclusive information society should be paid to the creation, circulation and preservation of content in diverse languages and in varied formats. The development of local, regional, and ethno-specific content should promote social and economic development and participation, especially in rural, remote and marginal areas.

Let us now consider each of the above-listed contributors, and their corresponding targets and modes of action.

National and Local Authorities

National and local official policies and activities are prime. Effective policies include a combination of measures to promote the preservation, free expression and development of linguistic, ethno-cultural and religious identity of ethnic communities, through the preservation and development of their cultural values and traditions, folklore, via the practical application of principles of cultural pluralism, bilingualism and multilingualism. This goal demands the enactment of special laws, and monitoring and ensuring compliance with these and already existing laws.

Federal and regional language laws must stipulate that the acquisition of the state status by certain languages must not encroach on the linguistic rights and expression by all ethnic entities historically inhabiting a particular territory.

Programmes to implement these goals should be based on the value of mutual intellectual and cultural enrichment, by preserving minority languages, customs, traditions, values, and institutions reflecting ethnic cultural specificity.

Authorities should contribute to systemic language studies and promoting multilingualism in education, administration, law, cultural education, news media and cyberspace.

The attainment of those goals can be facilitated by:

- establishing a regulatory framework for the development of languages at the national level (the national constitution and federal laws, along with constituent entities' constitutions, statutes and laws);
- forming and implementing cultural and educational strategies, policies and programmes explicitly aiming to promote minority cultures and languages;
- targeting federal funding and soft taxation of both governmental and non-governmental programmes for language preservation and development;
- granting state or official status to the largest minority languages either at the national level, or within regions densely inhabited by users of those languages; whenever possible, language equality must be affirmed in law;
- affirming a given minority language's official status in the records of government and municipal authorities: using the language in governmental work, publishing federal and republican legal acts in it (and guaranteeing their equal legal force), and granting the language equal standing with the principal state language during elections, referendums and industrial, office and administrative activities;
- creating official document databases in the language;
- establishing councils on language policy within central and/or regional governments, and determining their rights and duties;
- guaranteeing social, economic and legal protection of the language in legislative, executive and judicial bodies;
- providing material incentives for experts to use both national and minority languages in their work;
- signing and ratifying international acts promoting multilingualism;
- promoting ethnic entities' interest in the development of their languages;

- establishing targeted regional programmes to preserve culture and language;
- helping and legally assisting the development of the language's body of literature through financial and other support of book publication and media dissemination, particularly that which is oriented to children and youth;
- forming and implementing strategies and programmes promoting reading in the native language;
- partnering with ethnic cultural associations outside the administrative territorial boundaries that are historically densely populated with members of the given ethnic;
- supporting libraries, museums, archives and other cultural agencies in the preservation and development of minority cultures and languages;
- establishing ethnic schools to intergenerationally transmit experiences, traditions, culture and ethics;
- promoting the ethno-cultural component of education and extending it wherever necessary and possible;
- equipping public schools with minority language and literature classrooms;
- contracting the governments of other regions densely inhabited by speakers of a particular language to assist in measures to preserve that language, for example by supplying literature to public and school libraries to enable the study of a given language, and participating in the graduate and postgraduate training of teachers for ethnic minorities; and
- creating graphic systems for non-literate languages.

To promote multilingualism in cyberspace, authorities can take both general and goal-oriented measures to create a multilingualism-friendly environment:

- designing and implementing ICT penetration programmes;
- drawing up action plans to promote public use of the internet, including information literacy programmes for both dominant and minority languages;

- providing telecommunication networks to remote areas;
- elaborating information resource development programmes in minority languages;
- promoting training in ICTs and information, especially in local languages;
- promoting the creation of local content, translation and adaptation;
- promoting the translation of world literary classics into minority languages, and of minority speakers into other languages, and posting these translations online;
- establishing integrated multilingual information resource networks;
- introducing electronic documentation and record management in at least two languages; and
- promoting the research and development of operating systems, search engines and internet browsers, online dictionaries and term reference books, and their adaptation to local demands¹¹.

Research Centres

Research centres provide the theoretical basis for governmental and non-governmental multilingualism promotion efforts and make fundamental and relevant applied research. Their duties may comprise:

- studies of ethnic cultures, traditions and quotidian life;
- studies of languages and their history;
- studies of the current linguistic situation and related issues;
- studies of language-promoting policy and practice in other parts of the country/world, display and dissemination of pioneer experience;
- elaboration of proposals on adapting cutting edge experience;
- elaboration and implementation of permanent monitoring tools to measure language use by social groups;

11 See in this book: Dwayne Bailey, *Software Localization: Open Source as a Major Tool for Digital Multilingualism*.

- elaboration and implementation of permanent monitoring tools to qualify and quantify the work of language-promoting institutions;
- proposals to the government for draft regulatory legislation on language protection and promotion;
- initiation and organization of theoretical and applied conferences addressing the various aspects of minority language preservation and development;
- establishment of minority linguistic and cultural research centres;
- training of relevant experts;
- popularisation of minority languages and cultures;
- elaboration of national reading promotion programmes, in particular for minority languages, in cooperation with libraries, educational institutions, media outlets, and book publishers/traders;
- elaboration of best practices guidelines for relevant offices and organizations;
- publication of bilingual dictionaries that include audio recordings of words;
- establishment of terminology and orthography commissions;
- creation of text corpuses and phonetic databases;
- linguistic and folklore field studies and expeditions;
- establishment of centralised archives, including electronic archives, for minority languages;
- acquisition of private archives of researchers and community activists engaged in minority language support, and entrusting those archives to state memory institutions;
- establishment of clear standards and guidelines for recording and representing texts, alphabets and graphic systems for non-literate languages;
- establishment of a unified literary language, if absent;
- documentation of minority languages;

- research and development of operating systems, search engines and information scanning systems¹²; and
- development of fonts in cooperation with relevant experts.

Education

Primary, secondary and higher educational establishments should cooperate with federal and regional executive and legislative bodies, as well as research and cultural institutions, to support and develop minority languages and multilingualism.

Their sphere of activity includes:

- participating in writing the regional/local component of national educational standards;
- training minority language teachers for schools and universities;
- training experts on languages, history and traditional culture of ethnic minorities;
- implementing postgraduate teacher training programmes;
- elaborating basic curricula;
- elaborating academic curricula and learning packages;
- elaborating language teaching and speech improvement methods;
- making recommendations to implement new language teaching technologies;
- establishing university classes in minority languages;
- using minority languages as educational tools in all places of learning, especially to improve native language speech habits;
- teaching minority language as part of core curricula for students who speak it as a second language in all educational establishments in areas where an ethnic minority makes up a considerable section of the population;
- organising specialist language and literature classes;

12 See in this book: Pann Yu Mon & Madhukara Phatak, *Search Engines and Asian Languages*.

- organising educational competitions on minority languages and literature;
- organising conferences and events on linguo-cultural and ethno-cultural issues;
- organising off-campus language courses, especially on interregional and international levels;
- organising summer camps conducted in minority languages; and
- organising online conferences in minority languages (on diverse topics).

CULTURAL INSTITUTIONS

Cultural institutions and activists are tremendously important in language support, not only those directly connected with preserving linguistic cultures, but also theatres and conservatories, folklore performers, art schools, cultural centres in remote areas, and individual artists and cultural workers.

Memory institutions : libraries, archives and museums

It is the duty of these institutions to preserve, store, popularise and offer for public use all essential testimony of a particular people's history.

Libraries and archives must search, acquire, describe, study, popularise and store all printed matter, sound and video recordings emanating from a language, both in the geographical area that is densely inhabited by its users, and other areas (even foreign countries) where those languages are used. Not only materials in minority languages but all information about them published in other languages is important.

The activities of memory institutions include:

- gathering, preserving and extending comprehensive and thematic collections of all a minority language's published and unpublished materials;
- creating full-text databases of periodicals in the given language;

- constructing an exhaustive bibliography of printed and written resources in the language;
- making available centralized catalogues of publications in the language (especially important for languages that have recently acquired a graphic system);
- including bibliographic descriptions of works reflecting the history and culture of an ethnic minority in electronic national catalogues of all libraries at both national and international levels;
- popularising these works, especially by organising readers' conferences, reader clubs, and meetings with writers, critics, publishers, illustrators, and others;
- digitising documents and museum exhibits that reflect an ethnic entity's history and culture, establishing corresponding electronic libraries, museums and archives, and granting public access to them;
- establishing electronic and other museum expositions in the given language or bilingual exhibitions using that language;
- creating electronic catalogues in museum systems in the given language; and
- preparing archives of electronic publications and exhibitions on cultural and linguistic diversity and memorable dates and events.

Together with other cultural, research and educational establishments, libraries, museums and archives can launch multimedia projects pertaining to the founders of ethnic cultures, folklore collectors, writers, artists, composers and performing musicians. Texts, photographs, digital copies of paintings and sketches, sound and video recordings can be recorded on discs for broad circulation, and their online versions be posted on the websites of cultural, research and educational institutions.

Mass Media

Federal, regional and municipal media outlets can become purveyors of cultural and linguistic diversity. The contemporary mass media should focus on:

- preserving and developing periodicals in minority languages and sections in those languages in other periodicals;
- organising television and radio broadcasting in minority languages, especially the release of programmes entirely or partly conducted in those languages, and topical to the original ethnic culture of their speakers;
- organising internet broadcasting in minority languages; and
- establishing information portals.

Book publishing and circulation

Book publishers and traders can make a tremendous contribution to the support of minority languages and development of multilingualism: a language without access to the book industry is a language excluded from intellectual community life. Unesco stresses the importance of translation in strengthening multilingualism, especially in book publishing, which promotes both the industry and the free circulation of ideas.

Publishers can promote minority languages through:

- effecting research, popular science and fiction books, periodicals and translations in a minority language;
- promoting literary work in a minority language and its emerging authors;
- assuring that libraries of educational institutions include books in minority languages; and
- helping minority language speakers to acquire books, especially in remote areas that are historically densely populated by the given ethnic.

PUBLIC ORGANIZATIONS

Non-governmental language promotion activities include:

- establishing weekend schools, clubs and ethno-cultural associations to provide supplemental linguistic and literary education;

- organising competitions, festivals and creative events to promote cultural and linguistic traditions;
- participating in language and culture days in and outside the traditional settlement areas of a given ethnic;
- participating in folk festivals; and
- communicating with and supporting a language's expatriate population.

PRIVATE INITIATIVES

Language should gain support not only from the state, its institutions or non-governmental organizations. Individuals and groups of individuals can also participate in language preservation and promotion by:

- establishing and supporting Wikipedia in minority languages;
- establishing and supporting websites, blogs, Twitter and other social networks¹³.

THE ICT INDUSTRY

The ICT industry is a key agent in promoting multilingualism, and a crucial participant in supporting and enhancing a language's status. The ICT industry can channel its energy into the following areas:

- articulating and promoting technical standards, taking into account ethnic minorities' demands¹⁴;
- creating complete computer fonts for minority languages;
- participating in the establishment of international Unicode standards and the implementation of the unified keyboard layout;
- localising software and creating free software to support local languages;

13 See in this book: Vassili Rivron, *The Use of Facebook by the Eton of Cameroon*.

14 See in this book: Stéphane Bortzmeyer, *Multilingualism and the Internet's Standardisation*.

- elaborating computer language models and machine translation systems¹⁵;
- supporting minority languages in e-mail, chat and other messaging utilities;
- uploading electronic study books and dictionaries in minority languages;
- establishing multilingual domains and e-mail addresses¹⁶;
- creating software for multilingual internet domain names and content;
- establishing localised, minority language retrieval systems;
- creating information and other websites and portals in bilingual versions;
- making information resources available electronically; and
- developing the non-textual sphere of the internet (such as voice over IP, data streaming, and video on demand)¹⁷.

The above measures can bring about their desired results only when the entire ethnos – not only its cultural, intellectual and ruling elite – makes major intellectual and emotional efforts, and displays goodwill, desire and interest in the survival and development of its unique culture and linguistic identity.

BIBLIOGRAPHY

[KUZMIN, EVGENY (ED). 2008] “Report by the Russian Federation to Unesco General Conference on Measures Taken to Implement the Recommendation concerning the Promotion and Use of Multilingualism and Universal Access to Cyberspace”, In: *Preservation of Linguistic Diversity: Russian Experience*. Moscow: Межрегиональный центр библиотечного сотрудничества (Interregional Library Cooperation Centre), pp. 8–36.

[KUZMIN, EVGENY AND PARSHAKOVA, ANASTASIA. (EDS). 2011] *Развитие многоязычия в киберпространстве: пособие для библиотек* (Promoting Linguistic Diversity in Cyberspace: A Handbook for Libraries). Moscow: Межрегиональный центр библиотечного сотрудничества (Interregional Library Cooperation Centre).

15 See in this book: Joseph Mariani, *How Language Technologies Support Multilingualism*.

16 See in this book: Stéphane Bortzmeyer, *Multilingualism and Internet Governance*.

17 See in this book: Tunde Adegbola, *Multimedia and Signed, Written or Oral Languages*.

[KUZMIN, EVGENY AND PLYS, EKATERINA (EDS). 2007] *Языковое разнообразие в киберпространстве: российский и зарубежный опыт* (Linguistic Diversity in Cyberspace: Russian and International Experience). Collection of analytical works. Moscow: Межрегиональный центр библиотечного сотрудничества (Interregional Library Cooperation Centre).

[— 2008] “Development of Multilingualism on the Internet as a New Field of Activity of the Russian Committee of the Unesco Information for All Programme and the Interregional Library Cooperation Centre”, In: *Preservation of Linguistic Diversity: Russian Experience*. Moscow: Межрегиональный центр библиотечного сотрудничества (Interregional Library Cooperation Centre), pp. 65-85.

[— 2008] *Представление языков России и стран СНГ в российском сегменте Интернета* (Presentation of Languages of Russia and other CIS Countries in the Russian Segment of the Internet). Collection of reports. Moscow: Межрегиональный центр библиотечного сотрудничества (Interregional Library Cooperation Centre).

[— 2008] *Preservation of Linguistic Diversity: Russian Experience*. Moscow: Межрегиональный центр библиотечного сотрудничества (Interregional Library Cooperation Centre).

[— 2010] *Языковое и культурное разнообразие в киберпространстве* (Linguistic and Cultural Diversity in Cyberspace). Proceedings of the international conference held in Yakutsk, Russian Federation, 2-4 July, 2008. Moscow: Межрегиональный центр библиотечного сотрудничества (Interregional Library Cooperation Centre).

[— 2011] *Linguistic and Cultural Diversity in Cyberspace*. Proceedings of the International Conference held in Yakutsk, Russian Federation, 2-4 July, 2008. Moscow: Межрегиональный центр библиотечного сотрудничества (Interregional Library Cooperation Centre).

[KUZMIN, EVGENY, PLYS EKATERINA, KISLOVSKAIA, GALINA AND TCHADNOVA, IRINA (EDS). 2008] *Многоязычие в России: региональные аспекты* (Multilingualism in Russia: Regional Aspects). Moscow: Межрегиональный центр библиотечного сотрудничества (Interregional Library Cooperation Centre).

MULTIMEDIA AND SIGNED, WRITTEN OR ORAL LANGUAGES

Writing systems were not developed for all languages at the same time, giving written languages a great advantage. As we move deeper into the information age, how can we ensure that the inequalities of the agrarian and industrial ages are not amplified in our information age? If literacy is a priority as a fundamental value of the modern world, how can we use the Internet to allow speakers of unwritten languages the expression, memory and expansion of their fields of knowledge?

Original article in English.



TUNDE ADEGBOLA is a researcher, consultant and cultural activist with long experience in the field of information and communication media. As Director General of the African Languages Technology Initiative, he leads a team of researchers in adapting human language technologies for use in African languages.

TUNDE ADEGBOLA

MULTIMEDIA
AND SIGNED,
WRITTEN
OR ORAL
LANGUAGES

The invention of writing was a very important milestone of human development as it facilitated the accurate and detailed documentation of human experiences and ideas outside the confines of the human brain. This development enhanced the mobility of experiences and ideas, freeing them from temporal and spatial constraints, and thereby making possible their sharing within and between various cultures of the world. However, because writing systems were not developed for all human languages at the same time, writing put written languages at an advantage, creating a relative limitation on the pace and extent of the sharing of experiences and ideas coded in unwritten languages. Recent developments in digital technology have now made it a lot easier to document information and knowledge without writing. Multimedia in modern information communication technology has emerged as one of the salient characteristics of the information age as it facilitates easy and enhanced communication in oral, written and signed languages¹. This holds a promise of making cyberspace a truly inclusive communication space.

As we proceed further into the information age, there is a need to ensure that the forms and levels of inequity that characterised the agrarian and industrial ages do not become entrenched in the information age. Whilst on the one hand we continue to quest for increased levels of literacy as a fundamental value of the modern world, there is on the other hand a need to ensure that the offerings of digital technology are not applied to the advantage of users of written languages alone. We need to consciously develop multimedia-based techniques that apply the advantages of modern

1 See in this book: Annelies Brafort & Patrice Dalle, *Accessibility in Cyberspace: Sign Languages*.

digital technology to making languages without writing systems available and useful in cyberspace and also give access to users of signed languages.

ORAL, SIGNED AND WRITTEN LANGUAGES

Language is a semiotic system in which rules relate symbols to meaning. As a system for communication it features the arrangement of a finite set of auditory or visual symbols according to a finite set of rules resulting in the possibility of the production of an infinite set of statements. This capacity to produce an infinite set of statements from a finite set of symbols and rules provides the basis for language to refer to a large set of simple objects, describe convoluted notions and express complex concepts.

In speech, language is activated by the use of auditory symbols. Sound is a consequence of the compression and rarefaction of air in time and so there is a strict temporal dimension to the traditional use of auditory signals to realise language. When speech is used to realise language, the auditory activities that are produced decay and become imperceptible within a very short time. Yet the information conveyed by these auditory activities, more often than not, remain valid for a long time, much beyond the extremely short lifespan of the auditory activities that were produced to represent them. This temporal limitation had consequences on the capacity for the documentation and reuse of the information and knowledge expressed in speech.

In writing, language is activated by the use of visual markings on some appropriate media. These markings may be codes that represent the concepts to be communicated directly or by the indirect representation of the sounds expressed in speech statements that describe such concepts directly. The lifespan of such visual markings is usually far in excess of the lifespan of the sounds produced in speech and this temporal advantage of the written word over the spoken word is an important motivation for the development of writing.

In sign languages however, visual symbols are produced in time to make up for the auditory deficiencies of the deaf and hearing impaired. Because the visual symbols of sign language are produced in time, they are also subject to temporal limitations similar to those of speech.

Apart from the temporal dimensions of the distinction between oral, written and signed languages, there are a number of other salient characteristics that distinguish these ways of communicating. From a spatial point of view, the portability of the media on which information is written makes it possible for written ideas and information to travel far away from the sources from which they originate. In contrast, exclusively oral languages depend mainly on the processes of memorisation and recitation which restrict their use mainly to the performance mode of presentation.

Many of the world's languages have developed writing systems while even more have remained mainly spoken. The cultures that use these written languages have taken due advantage of the temporal and spatial features of writing. They share information and knowledge by writing and reading texts and this frees such information and knowledge to travel as far as the medium of writing can reach. Within the cultures that still use unwritten languages however, people continue to learn merely by memorisation and recitation. Hence, the extent to which knowledge coded in such languages can spread is determined by the capacity for the spread of word of mouth. Furthermore, due to the limitations of the human memory for accurate recall, information and knowledge coded in unwritten languages are bound to suffer from the limitations of human memory.

LANGUAGE, INCLUSION AND THE CHALLENGE OF LITERACY

The advantages of literacy have given an edge to cultures that use written languages while cultures whose language remains unwritten have to play catch up. In addition, studies have shown a causative correlation between literacy and human development. To this end, literacy is recognised as one of the indices of human development and lots of efforts go into the objective of increasing literacy levels around the world.

Due to the causative correlation between literacy and human development, illiterate people living within cultures of written languages are helped to become literate while efforts are also made to develop writing systems for yet to be written languages. However, despite the best of these efforts to improve literacy levels, illiterate people still abound around the world and they continue to live with reduced capacity to participate fully in development processes.

If people are able to communicate their ideas only to people within their immediate physical environment, they lose the capacity to spread such ideas far and wide. There are two important consequences of this limitation. On the one hand, a limitation is imposed on the population of people that may be able to benefit from such ideas, while on the other hand, such ideas may be starved of the inputs and refinements from a sufficiently wide number of people that could have helped to improve and enhance them for the benefit of humanity. Human understanding of the natural environment leads to a capacity for the prediction and control of the behaviour of the elements and this capacity has important implications for human development. The level of understanding of the natural environment by humans depends fundamentally on the capacity to build ideas upon ideas and knowledge upon knowledge. To this extent, the capacity to share information and knowledge is fundamental to human development.

As the global community becomes more connected and the global economic system tends towards becoming a single, common market place, people whose languages remain unwritten will continue to suffer the consequences of the temporal and spatial limitations of their languages while literate people will continue to derive advantages from the fact that their languages are written. The global community will be the worse for such inequity.

MULTIMEDIA AND MULTIMODALITY

Developments in information and communication technologies (ICT) have changed and are continuing to change our world in unprecedented ways. They have had dramatic effects on the way we communicate, learn and manage knowledge. One important way in which ICT has affected the way we communicate, learn and manage knowledge is in the presentation of information in the form of multimedia. Multimedia refers to a combination of various content forms such as text, audio, still images, video, animation and interactive content. It contrasts with traditional media such as printed materials, in that it can engage the auditory, visual, tactile and other human perceptual modes either concurrently or sequentially.

Probably the most salient characteristic of multimedia in communication, learning and knowledge management is its capacity to provide

information in multiple media and various modes. This capacity of multimedia not only serves to enhance access to information but also helps to heighten understanding.

Books traditionally contained mainly written texts and still images sometimes served to illustrate the texts. The book as a medium for the management of information and knowledge is generally limited to these two media forms. With multimedia however, content can be presented in written texts complemented not only with still images, but also by sound, video and animation. The sound can be in the form of speech, music or sound effects and the video can present sequences of scenes as motion picture that may never be available to the reader/viewer otherwise. The animation can provide visual illustrations of phenomena that can only be otherwise imagined, giving deeper insights into the possible structures of phenomena that have never been experienced directly.

While written text, still image, video and animation engage the visual senses, speech, music and sound effects engage the auditory senses thereby providing enhanced capacity for recall and understanding on the part of the information consumer. Furthermore, interactive media engage the information consumer in exchanges, thereby giving them the capacity to contribute information, making them active rather than merely passive information consumers. Due to the multimedia and multimodal engagement of the human senses, information that may be seen as cryptically coded in one medium may be found to be more easily decipherable in another medium. Also, information that may not be easily accessible in one mode may turn out to be better accessed in some other mode. This has amounted to a revolution in the way we consume information and share knowledge.

DOCUMENTATION WITHOUT WRITING

Important as the revolution presented by multimedia and multimodal engagement of the human senses is, there is an even more important way in which multimedia should be perceived. At a more fundamental level, multimedia can be viewed basically as a means of information documentation. It is a means of information documentation that does not discriminate against auditory or visual presentation of information and it can be realised in a multiplicity of media. It has a sufficiently

wide scope that it is not biased against language or dialect. Writing as an information medium is a system of information coding while some other media in the multimedia composite provide direct and intuitive representation of the objects and concepts they are intended to represent. Pictures do not necessarily need to be deciphered, and recorded speech is directly accessible to listeners without the need to learn to decode the information as it is in the case of written text.

As a means of information documentation, a writing system needs to be developed specifically for a given language. Even though it has been known for one language to adopt and adapt the writing system of another language, the necessary adaptations bring about distinctions that make such a new writing system unique in some important ways. Hence, the writing system of one language may not be directly usable by another language. To this end, the development of writing systems for any language demand standards and such standards require deliberate and structured efforts. Any culture that is not able to organise to develop such standards may have difficulties in developing a generally accessible and widely useful writing system. Certain components of the multimedia composite are not susceptible to these sorts of impediment.

The deliberate efforts that go into the process of learning to read and write must not be overlooked. When started early in life, the process of acquiring literacy naturally becomes an integral part of the acculturation process. For adult illiterates however, learning to read and write in adulthood generally presents a steep learning curve. Even though adult literacy programmes abound around the world, despite the spirited efforts of national and international organizations such as Unesco, there are still significant literacy gaps in various parts of the world. Hence, even though literacy has become a very important part of human experience, the enhanced capacity to document human experience using multimedia should be properly harnessed as one of the benefits of the information age particularly for illiterates as well as users of unwritten and signed languages.

Of course the use of multimedia is not totally new to mankind. Most of human pre-history is derived from the quest of our predecessors to document their experiences and present them to generations yet unborn. Such quests for documentation for future generations are sometimes expressed consciously but most times they were expressed unconsciously, simply

by leaving traces that hint to us of their ways of life and their experiences. This was the only option available to them before the invention of writing. Cave paintings, ancient sculptural pieces and many other pre-historic artefacts can be validly conceived of as elementary efforts to document information in multimedia. Due to the limitations in the technologies behind these efforts however, they were limited mostly to the visual medium. Today however, digital technology has given us the capacity to exploit multimedia and multimodality in ways that were hitherto unimaginable.

Hence, whilst we continue to make efforts to increase literacy levels in relevant communities around the world and develop writing systems for various unwritten languages spoken in the world today we should also integrate into these efforts the possible roles of multimedia and the new capacities made available by modern ICT in the documentation of information and the spread of knowledge through multimedia.

FROM ILLITERACY TO E-LITERACY

The above arguments are not designed to suggest that multimedia is not being used to productive ends at present. Rather, they are advanced to encourage the use of multimedia in cyberspace in more creative ways that can widen the scope of its use by illiterates and users of unwritten and signed languages. The present preponderance of written information in cyberspace is due not only to the efficiency and portability of written texts but also to the popularity that writing had acquired as an important information medium in the pre-internet era. Because we had become inadvertently positively biased towards writing as a means of communication before the Internet era, cyberspace has had to grow within the ecology of the written word, and cyberspace therefore reflects this bias for the written word by default. Yet there are multimedia alternatives that can be used to complement writing in cyberspace.

We can approach the use of multimedia to make cyberspace more accessible to the illiterate as well as users of unwritten and signed languages from two primary points of views. From the first point of view, we can use multimedia in its basic form, for example to record and play back information in the form of speech or by drawing and display of images. From the second point of view, we can exploit more sophisticated

technologies such as human language technology, particularly speech technology to mediate literacy.

There are many ways in which the basic use of multimedia can benefit illiterates as well as users of unwritten and signed languages. With the popularity of low data rate audio recording techniques such as MP3, as well as the steady reduction in the cost of recording media and their increased reliability, it is now feasible to record many hours of speech on relatively small and cheap digital media with a high level of reliability. Such recordings are valid documents that can be used in many ways as written text is normally used. They can be played back whenever the contents need to be consulted and they can also be indexed and therefore made searchable to facilitate efficient information retrieval. In the same vein, still images, video and animation can be used to tell stories, either as complements to recorded speech or separately on their own.

Apart from the use of multimedia in its basic forms, it is also possible to use human language technology and advanced speech technologies to mediate literacy. Despite the efficiency and portability of writing, speech still remains the preferred means of human communication. The organization of expensive face-to-face international conferences to get authors to read their papers to and interact with audiences, the development of audio books and other speech related technologies all bear testimony to the importance of speech to mankind. Hence, investments in the development of advanced speech technologies are justifiable on the grounds that humans prefer speech. Such investments in speech technology will inevitably trickle down for use in mediating literacy. However, the trickles have to be deliberately collected and harnessed for the benefit of the illiterate as well as users of unwritten and signed languages. For example, Automatic Speech Recognition (ASR) can be used to convert ideas expressed by an illiterate person to written text and Text To Speech (TTS) synthesis can be used to read written text to an illiterate person. By so doing, such an illiterate person has been enabled to interact with literature even without the ability to read or write.

Literacy is defined as the ability to read and write and its importance is based on the unprecedented access it gives to information and knowledge. With present levels of availability of multimedia and developments in human language technology as well as speech technology, coupled with the ever lowering cost of accessing them, there is a need to reassess the

relative value we put on written text and thereby create more space for the illiterate and users of unwritten and signed languages in cyberspace.

The recording and use of speech, still images, video and animation as well as the use of human language technology and speech technology in these above described ways are not uncommon in cyberspace. Hitherto however, they have not been seen as valid means of communication for the illiterate due to the high costs associated with the use of such media in the pre-Internet era. With recent developments in digital technology however, such uses are becoming more and more feasible. Even though they still appear to be relatively expensive for use in mediating mass literacy, they are bound to become less expensive as we move further into the information age.

The present rate of penetration of mobile phones in both urban and rural areas of the developing world present a salient pointer to the possibility of the spread of the use of multimedia to reach the illiterate as well as users of unwritten and signed languages. We therefore should approach multimedia from such futuristic perspectives and plan to harness its full benefits for creating an inclusive cyberspace.

MULTIMEDIA FOR AN INCLUSIVE CYBERSPACE

Before the information revolution, the world had experienced an agrarian revolution and an industrial revolution. Each of these revolutions featured change at a fast pace and high intensity. Inevitably, in such situations, people are bound to be overwhelmed and some may be left behind with the result that many people will suffer inequity.

In order to reduce the levels of inequity that may become manifest as a consequence of the information revolution, cyberspace being one of the most important products of the information revolution must be made as inclusive as possible. It must be made inclusive to the extent that people from any part of the world would be able to contribute to and benefit from cyberspace regardless of whether they are literate or illiterate, whether they speak written or unwritten language or whether they use signed languages.

The digital divide has emerged as one of the important metaphors that describe the levels of inequity in the information age. According to agreed

definition, the digital divide is “*the gap between people with effective access to digital and information technology and those with very limited or no access at all. It includes the imbalance both in physical access to technology and the resources and skills needed to effectively participate as a digital citizen*”². Even though the above description of the digital divide accommodates a multiplicity of impediments to participation as digital citizens, there usually is a connotation of lack of physical access to technology whenever people talk about the digital divide. The impression is usually created that the digital divide is mainly due to a lack of physical access to digital and information technology in form of terminal equipment and bandwidth.

Important as the availability of terminal equipment and bandwidth may be, the thinking that lack of physical access to technology is the main issue in the digital divide is patently misleading. We do know that even if all technological impediments to participation as digital citizens were to be removed today, difficult cultural and linguistic impediments would still exist. Some of these impediments can be addressed by the use of multimedia as well as human language technology because they provide much wider scope for the communication of information.

Given that the information revolution is propelled by digital technology, it has the inbuilt capacity to ensure the level of inclusion that can help reduce most of the anticipated inequities. It however behoves us as humans to identify the needs and devise schemes to ensure the levels of inclusion that can combat the anticipated inequities.

2 http://www.africa4all-project.eu/index.php?option=com_glossary

CYBERACTIVISM AND REGIONAL LANGUAGES IN THE 2011 ARAB SPRING

The Web is not a piece of writing, it rewrites itself every day. Millions of people seize different and complementary tools for networking or on file-sharing platforms. What role do these technologies play in personal Web 2.0 pages? How do they promote individual expression and collective intelligence? What is the role of languages in this redeployment of digital links between human beings? Between globalisation and protection of local languages, is it necessary to sacrifice local language to be heard by the global community?

Original article in English.



ADEL EL ZAIM is the Director of the international relations department at Sherbrooke University (Québec, Canada) and the former director of the International Development Research Center (IDRC/CRDI, Canada), as well as and Project Manager for Connectivity Africa (Cairo). He has also been president of the Internet Society of Quebec (ISOC-Quebec).

ADEL EL ZAIM

ARAB SPRING
IN THE 2011
LANGUAGES
AND REGIONAL
CYBERACTIVISM

2010 and 2011 will go down in history as years of dramatic change in the Middle East and North Africa. But the “Arab Spring” didn’t start then; rather, 2011 saw the culmination and success of revolutions in Tunisia and Egypt that had been attempted several times in the first years of the XXIst century. Revolts now continue in several countries; some have morphed into massacres (Syria), civil war and international intervention (Libya), or hidden repression (Bahrain).

If youth are largely credited for these Middle East uprisings’ success, they are having a similar impact across the continents. It’s a constant, from China, Burma and Iran, to Morocco, Algeria, Tunisia, and Egypt, to Libya, Yemen, Bahrain and Syria in 2010-2011, not to mention in Spain, Greece, Senegal, and more recently the United States or Russia, that young people are at the front line. Revolt and repression continue to cost lives and ruin economies. In many countries, like in Senegal¹, the population is demonstrating against the corruption and/or the president’s plans for a neverending mandate.

The role of information and communication technologies (ICTs) and social media was brought up early on in Tunisia and Egypt. Researchers and journalists wondered if the uprisings should be called the “Digital Revolution” or “Revolution 2.0”. More serious questions were asked about their timing and location: why now, why in Tunisia, why in Egypt, what country would come next? And how to use ICT and youth ability to use them to reform a country, to create new political systems, new categories of politicians and political practices?

Dozens of books and studies about the revolution have been published in and about Tunisia and Egypt, investigating this new renaissance era

1 <http://www.senrevolution.com>

and its reforms, impacts, benefits, history, and models. Cultural products, including songs, music, movies, and photos continue to be produced and proliferated worldwide.

But what of language? The relationship between language and revolution is a broad domain that can be tackled from diverse perspectives. When I had the opportunity to witness the revolution in Egypt in spring 2011, I found myself particularly interested in the role of language in the events, specifically the use of language in Egyptian social media. I observed its use by protestors and political movements to communicate, mobilize, and organize activities, as well as to document and share key moments of repression, success and celebration. Based on my observations, I came to study the impact of the revolution on Arabic language presence on the web and in social media. I suggest that the revolutions in Arab countries have from their very beginnings precipitated a heightened production and use of online Arabic language content, particularly in the social media. This content is produced by individuals, organizations and media, and is enhancing the rank of Arabic language online. While certain research corroborates my conclusions, much more remains to be done.

THE USE OF SOCIAL MEDIA IN ARAB COUNTRIES

Social media, such as Facebook, Twitter, and YouTube, are among the most used internet tools in Arab countries. Facebook is such an effective mass communication device that it is seen as a threat, and has been consequently blocked and unblocked by the government several times in various countries from Tunisia to Syria during the last three to four years in an attempt to contain political and social dissent.

The first issue² of the Arab Social Media Report³ showed over twenty-one million Facebook users in the Arab world as of January 5, 2011. This number jumped to 27,711,503 users in the first quarter of 2011 a 30% increase. Gulf Cooperation Council (GCC) countries dominate the top five Arab Facebook users in terms of percentage of population. Egypt constitutes about a quarter of total Facebook users in the Arab region, and added more users in the Q1-2011 than any other Arab country, closing at two million new users between January 5 and April 5⁴.

2 <http://www.dsg.ae/NEWSANDEVENTS/UpcomingEvents/ASMRHome.aspx>

3 <http://www.dsg.ae/NEWSANDEVENTS/UpcomingEvents/ASMROverview2.aspx>

4 <http://www.dsg.ae/portals/0/ASMR2.pdf> page 9.

Governments are capitalizing on the opportunity offered by this huge concentration of their citizens on Facebook to become more present in social media. Egypt and Tunisia are good examples of this: immediately following the revolts' success in ousting the presidents, the new transitional governmental bodies reached out to Tunisians and Egyptians through Facebook. Despite each government already having sophisticated online presence, each ministry and national organization having its own web site, they created pages on Facebook and YouTube, and opened accounts on Twitter, to communicate their message and try to engage in discussion with their constituency, both those living at home and abroad. Even the government of the United Arab Emirates (UAE), which was not even facing a conflict at the time, felt the need to build a Facebook profile and presence to connect with the 45% of its population linked to that network. The UAE government is now encouraging its employees to use social media to interact with citizens. It has even offered trainings on responsible use and risks of Facebook, and offers documented policy guidelines for government bodies⁵.

DIVERSE APPLICATIONS

Given the demographics of Arab countries, where a full 30% are youth age 19 to 25, and those countries' political and economic situations, Facebook is *"being used in a wide variety of ways: rally people around social causes and political campaigns, boost citizen journalism and civic participation, create a forum for debate and interaction between governments and their communities, or to enhance innovation and collaboration within government"*⁶. However, the main use of Facebook was and remains that intended by its creators: social networking between individuals and groups of friends. Despite censorship and blocking, Facebook persists as the networking tool par excellence for young people who want to communicate, meet, share hobbies and dreams, and endorse celebrities.

In the first quarter of 2011, the role of cyber-activism in the revolts on the streets triggered a dramatic change in the use and perception of Facebook in Arab countries. Blocking Facebook and the internet in Tunisia, Egypt,

⁵ Guides by the government of UAE in arabic <http://www.emiratesegov.ae/web/guest/83>

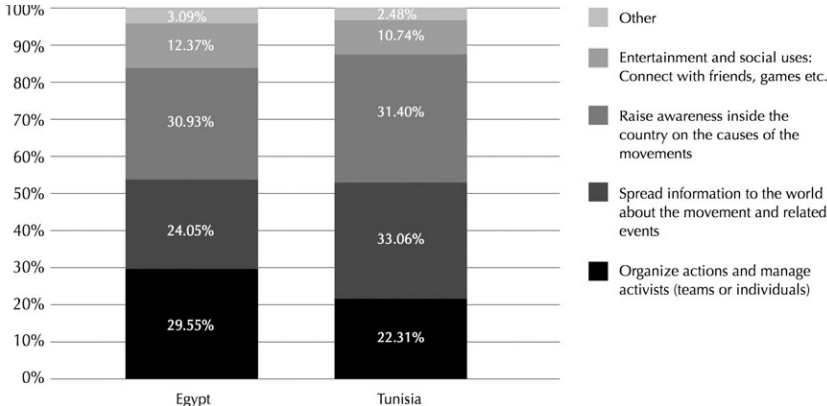
⁶ <http://www.dsg.ae/portals/0/ASMR2.pdf> page 1.

Libya, and Syria only gave these tools more credibility and impact, increasing their demand, and then their use.

In Egypt, since January 2011, YouTube and similar networks are used to document and share events, ranging from calls for meetings, to demonstrations, to attacks and massacres. In Syria, Libya, and Bahrain, dated videos have permitted protestors to prove the pacifist nature of their actions and the brutality of the authorities. Twitter provides a communication channel for rally notification, sos, quick instructions, and spreading and receiving news. The very structure of Twitter, whose short messages or “micro blog” entries are seamlessly integrated to mobile phones, make it a particularly able mass communication device.

The Arab social media report surveyed 126 people in Egypt and 105 from Tunisia regarding the main use of Facebook during the civil revolt. As represented in the figure below:

In both countries, Facebook users were of the opinion that Facebook had been used primarily to raise awareness within their countries about the ongoing civil movements (31% in both Tunisia and Egypt), spread information to the world about the movements (33% and 24% in Tunisia and Egypt respectively), and organize activists and actions (22% and 30% in Tunisia and Egypt respectively). Less than 15% in either country believed Facebook was primarily being used for entertainment or social reasons⁷.



⁷ <http://www.dsg.ae/portals/0/ASMR2.pdf>, page 6.

THE LANGUAGE OF FACEBOOK

Facebook offers its interface in dozens of languages, most of them localised by users themselves⁸. Users in Arab countries surveyed for the Arab Social Media Report “*vary in their preference of language interface*”⁹, the three main languages unsurprisingly being Arabic, French and English. The survey showed net preference for English in Gulf countries, with the exclusion of Saudi Arabia, and net preference for French in the Maghreb countries and Comoros. Egypt and Tunisia are worth paying special attention to because of the changes over the course of the revolution. In terms of preference of language interface, users in Egypt split evenly between the use of Arabic (49.88 %) and English (48.98 %) interfaces (similar to Jordan, Libya and Iraq). Tunisian users showed a net preference for French interface (94.60 %), then English (2.72 %) and finally Arabic (1.56 %).

However, the interface language setting preference cannot necessarily be used to conclude what languages users are actually interacting in¹⁰. Thanks to HTML and Unicode, browsers are now able to display text in virtually all languages. “*Facebookers practice a diversity [of languages that] challenges the conventional notions of multilingualism as a combination of two or more monolingualisms*”¹¹.

A LOCAL LANGUAGE REVOLUTION

It is not a surprise that language played an important role during the social and political uprisings in Tunisia and Egypt, as well as in other countries of MENA. Some slogans chanted by demonstrators became internationally known symbols and songs. “*Ben Ali, dégage!*” in Tunisia, or *al-sh’ab yureed asqat al-nitham*, (“the people want the regime to fall”), repeated in Tunis, Cairo, Damascus, Benghazi, and Sana’a, are intrinsically linked with the revolts. Signs held by protestors in Cairo were written mainly in Arabic, but also in English, French, and even Hebrew¹².

8 See in this book: Dwayne Bailey, *Software Localization: Open Source as a Major Tool for Digital Multilingualism*.

9 <http://www.dsg.ae/portals/0/ASMR2.pdf>, page 14.

10 As demonstrated in <http://www.languageonthemove.com/language-globalization/multilingualism-2-0>, blog posted on August 02, 2010 by Ingrid Piller.

11 *Ibid.*

12 See picture by glcarlstrom (Gregg Carlstrom) at <http://yfrog.com/h3fbsbj> translated as: *azov Mubarak* - “leave, Mubarak”.

On the social media front, linguistic creativity was positively impacted by the uprising. Perhaps the need to reach out to a larger community about burning issues led people to use the local language, thereby increasing the quantity of Arabic content published online both on social media and regular web sites.

This hypothesis is supported by my observation of 1) the number of new websites published in Arabic by newspapers, social movements, and government entities, and 2) the number of social media users who are now writing in Arabic.

SUPPORTING NUMBERS

This hypothesis is supported by the results of the second issue of the Arab Social Media Report, released in May 2011. If we compare the interface language to the language used by Egyptians and Tunisians to communicate during the civil movement of the first quarter of 2011, we see a huge difference. Below is the language distribution¹³:

Country	Arabic (% of FB users)	English (% of FB users)	French (% of FB users)
Egypt	49.88	48.98	0.39
Tunisia	1.56	2.72	94.60

However, when questioned¹⁴ about the primary language used to communicate on Facebook during the civil uprisings, respondents answered as per below:

Country	Arabic (% of FB users)	English (% of FB users)	French (% of FB users)
Egypt	75.40	25.60	0
Tunisia	51.43	0.95	47.62

In the assessment of this researcher, the revolutions in Egypt and Tunisia are behind the increasing disparity between interface language and the language of communication.

13 Arab Social Media Report, Issue 2, May 2011, <http://www.dsg.ae/portals/0/ASMR2.pdf> page 14.

14 Arab Social Media Report, Issue 2, May 2011, <http://www.dsg.ae/portals/0/ASMR2.pdf> page 7.

As an empirical way to test this hypothesis, I set out to examine the tweets for #CairoExplosion on July 6, 2011. Cairo and suburbs residents heard on Wednesday the 6th of July a big explosion whose origin remained unknown for hours. Rapidly, a hashtag was created on Twitter and users start sending messages asking or guessing or retweeting what they heard. In less than an hour, that hashtag generated enough attention to precipitate a certain level of panic among Egyptian Twitter users. I randomly copied 500 tweets with the hashtag #CairoExplosion, all sent within approximately two hours of one another, and classified them in terms of language to obtain the following distribution :

Arabic	Arabic in latin characters	English	Mixed	Not determined	Total
353	33	89	19	6 (smileys)	500

Despite the obvious interest of those numbers, the observation is limited and needs to be systematically examined and validated. Certain initiatives like R-Shief¹⁵ aim, among other things, to mine and visualise Twitter content and the public sphere of Facebook to study the language distribution and uses.

Increasing the content quantity doesn't necessarily mean enhancing its quality, however. Tweets are limited to short messages and microblogs. Social media content is characterised by its informal style, although it certainly makes important contributions to more institutional pages and networks, which together with the presence of government entities and organizations lend it more formal status. A huge number of international organizations, including the United Nations Development Program (UNDP), the International Fund for Agricultural Development (IFAD) and Canada's International Development Research Centre (IDRC) maintain Facebook pages and Twitter accounts, contributing to enhancing the style and credibility of these tools. And since Barack Obama's 2008 presidential campaign, almost all candidates of the developed and developing world use social media to try to reach voters, or at least the Facebookers among their potential voters.

15 <http://www.r-shief.org>

CONCLUSION

There are clear signals that around the world, social uprisings and unrest are making social media one of the preferred tools of communication, mobilisation, demonstration and voicing the concerns of the population. MENA countries are showing a serious increase in the number of social media users. These users are mainly communicating in their mother tongue, regardless of interface language preference. More research is needed to better document this phenomenon and build on the present research, to enhance its content quality, but more importantly to contribute to the quality of citizens participation via cyberspace, and to bring about increased benefits from the knowledge society.

MULTILINGUALISM, THE MILLENIUM DEVELOPMENT GOALS, AND CYBERSPACE

The “Millennium Development Goals” (MDGs) set by the United Nations in 2000 must be evaluated at a world summit in 2013. How can the use of languages in technology be a decisive help for their realisation? How can public language policies, local or multilateral, participate in achieving these shared goals?



ADAMA SAMASSÉKOU is the President of the International Council for Philosophy and Human Sciences (CIPSH) and the President of MAAYA, the World Network for Linguistic Diversity. Former Executive Secretary of the African Academy of Languages (ACALAN), a specialized Institution of the African Union based in Bamako (Mali), he served as the President of the Preparatory Committee (PrepCom) of the World Summit on the Information Society for the Geneva phase (WSIS/2002-2003). Previously, he was Malian Minister of Basic Education, initiator of the Rebuilding of the Educational System (1993-2000) and Spokesperson for the Government of Mali (1997-2000). Member of the Haut Conseil de la Francophonie from 2003 to 2006, he is today member of the ITU and Unesco International Broadband Commission for Digital Development.

ADAMA SAMASSEKOU

CYBERSPACE
GOALS, AND
DEVELOPMENT
THE MILLENNIUM
MULTILINGUALISM

In speaking of multilingualism, we often think of linguistic performance – the methods people have for communicating; and how they record, confront and share their cultures. Linking the language question to the so-called “development” question may appear irrelevant – the relationship may appear so natural and commonsensical that it needs no further highlighting! In many countries, however, this link is complicated by history and the long-term effects of colonization, to the point that the relationship between language and development is distorted or even denied. It then becomes a crucial topic for examination. As we enter into a new society under rapid construction, a society of shared knowledge and information, it is worth pausing to analyse multilingualism’s contribution to the Millennium Development Goals (MDGs). Cyberspace plays a major role in global re-organization and community projects, and the orientations of multilateral organizations have a major impact on online relationships. Multilingualism is no exception to this focus of the global future on communication technologies and information.

ELIMINATING POVERTY

The eight Millennium Development Goals (MDGs) were adopted at the Millennium Summit held on 6-8 September 2000, at the United Nations headquarters in New York.

“Eradicating extreme poverty continues to be one of the main challenges of our time, and is a major concern of the international community. [...] The Millennium Development Goals set timebound targets, by which progress in reducing income poverty, hunger, disease, lack of adequate shelter and exclusion—while promoting gender equality, health, education and environmental sustainability. [...] The Goals are

ambitious but feasible and, together with the comprehensive United Nations development agenda, set the course for the world's efforts to alleviate extreme poverty by 2015¹.

The eight goals aiming to significantly improve living conditions by 2015 include the following: 1. Eradicate extreme poverty and hunger; 2. Achieve universal primary education; 3. Promote gender equality and empower women; 4. Reduce child mortality rates; 5. Improve maternal health; 6. Combat HIV/AIDS, malaria, and other diseases; 7. Ensure environmental sustainability; and 8. Develop a global partnership for development.

As the year 2015 approaches, we can easily see that these goals are still works in progress, even if some of them have undergone significant improvements. In this light, the recent initiative of the International Telecommunications Union (ITU) and Unesco is worth noting: they have partnered to establish the Broadband Commission for Digital Development, to help accelerate the achievement of the Millennium Development Goals². Global development is propelling the integration of the MDGs with the UN Conference on Sustainable Development, to be held in Rio in June 2012. Environmental issues, sustainable development, interstate equilibrium, and planetary protection in the face of climate change and biodiversity loss, will be placed next to the goal of eradicating poverty, a goal approved by all countries in 2000. While the link to greening the economy may seem obvious, approaching it through cultural expansion and knowledge sharing remains too little understood.

Development is a complex concept, referring to history and to the global inequalities it has left us. The analysis presented in this article concerns only some of the world's countries – those for whom the international community identified the Millennium Development Goals, to be achieved by 2015.

For the sake of a better understanding, I have chosen to focus on Africa, where the problematic of the relationship between multilingualism, the MDGs and cyberspace is most urgent and thus most relevant. This strategy allows my argument to insist upon the linguistic dimension of the MDGs and the means of achieving them.

1 Ban Ki-moon, Secretary-General of the United Nations, *Committing to action: achieving the Millennium Development Goals*, 25 July 2008.

<http://www.un.org/millenniumgoals/2008highlevel/pdf/committing.pdf>

2 <http://www.broadbandcommission.org>

It is important to emphasize the specificity of the language question in formerly colonized countries in general, and in Africa in specific, a question marked by the continuing search for identity, a quest with well-known origins.

The Cultural Charter for Africa, adopted by the Heads of State and Government of the Organization of African Unity (OAU) upon meeting for their Thirteenth Ordinary Session in Port Louis (Mauritius), on 2-5 July 1976, reminds us that *“under colonial domination, the African countries found themselves in the same political, economic, social and cultural situation; that cultural domination led to the depersonalisation of part of the African peoples, falsified their history, systematically disparaged and combated African values, and tried to replace progressively and officially, their languages by that of the colonizer, that colonization has encouraged the formation of an elite which is too often alienated from its culture and susceptible to assimilation and that a serious gap has been opened between the said elite and the African popular masses”*³.

Proper management of the situation requires these countries to pursue a social project based on endogenous development, that guarantees cultural and linguistic diversity, which in turn ensures that all cultures and languages can find expression and put themselves forward. This will be truly possible only when African languages can truly be presented as working tools in all domains of public life, in partnership with the official languages inherited from colonization, and in all means of expression, communication and dissemination.

Hence the need and urgency to develop an African perspective on language planning. We know that “language planning” consists of using policy and technical measures to preserve or promote linguistic equilibrium in an area where several languages are spoken, in order to better serve the social project.

The adoption by these formerly colonized countries of an explicit language policy reflecting the social project is a necessary element of this methodology.

3 *Cultural Charter for Africa*. http://www.au.int/en/sites/default/files/CULTURAL_CHARTER_AFRICA.pdf

MULTILINGUALISM, A LIVED REALITY IN COUNTRIES COVERED BY THE MDGS

During the early years of African independences, culture guided and illuminated political choices, because it was seen as the basis for and the purpose of any social or economic development process. Today, however, it is undeniably the case that exogenous development programs increasingly – unfortunately – are the primary actors advancing projects to promote cultural values. As development projects cannot feasibly be designed outside the cultural context of those for whom they are intended, I feel it is urgent to reverse this trend.

However, in the vast majority of the states concerned, there is an unavoidable paradox impeding this reversal: we are dealing with multilingual countries that, as heirs of colonialism, already have a choice practically imposed on them in the form of the nation-state phenomenon, a construction privileging the monolingual perspective.

Because of this, in order to avoid pursuing a program that goes against these countries' realities, I advocate implementing a strategic approach that I call user-friendly functional multilingualism, which I define as *“a strategic approach to managing African linguistic pluralism, taking into account both the principle of linguistic equality and the reality principle of the various functions these languages perform. This approach goes hand in hand with administrative decentralisation and African integration, using a matrix of an identity language, a common language, and a language of international communication. It advocates linguistic usability and advocates the ‘delegation of linguistic sovereignty’ by way of the Subsidiarity Principle between local, regional, national and African levels”*.

MULTILINGUALISM, A SINE QUA NON FOR ACHIEVING THE MDGS AND AN INCLUSIVE CYBERSPACE

The vast majority of humanity lives in multilingual societies, where multilingualism is the norm. How, then, can we not take the language question and multilingualism into account when working towards the MDGs? A transversal issue par excellence, the language question determines

our ability to achieve each of the eight Millennium Development Goals. Let us attempt to measure the impact on each of them.

Eradicate extreme poverty and hunger

There will be no reduction, much less eradication, of poverty, so long as grassroots actors aren't empowered in their own language(s), which should be ordinary tools for working, learning, and transforming their environment in the broadest sense of the word. World hunger will be reduced by mobilizing peasant communities. But these communities speak local languages, which at most describe their land, their culture, and regional fruits and vegetables. The expert panel the International Assessment of Agricultural Knowledge, Science and Technology for Development⁴ indicated in its 2008 report that nearly one billion people go to bed hungry. The panel believes that the best fight this is to encourage small businesses, and to bring knowledge, technology, and credit to rural communities⁵. Knowledge must be expressed in languages these peasant communities can understand, assimilate, and enrich with their own knowledge set. Here we see language's direct impact on the reduction of world hunger.

Achieve universal primary education

There can be no high-quality universal primary education without the implementation of bi- and multilingual education, with the mother tongue as its basis⁶. We are beginning to see the capacity for cyberspace to expand education for all. But a great risk exists, insofar as the manuals, examples, and models are conceived and written in dominant languages and for the cultures of the most developed countries. Accessing scientific concepts in one's native language, and translating them only later into other languages, allows for improved internalisation of ideas and a genuine exchange of knowledge.

Teachers often learn multiple languages, both for their university coursework and for interacting with their pupils. To this effect, teacher trainings

4 International Assessment of Agricultural Knowledge, Science and Technology for Development/IAASTD. <http://www.agassessment.org>

5 Mark Kinver, "Global Food System Must Change", *BBC News*, 15 April 2008. <http://news.bbc.co.uk/2/hi/science/nature/7347239.stm>.

6 See in this book: Marcel Diki-Kidiri, *Cyberspace and Mother Tongue Education*.

may benefit greatly from ICT. The same goes for sharing curricula and educational resources. The *Open Educational Resources*⁷, which allow for their translation and re-use in other contexts, constitute powerful tools for achieving this goal by way of a multilingual cyberspace.

Promote gender equality and empower women

The UN indicators for reporting on progress in this domain are⁸:

- the ratio of girls to boys in primary, secondary and higher education ;
- the literacy rate of women ages 15 to 24 compared to men of the same age group ;
- the percentage of women employed in the non-agricultural sector ;
- the proportion of seats held by women in national parliament.

For all these criteria, the issue of the number of known languages seems crucial. While these criteria are far from being met, the education of girls and women is steadily spreading. But many obstacles persist when it comes to obtaining an education beyond the primary level. For this, distance learning is a potential tool. However, education has to be taught in the languages of its learners, and to offer the possibility of translation.

Particular attention should be paid to ICT education for women, either in formal education or in women-oriented technology learning groups. To promote this type of collective work, women must be able to interact in their own language and use the tools in either their own or a regional language. But most software and services come in so-called major languages, creating further barriers. A localizing effort for software⁹ will increase the capabilities of ICTs and cyberspace to support this goal that is central to world equality.

Reduce child mortality rates ; Improve maternal health

These two goals are interrelated. Welcoming children into the world should be one of society's greatest achievements. But women in developing

⁷ *Capetown Open Education Declaration*, <http://www.capetowndeclaration.org>.

⁸ <http://mdgs.un.org/unsd/mdg/Host.aspx?Content=indicators%2fofficiallist.htm>

⁹ See in this book: Dwayne Bailey, *Open Source and Free Content: Powerful Tools for Language Activists*.

countries are at risk during the period from pregnancy through childbirth ; and a child born in a developing country is 13 times more likely to die in its first 5 years than a child born in an industrialized country.

In this case as well, much progress has been made, particularly in having caregivers present during labour, and in preventing treatable diseases that can be particularly severe in young children, such as pneumonia, diarrhea, and measles.

Reducing mother and infant mortality is closely linked to children's and mothers' awareness of public health messages, which is facilitated when those messages are written in their own languages. The community's ability to regain confidence in its own endogenous set of knowledge, skill, and wisdom, which goes back centuries and is embedded in its language, is similarly crucial.

There is room for improvement in the use of multimedia¹⁰ to enhance the dissemination of public health messages tailored to the real conditions women and children in the poorest cities and villages. Overdubbing and using a country's national language are two examples of this.

Combat HIV/AIDS, malaria, and other diseases

Similar comments can be made about stopping and reversing the spread of HIV/AIDS, malaria, and other epidemic diseases. Despite advances in medicine and antiretroviral drugs, AIDS remains a wound on developing countries. HIV prevention messages, the use of contraception, and sex education remain the essential means of countering the disease's spread. The cultural dissemination of this mindset, at an unprecedented scale, is facilitated by cyberspace. Connecting community radio stations facilitates the exchange of statements, interviews, and public service announcements with musicians and cultural figures can help lead to prevention. In Africa, radio remains the most widely used medium for communicating information, but the internet can be used to build community radio stations if care is simultaneously taken to develop multilingualism.

For all infectious diseases, cyberspace is a very effective tool for prevention and treatment because of its ability to widely distribute cutting-edge medical research. Medical teams in developing countries must be able to

10 See in this book : Tunde Adegbola, *Multimedia and Signed, Written or Oral Languages*.

access to the latest research and study results. The entire chain of health care, from the university to the clinic, need to have medical knowledge sharing. Free access to medical literature¹¹, and the translations of key results into the greatest possible number of languages, transforms multilingual cyberspace into a catalyst for improved medical practices worldwide, including in those places that are farthest from research centres and advanced hospitals.

In general, advances in telemedicine in a country like Mali open up excellent prospects for the remote management of disease, and for patients' ability to more comfortably explain their illness in their own languages, especially rural areas poorly equipped with medical facilities.

Ensure environmental sustainability

The environment is the product of a cultural relationship to nature. Each society is involved in the protection of biodiversity, and unfortunately its destruction at times, through physical action onto the surrounding world. But this relationship also occurs through myths and legends, as well as research and writings, that shape people's minds. The use of traditional practices, and their related worldviews based on harmony and respect for nature, have survived for centuries via oral tradition in local languages, and can make environmental issues more familiar and quickly appropriated.

By allowing these worldviews to cohabit, which resurrects for urbanites the long wire connecting them to nature, cyberspace becomes a conductive thread making society aware of the dangers posed by global climate change and an eroding biodiversity.

It is also by documenting environmental damage, and spreading the information beyond the places directly affected, that multilingual cyberspace can catalyze a global environmental consciousness.

11 Leslie Chan, Subbiah Arunachalam & Barbara Kirsop, The chain of communication in health science: from researcher to health worker through open access, *Open medicine*, 2009, 3(3), p. 11-19.

Develop a global partnership for development

How can we develop multi-stakeholder partnerships that are fruitful for global and local development without sharing the language of communication?

The new knowledge society is being built, with ICTs as its cement. How can we ensure that it is fully inclusive, leaving out no stratum, community, or society?

CONCLUSION

The primary societies covered by the MDGs are also those that have not yet entered digital culture. This is largely because the digital divide is repeated as a language gap: their languages have not yet been established online.

How can we ensure that every language can be used in cyberspace as a means of communication, production, and information and knowledge sharing, as recommended by the World Summit on the Information Society (WSIS)?

In light of the transversal nature of the language question for achieving the MDGs, integrating multilingualism as a new MDG is essential:

- Multilingualism, strengthening the capacity of plural societies;
- Multilingualism, enabling the consolidation of multiple identities and reinforcing societal development;
- Multilingualism, building confidence in populations, building a true dialogue between cultures, guarantor of social cohesion, peace, and mutual understanding.

If multilingualism can accelerate the rate at which we achieve the MDGs, then it must itself become the ninth Millennium Development Goal.



**MULTI-
LINGUALISM
ON THE
INTERNET:
A MULTI-
LATERAL
ISSUE**

PART 4

The question of languages in cyberspace is a political issue at the multilateral level. From defining standards of network governance up to the relationship between the defence of multilingualism and Human Rights, languages represent major benchmarks in the Information Society.

DESCRIBING THE WORLD: MULTILINGUALISM, THE INTERNET, AND HUMAN RIGHTS

Promotion and protection of multilingualism on the Internet, how does it figure in the international architecture of Human Rights? To develop the legal issues of multilingualism in cyberspace, it is necessary to assess possible links between the philosophy of law, the theory of speech acts and transcultural universals in relation to the uses of the internet and the web. What would be the nature of a linguistic human right? How can we define its nature and scope? How can we change the legal regimes, public or private, local or international, that have effects on cyberspace?

*Celui qui a voulu que l'homme fût sociable
toucha du doigt l'axe du globe et l'inclina sur l'axe de l'univers.*
Jean Jacques Rousseau
Essai sur l'origine des langues, Œuvres, Belin, Paris, 1817.

Original article in English.



ISABELLA PIERANGELI BORLETTI holds a doctorate in International Law and European Union Law from the University of Pisa. She is a visiting scholar at the Social Science Institute of the Ankara University (Turkey) where she teaches Public International Law, International Organizations and European Union Law.

ISABELLA PIERANGELI BORLETTI

DESCRIBING
THE WORLD:
MULTILINGUALISM,
THE INTERNET, AND
HUMAN RIGHTS

This article considers the potential of national, regional and international legal trends promoting and protecting multilingualism on the internet within a human rights framework. A full exploration requires understanding the links between public and private international law, together with social norms and cultural universals regarding internet use worldwide.

Online development's increasing speed brings into consideration a number of human rights issues, including that of multilingualism¹ as an intrinsic element of cultural diversity. If multilingualism is considered a human right, then protecting the individual and collective rights of linguistic minorities in cyberspace is a means through which other internationally recognized human rights can be guaranteed, including the rights to education, development, and freedom of expression, and the right to seek and receive information. Failing to fully recognize online multilingualism hampers enjoyment of these rights.

An analysis of current thinking and of the national and international legal tools adopted by states and international organizations, can shed light on the extent to which human rights, the Internet and multilingualism are related or disconnected concepts, and provide a basis for future framing of the issue.

1 For the sake of clarity it is better to define how multilingualism differs from plurilingualism as defined by the Council of Europe in the *Language Education Policy*, http://www.coe.int/t/dg4/linguistic/division_EN.asp. Multilingualism refers to the presence in a geographical area, large or small, of more than one variety of language, that is, a social group mode of speaking whether or not it is formally recognized as a language; in such an area individuals may be monolingual, speaking only their own variety. Plurilingualism refers to the repertoire of languages used by many individuals. It is therefore the opposite of monolingualism, including a 'mother tongue' or 'first language' and any number of other linguistic varieties. In multilingual areas, some individuals are monolingual and others plurilingual.

THE INTERNET AND HUMAN RIGHTS

Email, blogging, search engines, online shopping and podcasts do not necessarily implicate all human rights, but the internet rapid development in society raises numerous human rights concerns, including civil, political, economic and social rights on both individual and collective dimensions. This discussion thus necessitates a transversal approach.

We have recently witnessed a correlation emerge between the Internet and human rights, particularly in circumstances where freedom of expression and privacy rights are jeopardized by policies of state control and censorship. Turkish Law 5651 adopted on 4 May 2007 limits freedom of expression and narrows citizens' access to information², and is a contemporary example of how state policies limiting internet access threaten other rights. International legal language furthermore considers human rights to be "*universal and inalienable, indivisible and interdependent*".

The need for a rights-based conceptual framework in the discussion of Internet access and use is increasingly evident. This approach should be promoted at the highest political level, with special attention given to notions of sustainable development, the digital divide, multilingualism and ethics in cyberspace.

THE INTERNET AS A COMMON GOOD

The Internet and the Right to Development

The right to development, and academically and politically disputed concept, was selected as the overarching theme of the Internet Governance Forum³. Its status as a human right is pending⁴, and a the legal obligation

2 In the OSCE Representative on Freedom of the Media Report on the Turkish Internet Law (January 11, 2010) Mrs. Miklos Harszti urged the Turkish authorities to amend or abolish the Internet Law.

3 The Internet Governance Forum's (IGF) purpose is to support the United Nations Secretary-General in carrying out the mandate from the World Summit on the Information Society (WSIS) to convene a new forum for multi-stakeholder policy dialogue. It focuses on maximizing the Internet's potential to boost social and economic well-being in developing countries.

4 A. Esterhuysen, R. Greenstein, "The Right to Development in the Information Society", In: *Human Rights in the Global Information Society*, MIT Press, Boston, 2006, p. 285.

of developed countries to provide development assistance to developing countries is being debated within the United Nations (UN)⁵. A rights-based approach to development includes four elements: *accountability, empowerment, participation and non-discrimination*.

- *Accountability* refers to the identification of specific duties and duty-bearers, which shifts development cooperation from the domain of charity to that of obligation, thereby requiring duty bearers to ensure access to a social good;
- *Empowerment* means that development activities should facilitate and assist community efforts to improve their conditions and assume influence over their own destinies, by providing power, capacity and access. As an example, the Global Information Society Watch⁶ (GISWatch) recently studied fishermen from coastal villages in southern India who gained access to information on weather conditions and the market in their own language via mobile phone⁷;
- *Participation* implies a high degree of inclusion of community, civil society, minority groups, indigenous peoples, women and others;
- *Non-discrimination and attention to vulnerable groups* highlights the need to guard against reinforcing preexisting asymmetries of power and resources by prioritising disadvantaged groups.

These principles should be analyzed in light of the major themes put forward by the Internet Governance Forum. The goal would be to maximize the Internet's potential to boost social and economic well-being for the greatest number of people in developing countries, through access, diversity, openness and security. Among these four principles, the issue of the access is of particular importance, considering its relevance – in relation to the internet's potential – to the right to development.

In practical terms, access means that all individuals should be able to participate in the progress and benefits of technological development (Articles 26 and 27 of the *Universal Declaration of Human Rights*), and

5 The legal basis for the right to development can be found in the following: Universal Declaration of Human Rights, Articles 2(1) and 2(3); African Charter on Human and People's Rights, Article 2(1); UN Declaration on the Right to Development (1986); Vienna Declaration and Programme of Action (1993), adopted at the World Conference on Human Rights; and the UNDP Human Development Report for 2000.

6 See <http://www.giswatch.org/about>

7 S. Arunachalam, "Information and Livelihoods", in *Global Information Society Watch 2009 – Advancing Human Rights and Democracy*, available online at <http://www.giswatch.org>

that for developing nations to have internet access is a key economic and educational development tool. This objective, which can only be achieved through infrastructure, equipment and capacity-building towards universal web access, has been advocated under a number of slogans such as “broadband for all”, “universal service”, and “services of general interests”.

A number of national and international actors consider wide broadband coverage as crucial to fostering economic growth, especially in less-developed areas. Europe’s recognition of its obligation to provide all citizens with Internet access only highlights its disparity with developing nations. The 2010 Millennium Development Goals Report – particularly Goal #8, To Develop a Global Partnership for Development – renders conspicuous the internet access gap, or “digital divide”, between developed and developing nations⁸. In 2008, 23 percent of the world population used the Internet; in the developing world it was only 17 percent. The digital divide also remained wide within developed countries. Furthermore, the report highlighted the “broadband gap” between those with high-speed connections and those with dial-up modems⁹.

When the right to internet access is not granted and protected through the elimination of both structural and technological barriers, then the right to development is also threatened, especially now that internet access is perceived as a “fundamental right of all people” by 87 % of those who use the Internet and 71 % of non-users¹⁰. In other words, the right to Internet can be considered part of the right to development.

8 See in this book: Adama Samassékou, *Multilingualism, the Millenium Development Goals, and Cyberspace*.

9 The report also explains that: “[a] challenge in bringing more people online in developing countries is the limited availability of broadband networks. Many of the most effective development applications of ICT, such as telemedicine, e-commerce, e-banking and e-government, are only available through a high-speed Internet connection. But a significant divide exists between those who enjoy fast access to an online world increasingly rich in multimedia content and those still struggling with slow, shared dial-up links. By the end of 2008, fixed broadband penetration in the developing world averaged less than 3 per cent and was heavily concentrated in a few countries. China – the largest fixed broadband market in the world – accounts for about half of the 200 million fixed broadband subscriptions. In most least developed countries, the number of fixed broadband subscriptions is still negligible; service remains prohibitively expensive and inaccessible to most people. However, the introduction of high-speed wireless broadband networks is expected to increase the number of Internet users in developing countries in the near future”.

10 A poll for the BBC World Service (March 8, 2010) suggested that almost four out of five people worldwide believe that access to internet is a fundamental right. <http://news.bbc.co.uk/2/hi/8548190.stm>

The Internet and the Right to Education

Article 10 of the *Universal Declaration of Human Rights* recognises the right to education, a right reaffirmed in numerous other documents, including the 1990 *World Declaration on Education for All*, Articles 13 and 14 of the *International Covenant on Economic and Social and Cultural Rights* and Article 2 of the *First Protocol to the European Convention on Human Rights*. The international community also accepts that the right to education may be apply even outside the framework of these treaties.

Not only does education provide the skills needed to encourage innovation and boost productivity, it also increases economic growth. Knowledge is no longer simply a device, but an economic tool, and an increasingly knowledge-driven global economy means its importance will continue to increase. This has implications for wealth distribution on both global and individual levels: the growing gap between education-rich and education-poor countries is being replicated at the level of household which evince a growing gap between skilled and unskilled individuals¹¹.

The right to build information technology (IT) skills is subsumed with the right to education. This creates positive obligations on states and is closely linked to economic development and poverty eradication. These obligations are outlined in Article 2(2) of the *International Covenant on Economic, Social and Cultural Rights*, including the provision of primary education to all on a non-discriminatory basis. The *Unesco Convention against Discrimination in Education* (1960) also outlines these obligations.

The internet is also one of the most effective tools in ensuring that MDG #2 may be met: that by 2015 all children will complete a full course of primary schooling, possibly through distance-learning, which expands educational opportunities to a maximum number of people at minimum cost. Distance-learning is evolving, and the internet now constitutes a virtual classroom with intense interactivity and resource and information sharing, but for most developing countries, lack of infrastructures – power, equipment and literacy – means this technology remains out of reach¹². In addition, most educational content currently available online was designed in Europe or North America, and is not necessarily appropriate, content-wise and language-wise, for students in other

11 A. Clapham, *International Human Rights Lexicon*, Oxford University Press, 2005.

12 <http://www.itu.int/newsarchive/wtd/2001/FeatureEducation.html>

contexts. But the fact that many universities are now shifting their existing distance-learning programs to the Internet shows its potential as a tool for expanding education.

The Internet and the Right to Cultural and Linguistic Diversity

IT evolution should offer new potential for all cultures and languages. Yet its convergence with market globalisation instead imperils cultural and linguistic pluralism.

The UN *Convention on the Prevention and Punishment of the Crime of Genocide* was the first legal instrument designed to protect minorities.. Subsequent codification of minority rights includes the *International Covenant on Civil and Political Rights* (Article 27), the UN *Declaration on the Rights of Persons belonging to National or Ethnic, Religious and Linguistic Minorities*¹³, the *Framework Convention for the Protection of National Minorities*, the *European Charter for Regional or Minority Languages* (adopted by the Council of Europe), and the Organization for Security and Co-operation in Europe (OSCE – Copenhagen Document of 1990). In addition, many countries have created relevant laws, commissions and institutions to protect minorities, in some cases including the linguistic minorities.

Unesco's *Universal Declaration on Cultural Diversity* (2001) made cultural diversity, or “common humanity heritage”, a pressing contemporary issue. The declaration rendered cultural diversity a concrete and ethical imperative that was inseparable from respect for human dignity.

Since the internet is today's globalisation tool *par excellence*, it should reflect the broader effort to promote real cultural and linguistic diversity. All languages, particularly of minority communities, should be vehicles for online culture and communication.

Although languages are disappearing, the principle of promoting linguistic diversity must be maintained and include all languages. The promotion of linguistic diversity in cyberspace should encourage content developments in all underrepresented languages, should promote the use of those languages as working languages, and should assist cyber-communities in

13 UN A/RES/47/135, 18 December 1992.

using them to communicate¹⁴. In other words: as plurilinguism has no self-contained limit, the internet's potential must be put to its fullest use.

Individual Rights and Internet Access: The Freedom of Opinion and Expression and the Freedom of Information

In his 1998 report to the UN Commission on Human Rights, the UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression outlined the case against government regulation of internet access and content:

“The new technologies and, in particular, the internet, are inherently democratic, provide the public and individuals with access to information and sources and enable all to participate actively in the communication process. The Special Rapporteur also believes that action by States to impose excessive regulations on the use of these technologies and, again particularly the internet, on the grounds that control, regulation and denial of access are necessary to preserve the moral fabric and cultural identity of societies is paternalistic. These regulations presume to protect people from themselves and as such, are inherently incompatible with the principles of the worth and dignity of each individual”¹⁵.

Now more than ever, power is in the hands of the informed. Citizens revolt and regimes are overthrown by the flow of information within and between countries; the role of the Internet both as disseminator and as advocate is increasingly significant. Information has become not only more powerful, but also more accessible¹⁶; what was historically the prerogative of the few is now potentially the tool of the many.

The UN Special Rapporteur recently issued a press statement expressing concerns regarding internet-specific legislation adopted by South Korea, in particular the *Framework Act on Telecommunications*, and the *Act on Promotion of Information and Communications Network Utilisation and*

14 M. DIKI-KIDIRI, *Towards real linguistic and cultural diversity in cyberspace*, <http://www.portal.unesco.org>

15 UN Doc. E/CN.4//1998/40IIC4.

16 According to J. Baudrillard (S.F. Glasner), *Simulacra and Simulation*, Ann Arbor: University of Michigan.

Information Protection. The former was used in the arrest of a blogger whose posts criticised the government's economic policy during the financial crisis; the latter was used to delete online posts and to punish individuals who initiated online campaigns for a consumer boycott¹⁷.

To establish a globally open society, states and leaders cannot retain control over information; a self-regulating mechanism must be established and social awareness generated through democratic and open processes.

The right to freedom of expression is recognized in the *Universal Declaration of Human Rights*, Article 19:

“Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless frontiers”

This provision includes the rights to seek and impart information and ideas; to inform and be informed. The right to freedom of expression is similarly protected by Article 19 of the *International Covenant for Civil and Political Rights*, Article 10 of the *European Convention on Human Rights*, and under generally applicable norms of International Human Rights Law. Where web access is hampered, these rights are violated. State censorship is one of the most basic forms of content constraint, but the underrepresentation of minority languages also constitutes a restriction on these rights. As a democratic new media, the internet should be an environment that protects human rights.

Beyond using human rights law to expand participation in new media, the Internet can protect human rights by providing access to information. Improved communication can also contribute to the organization of enforcement activities, and to mobilizing campaigns for change. On the other hand, with greater opportunity increased responsibility: the internet

17 UN Doc. A/65/284, par. 76, 11 August 2010. The report also stresses that emergency or national security laws are often used to justify restrictions on citizen journalists' expression of views or dissemination of information via the internet, often on the basis of protecting vaguely defined national interests or public order. It is especially in this regard that on February 27, 2004, the UN Special Rapporteur, the Chairperson of the Working Group on Arbitrary Detention and the UN Special Rapporteur on Torture, sent an urgent appeal to the Syrian government regarding a person for distributing articles by email. The Syrian authorities were quoted as saying the articles were “*detrimental to the reputation and security of the nation*” and “*full of ideas and views opposed to the system of Government of Syria*”.

is not only a great tool for democratisation, but may also be used for hate and abuse. Nevertheless, freedom of expression must be maintained.

The last issue is that of privacy as a human right. The right to privacy entered the arena of internationally recognised and protected human rights in Article 12 of the *Universal Declaration of Human Rights*, which prohibits arbitrary interference with “privacy, family, home or correspondence”, as well as Article 17 of the *International Covenant on Civil and Political Rights*, and Article 8 of the *European Convention on Human Rights*, which uses slightly different language. The Council of Europe’s *Convention for the Protection of Individuals Regarding Automatic Processing of Personal Data* (1981), also known as *Convention 108*, remains to this day the only international document that is binding on every country including non-members; and many cases heard by the European Court of Human Rights relate to the right to privacy.

Storage of personal data also falls under the rubric of privacy protection; the UN Human Rights Committee has specified that whether information is gathered and held by public authorities or private parties, everyone should have the right to ascertain whether information about them is stored, what specific information is stored and for what purpose (General Comment, Article 17, *Civil and Political Rights Covenant*)¹⁸.

MULTILINGUALISM IN ACTION

A language is considered threatened if it has few users and a weak political status, and when it is no longer transmitted through formal education¹⁹. Languages are disappearing faster than ever before; increasing internet use may contribute to the loss of “minority languages” and thus of cultural/linguistic diversity.

As of December 2009, the top ten languages on the internet were English (495.8 million users), Chinese (407.7 million users), Spanish (139.8 million users), Japanese (96.0 million users), Portuguese (77.6 million users), German (72.3 million users), Arabic (60.3 million users), French

18 See also the judgment of the European Court on Human Rights on the case *Rotaru v. Romania* (May 2000).

19 T. Skutnabb-Kangas, *Linguistic Human Rights in Education: Western Hypocrisy in European and Global Language*, paper presented during the 5th plenary session of the 5th International Congress of Hungarian Studies.

(57.0 million users), Russian (45.3 million users) and Korean (37.5 million users). Equal online linguistic representation is essential given the Internet's increasing role as the primary information source, knowledge repository, communication interface and business medium.

Enabling people to communicate online in their own language and script has great importance in providing diverse populations with a presence in the global information economy²⁰. However, ten years after the adoption of the multilingual architectural model²¹, minority languages and scripts are more threatened than ever. Multilingualism is even less of a force online than it is in the “real world”²². While the internet is no longer monolingual, few of the world's close to 7000 languages have a significant internet presence²³. The internet began as a predominantly monolingual medium in which English was essentially the only language; despite some increase in diversity, English remains its dominant language for correspondence and communication²⁴.

IT has made advances to accommodate less dominant languages and scripts, but even if technical limitations impeding online language diversity have been largely resolved, online presence is still an issue²⁵. The *European Charter for Regional or Minority Languages* states that,

“*regional or minority languages*” means languages that are: 1. *traditionally used within a given territory of a State by nationals of that*

20 See in this book: Michaël Oustinoff, *The Economy of Languages*.

21 The first document that provides a global architecture for online multilingualism dates to 1996 during an IAB (Internet Architecture Board) workshop at the Information Sciences Institute (ISI) in Marina del Rey, California.

22 It is worth mentioning the three categories of languages in cyberspace as defined by M. Diki-Kidiri: (i) *working languages* used to communicate; (ii) *languages objects* which are only mentioned or studied; and (iii) *absent languages* that are never used nor mentioned on internet.

23 See in this book: Daniel Prado, *Language Presence in the Real World and Cyberspace*.

24 P. Lacour, A. Freitas, A. Bénl, F. Eyraud, D. Zambon, *Translation and the New Digital Commons*, <http://odel.irevues.inist.fr/tralogy/index.php?id=150>

25 WSIS-05/TUNIS/DOC/7-E, November 18, 2005, Article 14, “*recognize[s] that in addition to building ICT infrastructure, there should be adequate emphasis on developing human capacity and creating ICT applications and digital content in local language, where appropriate, so as to ensure a comprehensive approach to building a global Information Society*”. In Article 32, states commit “*to promote the inclusion of all peoples in the Information Society through the development and use of local and/or indigenous languages in ICTs*” and “[*to continue [their] efforts to protect and promote cultural diversity, as well as cultural identities, within the Information Society*”.

State who form a group numerically smaller than the rest of the State's population; 2. and different from the official language(s) of that State.

In a few cases, these languages are afforded some form of state support by domestic legislation or national constitution²⁶. But if official and minority languages cohabit in certain states, online cohabitation poses a greater challenge given the fact that a few “big” languages and English in particular hold a dominant position, that transforms most languages into minorities.

However, not all underrepresented languages may be considered minorities. First, no official internet language has yet been identified or legally recognised. Second, categories of minority languages must be established; it would be misleading, for example, to place Aramaic, a language spoken by 2 million people, on the same level as Wu, spoken by over 77 million Chinese, even if they have similar online presence.

The Consequences of a Restricted Set of Characters on the Web

Online multilingualism is important for: 1. economic efficiency and productivity; 2. the establishment of a democratic society; 3. the promotion of education; and 4. the respect and promotion of human rights²⁷. It is therefore an essential means for collapsing the “linguistic divide” that accompanies the digital divide.

Accepting the dominance of certain languages on the Web has far-reaching consequences, not only for other languages and their marginalised speakers, but for Internet development itself. Attaining a critical mass of content in a given language is an essential precondition for attracting Internet users; inadequate online linguistic representation inhibits many from using the Internet even if physical access is available. As much of the world's communication and business migrates to the Internet, ensuring straightforward access for the next billions of users is critical.

The linguistic divide is now an unavoidable disadvantage for much of the world's population. Unconvincing shortcuts, like the use of Esperanto or

26 See in this book: Evgeny Kuzmin, *Linguistic Policies to Counter Languages Marginalization*.

27 Internet Society, *Multilingualism and the Internet* (Briefing Paper), May 14, 2009. http://www.isoc.org/internet/issues/docs/multilingual-internet-issues_20080408.pdf

Latin as global *lingua franca*, are unrealistic solutions that could further complicate the issues.

Increasing multilingualism and promoting cultural diversity online is thus a multifaceted problem requiring a multiplicity of considerations to attain the objective of making the internet accessible to all.

PROMOTION AND PROTECTION OF MULTILINGUALISM IN THE “REAL WORLD”

Any serious exploration of the right to multilingualism online necessitates asking whether multilingualism is considered a human right outside of the virtual world – otherwise put, in the “real word”.

Due to political sensitivities that are interwoven into power structures, linguistic and human rights are seldom presented together for discussion.

Still, a look at the abundant production of relevant legal texts since the second half of the twentieth century²⁸ shows that most States recognize the rights of linguistic minorities and accept the charge of protecting the right of persons to their own language, a trend that largely stems from the widespread belief that guaranteeing linguistic particularism assures a population’s linguistic and cultural identity.

Promotion and Protection of Linguistic Diversity at the International Level

While domestic legislation has tended to tolerate (at best) or disregard (at worst) cultural diversity, international law and International Human Right Laws have evolved towards inclusivity. International Human Rights Law plays an important role in setting standards for linguistic rights, as well as providing a normative framework for developing principles of

28 As pointed out by F. De Varennes in *To Speak or Not to Speak. The Right of Persons Belonging to Linguistic Minorities*, a working paper prepared for the UN Sub-Committee on the Rights of Minorities, March 21, 1997, <http://www.unesco.org>, international treaties and conventions with provisions related to the use of minority languages have been adopted since the 16th century. The 1516 Treaty of Perpetual Union between the King of France and the Helvetic State contained a provision identifying those who would receive certain benefits, such as the “Swiss who speak no language but German”.

democratic governance and multicultural policies aimed at managing potential ethno-linguistic conflict.

The *Universal Declaration of Human Rights*, Article 2(1), provides that “Everyone is entitled to all rights and freedoms set forth in this Declaration, without distinction of any kind, such as [...] language”. The *International Covenant on Civil and Political Rights*, Article 27, contains the most far-reaching binding protection for linguistic minority rights, stating that “In those States in which [...] linguistic minorities exist, persons belonging to such minorities shall not be denied the rights, in community with the other members of their group, [...] to use their own language”. Article 2(2) of the *International Covenant on Economic, Social and Cultural Rights* puts language on par with race, color, sex, and religion.

Within the framework of the Council of Europe, the European Union and the Organization for Security and Cooperation in Europe, the rights of persons belonging to regional and minority language groups have been addressed in multilateral treaties and conventions.

Both the Council of Europe’s *European Charter for Regional or Minority Languages* (1998) and the *Framework Convention for the Protection of National Minorities* (1998) contain monitoring bodies that seem to be working effectively at their attempts to stretch states’ willingness to follow more than minimal requirements.

The non-binding UN *Declaration on the Rights of Persons Belonging to National or Ethnic, Religious and Linguistic Minorities*, the *Universal Declaration of Linguistic Rights*, and other Unesco legal instruments like the *Unesco Declaration on Cultural Diversity*, are also relevant here, along with the *Draft Recommendation on the Promotion and Use of Multilingualism and Universal Access to Cyberspace* (2001), which for the first time promoted language preservation and diversity through access to electronic services and resources.

Since 1950, the rights of regional and minority language groups have been addressed at the European level. The *European Convention for the Protection of Human Rights and Fundamental Freedoms* was followed by other multilateral treaties, conventions, and a consistent corpus of soft law comparable to the one codified by the UN, including the *Charter of Fundamental Rights of the European Union* (2000), the *European Charter*

for *Regional of Minority Languages* (1992), and the above cited *Framework Convention for the Protection of National Minorities* (1994).

The EU's ratification of the *Unesco Convention on Cultural Diversity* (2001) provides an eloquent response to critics who allege that the EU wishes to erase national or regional characteristics to impose uniformity. Language diversity lies at the heart of the EU.

Promotion and Protection of Linguistic Diversity at the National Level

At the national level, linguistic diversity is increasingly considered a protected human right. To better understand the opportunities and challenges of multilingualism in a multicultural society, South Africa provides a fascinating example. Its 1996 post-Apartheid constitution is probably more generous to multilingualism than any other constitution in the world, granting official status to the country's eleven indigenous languages²⁹. Article 6 states that:

4. The national government and provincial governments by legislative and other measures, must regulate and monitor their use of official languages. Without detracting from the provisions of subsection (2), all official languages must enjoy parity of esteem and must be treated equitably. 5. A Pan South African Language Board established by national legislation must: (a) promote and create conditions for the development and use of (i) official languages; (ii) the Khoi, Nama and San languages; and (iii) sign language; and (b) promote and ensure respect for (i) all languages commonly used.

Other measures adopted by different countries include the following:

Under Austria's Ethnic Groups Act, the federal administration is obligated to promote measures that ensure linguistic minorities' identities.

In the Philippines, the Office of the Northern Cultural Communities and Office of the Southern Cultural Communities have been created to promote and protect the rights of persons belonging to linguistic and

29 During Apartheid (1948–1994), English and Afrikaans were the only officially recognized languages despite the country's wide variety of other spoken/studied languages. The myth of South Africa as a bilingual English-Afrikaans country persisted for years.

other minorities. Similar entities exist in many countries, including Russia, Canada, Australia, China, and India.

India's constitution includes broad recognition of minority rights to preserve their language, script and culture (Article 29), the right to establish and administer educational institutions in the language of their choice (Article 30), and the right for a minority language to be officially recognised and used by public authorities wherever the minority represents a substantial proportion of the population (Article 347).

Hungary's constitution contains a number of provisions to guarantee minority rights. The *Law of July 7, 1993*, includes respect for the human rights of members of minorities.

Elsewhere in the EU, the neighbouring Italy and France are examples of two opposite, or at least substantially different policies concerning minority languages protection. Italy is much more committed to safeguarding multilingualism; France, has a more advance position on online multilingualism that aims to defend "francophonie", but does not give legal status to minority languages, which it considers *langues régionales*.

In France several legislative provisions and regulations define the role of language for culture, education and the media: the *Deixonne Law of 1951*, the *Haby Law of 1975* and the *Law of August 4, 1994 (Loi Toubon)*. The latter specifies that, "[T]he provisions of the present law apply without prejudice to the legislation and regulations relative to regional languages in France and are not against their use" (Article 21). A constitutional law adopted on July 23, 2008, provides that "regional languages belong to France's heritage" (Article 75-1). The *European Charter for Regional or Minority Languages* was signed by the French Government but has not yet been ratified.

Italy meanwhile has always privileged the preservation of cultural diversity within its territory. Although Italian is the country's official language, Italian legislation (laws 482/1999 and 38/2001; the effective decree of the President of the Republic 345/2001) and Article 6 of the Italian constitution valorise non-dominant languages³⁰. *Law 482/1999* declares that these languages and cultures can be taught in schools, that official documents

30 According to the law, the following languages and cultures are preserved and promoted: Albanian, Catalan, Croatian, French, Franco-provençal, Friulian, German, Greek, Ladin, Occitan, Slovene, and Sardinian (total 2,428,770 speakers).

and acts should be bilingual, and that a local language may be used for territorial broadcasting information. However, immigrant languages like Arabic and Chinese, are not taken into account.

Some of Italy's protected languages are more commonly spoken than written; bilingual web sites are uncommon. The exceptions are borderland populations that speak Italian as second language, as with the German-speaking Trentino-Alto Adige region and the French speaking Val d'Aosta region, whose official websites are frequently bi- or multilingual, especially those of public bodies that are legally required to be bilingual. To date, only four of Italy's linguistic minorities have programmes on the national public broadcaster RAI: the French speakers of the Aosta Valley, the German speakers of South Tyrol, the Ladin speakers in the Dolomites and the Slovenian speakers of Trieste³¹.

As for public media, there exists some positive examples of measures protecting linguistic minorities. Although not necessarily inspired by the *International Covenant on Civil and Political Rights* (Article 27), they use similar language.

The *European Charter for Regional or Minority Languages*, Article 11(1) (a), generally addresses state conduct in relation to public media and minority languages. Many states explicitly recognise that the needs of linguistic minorities are not satisfied by the exclusive use of an official/majority language in public media. In these cases, the degree of minority language use in public media adequately reflects their respective linguistic populations' demographic weight, needs and interests.

ONLINE MULTILINGUALISM

To help bridge the digital divide, multilingualism must be fostered online. The UN stands at the forefront of linguistic diversity promotion towards universal access to political, economic and cultural goods³². The main concept being promoted is that all have the right to fully inhabit the web

31 For more information, see the Minerva Plus Survey, *Final Plan for using and disseminating knowledge and raise public participation and awareness report on inventories and multilingualism issues: Multilingualism and Thesaurus*, <http://mek.oszk.hu/minerva/survey/wp3multilingua.doc>

32 *Top UN official stresses need for Internet multilingualism to bridge digital divide*, December 14, 2009, UN News service, <http://www.un.org>

without limitations or obstacles. Internet as a common or semi-common good is thus internationally recognised.

An analysis conducted by the Internet Society shows that the reasons behind the slow progress of online linguistic diversity may be classified into technical, economic, social and political categories.

Technical Reasons. Although IT has the potential to catalyze a massive increase in the free flow of information, it has also strengthened existing inequalities, particularly due to the predominance of Latin script. The digital divide that stems from this trend evince effects on three levels: (I) the impossibility of using IT due to the lack of knowledge of code languages; (II) the impossibility of adapting technology to local contexts; (III) the difficulty for users with lower education levels to use technology. Technology is not neutral, and IT is largely dominated by the societies that created it, resulting in a predominance of Occidental/Latin language tools. For these technologies to become culturally global, they must be reconceived to be conducive to other languages.

Standards used for hardware and software implementation must also shift towards inclusivity. All new technical standards must be subjected to a rigorous evaluation of cultural impact to assess their impartiality towards linguistic and social groups.

Promoting open source and free software could be another efficient way to counterbalance the problem's technical elements. Since third parties may directly access and modify code, they are excellent vehicles to localize programs³³.

These challenges are much less significant now than they were a decade ago. The remaining technical challenges concern standards, tools and technical capacity. Most internet standards are evolving but must go further. Encoding problems, the most significant impediments to using diverse scripts a decade ago, have been partly solved by the introduction of Unicode³⁴, but this approach is partial and insufficient. Moreover, in many countries, especially in developing nations, standardisation has been slow. As a result, certain standards, like keyboard layouts, continue

33 See in this book: Dwayne Bailey, *Software Localization: Open Source as a Major Tool for Digital Multilingualism*.

34 See in this book: Stéphane Bortzmeyer, *Multilingualism and the Internet's Standardisation*.

to hamper multilingualism. As for tools, very few software tools are truly multilingual.

The biggest challenge is in the area of content development. As stated above, unless sufficient content exists in a given language, there is little incentive for its speakers to use the internet. Furthermore, most organizations prefer to publish information in so-called international languages and rarely develop content in local languages.

POLITICAL, ECONOMIC AND SOCIAL CONSTRAINTS.

Insufficient resources are another likely reason for the underrepresentation of the languages of developing nations. Low demand creates low economic incentive for software developers to produce tools in these nations' languages, even if those languages are spoken by tens of millions of people. Local governments and other actors also lack funding to support activities targeting the development of online multilingualism. Platforms for mobile applications are also extremely limited.

International decision-makers have long been aware of this problem. In 2002, former UN Secretary-General Kofi Annan challenged Silicon Valley to unleash its creative energies to close the digital divide. The UN is in a unique position to harness energy in the private sector and in governmental and non-governmental organizations. Defining clear goals, roles, resources and cooperation is paramount to bringing these different organizational cultures together.

MULTILINGUALISM, CYBERSPACE'S NEXT FRONTIER

What are the next steps towards a multilingual cyberspace?

First, given the transversal character of multilingualism, an interdisciplinary approach is crucial.

One of the first areas for development should be establishing a common practice for international, public and private law. This means identifying, granting and protecting new rights relevant to the information society.

Academics and legal practitioners should develop alternative arbitration and mediation mechanisms, taking linguistic diversity into account in light of the *Universal Dispute Resolution Policy* set up by the World Intellectual Property Organization for the resolution of domain name disputes. These principles should be developed to improve multilingualism in technological applications, including web crawlers, search engines, database indexes and cloud computing³⁵.

If we assume as evident the trend of domestic and international law toward exerting positive obligations on states to protect and enhance online multilingualism as a human right related to other human rights, then we admit the obligation to create online content in minority languages.

Nevertheless, recognition of the internet as an object of domestic and international law with specific rights and obligations is still unclear. Questions remain, like who must fulfill obligations related to internet multilingualism, and what authorities are to judge or sanction violations of online rights.

It appears more urgent than ever that the question of language assumes its natural place at the heart of the online evolution. This requires first and foremost the establishment of principles, instruments, and legal norms, as well as maintaining online linguistic diversity as a major issue on the international agenda.

Note of the author:

I wish to express my deep gratitude to Mr. Richard Delmas for giving me the opportunity to write about such a challenging subject as multilingualism, the Internet and Human Rights. Without his support and substantial contribution to the drafting process, this article wouldn't be published. Likewise I am grateful to Mr. Louis Pouzin who offered a precious assistance during the drafting process. Last, but not least, a big "merci" to my husband Erwan Marteil, who inspired much of the content in this article with his ability to identify and probe important issues with a unmistakably sharp critical thinking.

35 For further details <http://www.wipo.int/amc/en> and <http://www.wipo.int/amc/en/domains>

See in this book: Pann Yu Mon & Madhukara Phatak, *Search Engines and Asian Languages*.

STÉPHANE BORTZMEYER

MULTILINGUALISM AND INTERNET GOVERNANCE

Who decides on multilingualism on the internet? In which office do we authorise domain names in Chinese or Arabic? Who wanted the majority of web pages to be in English? This article takes you through the corridors of Internet governance.

Original article in French.
Translated by John Rosbottom.



STÉPHANE BORTZMEYER is a computer engineer, particularly specializing in TCP/IP networks. He works for AFNIC and maintains a blog where he talks from his own personal view mainly about technology but sometimes also of culture or politics. <http://www.bortzmeier.org>

STÉPHANE BORTZMEYER

MULTILINGUALISM
AND INTERNET
GOVERNANCE

One of the internet's quintessential features is its organization : as a multi-national and especially a multi-stakeholders network, the internet has no centre. There is no President of the Internet, no Council of Wise Internetizens who can make binding decisions. Hence, changes that are subject to broad consensus, like the migration to the IPv6 network protocol, have been delayed considerably because changes cannot be executed from the top down. No authority can say, *“On 30 April 2011, all connected networks must be IPv6”*, or, *“On 15 May 2011, for security reasons, only reliable secure operating systems may be connected to the internet”*. Migration depends on a critical mass of individual decisions and, as with economics and ecology, the accumulation of individual decisions is often very inefficient and does not allow changes that would benefit all, but necessitates that each player make an effort or incur expenses.

WHO ARE THE DECISION MAKERS FOR MULTILINGUALISM ON THE INTERNET ?

This lack of Centre or General Management doesn't mean the internet is a perfect anarchy, completely absent of power dynamics. Rather, there exists several power centres that possess variable legitimacy. Furthermore, there is no equality between these centres ; some are much more powerful than others ; obviously Google, Microsoft, the U.S. government, Apple, Baidu, Level 3, Cisco, France Telecom or Facebook hold more weight than Mr. Michu, M. Ali, M. Li, a small web design agency based in Charleville-Mezieres, or the Government of Mali.

This situation has its advantages ; first and foremost, it prevents a single group from taking over the internet and imposing its will. Certainly, such a decision-making mechanism could be useful in some cases (as

in the case of migration to IPv6 cited above). But there also exist huge risks if a hypothetical Centre or Director made incorrect decisions. If the internet was run by a group resembling the current French government, for example, the Network of networks would soon be sterilised, its role reduced to that of audio-visual content distribution, rendered innocuous to the powers that be. It is thus for the best that this situation continues not to present itself, even if every internet user/developer has probably, at least once in her life, wished that a seemingly crucial decision could finally be taken, once and for all!

In the case of multilingualism, the absence of a centre has often been felt: for example, in the difficulty of getting the 85 % of software properly internationalised to move to 100 %, since no one can force programmers to do their work correctly, or compel system administrators to deploy the relevant software.

Many organizations manage sections of the internet, small parts of a gigantic system. Obviously in the context of this chapter, there is no way to cite them all, but we can at least provide an overview of their categories:

- Network operators and service providers such as Tata or Comcast. Their role in the multilingualism question is rather low;
- Software producers such as the Mozilla Foundation, Apple, Wordpress developers, and so on, as well as managers of large and widely used internet services (search engines, platforms, blog hosting, etc.). Their role is crucial because software that would only allow, for example, text production and publication using the Latin alphabet, would seriously limit multilingualism;
- National governments; various official regulators. The obligations they put forth may play a role in the development (or lack thereof) of multilingualism. They can also play a role of incentive and encouragement;
- Standards bodies and technical standards are addressed in the chapter of this book devoted to normalisation¹;
- Let us not forget, of course, the users. Unlike many other communications systems, such as television, the internet is largely created by its users. It is they who produce most of the internet's multilingual

1 See in this book: Stéphane Bortzmeyer, *Multilingualism and the Internet's Standardisation*.

content, the classic exemplar being the user-produced encyclopedia, Wikipedia.

THE ROLE OF ICANN

When someone searches despairingly for a Centre of the Internet, the search generally indicates ICANN (Internet Corporation for Assigned Names and Numbers). As an international organization in charge (at least ostensibly) of a unique resource – the root of the domain names system – it seems well-placed for this task. This reassures anyone worried about the internet’s lack of central management, and as a result, ICANN is often a stakeholder in political struggles, while much more vital areas of internet governance, like management of IP addresses, arouse much less interest.

However, ICANN is in no way “the internet regulator” (a common journalistic charge, but nevertheless quite false). It is not even “the DNS regulator”, as is shown in the following chapter on domain name registries. Let us now wring the neck of myths:

- ICANN does not manage domain name root. This work, considered strategic, remains the sole property of the U.S. government², through its technical arm, the company Verisign;
- *A fortiori*, ICANN does not manage the rest of the Domain Name System (DNS) (standards are set by IETF; policies are decided by each registry), and even less so other online activity. ICANN has a certain role (at least in theory) in the allocation of IP addresses;
- The only real power of ICANN is in managing the American Top Level Domains (TLD) .com and .org. This is the only area where the term “regulator” is appropriate.

It follows that ICANN’s actual role in promoting or suppressing multilingualism is actually quite limited. The most oft cited case is the introduction of Internationalised Domain Names (IDN) in the root (effective since 5 May 2010). But the decision was taken by the U.S. government, as sole manager of the root. ICANN acted solely in the roles of retardant (through long and utterly unnecessary studies³) and in charge of public

² <http://www.ntia.doc.gov/ntiahome/domainname>

³ These studies were unnecessary because, technically, a name server does not differentiate between an IDN – a name in Unicode – and a traditional name. Due to the use of

communication. Long before IDN's official incorporation, the technical element had been settled, and several registries had already introduced it.

Today, ICANN retains the role of considering applications for new TLD roots, from which stance it has rejected applications from Bulgaria⁴ and Greece⁵, citing “confusion risk” – in the case of Greece, with a TLD that doesn't even exist!

DOMAIN NAME REGISTRIES

But if ICANN doesn't decide IDN deployment, who will? As often happens online, the decision is widely distributed. Consider the example of `.fr` (France's TLD). At the time of writing, IDNs are not yet available (but are expected by summer 2012). Who decides? The answer is not entirely simple. France has a law regarding domain name management, but it doesn't specify a complete registration policy. Overall, the role comes back to the AFNIC (Association française pour le nommage Internet en coopération). They kicked off the debate in 2003⁶, with no immediate result. During a public consultation on `.fr`'s management in 2008, no discussion was recorded on the issue of IDNs (though many actors in France's internet governance liked to criticise ICANN for being slow in deploying IDN). It was only in 2010 that a report of a semi-official organization, the FDI (Forum des Droits sur l'Internet), called for an acceptance of the IDN `.fr`⁷.

In the same way, each TLD registry decides if and when to open up to IDNs. Canada's TLD (`.ca`) doesn't have IDN. In Europe, the German TLD (`.de`, with special reference to the famous character ß⁸), Switzerland

the Punycode algorithm, the two are in ASCII in the name server memory. See <http://www.icann.org/en/announcements/announcement-28oct07.htm> for an example of one such study, whose main purpose was to provide a justification for delayed deployment of IDN TLDs.

4 <http://www.centri.org/main/6079-CTR/version/default/part/AttachmentData/data/Daniel%20Kalchev%20-%20bgidn20110202v3.pdf>

5 <http://svsf40.icann.org/meetings/siliconvalley2011/presentation-swords-confusing-gr-segedakis-15mar11-en.pdf>

6 <http://www.afnic.fr/fr/l-afnic-en-bref/actualites/actualites-generales/2511/show/avertissement-de-l-afnic-sur-les-noms-de-domaine-internationalises-idn.html>

7 <http://www.foruminternet.org/institution/espace-presse/communiqués-de-presse/la-langue-et-internet-le-forum-des-droits-sur-l-internet-publie-une-etude-inedite-2984.html>

8 <http://www.denic.de/en/faqs-about-idns-ss.html>

(.ch) and Austria (.at) were among the first to have one (the three German-speaking countries launched together on 1 March 2004). On the other hand, the TLD of countries using non-Latin scripts are often much faster to register IDNs, as was the case in China (.cn, which didn't even wait for the official release of the technical standard). Countries using a right-to-left graphic system, such as Saudi Arabia (.sa), often wait for long periods of time to have a domain name (TLD included) entirely in Unicode.

PROGRAMME INTERNATIONALISATION AND LOCALIZATION

A user interacts with other users by way of software that, in the vast majority of cases, she hasn't written. These programmes' capabilities or omissions therefore directly influence the multilingual experience, and it is no exaggeration to say that software vendors have had a more direct influence on this issue than most official organizations.

Even as of early 2011, the Android operating system⁹ was not yet managing Arabic script or its right-to-left writing system. Other programmes, even those developed by the French¹⁰, are still struggling to integrate accented characters (and display error messages like “unsafe characters” in response to them).

Why these weaknesses? First off, managing all the world's scripts clearly creates more work. If we remade all the world's languages to make life easier for IT professionals, it certainly would be limited to ASCII. But there is also a very clear lack of awareness among programmers. In the vast majority of programming courses, Unicode is not mentioned, or is mentioned given a couple of hours of lip service towards the end of the year in a class presented as “extra” – when, in fact, Unicode string processing should be part of core coursework.

Because of this, programmers have great practical importance in the context of multilingualism. But who writes and who decides to handle (or not) all the possible scripts? Many entities are developing programmes, to summarize:

9 An operating system for smartphones, developed by Google.

10 When the author of this text begins the day in his company, his badge reads “Stéphane Bortzmeyer”, without accent, or even an e, in his firstname.

- Large commercial companies, such as Microsoft, Google, Apple, Oracle or IBM, producing widely used products that are developed for profit. Most of the planet’s languages are spoken by minorities with low purchasing power ;
- The developers of free software¹¹ often isolated individuals, but also include some of the big companies mentioned above (Google and IBM, notably, are important players in free software development). The reasons may vary. *A priori*, source code availability and the freedom to improve it, regardless of the business strategies and priorities of individual companies, makes it easier to take multilingualism into account in the software. However, this point is offset by the fact that people who use less “central” languages have no more technical expertise than they do money. And the isolated developer of free software does not necessarily possess the means to know all the details of multilingual management. An effort to raise support and awareness for free software would certainly be helpful ;
- Finally, there is also a cloud of locally-developed software, created by software houses, students, or small service companies. The authors of this software are rarely competent in multilingualism¹². They lack the level of access to resources possessed by a company like Microsoft. Their software is not freely distributed, and cannot be improved by anyone but the author. Such software is, therefore, often the bane of multilingualism promoters.

CONTENT PRODUCTION : NOT ONLY IN ENGLISH

The previous section dealt with the role of programmers and software companies. If they worked effectively, and if all the world’s software was fully internationalised and ready to handle any language and writing system, would all be well? No. Content authors must use and integrate this software. If instant messaging software enables the exchange of texts in Urdu but correspondents prefer to use English, multilingualism will suffer. Likewise, if a Content Management System (CMS) allows for writing and publishing texts in all languages, this doesn’t mean the speakers of all

11 See in this book : Dwayne bailey, *Software Localization : Open Source as a Major Tool for Digital Multilingualism*.

12 This is undoubtedly so in the case of the badge reader cited above.

languages will use them. Many obstacles present themselves because of constraints of time and money, the momentum of habit, resignation to the *status quo*, lack of internet culture and/or expertise, a socio-political culture that encourages passivity rather than creativity, and so on.

This non-use phenomenon of existing technologies is even more pronounced because many institutional stakeholders prefer discourse to action: at a conference on multilingualism, speakers line up to deplore the lack of online content for language X or Y. The time and money could have been better spent producing some content. To cite just one example among many, the round table organised by the French Senate on 26 January 2011 on the theme “*The Creation of Cultural Content in the Digital World*” did not include a single content creator. Even the Wikimedia Foundation was not invited. Leafing through a recent brochure on the defence of French in a Paris bookshop, an author strongly criticised the dominance of English and lack of French content in online media. Out of curiosity, I searched the aforementioned network for what the author had done to fight against this trend: nothing. No content from him (not even his brochure) was available online.

Let’s try not to follow the sad example of the Senate’s round table and consider the projects that are successful. The leader is evidently Wikipedia, which represents one of online multilingualism’s greatest successes. As of May 2011, the collaborative encyclopedia counted 269 active editions¹³, including in languages like Alemannic, Uzbek, Kurdish and Dhivehi. No other website in the world makes so many languages available (and they are not translations, as is often seen on commercial sites, but often original content). In this context, proper management of Unicode in the underlying software has been used: the speakers of many languages have the opportunity to *defend* and *illustrate* their language by practising it. Since contributing to Wikipedia does not require significant technical skills, the barrier is relatively low. Everything does not necessarily proceed in joy and good humour (English-speaking Wikipedia contributors argue regularly about whether to write *color* or *colour*), but this is no different from any other human project.

13 A list is available at http://en.wikipedia.org/wiki/List_of_Wikipedias

Another good example is the Debian website¹⁴, a free software based on a dynamic, global community that ensures the site's availability in over twenty languages.

LANGUAGES IN GOVERNANCE

One of the problematic aspects of online multilingualism is the preponderance of English. Indeed, whether among teams of multinational programmers¹⁵, involving free or private software, whether in standards organizations or in governance, English is effectively the only working language. Certain presentations, like a conference's opening session, are translated. But the real work is done in English.

Is there an alternative to monolingualism that avoids the incredible burden and cost of inefficient organizations like the European Union? As of right now, unfortunately, it seems not.

A LITTLE BIT OF HISTORY

The story of the internet can help in understanding the situation of its governance. But beware! Like any human organization, the internet also has its origin myths, which are no more realistic than other creation stories. So. At the very beginning, the network that would become the internet – ARPANET – was highly centralised. It lacked one essential characteristic of the internet: its multi-organizational character. At the time of the ARPANET, it was still possible for one authority to select a D-day when all the machines had to adopt a certain technology. The last of these D-days took place in 1983, with the introduction of the family of protocols TCP/IP (IPv4 and TCP in its current state).

The multi-agency internet was born at about this time. At this time, all the governance structures, somewhat formalised, were present in one man, Jon Postel, who performed the role of de facto benevolent dictator. The growth of the internet, its qualitative changes, gradually rendered this

¹⁴ <http://www.debian.org>

¹⁵ See, for example, an interesting discussion on the language to be used in the programmes <http://stackoverflow.com/questions/1046419/variable-naming-and-team-members-who-speak-another-language> As one participant in the discussion notes "Something's Gotta Give".

situation untenable. The U.S. government, which funded almost all of the internet, then launched a showdown with those responsible for internet management. This dispute, which was ideologically represented in two documents, the *Green Book* and the *White Book*, peaked in January 1998 with the attempt by Postel to return control of the root servers to the collective IANA (Internet Assigned Numbers Authority), an attempt that failed, leading to the strict control of the root by the U.S. Government.

This control materialised through the creation, in the same year, of ICANN (an organization that is therefore relatively new to the internet), which then took over the core tasks from Jon Postel. But, as noted above, these tasks represent only part of the internet; even in the time of Postel, it was not he who decided, for example, if the Web would be deployed (the Web really began around 1991, without any involvement of a central body).

On questions of standardisation, the IETF, in its existing form, dates from 1986 and sets standards (RFCs) that actors may choose to follow or not. But it has never had any real deciding power.

We can see that most actors and organizations involved are American. This helps to explain the historical dominance of English and its script, the Latin alphabet. Today, though technicalities are ready, it is above all the dominant role of the United States in the world that helps perpetuate the preponderance of a particular language.

GLOSSARY

Domain Name System (DNS)

This term refers to both the system of domain names, the tree structure for creating identifiers such as `cooptel.qc.ca` or `véliplanchiste.com`, and protocol permitting the retrieval of information like the IP address, mail server name, etc., by way of such a name.

Internationalized Domain Names (IDN)

The term IDN denotes domain names expressed in Unicode, for example the tunisian TDL, `سنوت.تيرح`. Sometimes the acronym IDNA (*Internationalized Domain Names in Applications*) is used for the specific technology that is currently being used, which passes through a local conversion to ASCII before sending to the DNS.

As domain names are highly visible identity markers that are widely used for communication, it is crucial to be able to express them in their original script. This explains why the question of IDNs has been so hotly contested.

Internet Corporation for Assigned Names and Numbers (ICANN)

A private American organization, founded in 1998 by the Clinton administration to perform the functions previously assigned informally to Jon Postel. The management of the root finally escaped ICANN (it is managed now via direct contract between the U.S. government and Verisign). ICANN manages the IANA function (records other than the root), and serves as a regulator of certain TLDs, notably .com <http://www.icann.org>

Internet Engineering Task Force (IETF)

The internet's main standards organization, notably responsible for layers 3 (for routing) to 7 (for applications). It is known for its openness and debates, and for making its standards (the famous RFCs) publicly available. <http://www.ietf.org>

Top-Level Domain (TLD)

The domain at the head or root of a domain name – the part furthest to the right. In the name `google.com`, for example, the TLD is `.com`.

BIBLIOGRAPHY

[MILTON 2002] Milton Mueller. *Ruling the root. Internet governance and the taming of cyberspace*. MIT press. 2002.

ETHICAL PRINCIPLES REQUIRED FOR AN EQUITABLE LANGUAGE PRESENCE IN THE INFORMATION SOCIETY

There is a de facto inequality among the world's languages, although they are equal in dignity and rights. This inequality is the result of different living conditions and often inequalities created by history. The 70 most common languages, and the most spoken in the world, represent only 1% of all languages. It is therefore fair to assume that the information and knowledge available in these 70 languages, should be within reach of the majority of the world population. Should we, however, leave out the 360 million men and women who represent 6% of the world's population who do not speak any of these "big" languages?

Original article in French.
Translated by Laura Kraftowitz.



MARCEL DIKI-KIDIRI, Central African Republic, is now Consultant in Applied Linguistics. Before he retired in 2010, he was senior researcher at the CNRS in the unit Language, Languages and Black African Cultures (LLACAN : CNRS, INALCO).

MARCEL DIKI-KIDIRI

ETHICAL
PRINCIPLES
REQUIRED FOR
AN EQUITABLE
LANGUAGE
PRESENCE IN THE
INFORMATION
SOCIETY

Social life can only be harmonious and beneficial to everyone if it is organized so that conventional limits circumscribe individual freedoms. The codification of these limitations includes both individual and collective behaviour, in terms of both rights and duties, prohibitions and obligations, good and bad. The dividing line, inscribed in the conscience of each person, lets us know when we are acting honestly or in bad faith, even if it means facing the consequences. Indeed, ethics is a matter of conscience more than of law. Laws may vary in space and time. There even exist bad laws that assault and outrage the consciences of all who have the ability to distinguish right from wrong, and the courage to denounce evil, sometimes even risking their lives to do so. Wherever laws are not based on ethical values, the strong grab what they can, without hesitating to codify their immoral practices into law. Such actors can only be limited by others as strong as them. Peace is precarious: anytime someone feels strong enough to impose their will onto others, they don't hesitate to start a war. The weakest are poor and enslaved without any recourse except the awakening of the international conscience of all people of good will, who are courageous enough to express their disgust and to revolt against the unjust and intolerable situations created by potentates. The establishment of laws based on the common good, and on equal rights for all, is an act of civilisation that recognizes and protects the fundamental and inalienable rights of individuals and nations.

Respect for these rights is the primary duty of us all. In light of this principle, the following ethical principles are required for equitable language presence in the information society.

All the world's languages have equal rights and dignity, as do those who speak them

In the information society, access to information and thus to shared knowledge is a fundamental right, as unequal access creates unacceptable and cascading inequalities in the human condition. But for everyone to enjoy this basic right, all public information and accumulated knowledge of the entire human race must be made available in every world language and reach every linguistic community. Considering that world languages number around 6800-7000, this may sound utopian. But it responds to an essential first principle, that *all the world's languages have equal rights and dignity* and can be manipulated to express the whole planet's knowledge, from the moment the demand is there. To accept the inequality of languages as a natural fact is not far from assuming inherent inequality between their speakers. This constitutes a form of discrimination and is wholly unethical.

Each language is a treasure of humanity

There are certainly *de facto* inequalities between the world's languages despite their equal rights and dignity. But these inequalities result from the different and often unequal living conditions historically created by conflict between certain communities and the isolation of others. We now know that the 70 most widely spoken languages account for only 1 % of those in the world. For 94 % of the world's population, these languages are spoken not only as first languages, but also as second, third, or fourth. It is thus reasonable to consider that if information and knowledge were available in these 70 languages, it could be accessed by the majority of the world. This would represent a huge step towards a worldwide multilingualism reaching the vast majority of people. Must we, however, leave out the 420 million women and men, 6% of the world, who do not speak any of these "major" languages? Of course not. Accounting for this segment of the global population is of great importance since that 6 % speaks 99 % of the world's languages, each language its own coding system of thought, knowledge and human experience. It's not out of condescending generosity towards minorities that we must take care to preserve undervalued, marginalized, threatened, or endangered languages, but because these languages are *treasures of humanity*, each one yielding

rich and complex knowledge about the functioning of the human spirit, each a unique system for codifying knowledge, and a unique vision of the world. Each language that is protected from extinction in turn protects all humanity against ignorance and amnesia.

The mother tongue has inalienable identity value

Respecting a mother tongue is part of respecting its native speakers. The truth of this is such that wherever peoples have been dominated, rulers have systematically disregarded, marginalized and devalued their languages. “*These people do not talk, they make guttural sounds*”, or at best, “*They have no language, only a dialect*”¹. In the context of such contempt, erected in the place of objective truth, the rulers, in the spirit of “civilisation”, impose their own language. From Dhaka to Soweto, subjugated peoples have paid with their lives for the right to speak their own languages. And even in extreme cases, when people have been deported, transplanted, uprooted, and enslaved, forced to lose their language over the course of their tragic history, they have managed to create, on the sites of their *transplantation*, new roots and new identity, with a new language, a Creole built from the ruler’s imposed linguistic material. The link that binds humans to their mother tongue is as strong as that which attaches them to their home country. “*My language is the house in which I live, yours is to me like the window that allows me to look outward*”². And it is clear that no one has the right to deprive another of their language.

The mother tongue is not only important for individuals, but also for linguistic communities. Language is a powerful identity marker for human communities, regardless of the type of social organization: clans, tribes,

1 It is fundamental to remember that “*the term ‘barbarian’ was applied by Greeks in ancient history to anyone who didn’t speak their language*” <http://fr.wikipedia.org/wiki/Barbare>
In 1912, Carl Meinhof published *Die Sprachen Der Hamiten* (The Languages of the Hamites.) He used the term Hamitic. Meinhof’s system of classification of the Hamitic languages was based on a belief that “speakers of Hamitic became largely coterminous with cattle herding peoples with essentially Caucasian origins, intrinsically different from and superior to the ‘Negroes of Africa’. However, in the case of the so-called Nilo-Hamitic languages (a concept he introduced), it was based on the typological feature of gender and a “fallacious theory of language mixture”. Meinhof did this in spite of earlier work by scholars such as Lepsius and Johnston. http://en.wikipedia.org/wiki/Carl_Meinhof. See Meinhof, Carl. 1906. *Grundzüge einer vergleichenden Grammatik der Bantusprachen*. Berlin: Reimer.

2 Quote attributed to Ogotemmêli in an interview with French anthropologist Marcel Griaule (*Dieu d’eau, entretien avec Ogotemmêli*, 1948).

ethnic groups, or nations. Language is a primary instrument of integration of newcomers in a society, whether for migrants or simply a new generation. Language is an element of social cohesion that expresses and carries culture, knowledge, aspirations, and the values of an entire community. It is via education in the mother tongue that these values are transmitted to future generations, allowing the continuity of a community. It is because every human community's mother tongue has inalienable identity value that every effort must be made to enable all to carry out life in their mother tongue and to enable their language to live, including in the realm of cyberspace.

Access to information and knowledge is a fundamental right that must be guaranteed to all language communities to ensure equitable knowledge sharing

The mass of information and knowledge in the world is huge, and the information contained in cyberspace is growing exponentially. It is rather illusory to try to make each element of knowledge and information contained in all 7000 world languages available online, and this is not being advocated here. Rather, the vision contained within these principles is far more pragmatic and realistic. In cyberspace, all the peoples of the world are “neighbours”, because neither time nor geographical distance can impose boundaries. Everything shared freely is accessible to anyone who finds it of interest. Free knowledge pooling, as with Wikipedia and other online databases, has long been, and remains broadly, what characterizes and distinguishes the global network. As electronic commerce grows, the issue of copyright for culture producers (musicians, filmmakers, designers, writers, etc.) is becoming crucial and is not adapting well to the free download of cultural and intellectual products. Hence the need for regulation of both free and paid downloads, to simultaneously protect the legitimate rights of authors and artists, while at the same time allowing for the free market that characterizes cyberspace. But if we are not careful, it's a slippery slope from regulation to reduction in free space, thus depriving the poor of the means to participate in the life of the information society. One thing is certain: to use a computer, it is no longer necessary to understand English or have a computer science degree. Not only is English no longer required for computer use, there is no truly

required language, but only the computer user's personal language. Since access to information and knowledge is a fundamental right that must be guaranteed to all language communities to ensure equitable knowledge sharing, it is vital that with respect to intellectual property, each linguistic community have the right to tap into as much of this knowledge as they need and consider necessary to render in their own language.

Technical solutions that facilitate multilingualism should be favoured at all levels of intervention into all means of communication that involve language

The information society is characterized by the massive and global use of ICTs. The structure of these technologies must evolve to enable any language to be used as a medium of communication, in both oral and written forms. In addition, it is of great importance that international standards be established to facilitate the use of languages on ICTs³. The recently developed Unicode coding system is the best example of this. If the 8-bits ASCII system could code up to 256 characters, Unicode can encode over 110,000 unique graphic characters. It is thus theoretically capable of handling all the world's computerized writing systems. Another significant example is that of language coding (ISO 639-n), where n is a number from 1 to 6 that signifies different versions of a standard. This coding system language names and groups is in strong competition with the world languages referencing system LS 640, developed by David Dalby of the *Linguasphère* observatory and compatible with ISO 11179. Finally, ISO 15924 allows the referencing of all the world's writing systems along with their diachronic variants. All these standards of global reach constitute a foundation for the inclusion of all world languages in the process of ICT development. Technical problems are, of course, always complex, but the ethical principle that should prevail here is that at all levels of intervention, we must focus on technical solutions that provide greater flexibility and openness to multilingual applications. As Jean-Louis Garçon has said:

The instruments are now more or less in place. They are not yet perfect, but we can henceforth surf the Web using Chinese, Japanese, Korean, and many other languages that do not use the western alphabet. As

3 See in this book: Stéphane Bortzmeyer, *Multilingualism and the Internet's Standardisation*.

the internet extends to parts of the world where English is rarely used, for example China, it is natural that Chinese and not English be used. The majority of internet users in China have no other choice than their mother tongue⁴ (p. 85).

The development of communications infrastructure in the poorest communities is an obligation of solidarity and equity

Among the fundamental issues, inadequate or nonexistent communications infrastructure is a key factor making it impossible for underserved language communities to use ICTs. In industrialized countries, the gap between city and country in communications infrastructure is generally quite significant. This disparity is even more dramatic in developing countries, because it is combined with other equally serious gaps in public services.

The deployment of heavy infrastructure (particularly optic fibre) in developing countries and rural areas of developed countries is an obligation of solidarity allowing the most disadvantaged communities to access communication, and thus information and knowledge sharing. Thanks to such solidarity, language communities can develop the tools they need to use their language(s) in the information society's communication space. Solidarity with disadvantaged communities, on both national and international levels, is one of the necessary ethical principles ensuring the continued presence of mother tongues in the information society.

Nationally, it is the responsibility of government and legislature to guarantee that national solidarity weighs in on the choice to develop communications infrastructure not only in urban centres but in rural areas. In a multilingual country, such infrastructure takes on added importance because it is a basic condition for regional languages to be locally accessible in cyberspace.

Internationally, the very fact that industrialized countries, international organizations and foreign entrepreneurs invest in or advocate investments in developing countries to help build their national communications

4 Jean-Louis Garçon, *Ntic & Éthiques... Quelle valeur possède l'information en ligne?* Masters thesis in Multilingual Engineering (Paris : Centre de Recherche en Ingénierie Multilingue, INALCO, 2005).

infrastructure contributes to the principle of solidarity, as long as the conditions of these investments take into account the real interests of developing countries. It must be stressed very clearly that all such cases require a significant economic investment and thus cannot be envisaged solely as humanitarian acts without economic counterparts. Indeed, such operations remain highly economic and thus profitable to investors. And that is why the ethical principles of solidarity and equity must be kept clearly in mind, so that the logic of economic profit does not predominate political decision-making with regards to the development of communications infrastructure. It must also be clear that a communication network cannot be built across a country or continent without extensive training of technicians, managers, and networks, as well as of the consumers who will use these new modes of communication.

Training and capacity building in local communities

Capacity building in language communities requires significant investment and long-term programming. As a network expands or is planned across a country, new staff must be trained. Facilitating access to terminals for computers, televisions, and mobile phones can encourage the evolution of network users' communication habits. Without this market, the use of networks is not high enough to make them profitable. The empowerment of experts and users alike is a condition for success, in terms of both economic necessity and human rights. Therefore it is vital that this be strongly if not fully supported by vigorous government action. Businesses have an interest in developing internet cafes, which are good places to access the internet when one has no computer at home. The more people are able to access and use the communication terminals, the more conditions will be favourable to the use of local languages in ICTs. We must therefore constantly solicit and seek out the commitment of governments to a language policy that favours multilingualism, so that every language spoken in their territory is valued, and more specifically, used in communication networks.

Governmental authorities must play a vital role in promoting multilingualism, involving local languages and ICTs

On January 11, 1993, then Vice President of the United States Al Gore launched the global project of the information superhighway. The internet was first developed in North America and Europe, and then spread across all industrialized countries. Only after that did the internet slowly reach the southern hemisphere's developing countries. Without government involvement, this network of networks would probably have never developed.

The installation and deployment of a communications and information networks across a country or continent opens limitless opportunities to enhance economic and cultural activity and to accelerate social change. Appetites sharpen. Large investors, both private and public, usually based in industrialized countries, don't shy away from attempts to supplant the governments of developing countries from controlling the investments that they accept to make to install communication networks in their countries. Where investors succeed, they dictate prices without taking into account the purchasing power of the local population, their sole concern being a quick return on investment. This creates additional social inequalities between those who can afford to access ICTs and those who can't. Such monopolies obviously raise serious ethical concerns.

Only governments can make policy decisions regarding how to integrate ICTs into national plans for economic, cultural, educational, and social development, because only governments carry the ultimate responsibility for the fate of their country. Only governments can intervene to regulate prices to sustainably ensure the purchasing power of consumers. Only governmental policy choices can effectively guide economic, educational, cultural, social, and communicative actions in a way that promotes user-friendly multilingualism, to ensure the flourishing of all language communities in their countries. Indeed, a well-designed policy that gives a fair place to every language spoken in a given country is a policy that will ensure sustainable linguistic peace, which contributes to permanent social peace.

The fact remains that all forces acting as power centres, namely, national and international organizations, international governing organizations,

and cultural and linguistic NGOs have a moral duty to bear their share of commitment by mobilizing wherever necessary to promote an inclusive multilingualism, the development of relevant technologies, and the capacity building of local language speakers.

National and international, governmental and nongovernmental organizations, all have an important role to play in the promotion of multilingualism in the information society

International organizations have played a paramount and avant-garde role in the fight against human afflictions. After XXth century wars, both the League of Nations and the United Nations had as their fundamental mission to prevent conflict by tackling the core issues that could lead to conflicts between states, including poverty, hunger, illiteracy, endemic diseases, economic inequality, crimes against humanity, genocide, and so on. Hence, the United Nations Development Programme (UNDP) supports States in the fight against poverty; the United Nations Food and Agriculture Organization (FAO) fights world hunger, the United Nations Educational, Scientific and Cultural Organization (Unesco) works against illiteracy and toward the worldwide development of science, education and culture. The World Health Organization (WHO) and World Trade Organization (WTO) respectively seek to curb major endemic diseases and regulate trade. And if the biggest criminals cannot be prosecuted in their own countries, the International Criminal Court (ICC) is the last resort for victims seeking justice. All this shows the extent to which multilateral organizations are important in practically every field to work toward universal ethical values by fighting social illness.

It is thus natural that we consider the international organizations whose mission it is to promote the world's languages and cultures as essential partners in the protection of linguistic diversity, in both the linguasphere and in cyberspace. Just as biodiversity is essential for the development of life on earth, linguistic diversity is essential to human life, for it is through language that every community expresses its culture and identity, and adapts and learns to survive in its environment. Also, in the era of globalisation, with ICTs spreading to all human communities regardless of their environment, numerous weakened cultures are in serious danger

of extinction, along with the languages that express them and, of course, the identity of those who speak them. This is the existential question posed, for example, by some residents of the Tuvalu archipelago, who are concerned about the threat of flooding that global warming poses to their country. If the archipelago's hundred thousand inhabitants should be evacuated to various countries like New Zealand, Australia, the United States, Canada, France, England, and others, where they will constitute small minorities destined to dissolve within a few generations in their diverse host societies, what will become the Tuvaluans' collective identity, ancestral culture, language, and skills? This question can just as well be posed by the speakers of the world's 2,400 endangered languages as listed in Unesco's Atlas of Endangered Languages. International organizations are the last defence against the attitude of indifference in which 99 % of the world's languages, spoken by only 6 % of world population, may die.

In a Unesco report from the early 1950s, mother tongue education was considered a fundamental right of the child. Mother tongue internet surfing could very well be its equivalent in the information age. If the internet is to become the global network it promises to be, all users must enjoy access to it, regardless of language. To regard it as the preserve of those who happen to speak English by historical accident, practical necessity, or political privilege is unfair to those who do not know this language.

Launched in January 1999 by the European Commission, the website HLTCentral ("Human Language Technologies") provides a brief definition of language engineering⁵:

"Language engineering permits us to live comfortably alongside technology. We can use our knowledge of language to develop systems capable of recognizing both spoken and written text, to understand a text in sufficient depth to be able to select information, translate it into different languages, and generate both an oral speech and a printed text. The application of these technologies allows us to push the current limits of our language use. Voice-activated systems are expected to play a leading role and become an integral part of our daily lives"

5 Information and Quotation from Marie Lebert (1999, updated in 2009) "The Internet and Languages", reported by *Le Net des Études françaises*: <http://www.etudes-francaises.net/entretiens/multi.htm>

Since technical solutions exist, it is possible to apply them to all languages of the world if the ethical principles put forth and discussed here are applied by all social actors in every sphere of activity.

THE INTERNET IN CHINA

The rift between the Chinese Internet, and the American Internet, or more generally Western Internet, is one of those that experiences regular upheavals, sometimes like an earthquake, as media and political policy causes reactions in public opinion and at the highest state level. Differences and similarities between the two worlds meet on, and through, the medium that will greatly influence the future organization of the internet.

Original article in French.
Translated by Laura Kraftowitz.



STÉPHANE GRUMBACH, senior scientist at INRIA (Institut National de Recherche en Informatique et en Automatique - a French national institute for research in computer science and control), is a specialist of databases. He has held various positions in the international cooperation, as head of international relations of Inria, science counselor in the French embassy in China, and head of a joint lab in the Chinese Academy of Sciences, where he has been the first foreigner habilitated to take PhD students.

STÉPHANE GRUMBACH

THE INTERNET IN CHINA

The internet brings individuals together as no media before it, allowing people to communicate both directly and by accessing posted information. Even language differences no longer present an untraversable barrier thanks to online translation tools. But the internet is not a uniform space with a free flow of information, but a fractured space whose fault lines reveal both diversity, a plurality of uses and more generally cultures, but also rivalries between different actors for control of the internet and its content, and for market domination.

The rift between the Chinese and American/Western internet undergoes regular storms, which at times impact both the media and politics, prompting reactions both in public opinion and at the highest level of government. Balancing these forces could help direct the future organization of the web, its cultural, technological and economic options, and the degree of openness or closure/balkanization of the global network.

If fundamental differences have emerged in network philosophy, particularly around issues of neutrality and freedom of expression that are closely followed in western media, it is actually the less publicised subjects of economics and strategy that dominate movements of digital divergence and convergence between these two major players.

With its economy growing steadily since Deng Xiaoping's policy of openness was introduced in 1978, and becoming the second largest economy after overtaking Japan in 2010, China now owns a third of global foreign exchange reserves. As many students enter universities in China as in the United States and Europe combined, and in certain areas of engineering China's researchers now publish more than American researchers. The emergence of China as a leading world power ensures the country's very

high and for too long underestimated capacity for international action, especially with regards to innovation.

In the digital sphere as well, China is not far behind. Of its more than 600 million mobile phone users, nearly half access the internet via their mobile; of its 420 million internet users as of June 2010 according to the CNNIC¹, 60 % of them are under 30 and eager for new technology. China thus constitutes the world's largest IT market, and multinational corporations are prudent to build the majority of their new research labs in China.

China also carries significant weight in the domain of culture. With 450 million people online, Chinese is the second largest language group after English (536 million speakers)², but a penetration rate of only 32 %, lower than English (42 %), which promises a superior growth potential. Chinese is thus becoming one of the web's predominant languages.

In January 2010, the debate on censorship led to a serious showdown between China and the United States, with Google as the main issue. Following a slew of cyber-attacks from China that it posed as the victim of, Google decided to end the censorship it had previously imposed on its search engine in the Chinese territory by redirecting google.cn to google.com.hk, its site in Hong Kong, which offers uncensored searching in Chinese. It is easy to see that beyond freedom of expression, trade issues are far from marginal in such shows of force.

While Europeans have become accustomed to relying heavily on online services offered by U.S. companies (Google, Facebook, eBay, Twitter, etc.), the Chinese have developed their own national champions of the web that hold their country's largest market share, and occupy four of the Alexa top twenty world rankings³, competing with leading American companies. This is particularly true of the search engine Baidu, which holds over 60 % of China's market share and comes in 6th in the Alexa ranking. Baidu, which is being developed in Japan, offers many services, including an encyclopedia built in a cooperative manner like that of Wikipedia. Ironically however, the site itself (baidu.jp) is inaccessible in China.

1 <http://www.cnnic.net.cn>

2 <http://www.internetworldstats.com>, June 2010.

See in this book: Daniel Prado, *Language Presence in the Real World and Cyberspace*.

3 <http://www.alexa.com/topsites>

The dynamism of forums, blogs, social networks, instant messaging and online business reflects both the popularity of the web among Chinese youth, and its effectiveness in the domestic industrial sector. Among the most prominent companies are Tencent which, with its famous chat service QQ, games, and virtual worlds (an entire culture in China!) comes in 8th in the Alexa rating. Alibaba and its subsidiaries in electronic commerce, including Taobao, come in 13th; Sina.com, with information on China, comes in 16th.

Suspicion between the two regions is not limited to content or to the degree of market opening to foreign companies, but also concerns how data travels over the network. In spring 2010, when some traffic from the U.S. government and private companies passed through China for a few minutes due to errors in the routing information issued by a Chinese internet service provider, an investigation was launched by the United States on the causes of the diversion.

The example of China shows the relationship between development policy, internet presence using domestic language and script, and national economic and geopolitical issues. Network globalisation has become reality, and as traditional issues of international relations have entered the digital world, so has a new articulation of cultures and lifestyles.

THE ECONOMY OF LANGUAGES

It was not until the late twentieth century that economists took a serious interest in languages as a field of study in their own discipline. Why so late? Because it is only with globalisation and the growing power of multilingualism that languages have truly merged for what they are: a major strategic factor in global communication, not only on the political, cultural or societal planes, but also the economic. A product is no longer marketable on a global scale without taking into account this dimension: *“No translation, no product”*.

Original article in French.
Translated by Laura Kraftowitz.



MICHAËL OUSTINOFF is Associate Professor (Habil.) in Translation Studies at the Institute of the Anglophone World, University of Paris 3 Sorbonne Nouvelle and currently on sabbatical leave at the ISCC, the Institute for Communication Sciences of the CNRS. His third book *Traduire et communiquer à l'heure de la mondialisation* (Translating and Communicating in a Globalized World) was published by CNRS Éditions in 2011.

MICHAËL OUSTINOFF

OF LANGUAGES THE ECONOMY

Not until the late xxth century did economists take a serious interest in language as a field of study within their discipline. Why so late? Several explanations can be pointed out, but we'll start with the most obvious and important: globalisation made it impossible to ignore language. From the economic viewpoint, languages became a considerable and strategic industry in their own right.

In a recent study on the size of the “language industry” [RINSCHKE 2009], the Directorate-General for Translation (DGT) of the European Commission assessed its value at 8.4 billion euros as of 2008, noting that it was one of the sectors that had best weathered the economic crisis, with an impressive growth rate of 10% (pretty unbeatable), and conservative estimates predicting growth attaining 16.5 billion euros in 2015, while high estimates predicted up to 20 billion. As such growth is comparable planet-wide [DWYER 2010], we can see why economists feel compelled to take a closer interest, across all levels of economic analysis. A fundamental question that the United States is now taking seriously, along with the United Kingdom, through institutions like the British Council [GRADDOL 1997] and the British Academy [BRITISH ACADEMY 2009].

Nevertheless, the emergence of a new industry, whose spectacular expansion is linked to the ICT (Information and Communication Technologies) industry within the context of globalisation, is not limited to an analysis of its economic weight. As essential as that issue is, the language economy is inseparable from its political dimension at the level of global governance.

Viewed from this perspective, language can no longer be considered merely in terms of cost – a consideration that until now has been about minimizing, a function that English-only policies were supposed to help

meet¹ – but must be also viewed in terms of investment, just as much economic as it is cultural and geopolitical, and whose interest appears all the more starkly against a backdrop of the worst global financial crisis we have known since the Great Depression, which was also born in the United States.

NO TRANSLATION, NO PRODUCT: GLOBALISATION'S NEW GIVENS

The current slogan of globalisation from an economic point of view, in the words of Suzanne Topping, is *no translation, no product* [TOPPING 2000: 111]. This formula may appear expeditious or even brutal to some, but it does reflect the contemporary reality: from one end of the world to the other, it is unthinkable to launch a product without ‘localizing’ [FOLARON 2007] its accompanying advertising campaign in all aspects, up to and including the product’s documentation². In this sense, “localization” essentially means: “*All processes by which digital content and products developed in one place (in terms of geography, language, culture) are adapted for sale and use in another location. ‘Localization’ therefore includes (a) the translation of verbal content according to the textual and linguistic conventions of the receiving language, and (b) adaptation of non-verbal content – from colours, icons, and bitmap format, to packaging and other formal constraints. It also includes all data and parameters relevant to the consideration of cultural, technical and legal requirements of the target area. In short, localization is less about specific tasks than a process of product adaptation*” [DUNNE 2006: 4]. However, we can expand further on this perspective: it is not only true for digital space, it applies to all economic spaces. In fact, cyberspace is only the tip of the iceberg. It’s not because cyberspace has become massively multilingual that multilingualism is gaining traction globally; it is because multilingualism is a major economic force in today’s world that it has emerged in cyberspace as a new element of the globalised market economy.

A recent study at Stanford University about the state of California – which, by itself, ranks among the top ten world economies – constitutes a case

1 See in this book: Michaël Oustinoff, *English Won’t Be the Internet’s Lingua Franca*.

2 See in this book: Dwayne Bailey, *Software Localisation: Open Source as a Major Tool for Digital Multilingualism*.

in point. The overall argument is easily summarised. First, it is urgent to invest in languages because of the current risks of global recession: “*Economic, geographic and social realities in California create a pressing need for investing in foreign language education now. As we face a global recession, California (one of the world’s ten largest economies) has a \$42 billion deficit and a critical need for long-term solutions*” [STEELE 2009:1]. The economic centre of gravity’s shift to Asia and South America must be taken into account: “*In Section I of this report, we will argue that these solutions require reaching out to the rest of the world. California’s geographic location is ideal for capitalizing on growth in Asia-Pacific and Central and South American markets*” (*ibid.*). California must therefore call on its exceptionally diverse multilingual population: “*California also has an exceptionally diverse and linguistically skilled population. The millions of people in California who speak foreign languages such as Spanish, Chinese and Korean position our state to excel in international business, trade and diplomacy*”. And the conclusion: “*Further increasing the value of our workforce by emphasizing foreign language education will use our strategic position and social resources to bolster the economy*” (*ibid.*). The study does not shy away from proposing a short-term “stimulus plan”, in terms that Barack Obama wouldn’t probably disown:

Even in an economic decline, opportunities for California’s financial and job markets can grow. Now is the time to make the most of California’s unique multicultural strengths: a large and diverse population with millions of bilingual and multilingual, cross-culturally fluent people, and an ideal location to reach markets across the Pacific and in North, South and Central America. At a time of financial crisis, a large investment in education for world language and cultural competencies is a necessary component of a successful economic stimulus plan. (ibid.: 3)

Nevertheless, we shouldn’t expect things to fall into place themselves, as if by magic. As demand grows, “*California’s growing economic stake in international trade and the diversity of California’s trading partners create a demand for Californians who can apply their knowledge of world languages and cultures to business*” (*ibid.*: 4). But an adequate response must be taken: “*International markets are tough, and even large, successful companies have floundered without sufficient linguistic and cultural competency*” (*ibid.*).

The disappointment of KFC (Kentucky Fried Chicken) in China is a shining example. Initially, the KFC slogan “*finger lickin’ good*” was translated literally into Chinese for a message as sinister as it was funny: “*Eat your fingers off*”. It’s easy to imagine the disastrous results that can ensue whilst trying to establish a foothold in a local market of that size. The financial losses of this blunder in terms of market share and image were assessed in the millions. Today, KFC is the number one fast food chain in China. But it took some backpedalling and a swift reaction :

The company quickly produced a slogan that made sense in Chinese, replaced coleslaw with familiar local dishes, and rose to become the number one American fast food chain in China. Two million Chinese eat at KFC every day, and the company opened new stores at a rate of 250 per year in the early 2000s – a huge market victory that depended on both linguistic and cultural acumen. (ibid.: 4).

Many other examples could be cited, but the bottom line is that this kind of localised business model was simply unthinkable in the United States only twenty years ago. After the fall of the Berlin Wall and the collapse of the USSR, the world was declared unipolar, with the United States as the sole remaining superpower of the “space age” [DOLLFUSS 1994]. Under the effects of “soft power” [NYE 2004], all languages were arranged equally around the “hypercentral” language [CALVET 2007]. The economy of languages was limited to that of their relationship to Global English, which was by definition an unequal relationship. Wasn’t the economic benefit of a universal lingua franca precisely that it would bypass other languages, which were seen as so many additional costs?

THE LANGUAGE ECONOMY IN A MULTIPOLAR POST-AMERICAN WORLD

The example of California, a state of exceptional multicultural wealth, of which the economic potential hasn’t ever really been exploited, reveals a paradox that Fareed Zakaria has highlighted: “*Generations from now, when historians write about these times, they might note that by the turn of the XXIst century, the United States had succeeded in its great, historical mission – globalizing the world. We don’t want them to write that along the way, we forgot to globalize ourselves*” [ZAKARIA 2008:95].

And one might add... Europe! Certainly, the European Union is founded on the protection of linguistic plurality, the EU-27 counting 23 official languages, and intends to promote multilingualism in the EU, as former European Commissioner for Multilingualism, Leonard Orban, has underscored: “*Employees should master, for the benefit of their employers, at least three languages: that of their country of origin, English of course, and a third from among the most widely spoken in the eu – German, French, Spanish or Italian. Without neglecting Russian, Arabic or Chinese*” [RICARD 2007]. Nevertheless, the effect of such policy is slow to be felt in companies where the shortfall in foreign language mastery, starting with Britain, costs 100 billion euros a year:

Orban cites a study undertaken in late 2006 by the UK’s governmentally-approved language expertise organization. According to the study, 11 % of 2000 small and medium enterprises (SMEs) surveyed in the European Union (EU) have already lost a contract for export due to insufficient language skills. The resultant loss is estimated at 100 billion euros per year. “*The idea that English is the universal language is sinful in its simplicity*”, the study observes. Preferred by multinationals, the language of Shakespeare could certainly be used primarily to negotiate twenty markets. But German would facilitate exportation to fifteen markets, including Germany and Austria. As for French, it is used in eight markets, including France, Belgium and Luxembourg. [RICARD *IBID.*]

Despite the obvious differences between the cases of California and Britain (the effects of globalisation in the first case; trade within the European Union in the other), one key similarity is just as readily evident: both involve market deregulation between nation-states that have been “*more or less protected until now*” [DAVEL 2008:6], whether on a local, global or “glocal” scale. But this explanation is insufficient: what is spectacular is the abandonment of the “English” standard (as we saw the abandonment of the gold standard), or, more precisely, the use of other standards. In the past, neither Great Britain (hyperpower of the nineteenth century until the US took over beginning with World War II), nor the United States, have cared to reciprocate with language, and English has imposed itself as the *lingua franca* of reference.

Today, the situation is radically different, for three basic reasons. First, the failure of English-only policy, which was supposed to allow us to “communicate” on a global scale through a single language. General

linguistic mastery has clearly proceeded much more slowly than market deregulation. Second, and this relates directly to the previous reason, is the realisation that it is in the language of the Other that selling must take place, not in a foreign language, even English. The third reason is relatively new and probably more important in the longer term: we live in an increasingly multipolar world where the US will eventually cease to be the sole master of the game. Whether the US is in decline [TODD 2002] or simply facing what Fareed Zakaria calls “*the rise of the others*” (including the BRICs: Brazil, Russia, India and China), these “emerging” powers cannot help but be aware of the importance of the soft power – of any power at all – that their respective languages represent. The current enthusiasm for Chinese, despite its reputation for being extremely complicated, cannot be explained differently: as the language of the second world power, it hopes for economic gain, even if other motives (particularly cultural) cannot be excluded.

THE EMERGENCE OF A NEW FIELD OF ECONOMIC ANALYSIS: LANGUAGES AS VALUE

In light of the above, it is unsurprising how little economists have paid attention to languages until recently. In fact, the earliest studies date back to the 1960s. Three sets of reasons, it seems, can explain the disinterest. The first set links the Western world to the Greco-Roman tradition, inherited from antiquity: the Greeks had only one language, their own. The Greek world was fundamentally monolingual. They transmitted this model to the Romans, who in turn transmitted it [CASSIN 2004]; foreign languages had at best a secondary, purely instrumental role with respect to the highest mental processes and human activities. It would not be until the late nineteenth century that language would become a subject of scientific study, specifically, linguistics.

The second set of reasons is mentioned above: only recently did market deregulation, globalisation, the emergence of a multipolar world, not to mention the ICT, which are undergoing a real revolution in their field, appear. For both theoretical and empirical reasons, languages could not previously be considered objects of real reflection for economists. Moreover, to combine these first two sets of explanations, languages were essentially perceived as a cost (as an obstacle to direct access to “concepts” or acting in the world).

But there is a third perspective to consider that is more strictly methodological, to understand why “*the economics of language remains a relatively marginal area of specialisation, on the edge of the economy*” [GRIN 2002:14]. Two factors should be taken into account: “1. *the economics of language is necessarily an interdisciplinary endeavour and [...] as a discipline, the ‘orthodox’ economy is notoriously closed to interdisciplinarity*; 2. *economic modelling, and the set of theoretical concepts on which it rests, require the use of quantitative variables, or at least variables that lend themselves to interpretation in terms of “more” and “less”. Yet the study of linguistic questions routinely requires taking into account variables that hardly lend themselves to quantitative interpretation, rendering the modelling as well less relevant. Given the popularity of modelisation in economics departments at universities, any area of specialisation that does not lend itself to the use of algebraic tools is likely to be neglected*” [GRIN *IBID.*:14-15].

Be that as it may, we note a major distinction in the definition of a language’s “value”. It can be assigned, first of all, a true definite value: “*The market values are reflected in the prices or in another indicator of this type. Suppose, for example, that to speak language X facilitates the sale of goods to the language X-speaking public, and thus allows for higher profits; or an language X-speaking employee earns more, all else being equal, because s/he knows language X, in which case, language X has a market value*” [GRIN *IBID.*:21].

The second fundamental facet of value is non-market value: “*For example, knowing language X gives access to the culture in language X, facilitates social contact with members of the community speaking the language X, etc. This value is generally not reflected in market prices [...]*” [GRIN *IBID.*:21].

Moreover, the value should be considered on two levels: the individual, or “private”, level; and the collective, or “social”, level. Therefore, we are left with a table of four entries.

If you start to combine these four dimensions (social vs. private market value, private non-market value vs. social) with the above, we see the complexity that must be taken into consideration. Hence, with the current state of knowledge, the difficulty of accurately assessing a language’s exact value: “*In short, it is not possible, for the time being, to truly calculate (I) the ‘value’ of a language, (II) the ‘value’ of one linguistic area in relation to another, (III) the ‘benefits’ (market and non-market) to expect from a*

particular policy; (IV) many of the costs, direct and indirect, associated with such an initiative” [GRIN IBID. :21].

This does not invalidate all economic analysis, quite the contrary. In particular, François Grin reveals a major drawback to English-only policy that David Graddol [GRADDOL 1997 AND 2006] also signals when it enters into competition with other languages: *“It can be rightly argued that even if English is cost-effective, to promote learning this language will render knowledge of it more banal, and the salary benefits enjoyed by Anglophones will gradually erode [...] This trend will in all likelihood affect most of the countries where English is (increasingly) taught as a second language and judged as essential to economic activity. In other words, convergent evidence suggests that knowledge of other languages will gradually become more profitable as English spreads, implying that language teaching policy should not be focused solely on English as a second language, but also on other languages” [GRIN IBID. :34].*

In other words, the first collateral victims of the new circumstances of globalisation will be the Anglophone monoglots, but not only they. What is true for them might well, eventually, apply to the others as well. The following analysis can thus be generalised: *“The cost of learning English by non-Anglophones keeps falling while the cost of learning other languages for Anglophones (and everyone else) keeps increasing. One consequence of the universal spread of the lingua franca would then be that Anglophones will face competition on their home labour markets with everyone else in the world, while having no real access to those labour markets in which another language remains required” [VAN PARIJS 2004].* In a post-American world, the marginal value (in the economic sense of the word) of English declines gradually as that of other languages correspondingly increases.

Given the importance of the language industry [RINSCHÉ 2009] and its growth rate that is predicted to hold steady at 10 % for quite some time, it doesn't take a genius to understand the economic dynamics at play of languages in the contemporary world – issues that play out at the State level as well as the individual. It is thus also in political terms that the language question is posed, even in terms of international governance in the time of globalisation.

CONCLUSION

English-only policy was supposed to be the most practical solution and, consequently, the most economical [GRIN 2004]. Today, we are realizing that this is just not the case. Far from advantageous, this model actually represents a considerable loss, something that Anglophone countries are themselves realising, starting with Great Britain and the United States. Once considered a question of secondary importance, the linguistic potential of nations now constitutes a major strategic asset in the era of economic globalisation. Suzanne Topping's formula, "*No translation, no product*", certainly requires some qualification, but it translates into no less than a major paradigm shift, which doesn't consider languages as expendable, but lends them major market value, not only in terms of cost but in terms of profitable investment [GRIN 2009].

The language economy cannot be reduced to its linguistic dimension, in the traditional sense of the term. Its cultural dimension must be considered. The most prestigious Anglophone business schools have understood this since the early 2000s [EARLEY 2003; EARLEY, 2004; CHUA 2009], emphasizing "Cultural Intelligence" (Cultural Intelligence, or CQ, "Cultural Quotient") that Cultural Studies and Unesco have long recognised the importance of:

[...] The global village is a multicultural mosaic. Different countries have different cultures. It is therefore necessary for companies and managers to exercise caution in trying to establish the best way on a global scale. When such differences are not taken into account, companies adopt inappropriate strategies and policies. As managers, they take bad decisions, clash with local cultures, experience major cultural maladjustment and fail to carry out their initiatives. [REGO 2009: 34]

Unfortunately, it is less certain whether States are taking necessary measures. There are considerable delays, because the language policies introduced since Second World War have made the English teaching the primary objective. In addition, the budget cuts currently affecting the whole of civil service are crashing down on languages other than English. This runs the major risk of aggravating what we call the "linguistic divide" over and above the digital divide: by relying upon the market's "invisible hand", the disparities related to fully operational multilingual access grow, with such access remaining confined to elites. However, in a global economy and in the era of mass

education, training only the elite is insufficient [OUSTINOFF 2011]. It is an obsolete model with disastrous long-term social and economic consequences. This is what the language economy teaches us; will policy makers listen?

Still, there is a promising area which is cause for some optimism, namely the digital space in which costs may be brought down drastically if not eliminated altogether thanks to the new information and communication technologies (ICTs). Wikipedia is a case in point. All you need is a computer and Internet connection to have access to a wealth of free online sources in the most varied languages. This is how Wikipedia defines itself online: “*Wikipedia is a multilingual, web-based, free-content encyclopedia project based on an openly editable model [...]. Anyone with Internet access can write and make changes to Wikipedia articles*”. The model has been emulated by websites such as *Global Voices*, founded by Harvard’s Berkman Center for Internet and Society, where the following explanation can be found: “*Global Voices is a community of more than 500 bloggers and translators around the world who work together to bring you reports from blogs and citizen media everywhere, with emphasis on voices that are not ordinarily heard in international mainstream media. [...] Global Voices is translated into more than 30 languages*”.

Such initiatives are bound to multiply and their interest is self-evident. It lies in making available to as many people as possible the countless applications of tics and allowing everyone to use them in his or her own language (not to mention language teaching and self-education). Even free use has a cost, however, as economists for whom “there ain’t no such thing as a free lunch” found it easy to demonstrate. This is further proof that relying on the law of markets alone would be the worst option.

BIBLIOGRAPHY

[BRITISH ACADEMY 2009] The British Academy. 2009. *Language Matters. Position Paper*. <http://www.britac.ac.uk>

[CALVET 2007] Calvet, L.-J. 2007. La traduction au filtre de la mondialisation, In : *Traduction et mondialisation*, ed Oustinoff, M. and Nowicki, J. *Hermès* 49. Paris: CNRS Éditions.

[CASSIN 2004] Cassin, B. 2004. *Vocabulaire européen des philosophies. Dictionnaire des intraduisibles*. Paris: Le Robert/Le Seuil.

[CRONIN 2003] Cronin, M. 2003. *Translation and Globalization*. London: Routledge.

[CHUA 2009] Chua, R.Y.J. and Morris, M. W. 2009. Innovation Communication in Multicultural Networks: Deficits in Inter-cultural Capability and Affect-based Trust as

Barriers to New Idea Sharing in Inter-cultural Relationships. *Harvard Business School Working Paper 09 (130)*, May.

[DAVEL 2008] Davel, E., Dupuis, J.-P., and Chanlat, J.-F. 2008. *Gestion en contexte interculturel. Approches, problématiques, pratiques et plongées*. Quebec: Presses de l'Université Laval.

[DOLLFUSS 1994] Dollfuss, O. 1994. *L'espace monde*. Paris: Economica.

[DUNNE 2006] Dunne, K. (ed). 2006. *Perspectives on Localization. ATA Scholarly Monograph Series 8*, Amsterdam and Philadelphia: John Benjamins Publishing.

[DWYER 2010] Dwyer, T. 2010. Traducteurs, associations professionnelles et marché: approches empiriques, In: Oustinoff, Nowicki, and Machado Da Silva (eds) 2010. *Traduction et mondialisation, Volume 2. Hermès 56*. Paris: CNRS Éditions.

[EARLEY 2003] Earley, C. and Ang, S. 2003. *Cultural Intelligence: Individual Interactions across Cultures* Palo Alto: Stanford University Press.

[EARLEY 2004] Earley, C. and Mosakowski, E. 2004. Cultural Intelligence. *Harvard Business Review* Oct.

[FOLARON 2007] Folaron, D. and Gambier, Y. 2007. La localization: un enjeu de la mondialisation, in Oustinoff, M. and Nowicki, J. (eds). 2007. *Traduction et mondialisation*. Hermès 49. Paris: CNRS Éditions.

[GRADDOL 1997] Graddol, D. 1997. *The Future of English? A Guide to Forecasting the Popularity of the English Language in the 21st Century*. The British Council & The British Company (UK) Ltd. (new ed. 2000). <http://www.britishcouncil.org/learning-elt-future.pdf>

[GRADDOL 2006] Graddol, D. 2006. *English Next. Why Global English may Mean the End of "English as a Foreign Language"*. The British Council & The British Company (UK) Ltd, 2006 (new ed., 2007). <http://www.britishcouncil.org/learning-research-english-next.pdf>

[GRIN 2002] Grin, F. 2002. *L'économie de la langue et de l'éducation dans la politique de l'enseignement des langues*. Strasbourg: Language Policy Division, Council of Europe.

[GRIN 2004] Grin, F. 2004. Coûts et justice linguistique dans l'élargissement de l'Union européenne. *Panoramiques*, n°69, 4^e trimestre 2004, 97-104.

[GRIN 2009] Grin, F. 2009. *The Economic Value of Multilingualism: Private, Social and Macroeconomic Perspectives*. Brussels: Directorate-General for Translation, European Commission. 27 Nov. http://webcast.ec.europa.eu/eutv/portal/_v_fl_300_fr/player/index_player.html?id=8095&pId=8093

[NYE 1997] Nye, J. 1997. *Soft Power: The Means to Success in World Affairs*. New York: Public Affairs.

[OUSTINOFF 2011] Oustinoff, M. 2011. *Traduire et communiquer à l'heure de la mondialisation*. Paris: CNRS Éditions.

[REGO 2009] Rego, A. and Pina e Cunha, M. 2009. *Manuel de gestão transcultural de recursos humanos* (Handbook for Transcultural Management and Human Resources). Lisbonne: Editora RH.

[RINSCHÉ 2009] Rinsche, A. and Portera-Zanotti, N. 2009. *The Size of the Language Industry in the EU*. Brussels: Direction générale de la traduction, Commission européenne.

[RICARD 2007] Ricard, P. 2007. Une étude britannique prône le multilinguisme en affaires. *Le Monde*, 25 Sept.

[STEELE 2009] Steele, T., Oishi, L., O'Connor, K., and Silva, D. M. 2009. *Learning World Languages and Cultures in California: A Stimulus for Academic and Economic Success*. Stanford: Stanford University School of Education.

[TODD 2002] Todd, E. 2002. *Après l'empire: Essai sur la décomposition du système américain*, Paris: Gallimard.

[TOPPING 2000] Topping, S. 2000. Shortening the Translation Cycle at Eastman Kodak, in Sprung, R. C. (ed). 2000. *Translating into Success: Cutting-Edge Strategies for Going Multilingual in a Global Age*, Amsterdam and Philadelphia: John Benjamins.

[VAN PARIJS 2004] Van Parijs, P. 2004. Europe's Linguistic Challenge, *Ce*, *European Journal of Sociology*, 45 (1).

[ZAKARIA 2008] Zakaria, F. 2008. *The Post-American World*. New York: Norton.

**DANIEL PRADO
& DANIEL PIMIENTA**

PUBLIC POLICIES FOR LANGUAGES IN CYBERSPACE

Many recommendations and statements from summits, international organizations, meetings, etc. suggest multiple actions in the matter of the presence of languages in cyberspace in order to promote linguistic diversity in the Knowledge Society. We will review the appropriateness and feasibility of the proposals made, as well as gaps identified, and make a coherent synthesis.

Original article in French.
Translated by Laura Kraftowitz.



DANIEL PRADO is the former head of the linguistic unit of Union latine, an intergovernmental organization composed of 35 states whose mission is to disseminate and promote the Latin languages and cultures. He is the current Executive Secretary of Maaya, World Network for Linguistic Diversity.



DANIEL PIMIENTA was born in Casablanca. He studied Applied Mathematics and Computer Science at the University of Nice. In 1975, he joined IBM France as a telecommunications systems architect. In 1988, he became Scientific Adviser to the Union Latine in Santo Domingo, where he manages a regional network project. In 1993, he founded the Networks & Development Foundation (Funredes). In 2008, he received the Namur Award (IFIP WG9.2) for his actions working toward a holistic vision of the social impact of ICTs.

DANIEL PRADO
& DANIEL PIMIENTA

IN CYBERSPACE FOR LANGUAGES PUBLIC POLICIES

Language and culture play an increasingly important role in contemporary political projects, at regional, national and international levels.

Our world has changed considerably since the partition of Yalta in the aftermath of World War II, when two different visions for the future confronted one another, and international relations were based on military, political and economic criteria. In that bipolar schema, culture's value as the basis for society was completely ignored. Nor is the world any longer that of the late eighties, when under the misleading headline "the end of history", a political and economic model was imposed that put forth a unitary culture and sought hegemony.

Now in the XXIst century, the world has become multipolar. Supranational alliances based on intercultural respect have grown stronger. Alliances between regions and nations that share a language or culture are born each day. If political and economic issues are ever present in international relations, cultural elements now play an increasing role. Religion, ethnicity, custom and language participate in international policy making, even if at times they can also be a divisive factor. The cultural factor is increasingly one of the elements of sustainable development and equitable growth, facilitating harmony between peoples and a common respect for dignity.

Language is now a presence in international issues, both political and commercial. In addition to English, international and regional bodies recognize other great languages as co-official: French, Spanish, Russian, Chinese and Arabic at the United Nations; 22 languages for the European Union; Portuguese, French, Arabic and Swahili for the African Union; French, Spanish and Portuguese in the Americas. These languages demand respect and equal treatment.

Other languages without official status (for example, Portuguese in the UN, regional languages in the European Union, Guarani in South America) are demanding their right to be recognised as co-official. Initiatives by various agencies (the Council of Europe, Linguapax, Unesco, and others) tend to give an increasing role to languages without national or regional official status.

This regionalization of relations renders more powerful and present the languages that lost ground decades ago, which are presently returning to education and commerce. Economic globalisation is accompanied by a proliferation of the languages used by industries. According to a well-publicised survey by the government, conducted to promote the language industries¹, 60% of consumers in developed countries never buy a product that is not labelled in their language. Companies have been slow to follow up, but they are increasingly localising², while national and international administrations, suffering from bureaucratic inertia, take a bit longer to react.

The adoption of the *Convention on the Protection and Promotion of the Diversity of Cultural Expressions*³, put forth by Unesco, was the symbol in itself of culture's changing role in international relations. Beyond simple economic value brought about by specialized industries, culture is primarily an essential precondition for human existence, and a driving force for a mutually respectful development of the planet's future.

This Convention addressed a gap that multilingualism activists were well aware of, because it was time to consider the scientific study of linguistics as intimately associated with free expression, self-empowerment, equal opportunity, and promotion of international understanding on a just and equal basis.

1 *Report on Global Consumer Online Buying Preferences, Showing the Impact of Language, Nationality, and Brand Recognition*, Common Sense Advisory, 2006. Accessed freely on 09/11/2009, but access is currently restricted:

http://www.commonsenseadvisory.com/Media/PressReleases/tabid/98/Default.aspx?udt_1084_param_detail=964&zoom_highlight=60+

2 See in this book: Michael Oustinoff, *The Economy of Languages*.

3 Adopted by Unesco 20 October 2005. <http://www.unesco.org/new/en/culture/themes/cultural-diversity/diversity-of-cultural-expressions/the-convention/convention-text/>

Language is certainly implicit in the Convention, but “implicit” does not mean “obvious” to everyone. The Millennium Development Goals⁴, in neglecting to give culture its own goal, also failed to address the languages of persons. Certainly, the *Convention on Intangible Heritage*⁵ makes careful mention, and both the *Recommendation on Multilingualism in Cyberspace*⁶ and the *Declaration of Principles of the World Summit on the Information Society*⁷ embody the idea of respect for linguistic diversity and multilingualism. But it is also necessary that these instruments be followed with concrete achievements. And we know that we are far from affording all individuals the opportunity to develop freely in their own language.

In the absence of international conventions on currently existing languages, and lacking reliable indicators on their place and impact in the developing world; given the probable disappearance of nearly half of them, and the high number of injustices linked with speaking an unrecognized language, action is still needed. These actions include indicators, public policies, and the promotion of multilingualism and linguistic diversity, not to mention legal instruments.

While fewer than one in three individuals currently has internet access⁸, the internet’s evolution is continuous and tends towards universalisation. We also know that cyberspace and its associated technologies are gradually replacing our old modes of communicating, expressing ourselves, transmitting information, sharing knowledge, and connecting with others. Languages that cannot circulate such information, knowledge or dialogue, risk losing value in the eyes of their speakers. But migration and urbanization, as well as universal access to the internet and media, provoke inter-linguistic tensions, in which only those languages that are highly valued by their speakers will survive.

4 See in this book: Adama Samassékou, *Multilingualism, the Millennium Development Goals, and Cyberspace*.

5 Adopted by Unesco 17 October 2003.

<http://www.unesco.org/culture/ich/index.php?lg=en&pg=00006>

6 *Recommendation Concerning the Promotion and Use of Multilingualism and Universal Access to Cyberspace*, adopted by Unesco 15 October 2003.

7 Adopted in Geneva in December 2003.

<http://www.itu.int/wsis/docs/geneva/official/dop.html>

8 <http://www.internetworldstats.com/stats.htm>

Despite some advances in the level of multilingualism on the internet, only a handful of the world's languages enjoy a significant online presence, and English remains the language that is most commonly used, although its relative presence is shrinking and could fall to 30 % of web pages⁹, matching the percentage of internet users who are comfortable in English¹⁰. Jean-Claude Corbeil announced in 2000 that, “*Very shortly, the presence of English is expected to decline to roughly 40 %, as various countries create websites, and as they connect to the net*”¹¹.

The reality is that of all the languages in the world, only a small number will benefit from institutions ensuring their protection, development and equipment, and only five hundred¹² will have a presence in cyberspace (on the other hand, the number increases if you count audio-visual resources). A language develops not only through an institution's willingness and capacity to ready it for use in every context, but rather, it is often the desire of individuals, faculty, or small organizations working in the field who participate in this development¹³.

Whether it concerns a solid institution or lighter initiatives, there is no rule, no science, and no exemplary praxis for policy language. While many works are devoted this matter, and despite numerous conferences dedicated to it (sometimes explicitly, but more often implicitly), language policy (or language planning) has only been an object of study for a few decades, and because of the disparity of sociolinguistic situations, they cannot effectively guide practitioners, apart from the most-studied languages (mostly European and Asian languages).

So it is difficult to find two consistent models of language policy in a given country or territory. Under the heading “language policy”, we find

9 At the time of writing, the estimated figure is that 26.8% of internet users speak English, closely followed by Chinese with 24.6% and Spanish with 7.8%
<http://www.internetworldstats.com/stats7.htm>

10 See in this book: Michael Oustinoff, *English Won't Be the Internet's Lingua Franca*.

11 Jean-Claude Corbeil, “I comme informatique, industries de la langue et Internet”, In B. Cerquiglini et al., *Tu parles!?, le français dans tous ses états*, Paris, Flammarion, 2000, p. 129. My translation.

12 This is the number of languages identified by Unicode as having online representation
http://unicode.org/repos/cldr-tmp/trunk/diff/supplemental/languages_and_scripts.html
Of those, an estimated 300 are actually used in cyberspace. By far, the most multilingual application is Wikipedia, with 269 different languages as of 2011.
<http://stats.wikimedia.org/EN/Sitemap.htm>.

13 See in this book: Evgeny Kuzmin, *Linguistic Policies to Counter Languages Marginalization*.

a number of fairly heterogeneous measures, ranging from policies that promote, protect or help regenerate one or more languages, to policies for eradicating them! We find very elaborate and explicit policies aimed at reviving a minority or rarely used language (Quebec, Catalonia, Israel), and we find policies valorising indigenous languages (Mexico, Bolivia, Paraguay, Mali, Benin, and others), as well as Equity Policies for citizens vis-à-vis their government (Switzerland, Luxembourg, Aruba, and so on). There are also, unfortunately, much less laudable policies for linguistic hegemony, giving absolute priority to the state language and banning the use of others.

A language policy¹⁴, for example, could impose itself in the adoption of a writing system, or grammar and spelling rules and vocabulary development, by determining the status of one or more languages (granting official status, judicial and administrative use, status as a teaching language, a regional or national language, and so on), or teaching it at the international level.

For every language policy, whether it advocates linguistic hegemony, or conversely, promotes equal rights of citizens by allowing them to participate fully in society through the use of their first language, indicators are a fundamental orientation tool. They can objectively diagnose a situation, measure trends, and evaluate the effects of policies.

However, although many national or international bodies (Unesco, OIF, Union Latine, SIL, University of Laval, British Council, CPLP, and others) provide statistics, follow-up studies, surveys and studies, and seek to establish more or less stable parameters for the observation of one or more languages, the indicators derived from them differ according to the home institute, showing glaring inconsistencies and allowing for media hype that at times just feeds uncertainty rather than clarifying the situation.

From the simple question of how many languages exist in the world (which raises the thorniest of borders between dialects), what methods to use for measuring the number of speakers (in terms of their ability or proficiency level in the language), other questions are emerging around the development of indicators, such as the size of the available corpus,

14 On this topic, see Leclerc's page (CIRAL) on language policy.
http://www.tlfq.ulaval.ca/axl/monde/index_politique-Ing.htm

use in everyday life, and in education, administration, health, media, and scientific and technical information.

New parameters have emerged, strengthening a rationale that allows more room for one language over others, or balancing the presence of different languages: namely, how much a language “weighs” or how much a language is “worth”. Various studies by Grin¹⁵, Graddol¹⁶, Lopez Delgado¹⁷, Hope¹⁸, and others on language value (specifically, English, French, Spanish and Portuguese), as well as research conducted by Calvet¹⁹ and others²⁰ on language weight, attempt to provide verifiable parameters to decide whether to learn or teach a language, the need to promote the presence of a language in sectors of society where it is less present, or simply to help it position itself in the labour market or sell a better product.

However, only the languages of people who are aware of this opportunity or challenge will flourish, while others may attend their own inevitable death in a globalized world where digitisation is imposing itself.

In proposing readjustments that respect linguistic diversity, we show our lack of real vision for the language situation. The statistics are incomplete, the indicators distorted, the studies biased, and above all, the metric studies on language are in their infancy.

The observation of a linguistic evolution in cyberspace is no exception to this general rule; on the contrary, it rather highlights the existing gaps

15 Grin, F., *Compétences et récompenses. La valeur des langues en Suisse*, Fribourg, Éditions Universitaires Fribourg, 1999.

Grin, F., « English as economic value: facts and fallacies », *World Englishes*, n°20, 2001.

16 Graddol, D. *English Next. Why Global English may Mean the End of “English as a Foreign Language”*. The British Council & The British Company (UK) Ltd. 2006 (nvlle éd.. 2007). <http://www.britishcouncil.org/learning-research-english-next.pdf>

17 García Delgado, JL et al, *Economía del español. Una introducción* (2ª edición Ampliada), Madrid, Editorial Ariel, 2008.

18 Esperanca, Jose Paulo. *O Valor Económico Língua Portuguesa da Pode Ser Potenciado*. <http://www.portalingua.info/fr/actualites/article/valor-economica-portugues/>

19 Calvet, Louis-Jean (2002). *The Language Market (Le marché aux langues)*. Paris: Plon. We can also use the same author’s language weighting assessment tool, based on several criteria that can be weighted:

Alain Calvet and Louis-Jean Calvet, The Calvet World Language Barometer. <http://www.portalingua.info/fr/poids-des-langues/>

20 Maurais, Jacques and Morris, Michael A. (eds.) (2003). *Languages in a globalising world*. Cambridge: Cambridge University Press; Wallraf, Barbara. “What global language”. *The Atlantic Monthly* 286: 5 (2.000): pp. 52-66.; Weber, George. “Top Languages”. *Language Today*. December (1997): 12-18.

in the material. It is precisely the emergence of the internet that poses the new question of “opportunity” and “challenge” to languages. Indeed, cyberspace is a challenge for every language that faces new competition for providing access to a wealth of information, failing which its own speakers abandon it gradually, preferring a language that in their eyes is more “prestigious”, or at least provides more information in a given domain. But it is also an opportunity because as a medium, it provides easier and less expensive expressive capabilities than traditional media (including paper publishing and terrestrial broadcasting), and therefore can become an ideal way for linguistic resurgence.

So is the internet language’s new hope? Probably, because cyberspace allows doors to open for forms of expression that don’t interest the traditional publishing circuits. Scientific publications in languages other than English can find a place, for example, albeit modest, thanks to the ease and low cost of online publishing; whereas academic journals published by traditional print publishers concern a minimal number of readers, a risk that few publishers are willing to take.

Without a doubt, the internet has permitted minorities absent from traditional publishing to speak, but let us not think that the game is over. The relationship between the internet and global linguistic diversity is inversely proportionate, as we demonstrated in 2008²¹; and the linguistic divide follows the digital divide.

To our knowledge, no comprehensive study provides real insight into the place of languages worldwide on the web, social networks, messaging, chat, and other online services. However, we can gather from information fragmented across various studies that the production of social networks is in straightforward acceleration compared to the production of traditional web pages, even if content is often ephemeral²². Studies conducted by Semiocast²³ on Twitter in 2010, for example, showed that Malay and

21 Daniel Prado, “Languages and Cyberspace: Analysis of the General Context and the Importance of Multilingualism in Cyberspace”, In: *Proceedings of the International Conference Linguistic and Cultural Diversity in Cyberspace*, (Yakutsk, Russian Federation, 2-4 July 2008).

http://www.ifapcom.ru/files/Documents/multiling_eng.pdf

22 The observatory site Portalingua <http://www.portalingua.info>, created by the Union Latine, attempts to address this problem by compiling and paralleling existing studies and statistics on language presence in various spheres of the Knowledge Society.

23 http://semiocast.com/static/downloads/Semiocast_Half_of_messages_on_Twitter_are_not_in_English_20100224.pdf

Portuguese were used much more than Spanish, German, Russian and Italian – languages with much more presence on the traditional web, with traditional translation policies and a more solid book digitization. This study has not been repeated since, but languages spoken in Indonesia, for example, might be even more prevalent today because their country has the planet’s third largest “Generation 140”²⁴. We are seeing similar phenomena on Facebook. A study from the advertising industry (to be taken with a grain of salt for this type of census) described an increase of 175% in Arabic speakers on the social network, compared to a stagnating growth of Anglophones (45% increase in 2011)²⁵. Which is not without social, political and generational consequences in the relevant parts of the world²⁶.

That is why it is important to emphasize the aspects of infrastructure, technology ownership and content production (both text and multimedia). The promotion of language use in the relevant national or regional entities should occur at all levels: educational, administrative, scientific, technical and recreational. But it must be undertaken at the foundation, at the level of access to technology by a given language, and on the basis of reliable data.

However, documented studies and systematic observation of languages on the web are rare. We can’t really identify any, apart from the initiative of the Union Latine and Funredes, conducted from 1998 to 2007²⁷, and the Language Observatory Project²⁸. Yet we are faced with various new factors; specifically, the internet’s extreme size and search engines’ latest developments that have lost their seriousness about information retrieval and the possibility of covering a large proportion of content, and that can no longer obtain reliable and universal statistics. Social uses (Twitter,

24 The name bestowed on the Twitter-using generation, since Twitter doesn’t allow posting messages with over 140 characters.

25 Mathieu Olivier, “Facebook: Arabic Will Soon Overtake English” (“Facebook: l’arabe supplantera bientôt l’anglais”), *Jeune Afrique*, 22 July 2011. <http://www.jeuneafrique.com/Article/ARTJAWEB20110722101339/algerie-liban-internet-facebookfacebook-l-arabe-supplantera-bientot-l-anglais.html>

Carrington Malin, *Rising Facebook Arabic*, 8 June 2011. <http://www.spotonpr.com/facebook-arabic-uprising/>

26 See in this book: Adel El Zaim, *Cyberactivism and Regional Languages in the 2011 Arab Spring*.

27 See the study online at: http://dti1.unilat.org/LI/2007/fr/resultados_fr.htm

28 See in this book: Yoshiki Mikami & Shigeaki Kodama, *Measuring Linguistic Diversity on the Web*.

Facebook, voice and video Instant Messaging) and multimedia sites often escape web-focused analysis. This deficiency in the crucial domain of online language presence indicators has pushed the network Maaya to propose the creation of an international consortium that would conduct an ambitious “language cybermetric” in a project called Dilinet.

The project involves developing a set of methods to produce sustainable indicators of online linguistic diversity that account for the web’s diversity, to provide a basis for public policy in all areas related to the information society at national and international levels.

This project takes an exploratory approach to research, taking into account existing methods by adding innovative approaches, including measurement systems reflecting user behaviour. To overcome the limitations created by the web’s size, the project uses optimal indexing methods based on mathematical approaches that are neither sequential nor random, while opening new avenues like voice identification techniques or automatic content characterization.

Dilinet is transversal, at once a field of research (with researchers from both public and private sectors) and tool for defining public policy. It begins with the motivation of a group of international agencies including Unesco, the International Telecommunication Union, the Organization Internationale de La Francophonie and the Union Latine, and draws on the experience of the aforementioned Funredes and the Language Observatory Project.

By providing verified and valid results for a set of indicators of linguistic diversity in cyberspace, Dilinet will be able to pick up on trends and gauge the results of implemented policies.

Effectively measuring linguistic diversity in the digital world will contribute to a paradigm shift in how we envision the digital divide, by substituting material measurement (networks and terminals) with the perspective of access to content. Indirectly, Dilinet will also open promising new perspectives for the production of impact indicators on the information society, and will create opportunities for languages to be considered important parameters of the digital economy.

BIBLIOGRAPHY

[PIMIENTA 2009] D. Pimienta, D. Prado, Á. Blanco, *Douze années de mesure de la diversité linguistique sur l'Internet: bilan et perspectives*, Unesco, 2009. <http://unesdoc.unesco.org/images/0018/001870/187016f.pdf>.

[UIT 2009] UIT, *Measuring the Information Society: the ICT Development Index*, ISBN 92-61-12831-9, 2009.

[PAOLILLO, PIMIENTA, PRADO ET AL. 2005] J. Paolillo, D. Pimienta, D. Prado, et al. (2005), *Mesurer la diversité linguistique sur Internet*, Unesco, 12/2005. <http://unesdoc.unesco.org/images/0014/001421/142186f.pdf>.

[SUZUKI, MIKAMI ET AL. 2002] Suzuki I., Mikami Y., et al. (2002), "A Language and Character Set Determination Method Based on N-gram Statistics", dans *ACM Transactions on Asian Language Information Processing*, Vol.1, n°3.

[PRADO 2010] Prado, D. (2010) « Languages and Cyberspace: Analysis of the General Context and the Importance of Multilingualism in Cyberspace » In : *Proceedings of the International Conference Linguistic and Cultural Diversity in Cyberspace*, (Yakutsk, Russian Federation, 2-4 July, 2008). http://www.ifapcom.ru/files/Documents/multiling_eng.pdf

**THE FUTURE
SPEAKS,
READS
AND WRITES
IN ALL
LANGUAGES**

CONCLUSION



ADAMA SAMASSÉKOU, President of Maaya.

ADAMA SAMASSÉKOU

THE FUTURE
SPEAKS, READS
AND WRITES IN
ALL LANGUAGES

Language is the vehicle through which we express our thoughts and communicate. It allows us to share our cultural experience. Our linguistic repertoire and choice of expressions determine who we are in a given moment's place and circumstances. Languages are thus the living expressions of individual and collective cultural identities.

Hampate Ba said that of all the elements that characterise the individual, from physique to clothing, language remains the most salient and obvious.

Because it transcends the individual in favour of the community, language belongs to us as much as it belongs to our culture. Through language, we acquire and transmit our knowledge and expertise, which enables us to exert a certain amount of control over our environment. Both the essence and the barometer of our development, language is identity's most fundamental component.

WELLSPRINGS OF CREATIVITY

At once the cradle of culture and matrix of creativity, language is the preferential tool for building knowledge and know-how. In this respect, it is undeniably one of the foremost expressions of a people's creative genius. Language is culture's receptacle and vehicle. It is the vector par excellence for the cosmic visions of human societies.

Evoking this crucial issue, Professor Joseph Ki-Zerbo, in his book *À Quand l'Afrique?* ("When Will Africa?"), underscores, "*The language issue is*

fundamental because it affects the identity of peoples. And identity is necessary for development and for democracy. Languages also affect culture, the nations' problems, the capacity to imagine and creativity. When we repeat in a language which is not originally our own, there is a mechanical and mimetic expression of ourselves, with some exceptions. (But do we govern for exceptions?) We are only imitating. Whereas when we express ourselves in our language, our imagination is released"¹.

For his part, Raymond Fox, in his essay, *Une éthique pour la francophonie. Questions de politique linguistique* ("Ethics for the Francophonie. Linguistic Policy Issues"), stresses the importance of safeguarding linguistic diversity: "*There is a reason, as fundamental as the identity's foundation, for wanting to safeguard linguistic diversity: it is through their own language that individuals view the world and interpret its meaning in their own way, which ensures their access to universal. All languages are involved in the interpretation of the universal, because every culture produces meanings of universal value. And as has been well demonstrated by Alain Touraine or Stephen Wurm, no language or culture can claim to represent the universal, but each provides its own contribution; it is through cultural and linguistic dialogue that we approach one another*"².

LINGUISTIC AND CULTURAL DIVERSITY

Linguistic and cultural dialogue is advocated by Unesco in "*The Universal Declaration of Unesco on Cultural Diversity*"³ – a founding text if there ever was one – which considers culture to be the whole of a society's or a social group's distinctive spiritual, material, intellectual and emotional features. Defined as such, culture remains at the heart of all debates about identity, social cohesion, and the development of any economy based on knowledge and know-how. That's why it is a common heritage of mankind.

Because every society and social group is called upon to live its culture, to preserve as much as possible its ancient cultural values, cultural and linguistic diversity is a reality at individual, community, national and

1 Joseph Ki-Zerbo, *À Quand l'Afrique ?* Éditions de l'Aube, 2003, pp. 81-82.

2 Raymond RENARD, *Une éthique pour la francophonie. Questions de politique linguistique*, Paris, Didier Erudition, Edition du CIPA, 2006.

3 Adopted by the 31st session of the Unesco General Conference on 2 November 2001. <http://unesdoc.unesco.org/images/0012/001271/127160m.pdf>

global levels. Accepting this reality endows every individual, community, and society with fundamental human rights, as stipulated in the Unesco Declaration, including :

- The right to express oneself, to create and disseminate one’s work in the language of one’s choice, especially in one’s native language ;
- The right to education and quality training that fully respect one’s cultural identity ;
- The right to participate in the cultural life of one’s choice ;
- The right to practice one’s own cultural practices within the limits imposed by respect for human rights and fundamental freedoms.

Preserving and promoting diversity and cultural and linguistic pluralism in all areas remains, in this view, an absolute necessity.

Rooting each social group or language in its own cultural values, far from being a source of division, actually reinforces the mobilisation of social forces. As the ancients said, there is no better knowledge than self-knowledge ; to understand the other, one must first know oneself. Mutual understanding, beginning within one’s society, leads to social cohesion. This type of learning, which is acquired in one’s original environment, also allows for a universal construct.

In Africa, this attitude towards coexistence remains common, due to the conservation of basic educational and social values that promote understanding, respect and mutual consideration between individuals. These community values, which include solidarity, sharing, moderation, consensus, mutual aid, welcoming the Other, and hospitality, allow for the integration of behaviours and concepts that can prevent and manage conflict.

The question of cultural and linguistic diversity, a philosophical and political choice firmly rooted in the African world view, is well summarised by the Malian, and more largely the African, writer and philosopher Hampate Ba, in this quote : *“The beauty of a carpet lies in the variety of its colours. If it is only white, it would be a white cloth, if it is black, it would be a mourning loincloth. The entire Universe is our homeland. Everyone is a page in the Nature register. In the vast human community launched to find a new equilibrium, each people must bring the note of his own genius,*

so that the whole can be enriched. Everyone must be open to others while remaining himself”⁴.

David Crystal wrote: “Diversity occupies a central place in evolutionary theory because it enables a species to survive in different environments. Increasing uniformity holds dangers for the long-term survival of a species. The strongest ecosystems are those which are most diverse. The need to maintain linguistic diversity stands on the shoulders of such arguments. If the development of multiple cultures is a prerequisite for successful human development, then the preservation of linguistic diversity is essential, because cultures are chiefly transmitted through spoken and written languages”⁵. Multilingualism is one of the best ways to preserve this diversity.

MAAYA : HUMANITUDE

This is why we must resolve to affirm the cultural and linguistic heritage of different countries, by taking inventory of and promoting their languages and cultural and artistic achievements. This will lead to fruitful research and discussion of cultural and linguistic values. It will enable us to entrench development processes in diversity, and in the endogenous strengths of our multifarious cultures and societies.

Indeed, making the choice to preserve linguistic diversity means accepting to substitute the destructive logic of market competition, with the logic of solidarity and complementarity, which can restore harmony between beings and species. We know that cultural and linguistic diversity is for human society what biodiversity is for nature: the ferment, the bedrock of what I call our humanness, our permanent opening to the Other, our relationship as human beings to be human, which requires a permanent relationship of solidarity, without calculation, a spontaneous impulse towards the Other... this “*humanitude*”, or humanness, that links each person to the next, as articulated in the beautiful expression of our dear Elder Aimé Césaire!

It is through this concept of *humanitude* that I translate what we call in Africa *maaya* (in the Mandenkan language), *nedd̄aaku* (in Fulfulde),

⁴ See Adama Samassekou, “Archives, Public Policy and Research Evidence: The Issue Of Language”, Human Sciences Research Council. <http://www.hsrb.ac.za/Document-3177.phtml>

⁵ David Crystal, “Millenium Briefing: The Death of Language”. *Prospect Magazine*, November 1999.

boroterey (in Songay), *nite* (in Wolof), *ubuntu* (in the Bantu languages)... and in Mande we say, “*I am a human being not because I think, but rather it’s your eyes fixed on me that make me a human being!*”.

Promoting multilingualism thus means preserving the vitality of human societies, developing the uniqueness of human communication, strengthening the dialogue between cultures, between people, opening ourselves to the Other, and fighting against violence and ethnic tension.

TOWARDS A MULTILINGUAL CYBERSPACE

Concerning the challenges of multilingualism in cyberspace, and the problematics of global governance, the language approach developed in these pages can serve as a guide and a goal.

Since each language reflects its speakers’ worldview, the world’s linguistic diversity reflected in a multilingual internet can save humanity from the mechanical abuses of a disembodied digital culture.

Promoting multilingualism in cyberspace can help slow the development of “individual communication cells” and group isolation, while strengthening the social ties that are threatened by the dominant development model.

The multiplicity of languages, and the ability to speak several languages, is a powerful factor bringing individuals and communities together. In the new knowledge society of shared knowledge that is under construction, ensuring a multiplicity of languages throughout cyberspace will facilitate the implementation of a multilingual and democratised digital culture for every citizen of world.

Multilingualism is to culture what multilateralism is to politics. One of the essential challenges of a multilingual culture is to contribute to the emergence of a new global governance partnership that will help preserve global peace and safeguard humanity.

Graphic Design and Page Layout: Nicolas Taffin and Kathleen Ponsard.

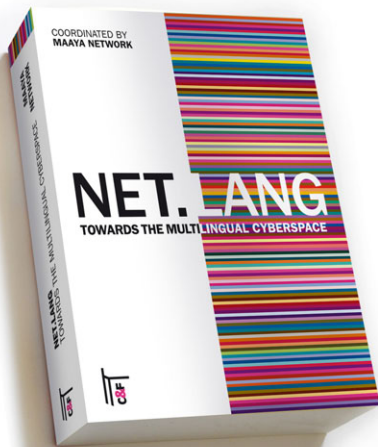
Work composed in *Minion* by Robert Slimbach and *Franklin Gothic* by Morris Fuller Benton. Two classics that are the “little black dresses” of typography for Latin script.

Arabic and Russian are rendered in *Times New Roman*. Devanagari writing is composed in *Devanagari Sangam*, Myanmar and Thai in *Zawgyi-One*, a Unicode font developed and maintained by Computer Alpha at Copyleft (<http://code.google.com/p/zawgyi/wiki/MyanmarFontDownload>).

Kannada writing uses *Kedage*, a Unicode font adapted by Nicholas Shanks from a font developed by the Indian Institute of Science in Bangalore (<http://web.nickshanks.com/fonts/kannada/>). This free font is still incomplete, and seeks online contributions from printers of goodwill.

Covering all the tables of the Unicode project with high-quality glyphs is an enormous task in and of itself, but to provide on top of that an aesthetic choice and a variety of styles is a collective challenge. Typographic design requires great precision, and a text’s readability ultimately depends on it, in all cultures, in all writing systems.

Dépôt légal 1^{er} trimestre 2012.
ISBN PDF edition 978-2-915825-24-4
<http://cfeditions.com>



NET.LANG IS AVAILABLE IN VARIOUS LANGUAGES AND FORMATS :

In print (French or English)

446 pages, 17 × 22.5 cm, softcover

Price : 34 euros

French : ISBN 978-2-915825-08-4

English : ISBN 978-2-915825-09-1

In bookstores or at <http://cfeditions.com>

eBook (French or English)

DRM free

French : ISBN 978-2-915825-25-1

English : ISBN 978-2-915825-26-8

<http://net-lang.net>

PDF (French or English)

DRM free

French : ISBN 978-2-915825-23-7

English : ISBN 978-2-915825-24-4

<http://net-lang.net>

**FOR OTHER VERSIONS AND TRANSLATIONS
VISIT [HTTP://NET-LANG.NET](http://NET-LANG.NET)**

NET.LANG TOWARDS THE MULTILINGUAL CYBERSPACE is an educational, political and practical guide to policy and practice to understand key issues of multilingualism in cyberspace. Multilingualism is the new frontier of the digital network. This book presents the issues, and provides suggestions for a cyber-presence that is equitable among languages in the information society.

This book is addressed to everyone involved in the political, economic, cultural or social life of their community: confronted by the growth of cyberspace and questioning the scope of linguistic diversity in the digital environment. Promoting one's language, or language policy, teaching or working in a language, translating or interpreting languages, creating content in multiple languages, contacting people from other languages... are all situations where this book offers guidance. The vitality of multilingualism is a force both for the development of the internet for building inclusive societies, sharing knowledge and working with the objective of living well together.

Maaya / the World Network for Linguistic Diversity is a multilateral network created to contribute to the development and promotion of linguistic diversity worldwide. In terms of the Bambara language, Maaya could mean the neologism "humanitude". The Maaya Network was established following the World Summit on the Information Society (WSIS), in which the cultural and linguistic diversity in cyberspace was identified as a priority. Maaya was founded by the African Academy of Languages (ACALAN), under the auspices of the African Union. <http://www.maayajo.org/>

This book is a production of the Maaya Network, with support from Unesco, Communication and Information Sector, Organisation internationale de la Francophonie, Union Latine, ANLoc, and IDRC/CRDI.



With the support of
**Communication and
Information Sector**

ISBN (PDF ENGLISH EDITION) : 978-2-915825-24-4

<http://cfeditions.com>