The Multilingual Internet

For the first time in human history, the thousands of languages that give meaning to human experience are going digital—changing the economics, politics, and culture of global exchange.

Even with all the changes wrought so far by the World Wide Web, we haven't seen anything yet. This decade will witness the most spectacular flowering of human knowledge in our history as, for the first time, hundreds of millions of people are able to digitize their spoken and written knowledge. Digital interactions will come closer to our various spoken languages. Vast amounts of human experience—much of it contained within marginalized groups who have never expressed their cultural heritage in print—will come online, creating a common pool of diverse ideas about human relationships to one another and to the natural world. Think of linguistic diversity as the cultural equivalent of biodiversity: the result of tens of thousands of years of small evolutionary steps, the literal encoding of human wisdom. This knowledge will be critical as we struggle to find a way to live sustainably on the planet.

—Lyn Jeffery

critical balances

DISINTEGRATION integration	Even as demands for the localization of content and commerce fragment the global marketplace, tools for transcending language barriers proliferate.
EXPOSURE accountability	Automatic language-to-language translation exposes local information, resources, and culture—as well as corruption—to global audiences.
CONTAGION isolation	Networking within and across local language cultures speeds the spread of socially contagious behaviors across formerly isolated locales.

language: the multilingual intern

CODING DIVERSITY:

THE END OF THE ASCII ENGLISH INTERNET

Despite its name, the World Wide Web has been anything but. Started as an American military project, global Internet use has nevertheless grown exponentially: today, almost 29% of the world goes online at least once a month. But what English speakers consider "the web" is neither intuitive nor friendly for hundreds of millions of users (and billions of potential users). Half of all current Internet users do not natively use a Latin script. For them, the URL addresses they need to navigate the web—which they must enter on keyboards that may not even have Latin scripts—have all been experienced with incomprehensible suffixes like *.com* and *.org*. If this seems like a trivial matter, consider what you would do if you were handed a business card with this email address: 江淋@未来研究所。中国

After 40 years, all this is changing. Internet activists representing a variety of non-Latin scripts have succeeded in their fight to force the Internet Corporation for Assigned Names and Numbers (ICANN)—the organization that regulates the core code that keeps the Internet interoperable—to enable non-ASCII scripts at the level of ".com" or ".org" extensions.

Known as International Domain Names (IDNs), the first four, from the United Arab Emirates, Saudi Arabia, Russian Federation, and Egypt, were introduced in 2010 as امارات., السعودية., рф, and مص, respectively. (Note that Arabic script domain names read from right to left). By June 2010, 21 countries had submitted applications to ICANN, representing 11 languages. These developments pave the way for the majority of the world to take advantage of the Internet in a way that has simply not happened yet. The growth of non-English content, user interfaces, and domain names will create more locally relevant electronic cultural identities and ultimately a more level playing field for self-expression on the worldwide stage.

> "We need to build cloud-based, multilingual collaborative spaces that will allow people to exchange information across borders, across cultures, and across languages. It's very, very important for software engineers to realize this and to start working towards creating virtual newsrooms and creating virtual spaces between languages—spaces that would allow instantaneous translation, not only between English and other languages, but between other languages themselves."
>
> **Paul Radu**
> **Cross-Border Journalist, Organized Crime and Corruption Project**

> "Text-to-speech and speech-to-text technology will allow colleagues that otherwise were unable to share their information to share their findings and to share their interests. So this is an amazing opportunity for science to really take a big leap forward. Just having access to the Internet was a massive breakthrough for scientists. If you were studying *nudibranch*—sea slugs—you could create a community around that, and just be sitting at your desk. Really, what we'll see is an extension of that process. Those networks will be bigger. They'll be more inclusive. There will be faster communication, much like what happened when the Internet just began to connect scientists."
>
> **Wallace "J" Nichols**
> **Ocean Scientist and Activist, Ocean Revolution**

RE-CODING LITERACY:
TEXT-TO-SPEECH CONVERSION

Creating new paths to non-Latin domains will transform how the Internet feels and what it means to people around the globe. But new ways of displaying text are only part of the story. If you can't read, it doesn't matter what script is on the screen.

The widespread adoption of technological tools that provide easy two-way bridges between spoken words and printed text is changing that, too. A cluster of innovations in voice recognition, speech synthesis, optical character recognition, and human-assisted or autonomous machine translation is driving speech-to-text and text-to-speech conversion to new levels of ease and accuracy.

In the digital world, oral and written communications will become more fungible, more capable of mutual substitution, giving oral communication the sharable and searchable qualities of print and making print audible for those who can't read it. In addition to helping people communicate in real-time across languages using voice or text inputs, these technologies will open up the world of knowledge stored in print to hundreds of millions of people who don't read well due to educational disadvantages, visual challenges, or computer illiteracy.

> "Increasing resilience worldwide will come largely from enabling platforms and tools to examine, proliferate, and preserve existing languages. We've seen examples like websites such as Facebook that have crowdsourced the ability to translate and localize their site into a given language by giving its users the tools to take that site and make it their own and translate it in a manner that fully captures the colloquialisms better than any mechanical tools. Once these tools exist, it almost intrinsically fosters happiness among that community by being able to find something that is locally yours and something that you can then build upon and have ownership and agency over."
>
> **Mike Lin**
> **CEO, Fenix International**

CODING EQUITY:
THE RISE OF LINGUISTIC HUMAN RIGHTS

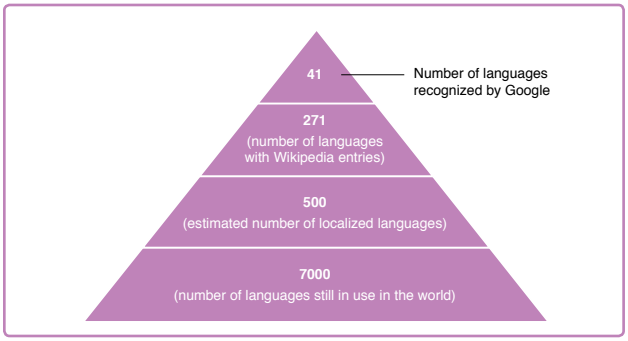
Linguists estimate there are between six and seven thousand languages currently spoken in the world, over three hundred of which have at least a million speakers each. The vast majority are so-called "minority languages" that differ from the dominant or official language of a nation. A substantial number have never been committed to paper. People who speak these languages are often forced to navigate a world that doesn't match the language they speak at home, encountering a variety of obstacles (both practical and often legal) to literacy, education, and professional mobility. According to the World Bank, 50% of the world's out-of-school children live in communities where the language of schooling is rarely, if ever, used at home.

The digitization of these languages, however, will be made possible in the next decade by recording devices on cheap mobile phones and lightweight video cameras, connected to cloud-served databases, semantic analysis, and human-assisted machine translation. In effect, the tools of social connection are also becoming the tools of language preservation.

Bringing minority language speakers into the digital world will empower community identity for people such as the millions who speak Quecha, the most common language of the indigenous people of the Americas. Eventually it will open up the kinds of cross-border sharing, collaboration, and business generation that English-language and other majority language speakers have benefited from over the past several decades. As these languages literally create new platforms for human exchange, they will create new avenues for participation in global and local economies.

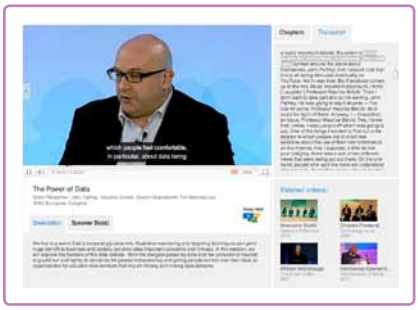
Along with the changes to the competitive landscape in the marketplace, these technologies will also highlight increasingly salient political and cultural issues by politically empowering marginalized linguistic and cultural communities, sometimes to the dismay of dominant groups. A new movement of linguistic human rights is likely to emerge, and we'll find growing protections—and tools—for the right to take advantage of new connective technologies and the knowledge they embody without ever having to use Latin scripts or majority languages.

itu.int/dms_pub/itu-d/opb/ind/D-IND-WTDR-2010-PDF-E.pdf



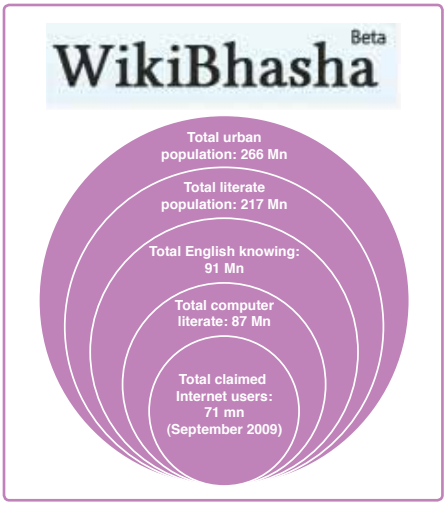
According to the International Telecommunication Union, 52 languages were available for translation in Google Translate in early 2010, and the Internet browsers Internet Explorer and Mozilla supported 63 and 70 languages, respectively. Furthermore, many "international" websites (such as Amazon, eBay, Google, and Facebook) now propose localized versions, often also in local languages. Thus, Facebook was available in 67 languages, Blogger in 50, YouTube in 19, Flickr in 8, Twitter in 6, and Linkedin in 4 languages. The 140 scripts that have thus far been digitally encoded have been used to support localized content in approximately 500 languages.

zeitgeistminds.com/videos/the-power-of-data



Google's online repository of oral presentations at its annual Zeitgeist conference showcases the company's rapidly evolving speech-to-text capabilities. Some of the videos come with a transcript whose text can be selected to correspond to the same moment of the video. The transcript also scrolls alongside as the video plays. Says Franz Och, Google's head of translation services, "We think speech-to-speech translation should be possible and work reasonably well in a few years' time."

iamai.in/Upload/Research/vernacularreport_44.pdf
research.microsoft.com/en-us/news/features/wikibhasha-101810.aspx



Data from the Internet and Mobile Association of India (IAMAI) illustrate the challenge of "getting everyone online" in India: a vast gap exists between Indians who are literate in non-English, non-Latin script language, those who know English, the computer literate, and Internet users. The technical solutions for supporting minority languages will almost certainly involve a combination of humans and machines: researchers at Microsoft Research India have developed a crowdsourced human-machine translation tool called Wikibhasha for languages that don't already have a large enough corpus online to translate. It uses machines to begin translations and an online community of experts to teach the machine without the benefit of a large database.

the quick list

- *The State of the World's Indigenous Peoples,* Department of Economic and Social Affairs. New York, United Nations, 2009. www.un.org/esa/socdev/unpfii/en/sowip.html
- "In Their Own Language…Education for All," *Education Notes*, Washington, DC: World Bank, June 2005, p.1
- "Kieren McCarthy [dot com]," by Kieren McCarthy, IT journalist and former general manager of public participation for ICANN, www.kierenmccarthy.com

Cover source: flickr user Firas