Imagining a multilingual cyberspace

We have been told that Facebook will “bring the world closer together” and that Google will “organise the world’s information and make it universally accessible and useful.” In 1996, digital rights activist John Perry Barlow dreamed that Cyberspace would be “a world that all may enter without privilege or prejudice accorded by race, economic power, military force, or station of birth.” Yet it’s now 2019, and it’s not difficult to imagine a world in which only ten per cent of our languages remain, the rest having gone extinct under a system where they are considered superfluous, antiquated, and not worthy of digital support.

Cyberspace exists in a world where linguistic diversity is on the decline. Linguists predict that out of the 6000+ languages spoken today, 50 to 90 per cent face extinction this century. Many factors contribute to this trend – such as imperialism, globalisation, and urbanisation – but digital technologies also appear to play a role. Although digital tools have the potential to support all languages and the scripts used to write them, currently only the most internationally and financially dominant languages enjoy robust digital support. Things you take for granted – fonts, keyboards, spellchecks, autocorrects, and voice activation tools – do not cater for everyone.

In his 2013 article *Digital Language Death,* the computational linguist András Kornai calculated that at most 5 per cent of languages will achieve digital vitality, or active use in the digital sphere. In 2015, Unicode president Mark Davis estimated that 98 per cent of languages are digitally-disadvantaged, that is, not supported on leading devices, operating systems, mobile apps, and browsers. As communication throughout the world becomes increasingly mediated by digital technologies, technology can impact whether or not a language survives the digital age at all. While cyberspace was built on the promise of serving all the people of the world, without a rapid course correction we may instead find that language diversity is, to some degree, a collateral damage of digital tech.

“Digital language extinction” will be the fate, computational linguist Georg Rehm stated in 2014, of languages that suffer from insufficient technological support. This is a three-pronged process. First, if digital technologies prevent or dis-incentivise the use of a language, its community will suffer a loss of “function” as other languages take over tasks such as email, texting, and search. Second, a simultaneous loss of “prestige” is associated with the absence of a language within the hip, cutting-edge technological realm. Thirdly, a the loss of “competence” occurs as it becomes increasingly difficult to raise a “digital native” competent in a language.

Declining language diversity comes out of asymmetries of power, and often histories of brutal oppression. Yet what is the best way for at-risk language communities to move forward? While it is important that all language communities (that wish to) have access to what literary scholar Mary Louise Pratt (in her lecture *Toward a Geolinguistic Imagination*) called “languages of power,” this fluency does not necessarily have to come at the expense of a community's mother tongue. And although many cultural forms can survive a language's extinction, language extinction is often accompanied by a loss of identity, inter-generational cohesion, and a wealth of knowledge to address future problems facing humanity, as documented by linguist K. David Harrison's in his 2007 book *When Languages Die*.

Some language communities may embrace the fact that their language will "sleep," to borrow language Pratt has heard in communities that have embraced this more open-ended and positive metaphor over the term "language death." Yet even in these cases, basic digital supports for the language can help the process of recording and archiving information about (and in) the language for future generations. An example of language revitalization that has drawn on such documentation can be found within the Myaamia or Miami language community. While the last native speaker of Myaamia died in the 1960s, in the 1990s, a team of researchers put together an online dictionary, enabling a mini-revival of the language.

Digital tools should serve the needs of all language communities, even if the future of their language is uncertain. A stark example is the Ebola outbreak, which linguist Suzanne Romaine has called not just a health crisis, but a communication crisis, in her 2018 lecture *Linguistic Diversity and Sustainability: Global Language Justice Inside the Doughnut*.

Educational posters distributed in affected communities in Liberia were printed in the country's official language, English, a language which only 20 per cent of the country speaks. In fact, there are 31 languages spoken in Liberia, some written in non-Latin scripts like Bassa, Vai, and N'Ko. Had awareness about the linguistic diversity in these areas been coupled with graphic design tools that support these languages and scripts, lives might have been saved.

So why don't digital technologies adequately support language diversity? Historically, the first computers were built in English-speaking contexts in the US and the UK around the middle of the 20th century. As demand grew for computers in Europe, it was relatively easy to expand support for the additional Latin characters used by other European languages. Over time, digital innovation and support for languages written in other scripts was developed in countries such as China, Korea, Japan, and India. This gradual expansion has continued, with milestones including Keyman and Google's smartphone keyboard apps now supporting more than 600 language varieties, and Google Translate and Voice Dictation now both available in more than 100 language varieties, according to Daan van Esch, a technical program manager at Google AI.

While these gains are important, the drive for corporations to expand support to more languages has mostly been to meet the needs of new "markets," a process which has largely excluded small and poor language communities.

As computational linguist and cultural anthropologist Benjamin Martin pointed out in his 2016 conference presentation entitled *Digital Language Diversity: Seeking the Value Proposition*, the gap in digital supports for other languages has been largely invisible to native speakers of dominant languages. We often take technological affordances for granted and don't consider how speakers of other languages struggle to use these technologies. Benjamin argues that, simultaneously, speakers of digitally-disadvantaged languages often do not imagine that the digital tools they use or aspire to use have the potential to support their native language and script.

So how do speakers of digitally-disadvantaged languages make do with the digital tools available to them? There are a few common patterns. For communities that are diglossic, meaning most speakers are bilingual, people may simply opt to use the "language of power" they know, such as English, since the technology is well adapted to it. For example, if they choose to send a text message in English, it's highly likely that a standard QWERTY keyboard is their phone's default, and that auto-corrector works in their favour.

For users that are not bilingual in a globally dominant language, or who are communicating with someone who is not, it is also common to sound out and type words phonetically in the Latin alphabet. This is called transliteration. For example, the Amharic word for peace is written ሠላም, but transliterated would be "selam." Transliteration can also be done on the basis of changing a script's characters into Latin letters that look similar. Transliteration is a workaround that allows speakers of digitally-disadvantaged languages written in non-Latin scripts to use devices and applications designed for the English language.

Transliteration into the Latin alphabet is exceedingly common, even among communities that enjoy fairly robust digital supports, such as Arabic-speaking communities. Transliterating Arabic into Latin characters is so common it is called "Arabish", "Araby", or "Arabizi". For example, تحكي عربي؟, or "do you speak Arabic?" becomes "ta7ki 3arabi?". The national language of Amharic, written in the Ethiopic script, is one of roughly 100 languages that enjoy support on Facebook, yet my dissertation analysis in 2017 revealed that more than half of Amharic comments on popular Ethiopian-themed Facebook pages are still transliterated into Latin.

As a workaround, transliteration has a number of disadvantages. First, it is very difficult to read one's language in a non-native script. Imagine trying to read English written in the Cyrillic script, in which case the word "English" would be written Енглисх. Second, transliterations rarely follow any spelling standard, and therefore are doubly difficult to read. Third, users have to continuously battle with functions such as auto-correct and spellcheck that try to "correct" their language into English. Fourth, it is virtually impossible to effectively "search" archived content that is transliterated, due to its non-standard spelling. This deprives users of the ability to dig back and find relevant emails or files, or search for information online.

The impact of these shifts extends outside the digital sphere. Transliteration can degrade a community's fluency in their own writing system over time. In this way, it fuels the logic that communities should abandon their own script in favour of the Latin alphabet. It is not just languages that are at risk of extinction, but scripts as well.

It is also essential to note that digitally-disadvantaged language users do develop their own digital tools. All over the world, digital pioneers, often self-taught, created mechanisms for digital use of their own language and writing system. One example is Fesseha Atlaw, an Ethiopian who lives in California, who developed the first word-processing software for the Ethiopic script and the languages that use it, such as Ethiopia's national language Amharic. Many of these "legacy" digital tools, such as Atlaw's, were created at the advent of the age of personal computing from the mid-1970s to the mid-1980s. These tools were often used to connect tight-knit linguistic communities.

Yet these legacy programs typically lacked interoperability with other systems since they didn't share a common "standard," that is, an encoding of one numerical identifier for each written character that all software, devices, and platforms share. Lack of a common standard means that a document from one software, opened in another, will appear as empty boxes or wingdings. As such, legacy systems, which played an important role at the time of their invention, have often been replaced by updated tools that meet the digital-linguistic needs of their communities in ways that are interoperable by using the Unicode Standard.

It is good news for digital language diversity that Unicode has largely solved the problem of interoperability by making remarkable progress towards a universal character encoding standard since it incorporated in 1991. The Unicode Standard, as well as its synchronized sister-standard ISO/IEC 10646, in its 12.0 version includes 150 scripts, 91 modern and 59 historic, as well as 2884 emoji characters.

The Unicode Standard is free to use and is accompanied by abundant information to make implementation easy for programmers. Unicode now underpins the vast majority of software, websites, and devices, making them interoperable and capable, at least in theory, of supporting the majority of the world's languages. Unicode support is foundational, but "full stack support," such as fonts, keyboards, spellcheck, and voice activation, must also be included in each application to make it truly accessible.

An important player in Unicode's expansion to support digitally-disadvantaged scripts and their languages is the Script Encoding Initiative (SEI) at University of California at Berkeley. SEI has played a major role in the encoding of 106 new scripts or additions to character sets in Unicode and ISO/IEC 10646 since its founding in 2002. SEI is currently working on an additional 146 modern and historic scripts and additions. SEI's secondary goal is to help make the scripts in Unicode accessible to users via fonts, keyboards, and software updates. SEI's website states, "For a minority language, having its script included in the universal character set will help to promote native-language education, universal literacy, cultural preservation, and remove the linguistic barriers to participation in the technological advancements of computing."

For those who share a common vision for a future in which cyberspace supports the world's language communities, what are positive steps that we can take? Different stakeholders each have a role to play. Here are a few ideas:

Community language advocates can:

- Contact the Script Encoding Initiative if your script is not included in Unicode
- Connect native speaker experts with tech innovators, so that digital supports accurately reflect the structure and richness of your language
- Raise awareness within your community about existing digital supports
- Develop ways to increase use and prestige of your language in the digital sphere through initiatives that appeal to youth, possibly led by youth, such as "Mother Language Twitter Day"
- Find and connect with digital pioneers through platforms like Indigenous Tweets
- Reach out to the internationalization (i18n) & localization desks of major tech companies to request better or new supports.
- Provide information to CLDR (Unicode Common Locale Data Repository), so that devices in your community will automatically sync to linguistically and culturally relevant defaults.
- Work with W3C projects to define text layout requirements, including work on Ethiopic, Hebrew, Indic, Arabic, Hangul, and Mongolian, among others.

Major IT companies should:

- Embrace multilingual support as a part of your social responsibility to serve all communities
- Find ways to reward the crowd-source volunteers you rely on to build multilingual tools. Rewards could be financial (for example, a chance at scholarship funds or a fellowship at your headquarters) or an award that would help contributors access new educational or professional opportunities at home.

Digital Designers should:

- Implement Unicode for interoperability and longevity of the tools you design
- Focus on both linguistic diversity and voice tools for maximum accessibility
- Design platforms and text-layout software compatible with Microsoft's Universal Shaping Engine, which helps support interoperability for texts written in complex scripts.

Academics should:

- Consider that academia is one of the most likely places developments for “non-market” languages can take place
- Bridge disciplines to introduce linguistic expertise into digital endeavors, and vice versa

Digital Governance Institutions (Unicode, ISO, ICANN, W3C, etc.) can:

- Lower barriers to participation for digitally-disadvantaged language communities by making your work processes open and participatory
- Convene in locations where community representatives can attend
- When necessary, work with a trusted third-party non-profit like SEI to create a bridge to communities

Governments must:

- Develop user-centric national standards for keyboards and other digital tools to provide a stable market for manufacturers and a consistent experience for users
- Procure products that implement free, global standards like Unicode
- Make digital supports for national and local languages a funding priority, including funding support for work being done in universities
- Create a regulatory environment that allows local tech firms to thrive

Font enthusiasts could:

- “Decolonize typography” by designing a font for a digitally-disadvantaged script, taking into account its cultural-aesthetic legacy, as well as other scripts with which it co-exists

There is also an important role to be played by institutions that can connect these varied stakeholders and bridge their common interests. This could be a free-standing non-profit, or located within a university or government institute. Such an institution could help pool intellectual resources, avoid duplication of effort, and maintain lines of communication between major IT companies and language experts. It could also advise governments, reward volunteer contributions, and educate the public about digital supports.

My vision for the future internet is one which is truly global and truly local, simultaneously. New tools such as machine translation mean more and more language communities will be able to access information in any language, translated into their own. Communities will be able to communicate without losing their mother tongue, and will also be able to make contributions to the global Internet in their own language.

But this will only be possible if we build foundational supports for language diversity now. Let us work together towards a vision of cyberspace that is both linguistically inclusive and just.

Bibliography:

Anderson, D. W. (2018). *Preserving the World’s Languages and Cultures through Character Encoding*. Presented at the Mellon-Sawyer Seminar on Global Language Justice, Columbia University, Institute for Comparative Literature and Society.

ArabicGenie. (2009). Do You Speak Arabic Chat? ta7ki 3arabi? Arabic Genie.

Barlow, J. P. (1996). *A Declaration of the Independence of Cyberspace*. Davos, Switzerland.

Benjamin, M. (2016). Digital Language Diversity: Seeking the Value Proposition. In *Collaboration and Computing for Under-Resourced Languages: Towards an Alliance for Digital Language Diversity* (pp. 52–58).

Harrison, K. D. (2007). *When Languages Die: The Extinction of the World’s Languages and the Erosion of Human Knowledge*. Oxford University Press.

Kornai, A. (2013). Digital Language Death. *PLoS ONE*, *8*(10), e77056. https://doi.org/10.1371/journal.pone.0077056

Leddy, M. (2018). Beyond “Graphic design is my Passion”: Decolonizing Typography and Reclaiming Identity. Explorations in Global Language Justice.

Leddy, M. (2018). Graphic Design Is My Passion, and Other Unassuming Gateways to Global

Language Justice. Explorations in Global Language Justice.

Leddy, M. (2018). the fedra™ project. lingothèque.

Loomis, S. R., Pandey, A., & Zaugg, I. (2017, June 6). Full Stack Language Enablement. http://srl295.github.io/

Pratt, M. L. (2019). *Toward a Geolinguistic Imagination*. Presented at the Mellon-Sawyer Seminar on Global Language Justice, Columbia University, Institute for Comparative LIterature and Society.

Rehm, G. (2014). Digital Language Extinction as a Challenge for the Multilingual Web. In *Multilingual Web Workshop 2014: New Horizons for the Multilingual Web*. Madrid, Spain: META-NET.

Romaine, S. (2018). *Linguistic Diversity and Sustainability: Global Language Justice Inside the Doughnut*. Presented at the Mellon-Sawyer Seminar on Global Language Justice, Columbia University, Institute for Comparative LIterature and Society.

Unicode, Inc. (2015, December 16). Unicode Launches Adopt-a-Character Campaign to Support the World’s “Digitally Disadvantaged” Living Languages.

van Esch, D. (2019). *Language Technology for the World’s Languages*. Presented at the Mellon-Sawyer Seminar on Global Language Justice, Columbia University, Institute for Comparative LIterature and Society.

Zall, C. (2016). How the Miami Tribe got its language back. *Public Radio International*.

Zaugg, I. (2017). *Digitizing Ethiopic: Coding for Linguistic Continuity in the Face of Digital Extinction* (Doctor of Philosophy in Communication). American University, Washington, D.C.