Digitally-disadvantaged languages

Isabelle A. Zaugg *Columbia University* iz2153@columbia.edu
Anushah Hossain *University of California Berkeley* anushah.h@berkeley.edu
Brendan Molloy *Techno Creatives*

Abstract: Digitally-disadvantaged languages face multiple inequities in the digital sphere including gaps in digital support that obstruct access for speakers, poorly-designed digital tools that negatively affect the integrity of languages and writing systems, and unique vulnerabilities to surveillance harms for speaker communities. This term captures the acutely uneven digital playing field for speakers of the world's 7000+ languages.

DEFINITION

Digitally-disadvantaged languages face multiple inequities in the digital sphere including gaps in digital support that obstruct access for speakers, poorly-designed digital tools that negatively affect the integrity of languages and writing systems, and unique vulnerabilities to surveillance harms for speaker communities. This term captures the acutely uneven digital playing field for speakers of the world's 7000+ languages.

ORIGIN & EVOLUTION OF THE TERM

The term originates with Mark Davis, president and co-founder of the Unicode Consortium, a nonprofit that maintains and publishes the Unicode Standard.[1] In 2015, Davis said, "The vast majority of the world's living languages, close to 98 percent, are 'digitally disadvantaged'—meaning they are not supported on the most popular devices, operating systems, browsers and mobile applications" (Unicode, 2015, n.p.). Computational linguist András Kornai (2013) similarly estimates that at most 5% of the 7000+ languages in use today will achieve "digital vitality," while the other 95% face "digital extinction". Gaps in language access are one facet of the *digital divide* (Zaugg, 2020).

Critical digital studies scholar and co-author Isabelle Zaugg utilises the term *digitally-disadvantaged languages* in her work on language justice in the digital sphere (2017; 2019a; 2019b; 2020; forthcoming). Zaugg (forthcoming) proposes that *digitally-disadvantaged language* communities face three primary challenges: 1) gaps in equitable access; 2) digital tools that negatively impact the integrity of their languages, scripts and writing systems,[2] and knowledge systems; and 3) vulnerability

---

1. The Unicode Standard is a character coding system designed to support interoperable exchange and consistent representation of text in the world's writing systems on digital devices, providing a foundation for a multilingual digital sphere.

2. A language is a shared means of communication, while a script is the collection of written characters used to write a language. A language's writing system incorporates a script and a set of rules regarding its use. Languages and scripts do not have a one-to-one or static relationship. Some languages, such as Kazakh, Mongolian, and Urdu, are written in multiple scripts. Many languages share a script, although the rules of their writing systems may differ. More than 1000 languages are written in the Latin script, including English, French, Czech, Kazakh, Nahuatl, Tagalog, Vietnamese, and Igbo; Hindi, Nepali, Marathi, Bodi, and Konkani are among languages written in the Devanagari

to harm through digital surveillance and under-moderation of language content.

Digitally-disadvantaged languages overlaps and extends upon adjacent terms used in geopolitics and computational linguistics, i.e., natural language processing (NLP). While the category of *digitally-disadvantaged languages* includes many if not all minoritised languages, Indigenous languages, oral languages, signed languages, and endangered languages, it also includes many national and widely-spoken languages that enjoy robust intergenerational transmission.[3] There is no sharp line that delineates whether a language is *digitally-disadvantaged*. Rather, the term captures a relative degree of disadvantage as compared to the handful of languages that enjoy the most comprehensive digital support and wider political advantages. That said, there are stark differences between the levels of support for languages such as English, Chinese, Spanish, and Arabic and even widely-spoken national and regional languages such as Amharic, Bulgarian, Tamil, Swahili, or Cebuano. However, *digitally-disadvantaged* is not a static state; it is possible for a language to "digitally ascend" (Kornai, 2013) through wide-reaching efforts to create digital support for the language and foster digital use among speakers. Cherokee, Amharic, Manding languages written in N'Ko, Fulani written in Adlam, and Sámi are a few languages whose digital ascent has been hastened by concerted advocacy efforts.

The term also overlaps with and contrasts against *low resource* or *under-resourced languages*, NLP terms that refer to languages with sparse data available for analysis. A language may be *digitally-disadvantaged* in part because digital corpora are unavailable to develop machine translation and search functions. Digital corpora often do not exist due to lack of basic digital support like fonts and keyboards that allow speakers to develop online content—a vicious cycle. By focusing on resource deficits, NLP terms shift focus away from how power has shaped the techno-social imbalances that have rendered the vast majority of languages *low resource* in the first place.

In contrast, the term *digitally-disadvantaged languages* captures how languages' digital marginalisation represents how wider linguistic power dynamics map onto the digital sphere. The fact that the earliest digital technologies were developed in the US and UK laid the foundation for English to become the best-supported and default means of digital communication in many contexts (Zaugg, 2017). Illus-

script; Bulgarian, Kazakh, Russian, Tajik are written in the Cyrllic script; while Chinese, Korean, Japanese, Vietnamese, and Miao are written in the Hanzi script.

3. Marked by a high EGIDS score (Ethnologue, n.d.)

tratively, the QWERTY Latin character layout remains the default keyboard all over the world, leading many to write even well-supported languages like Arabic in a transliterated Latin form such as "Arabizi" (Zaugg, 2019a). The global spread of digital tools and systems including QWERTY keyboards, ASCII,[4] ICANN oversight of the originally Latin character-only domain name system,[5] and default English auto-correct have all contributed to the "logic" that English is the global *lingua franca*, and the Latin alphabet the most modern, rational, and universal script.[6] This "logic" in turn builds upon US and UK imperial power that laid the groundwork for the "digital revolution" as well as first brought English and the Latin script to far flung corners of the globe.

Digital advantage for English and the Latin script - and to a lesser degree other dominant languages and scripts - has created a paradigm in which many bilingual or multilingual speakers of *digitally-disadvantaged languages* become habituated to consuming and sharing content in a dominant "bully" language or script.[7] Many *digitally-disadvantaged language* speakers do not imagine that the digital sphere could be equally hospitable to their mother tongue and native script as it is to English and Latin (Benjamin, 2016). Unfortunately, gaps in digital support and use may be contributing to many of these languages' extinction as speakers increasingly use "bully" languages on and offline. Shockingly, 50-90% of language diversity is slated to be lost this century (Romaine, 2015); inequities in the digital sphere appear to be a factor in this shift (Kornai, 2013; Zaugg, 2017; Zaugg, 2019a; Zaugg, 2020).

The route out of *digitally-disadvantaged* status is "full stack support"[8] (Loomis,

4. The American Standard Code for Information Interexchange, widely known as ASCII, assigned the Latin letters, numbers, and other characters common to American English to the 256 slots available in the 8-bit code. ASCII was the predominant character encoding standard pre-Unicode and is still used by many websites and devices today.

5. ICANN, or the Internet Corporation for Assigned Names and Numbers, is a U.S nonprofit and multistakeholder group that maintains the central repository for IP addresses and helps coordinate their supply while also managing the domain name system.

6. This digital "logic" perpetuates supremacist theories such as Jean-Jacques Rousseau's hypothesis in *On the Origin of Language* that "the depicting of objects is appropriate to a savage people; signs of words and of propositions, to a barbaric people; and the alphabet to civilised people" (1966, p. 17, as quoted in Lydia Liu, 2015, p. 380).

7. Poet Bob Holman calls dominant languages that push out mother tongues "bully" languages (Grubin, 2015).

8. "Full-stack support" is similar to Kornai's (2013) definition of "digital vitality," but the difference is that Kornai's definition encompasses both digital support and digital use. This is an important distinction because digital support does not necessarily lead to digital use of a language; long-standing lack of digital support may in fact incentivize bilingual/multilingual speakers to utilise a dominant, well-supported language for digital communication, such that these habits may be irre-

Pandey, and Zaugg, 2017). This term, used among technologists, designates comprehensive digital support for a language from basic levels like fonts and keyboards to sophisticated NLP tools. Achieving full stack support requires numerous steps, from documenting the language, submitting its script for inclusion in the Unicode Standard,[9] and designing fonts, to building input methods such as virtual keyboards (Loomis et al., 2017; *Indigenous Languages: Zero to Digital*, 2019). Text must be translated and interfaces localised so menu headers and dates follow the correct conventions. Advocates must lobby software vendors to include support for their language at the operating system and application levels.[10] High-level technical affordances require NLP research and include optical character recognition, spell-check, text-to-speech, and search capabilities. Developing full stack support can take years or decades, requiring the coordination of many stakeholders. Even under ideal conditions—a large speaker community with a base of committed language advocates and technologists—challenges in reaching full stack support abound due to commercial, technical, and political hurdles.

EQUITABLE ACCESS

Equity, versus equality, acknowledges that each language community has unique circumstances and requires an allocation of resources and efforts to match, including potentially refusal of digital support. Issues with equitable access can fall anywhere on the "stack," from fonts to support on popular social media platforms. For example, while Indic scripts are encoded within the Unicode Standard, disproportionately few Indic fonts exist, due in part to the technical difficulty of engineering such fonts and the historically low commercial interest in Indian markets. Support by major software vendors has also followed political and commercial interests, from prioritising national and "commercially-viable" scripts in early editions of the Unicode Standard (Zaugg, 2017), to the targeting by software localization vendors of Europe and Japan through the late 20th century (Oo, 2018).

Even for languages where typographic access is not a barrier, a major issue is a lack of integration methods through a "digital re-colonization" supposedly driven

---

versible even if digital supports for their mother tongue later exist. In this context, it is possible for a language to be *digitally-disadvantaged* while also being well-supported.

9. Unicode inclusion itself often requires extensive historical research, documentation, and resolution of differences in character representation, etc. (Zaugg, 2017; Bansal, 2021).

10. Users on the popular streaming platform Twitch complained, for example, about the lack of Indigenous language tags available to help them find other members of their language communities, e.g. Basque and Gaelic (Sinclair, 2021). One example of lobbying working is Apple's attempts to support the *nastaʿlīq* script used to write Urdu (Kohari, 2021).

by market conditions. Modern operating systems are becoming black boxes with limited extensibility and few supported languages. For example, Google's Chrome OS has no means to recognise languages beyond its pre-existing repertoire. For Sami students in Norway who are required to use Chrome OS laptops, a workaround had to be implemented to enable Sami keyboard access,[11] with no mechanism for enabling proofing tools. iOS and Android require manual maintenance of separate keyboard apps, with limited operating system integration. It is presently not possible to provide a high-quality user experience for *digitally-advantaged language* speakers on these platforms.

Many *digitally-disadvantaged language* communities include passionate advocates who have led grassroots efforts to develop fonts, keyboards, and word processing software for their languages and scripts (Zaugg, 2017; Zaugg, 2019a; Zaugg, 2020; Zaugg, forthcoming; Scannell, 2008; Bansal, 2021; Coffey, 2021; Kohari, 2021; Rosenberg, 2011; Wadell, 2016). The challenges of lobbying major software vendors for technical support have led some communities to embrace free and open-source software instead (Bailey, 2016). User communities have created fonts using free tools like FontForge and libraries such as Pango and HarfBuzz. Virtual keyboards are created using KeyMan or kbdgen, and content translated using platforms such as Weblate or Pontoon. In the absence of high-quality support within operating systems, some have localised Linux desktops and applications. A suite of advanced NLP tools is also available as free and open-source software, enlarging possibilities for decentralised efforts by communities (Littauer, 2018).

Peer production can assist with reinvigorating *digitally-disadvantaged languages*. Organisations such as Divvun[12] provide open source tools to enable spell- and grammar checking, keyboard layouts and additional necessities for high-quality digital functionality for Sámi and other Uralic languages. Once baseline tools exist, organic communities arise to create content on Wikipedia, Twitter and other platforms. Non-profit and international efforts, such as the University of California, Berkeley's Script Encoding Initiative, and UNESCO projects such as those associated with the 2019 UN Declaration of the Year of Indigenous Languages,[13] are also

---

11. The workaround was to add the keyboard as a variant under the majority language, as well as to write the necessary operating system extension to implement the actual keyboard functionality as well (i.e., the ability for a key press to input the necessary key input).

12. <https://divvun.org>, funded by the Sámi Parliament of Norway

13. For example, see the International Conference Language Technologies for All (LT4All): Enabling Linguistic Diversity and Multilingualism Worldwide held in December 2019. Furthermore, the UN proclaimed 2022-2032 as the International Decade of Indigenous Languages (IDIL2022-2032), with UNESCO the lead organizer; expanding digital support for Indigenous languages will continue to be a focus.

working to widen access; but it is an uphill battle, as what constitutes "full stack support" grows with each new digital innovation.

LANGUAGE AND SCRIPT INTEGRITY

While some efforts to support *digitally-disadvantaged languages* are well-grounded, others are based on superficial knowledge of languages and writing systems (Zaugg, forthcoming). A virtual keyboard is only useful if it includes all the characters a language utilises, and ideally has a layout optimised for the most commonly used characters, etc. A well-designed font that incorporates calligraphic traditions can elevate a script's readability and status; a poorly designed font can signal its devaluation compared to font-rich scripts such as Latin (Leddy, 2018). Tools such as auto-correct, spell-check, and predictive typing can speed input, but can also degrade a language's orthography, honorifics, and patterns of respectful address if developed without appropriate care.

A significant trend within NLP is reliance on "big data" approaches to solve language access issues, such as generating text-to-speech engines or automatic translation. This exacerbates the disadvantage of *low-resource languages*, as dominant languages receive better quality tools as the bulk of cultural discourse already exists in these languages. Optimistically, new approaches such as "transfer learning" may allow using higher-resourced languages to train models for lower-resourced languages. However, to avoid building linguistically-damaging or unwanted tools, computational linguists should commit to "decolonizing NLP" by only developing tools in partnership with and led by the interests of language communities (Bird, 2020).

SURVEILLANCE VULNERABILITIES

Even when *digitally-disadvantaged languages* achieve a baseline of digital support, knock-on challenges remain. For example, social media platforms do not adequately moderate content in these languages (Zaugg, 2019b; Fick & Dave, 2019; Martin & Sinpeng, 2021; Marinescu, 2021). Facebook in particular has failed to moderate hate speech and fake news in *digitally-disadvantaged languages*, leading to real world harms across the globe (Adegoke & BBC Africa Eye, 2018; Stevenson, 2018; Taye & Pallero, 2020).

Given that *digitally-disadvantaged languages* have a smaller mass of digitised content, data mining puts these communities at higher risk relative to dominant languages. The smaller the corpus, the higher the chance that individual privacy of

community members will be invaded. Finding the balance between technological solutions and social responsibility is challenging. Ensuring that users are not surveilled, while simultaneously improving language tool quality, requires consent-based measures significantly beyond those provided by laws and regulations like GDPR. Privacy-protections are critical for *digitally-disadvantaged language* communities; surveillance capitalism will likely lead to disproportionately negative outcomes in these communities, as many are uniquely vulnerable to state, NGO, and corporate harms (Zaugg, 2019b). For example, digital tools have been used to surveil the Rohingya in Myanmar and Bangladesh (Aziz, 2021; Ortega, 2021), while U.S. Customs and Border Protection surreptitiously collects migrants' cell phone conversations and social media posts, using them to inform asylum decisions at the US-Mexico border (Korkmaz, 2020).

Some *digitally-disadvantaged languages* are of "strategic interest" to governments, and tools such as machine translation are built through military-intelligence funding to aid surveillance. Amandalynne Paullada (2021, n.p.) reminds us that a push for militarised surveillance is "precisely what fostered the development of machine translation technology in the mid-20th century" and its deployment today extends this tradition of "exerting power over subordinate groups." Efforts towards digital justice for *digitally-disadvantaged language* communities must balance the fact that increased digital support for a language also increases its speaker community's legibility to surveilling actors, benevolent or malevolent. These languages require design solutions that maintain data privacy, sovereignty,[14] and safety within the digital sphere.

CONCLUSION

Digitally-disadvantaged languages face multiple inequities in the digital sphere, including gaps in digital support that obstruct access for speakers, poorly-designed digital tools that negatively affect the integrity of languages and writing systems, and unique vulnerabilities to surveillance harms for speaker communities. The term can bridge the work of a wide range of stakeholders who seek to study, discuss, and address language equity in the digital sphere, including scholars, NLP researchers, technologists, speaker communities, and language advocates.

---

14. For example, the Māori non-profit Te Hiku Media is working to build language tools for their community while keeping their annotated audio data, which can be used to develop automatic speech recognition and speech-to-text tools, out of the hands of corporate actors (Coffey, 2021).

REFERENCES

Adegoke, Y. (2018, November 13). Like. Share. Kill. Nigerian police say "fake news" on Facebook is killing people. *BBC News*. https://www.bbc.co.uk/news/resources/idt-sh/nigeria_fake_news

Aziz, A. (2021). A repertoire of everyday resistance and technological (in)security: Constructing the Rohingya diaspora and transnational identity politics on social media. *AoIR Selected Papers of Internet Research*. https://doi.org/1631738803

Bailey, D. (2016). Software localization: Open Source as a major tool for digital multilingualism. In L. Vannini & H. L. Crosnier (Eds.), *Net.Lang: Towards the Multilingual Cyberspace*. http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CI/CI/pdf/netlang_EN_pdfedition.pdf

Bansal, V. (2021). *Forget emoji, the real Unicode drama is over an endangered Indian script* [Report]. Rest of World. https://restofworld.org/2021/tulu-unicode-script/

Benjamin, M. (2016, May 23). *Digital language diversity: Seeking the value proposition*. 2nd Workshop on Collaboration and Computing for Under-Resourced Languages, Portoroz, Slovenia. https://infoscience.epfl.ch/record/222525?ln=en

Bird, S. (2020). Decolonising speech and language technology. *Proceedings of the 28th International Conference on Computational Linguistics*, 3504–3519. https://www.aclweb.org/anthology/2020.coling-main.313

Coffey, D. (2021, April 28). Māori are trying to save their language from Big Tech. *WIRED*. https://www.wired.co.uk/article/maori-language-tech

Fick, M., & Dave, P. (2019, April 23). Facebook's flood of languages leave it struggling to monitor content. *Reuters*. https://www.reuters.com/article/us-facebook-languages-insight-idUSKCN1RZ0DW

Grubin, D. (2015, January 25). *Language Matters with Bob Holman*. David Grubin Productions Inc. and Pacific Islanders in Communications.

Indigenous languages: Zero to digital: A guide to bring your language online. (2019). Translation Commons. https://translationcommons.org/impact/language-digitization/resources/zero-to-digital/

Kohari, A. (2021, February 9). *Meet the people fighting to keep a language alive online*. Rest of World. https://restofworld.org/2021/bringing-urdu-into-the-digital-age/

Korkmaz, E. E. (2020, December 8). Refugees are at risk from dystopian "smart border" technology. *The Conversation*. https://theconversation.com/refugees-are-at-risk-from-dystopian-smart-border-technology-145500

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

Language Status. (XXXX). *Ethnologue*. https://www.ethnologue.com/about/language-status

Leddy, M. (2018, May 29). Beyond "Graphic design is my Passion": Decolonizing Typography and Reclaiming Identity (II of II. *Explorations in Global Language Justice*. https://languagejustice.wordpress.com/2018/05/29/beyond-graphic-design-is-my-passion-decolonizing-typography-and-reclaiming-identity-ii-of-ii/

Littauer, R. (2018). *Open source code and low resource languages* [Master's Thesis, Saarland University

Department of Computational Linguistics]. https://raw.githubusercontent.com/RichardLitt/thesis/master/single-thesis.pdf

Liu, L. H. (2015). Scripts in motion: Writing as Imperial technology, past and present. *PMLA*, *130*(2), 375–383.

Loomis, S. R., Pandey, A., & Zaugg, I. (2017, June 6). Full Stack Language Enablement. *Steven R. Loomis*. https://srl295.github.io/2017/06/06/full-stack-enablement/

Marinescu, D. (2021, September 8). Facebook's Content Moderation Language Barrier. *New America*. http://newamerica.org/the-thread/facebooks-content-moderation-language-barrier/

Martin, F. R., & Sinpeng, A. (2021, July 5). Facebook's failure to pay attention to non-English languages is allowing hate speech to flourish. *The Conversation*. http://theconversation.com/facebooks-failure-to-pay-attention-to-non-english-languages-is-allowing-hate-speech-to-flourish-163723

Oo, M. T. (2018, September 19). A brief history and evolution of IT localization. *Translation Royale*. https://www.translationroyale.com/history-of-it-localization/

Ortega, A. (2021, March 23). Myanmar and the oppressive side of the digital revolution. *The Globalist*. https://www.theglobalist.com/myanmar-dictatorship-surveillance-technology/

Paullada, A. (2021, July 31). Machine Translation Shifts Power. *The Gradient*. https://thegradient.pub/machine-translation-shifts-power/

Romaine, S. (2015). The global extinction of languages and its consequences for cultural diversity. In H. F. Marten (Ed.), *Cultural and Linguistic Minorities in the Russian Federation and the European Union* (pp. 31–46). Springer International Publishing.

Rosenberg, T. (2011, December 9). Everyone Speaks Text Message. *The New York Times Magazine*. http://www.nytimes.com/2011/12/11/magazine/everyone-speaks-text-message.html

Rousseau, J.-J. (Ed.). (1986). *On the origin of language*. University of Chicago Press.

Scannell, K. P. (2008). *Free software for Indigenous languages* [Thesis, Saint Louis University]. https://cs.slu.edu/~scannell/pub/ili.pdf

Sinclair, K. (2021, June 16). The Twitch streamers fighting to keep minority languages alive. *The Verge*. https://www.theverge.com/2021/6/16/22533319/twitch-streamers-minority-languages-basque-gaelic

Stevenson, A. (2018, November 6). Facebook Admits It Was Used to Incite Violence in Myanmar. *The New York Times*. https://www.nytimes.com/2018/11/06/technology/myanmar-facebook.html

Taye, B., & Pallero, J. (2020, July 27). *Open letter to Facebook on violence-inciting speech: Act now to protect Ethiopians*. Access Now. https://www.accessnow.org/open-letter-to-facebook-protect-ethiopians/

Unicode. (2015, December 16). Unicode launches adopt-a-character campaign to support the world's "digitally disadvantaged" living languages. *The Unicode Blog*. http://blog.unicode.org/2015/12/unicode-launches-adopt-character.html

Wadell, K. (2016, November 16). The alphabet that will save a people from disappearing. *The Atlantic*.

Zaugg, I. A. (2017). *Digitizing ethiopic: Coding for linguistic continuity in the face of digital extinction* [PhD dissertation, American University]. http://gateway.proquest.com/openurl?url_ver=Z39.88-200

4&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&res_dat=xri:pqm&rft_dat=xri:pqdiss:10599782

Zaugg, I. A. (2019a). *Imagining a multilingual cyberspace*. Nesta. https://findingctrl.nesta.org.uk/imagining-a-multilingual-cyberspace/

Zaugg, I. A. (2019b, December 4). Digital surveillance and digitally-disadvantaged language communities. *International Conference Language Technologies for All*. https://lt4all.elra.info/media/papers/O8/188.pdf

Zaugg, I. A. (2020). Digital inequality and language diversity: An Ethiopic case study. In M. Ragnedda & A. Gladkova (Eds.), *Digital Inequalities in the Global South* (pp. 247–267). Springer International Publishing. https://doi.org/10.1007/978-3-030-32706-4_12

Zaugg, I. A. (forthcoming). Language justice in the digital sphere. In L. H. Liu & A. Rao (Eds.), *Lifeworld of Languages*. Columbia University Press.