How Can AI Support Language Digitization and Digital Inclusion?

Juan N. Pava
Thomas S. Mullaney
Caroline Meinhardt
Audrey Gao
Diyi Yang

Stanford University
Human-Centered
Artificial Intelligence

Stanford | SILICON

Authors

Juan N. Pava is a research fellow in the Tech Ethics & Policy Rising Scholars Program at Stanford University's McCoy Family Center for Ethics in Society. At the Stanford Institute for Human-Centered Artificial Intelligence (HAI), he works at the intersection of AI, the social sector, and the Global South. His research interests include the political economy of emerging economies and its relationships with political philosophy and ethics. He holds a bachelor's degree in philosophy and economics from New York University.

Thomas S. Mullaney is the director of SILICON (the Stanford Initiative on Language Inclusion and Conservation in Old and New Media). He is the director of the Program in Science, Technology & Society and a professor of Chinese history at Stanford University, the Kluge Chair in Technology and Society at the Library of Congress, and a Guggenheim Fellow. For the past 15 years, his research, publications, conference planning, and coursework have focused expressly on asymmetries in the global information and language technologies, with a keen focus on writing systems that have been systematically marginalized and excluded from the modern information age.

Caroline Meinhardt is a policy research manager at Stanford HAI, where she manages the institute's policy research initiatives. Her research focuses on the implementation challenges of AI regulation, the governance of large-scale AI models, and global AI governance approaches. Prior to joining HAI, she worked as a China-focused consultant and analyst, delivering in-depth research and strategic advice regarding China's development and regulation of emerging technologies, including AI. She holds a bachelor's degree in Chinese studies from the University of Cambridge and a master's degree in international policy from Stanford University.

Audrey Gao is the project manager of SILICON. She holds a bachelor's degree from Emory University in philosophy and political science. Prior to joining Stanford, she led a team of researchers analyzing policy change in West Africa. At SILICON, she manages project work with practitioners, students, and partner organizations. Much of her work also involves scaling global community engagement efforts that support participation and dialogue in language digitization.

Diyi Yang is an assistant professor in computer science at Stanford University and a faculty affiliate at Stanford HAI. Her research interests are in socially aware natural language processing, large language models, and human-AI interaction, with a focus on designing human-centered AI systems that are not only technically capable, but also meaningfully connected to how people think, interact, and collaborate. Her work has been recognized with best paper nominations or awards at ICWSM, EMNLP, SIGCHI, ACL, UIST, and CSCW. She was named to IEEE's "AI 10 to Watch" and received the Intel Rising Star Faculty Award, Microsoft Research Faculty Fellowship, NSF CAREER Award, and Sloan Research Fellowship.

Acknowledgments

The authors would like to thank Blue Tarpalechee, Claudio Pinhanez, Elena Cryst, Nay San, Sang Truong, Sanmi Koyejo, Tolúlọpẹ́ Ògúnrẹ̀mí, and Tracy Navichoque for their valuable comments and feedback; Maroua Bezzaoui and Tyler Grace Abernethy for their assistance with documentation and visual materials; and Carolyn Lehman, Chris Ellis, Jeanina Matias, Michi Turner, Nancy King, and Shana Lynch for their help preparing the publication. They also thank the many scholars, practitioners, and interns whose work informed this paper, including Toral Cowieson, Mark Davis, Debbie Anderson, Anushah Hossain, Steven Loomis, Conrad Nied, Andrew Glass, Anshuman Pandey, Arjun Raj, Erin Dai, Samantha Leventis, Daniel Argento, Diana Bernabe, Neev Seedani, Mathias Becerra-Sanchez, Christian Roy, and Alyssa Hoang.

Table of Contents

Executive Summary

- In the wake of rapid AI development, attention is increasingly being drawn to the fact that most AI systems fail to serve most of the world's linguistic communities. Data scarcity is often highlighted as a key reason, yet there are much more basic digital foundations that are prerequisites for building AI training datasets.

- Over 6,000 of the world's 7,000-plus living languages remain digitally disadvantaged, meaning that they are unsupported across mainstream devices, operating systems, browsers, and applications. Language communities excluded from digital systems can only participate minimally in a world increasingly mediated by technology and are at the same time unable to generate enough data needed to be represented in AI.

- Empowering digitally disadvantaged language communities to participate in today's digital world requires holistic progress on a set of foundational language tools (from script encoding to keyboard layouts) and supporting language tools (from grammar checkers to accessibility features).

- A global network of language practitioners, scholars, and grassroots groups have been working tirelessly to create and sustain these language tools. Yet progress is often slow and uneven amid chronic underfunding and a lack of coordination.

- AI has the potential to scale and accelerate language digitization. In recent years, scholars have begun leveraging AI — and especially natural language processing tools — to sidestep major bottlenecks in the field:

    - In the early stages of language digitization, AI tools such as grapheme-to-phoneme systems, morphological analyzers, optical character recognition systems, and image generation models can assist with script development and foundational language infrastructure tooling.

    - Once a language can effectively be rendered on devices, AI tools such as language identification models, optical character recognition systems, and automatic speech recognition systems can support language transcription and broader documentation and data collection processes.

    - In the final stages of language digitization, AI tools such as machine translation, grammar- and spell-checking systems, text-to-speech systems, forced alignment tools, and large language models are increasingly the foundation for supporting digital tools that help ensure true digital inclusion.

- It is important to note that not all language communities may choose to develop a writing system for their language. Technical approaches are emerging that enable the creation and use of digital tools for spoken-only languages.

- While all these nascent efforts are promising, AI alone cannot address the field's more fundamental research problems, workflow bottlenecks, and adoption challenges. Language digitization is also an inherently community-centric process that requires a deeply sensitive cultural and linguistic understanding. Much of the work in this field should thus continue to be driven by the language communities themselves, with AI as an accompanying tool.

- Additional work, time, and resources need to be invested in harnessing AI for language digitization in a way that centers communities and their individual needs and contexts. We outline detailed recommendations for different stakeholders to work together to advance language digitization and digital inclusion in the age of AI, including:

    - **Building trust and empowering communities** by fostering community-engaged convenings and collaborations, and building community-driven benchmarks and standards for digital language tools.

    - **Strengthening research foundations** by creating reliable resources to track progress on digitally disadvantaged languages, investing in, expanding, and evaluating effective AI tools for language digitization, and creating forums for interdisciplinary exchanges.

    - **Improving workflows** by moving to parallel workflows for language digitization and leveraging AI for organizational improvements.

    - **Forming coalitions** by implementing mechanisms to reform incentive structures surrounding language tool adoption and strengthening storytelling for general audiences.

    - **Ensuring cultural sustainability** by empowering culturally aware AI development, promoting contextualized adoption, and impact assessment.

1. Introduction

Over 6,000 of the world's 7,000-plus living languages remain digitally disadvantaged languages (DDLs) — those that are, to varying degrees, barred from full-scale participation in the digital age, with dire consequences.

Over 6,000 of the world's 7,000-plus living languages remain digitally disadvantaged languages.

Digital exclusion exists along a spectrum, ranging from instances of complete exclusion to partial but still compromised inclusion. It affects far more than simple access to digital tools and services. In the 21st century, language death and digital exclusion have become tightly linked in a mutually reinforcing cycle of marginalization and extinction. Languages that cannot be written digitally are less likely to be written at all, further driving multilingual communities toward preferring dominant or colonial languages. Linguists predict that 50% or more of the world's languages may become extinct this century.[1]

The gap that separates the world's top 100 digitally dominant languages and DDLs is steadily becoming a chasm. While over a third of the world's youth will live in Africa by 2050, not a single African language ranks among the top 34 used on the internet today.[2] Modern AI systems risk amplifying this divide. Large language models (LLMs), voice assistants, translation tools, and more are disproportionately trained on English and other highly resourced languages (HRLs). These models rely on massive datasets, which are almost entirely unavailable for under-resourced languages. As a result, most AI systems fail to serve most of the world's linguistic communities.[3]

The ramifications of this widening divide are profound.[4] When a language and culture are digitally disadvantaged, their community can, at best, participate minimally in a world where the written word is increasingly mediated by technology. Not being able to write a language digitally affects everything from daily exchanges (e.g., texting a last-minute grocery item or sending a quick love note) to life-threatening situations (e.g., sharing critical health bulletins or issuing localized evacuation orders).

While there are efforts to build AI systems that are better attuned to and represent DDLs, such efforts face deep, structural challenges. For the overwhelming majority of languages, some or all of the prerequisite digital foundations for AI tools — keyboards, digital fonts, and other language tools that enable people to communicate digitally in their own languages — simply do not exist. Beyond issues surrounding the language technologies themselves, many communities feel a deep sense of distrust for technology companies. This distrust must be understood from two wide-ranging perspectives: the multi-century historical context of exploitation and maltreatment and the far more recent and ongoing reports of ethical violations.[5]

Still other areas of concern pertain to the struggle of data collection and the tendency to resort to datasets that represent certain legacies of the past, such as the prevalent use of Bible translations among AI researchers and companies.[6] From their perspectives, these translations offer up widely available data

eminently suited to the creation of parallel corpora, while for many communities they serve as a reminder of colonial and missionary legacies.

This entrenches a feedback loop: Languages excluded from digital systems cannot generate the data needed to be represented in AI, and without AI, communities face further barriers to digital participation and language vitality.[7] Additionally, without the ability to access tools such as content moderation systems, DDL speakers are left more vulnerable to online harm, misinformation, and social exclusion. For Indigenous, postcolonial, and minoritized communities, digital invisibility reinforces historical patterns of marginalization. If the gap between HRLs and DDLs goes unaddressed, it will not only persist but widen.

In this white paper, we provide one of the first overviews of the varying ways AI tools and techniques can support language digitization work and digital inclusion efforts more broadly. We start by defining the scope and challenges of digital inclusion, discussing the foundational and supporting tools required to bring a language into the digital world. We then offer a schematization that identifies AI tools and techniques that can help scale and accelerate different, often extremely labor-intensive, stages of the language digitization process. Finally, we analyze the significant structural, informational, and procedural challenges that must be addressed and provide recommendations for how language digitization and AI researchers and practitioners can responsibly realize the full potential of AI in supporting the digital inclusion of under-resourced languages.

Languages excluded from digital systems cannot generate the data needed to be represented in AI, and without AI, communities face further barriers to digital participation and language vitality.

2. What Counts as Digital Inclusion?

2.1. The Spectrum of Digital Exclusion and Inclusion

The term "digitally disadvantaged language" refers to the majority of the world's languages that remain unsupported across mainstream devices, operating systems, browsers, and applications.[8]

Although the term includes many widely spoken languages, it frequently overlaps with other terms such as minority, minoritized, Indigenous, and endangered languages. While *minority* languages are those spoken by a small subset of a population (e.g., Basque in Spain), *minoritized* languages are those that have been marginalized relative to a dominant language, regardless of demographic size (e.g., Cantonese in China).[9] Many Indigenous languages and endangered languages — those whose intergenerational transmission is disrupted — fall here.[10] Yet minority status does not necessarily imply endangerment. Languages like Taiwanese Hokkien are digitally unsupported but continue to have tens of millions of speakers. Conversely, some endangered languages (e.g., Irish Gaelic) enjoy substantial digital infrastructure due to state investment.[11] Crucially, then, digital disadvantage is not a function of speaker numbers or vitality, but of a language's inability to operate across the full digital stack.

Because DDLs are defined by the scarcity of digital infrastructure and tooling,[12] the term is distinct from another set of terms commonly used in natural language processing (NLP) literature: "low-resourced" or "under-resourced" languages. These are defined by data scarcity; that is, such languages lack the volume and quality of labeled and unlabeled data to train NLP systems.[13] Hence, although digitally disadvantaged and low-resource status often coincide, they are not identical. A language such as Mongolian may have sizable corpora for NLP work yet remain digitally disadvantaged because key infrastructure — especially around script encoding — remains incomplete or inconsistent.[14] Conversely, some languages are low-resource but not digitally disadvantaged because strong institutional support provides robust tools despite limited data. Māori illustrates this through community-led, data-efficient NLP development,[15] and Welsh through government-backed infrastructure built on comparatively modest corpora.[16] In many cases, however, the lack of digital infrastructure (which defines DDLs) and the lack of data (which defines low-resource languages) reinforce each other, creating a self-perpetuating cycle: Without tools, data creation is difficult; without data, tool development stalls.

Digital disadvantage is not a function of speaker numbers or vitality, but of a language's inability to operate across the full digital stack.

These terms remain contested, characterized by an ongoing disagreement on where their boundaries lie and which dimensions should be prioritized.[17] For the purpose of this paper, we focus on the spectrum of DDLs (see Table 1) because it captures both sociopolitical realities (such as those faced by minoritized or endangered language communities) and the technical bottlenecks that parallel those illustrated by the low-resource NLP literature. This framing allows us to take a more holistic view of the socio-technical conditions shaping digital inclusion.

Table 1. Spectrum of digitally disadvantaged languages[18]

	Potential[19]	Emerging	Ascending	Vital	Thriving
Categories of digital language support	Language shows no sign of digital support.	Language has some content in digital form and/or encoding tools.	Language has some spell-checking, localized tools or machine translation.	Language is supported by multiple tools in all of the prior categories, as well as some speech processing.	Language has all of the previous tools plus virtual assis-tants.
Number of languages	3,996	3,304	401	95	33
Examples	Muruwari	Naxi, Kunwinjku	Hokkien, South Sámi	Basque, Cantonese, Irish, Māori, Urdu	English, Chinese, Latin

2.2. The Digital Inclusion Stack: Nine Key Tools

Digital inclusion is often framed as a question of access: getting people online, connecting communities, and bridging physical infrastructure gaps. While access to the internet and other digital tools is a crucial prerequisite, it is only meaningful when accompanied by foundational language tools that enable people to read, write, and interact in their own languages in digital spaces. These tools tend to be taken for granted in high-resource settings.

Access to telehealth and education applications, for example, only helps a community when the applications are capable of understanding that community's mother tongue. Without foundational language tools (see Table 2), billions of people remain excluded from, or markedly disadvantaged within, the 21st century digital age — even when they possess the necessary hardware and connectivity. Empowering DDL communities to build a digital world that serves their needs requires holistic progress on all these foundational language tools. Furthermore, since only the dominant form of a language is typically digitized, which can exclude large swaths of speakers, supporting regional varieties and dialects is equally crucial to digital inclusivity and for aligning language technology with actual usage patterns.[20]

Without foundational language tools, billions of people remain excluded from, or markedly disadvantaged within, the 21st century digital age.

Table 2. Five foundational language tools that establish digital inclusion

Foundational language tool	Explanation
1. Script encoding	A language's writing system must be formally encoded in the Unicode Standard to be stored, rendered, and transmitted digitally.[21] Without Unicode support, no amount of digital literacy or internet connectivity can make a language usable online.
2. Fonts and type design	Readable, culturally appropriate fonts ensure that a language is legible across devices.[22] Poorly rendered typefaces, or their nonexistence, can discourage use and lead to platform-level inconsistencies.
3. Keyboard layouts and input tools	Community-informed keyboard designs, input methods, and text entry systems are essential for typing in a language — especially for those with complex or newly invented scripts. Without intuitive ways to input text, digital communication remains inaccessible.
4. CLDR data contributions	The Unicode Common Locale Data Repository (CLDR) supports software localization through culturally specific formats for dates, times, numbers, and plurals.[23] DDLs are often absent from this system, making user interfaces impossible to translate or adapt appropriately.
5. Support for neographies and revitalized scripts	Many communities have developed new writing systems, also known as "neographies" — especially in contexts of cultural revival.[24] These neographies must be encoded in the Unicode Standard and supported just like other scripts in order to become usable in digital contexts.

Once a language can be digitally written, there is more work to be done. For many language communities who do enjoy digital and online access, the absence of higher-order language tools, such as content moderation systems, leave them more exposed to online abuse, harassment, misinformation, and mistreatment than their HRL community counterparts. Similar inequities prevent DDL communities from taking advantage of the many other benefits of the digital age — educational, civic, health-related, and more — enjoyed by HRL communities. Educational tools and accessibility features, for example, are crucial to ensuring the confident production and consumption of accurate text in the digital arena. Even widely spoken languages like Arabic and Indonesian have far less tooling available compared to English. Even when such tools exist for DDLs, they often don't perform well for these languages, and institutions and companies often hesitate to deploy imperfect tools due to reputational risk.

Without supporting language tools (see Table 3), communities that are most in need are thus the most likely to be exposed to harmful digital interactions with little recourse. Enduring and meaningful digital inclusion requires the creation or adaptation of these supporting digital language tools to ensure that digital participation can be sustained over the long term, and that it is safe, equitable, and empowering for all.

Table 3. Four supporting language tools that sustain digital inclusion

Supporting language tool	Explanation
1. Abusive language and hate speech detection[25]	Enables the identification of harmful content, targeted abuse, disinformation, and politically motivated harassment in online spaces.
2. Harmful language moderation,[26] **spam filtering, and protections against predatory practices**[27]	Shield vulnerable users, particularly youth and marginalized groups, from harassment, manipulation, and scams.
3. Grammar- and spell-checkers[28]	Support correct language usage, digital literacy, and users in confidently producing written content in their own languages.
4. Accessibility features	Make digital environments usable by people with disabilities and to users with no or low literacy (e.g., screen reader compatibility, text-to-speech in local languages, captioning for regional dialects).

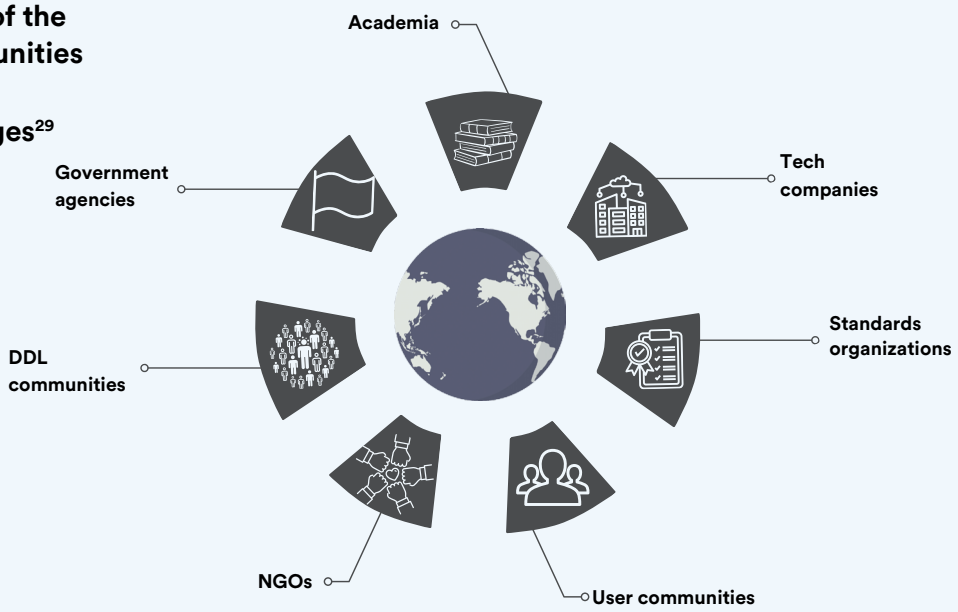
Two clarifications are needed. First, although the nine tools listed above are necessary for a language to digitally ascend, they are neither exhaustive nor individually sufficient; a language may have many of these tools yet remain digitally disadvantaged if even one foundational element is missing. Additional factors like the availability of digital content (e.g., Wikipedia pages) or tokenization (i.e., the way text is broken down into smaller units for LLMs) also influence a language's movement up the digitization pipeline or participation in AI development.

Second, the language digitization process that might take a language from potential to thriving (see Table 1) is rarely linear, even if the five foundational tools of Table 2 must logically precede the four sustaining tools of Table 3. Script encoding may advance in parallel with font and keyboard design, and spam or cyberbullying filters may be developed alongside hate-speech detection. In practice, progress is iterative and interdependent, with advances in one area regularly prompting revisions in others.

2.3. On the Front Lines of Digital Inclusion

A highly decentralized but also highly dedicated global network of community organizers, standards bodies, nonprofits, scholars, practitioners, and educational institutions has been working to make inroads on the above-mentioned fundamental and supporting digital tools (see Figure 1). However, while many of the technical tools needed to build digital language infrastructure are increasingly available, progress remains slow and uneven. Efforts to digitize low-resource languages face major challenges, including chronic underfunding and a lack of coordination. Most initiatives are driven by grassroots networks — volunteers, local communities, linguists, and small nonprofits — often working outside formal research pipelines without the support of major tech companies, governments, or academic institutions. These efforts are frequently isolated from one another, leading to duplicated work and missed opportunities for collaboration.

Figure 1. An overview of the institutions and communities advancing digitally disadvantaged languages[29]



3. The Promise of AI: Scaling and Accelerating Language Digitization

While technology alone cannot overcome deep, structural challenges, AI has begun to play a promising role in advancing digital inclusion.[30] Resource-constrained organizations and communities have started turning to AI-powered technologies — ranging from character and speech recognition to LLM-driven data annotation — to help scale and accelerate what are extremely labor-intensive efforts to bring DDLs into the digital world and sustain their safe and equitable digital use. Scaling and accelerating these efforts not only enables the preservation and revitalization of low-resource and endangered languages but also opens the door to these language communities participating in and benefiting from AI development.

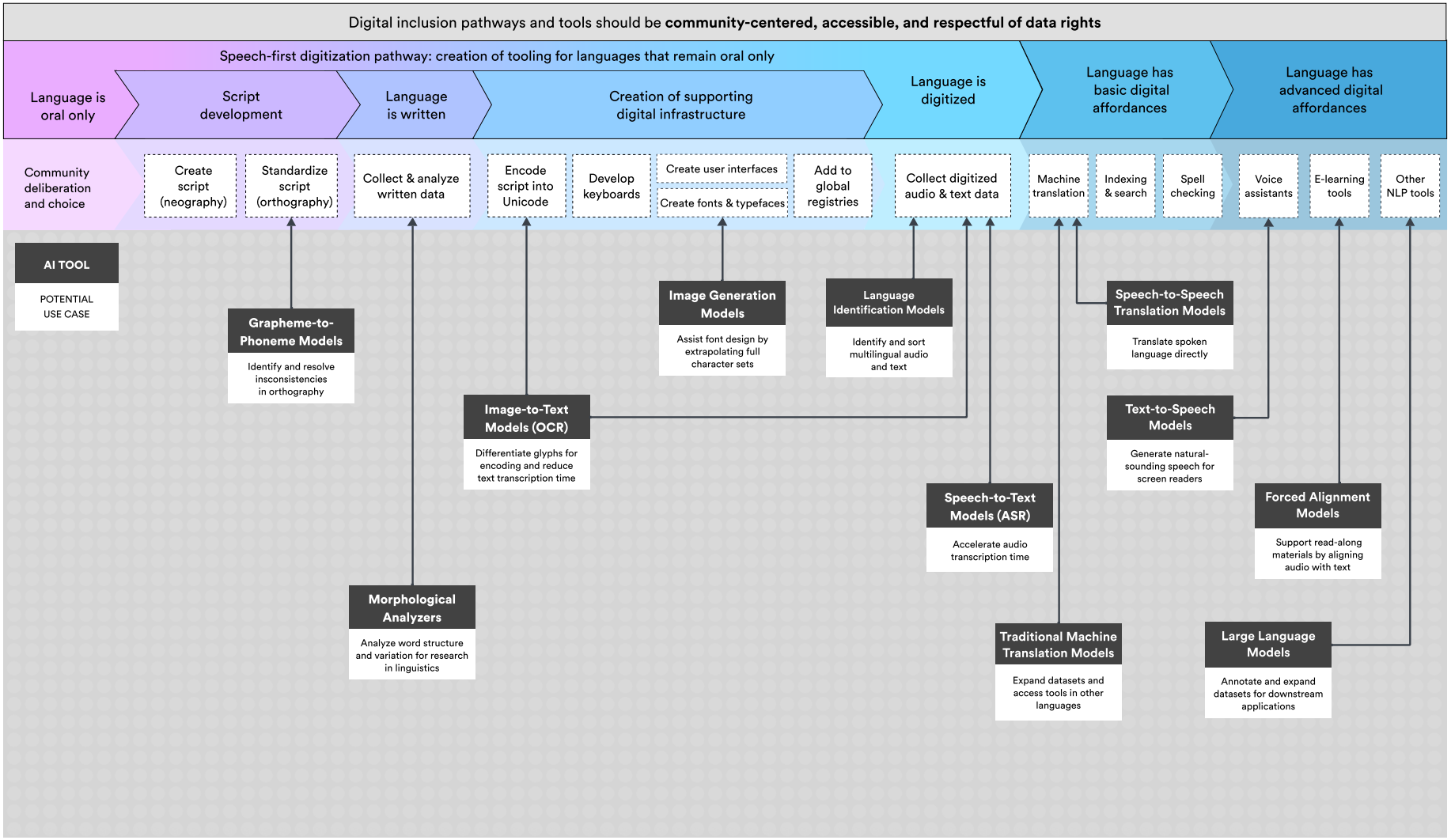
Digital documentation is generally considered the key to bringing languages into the digital sphere — landmark initiatives like The Rosetta Project and the Endangered Languages Project exemplify large-scale attempts to digitally archive and preserve DDLs, especially endangered languages.[31] In recent years, AI and especially NLP tools have shown promise in sidestepping major bottlenecks in this area, particularly when it comes to compiling, organizing, and reviewing digital records.[32]

However, a more comprehensive look at all the steps needed to create digital affordances for a language reveals a variety of places where AI can play a role. The use of AI applications is increasing throughout the digitization pipeline as scholars explore and leverage tools such as optical character recognition[33] and automatic speech recognition,[34] among many more.[35]

Below, we outline some of the most promising AI tools and techniques currently being applied to ongoing efforts to digitize DDLs. Though by no means exhaustive, these examples highlight promising areas that deserve attention and further research and resource investment. In Figure 2, we place these AI tools along the relevant stage(s) of the language digitization process (depicted horizontally). It is worth noting, however, that much of this work rests on painstaking, often unglamorous groundwork, and that real-world digitization is nonlinear, iterative, and full of unforeseen complications. As such, the figure offers a simplified vantage point on a far messier reality, highlighting the breadth of areas where AI tools can improve or scale workflows.

Moreover, when discussing the application of AI tools to language digitization, it is important to note that the decision to digitize a language is not a foregone conclusion. Before embarking on digitization — or even precursors such as script creation — it is essential to engage in community deliberation so that any choices made about language digitization remain firmly in the hands of speakers. Script and digital development carry profound implications for cultural identity, shaping how a community speaks, remembers, and projects itself. As a result, speaker consent and participation are not procedural niceties but the very conditions of legitimacy. Without this grounding, digitization risks stalling or, worse, undermining the communities it claims to support.

Figure 2. The potential role of AI tools in language digitization processes[36]



3.1. Regional, Local, and Community-Led AI Initiatives Focused on Digital Inclusion

While it is not a "tool" within and of itself, we begin with one of the most promising trends in AI: the uptake of AI by regional, local, and other community organizations to advance digital inclusion on their own terms. The digitization journey starts with community choice: Is there a desire to undertake the process of language digitization, and if so, to what ends and with which partners? This starting point — undertaken with community leadership, often in partnership with technical experts hailing from within and beyond that community — is multifaceted. It examines both the potential benefits and pitfalls of digital inclusion — along with deeper conceptual, legal, and even epistemological issues regarding, for example, the community's approach to and definition of key terms such as "data" itself.

The digitization journey starts with community choice.

For many Indigenous communities, this has led to the articulation of data sovereignty (e.g., the Māori Data Sovereignty Network[37]) and data governance frameworks that assert collective rights over linguistic data — governing how it is collected, stored, shared, and reused, including in AI systems.[38] These frameworks often challenge assumptions about "open data," emphasizing consent, stewardship, and long-term community control. Even for many of the most taken-for-granted terms within 21st century technology circles — terms such as "open access," "open data," and more — language communities need to analyze, accept, and perhaps reimagine such concepts so that the digital process itself aligns with community values.

3.2. AI Applications for Script Development and Foundational Language Infrastructure Tooling

The second step in the journey pertains to questions of orality and writing, namely: addressing whether a language is associated with one or more written scripts, and whether this script is standardized and encoded such that it can be stored and rendered digitally. In the case of communities who speak an exclusively oral language but are committed to the creation of a new writing system (or perhaps standardization), this early stage of the language digitization process requires extensive human resources and expertise. It entails devising a new writing system (also known as a neography), standardizing this system through the establishment of spelling systems and other rules (also known as orthography), encoding the resulting scripts into the Unicode Standard, and creating supporting keyboards, typefaces, and other user interfaces. Each of these sub-steps requires deep linguistic and cultural knowledge.

However, even in these early stages of language digitization, there are several ways in which AI can assist regardless of script.

Grapheme-to-phoneme systems: The history of script creation and standardization is inseparable from broader political dynamics, as decisions are frequently shaped by questions of identity, authority, and power. Yet AI tools can still play a useful supporting role once these decisions have been made. Grapheme-to-phoneme (G2P) systems, for instance, can serve as a companion tool to assist linguists and communities in the technical phase of standardizing a script. These

systems can help ensure that the written symbols of a language (i.e., letters or characters, also known as graphemes) consistently match the right speech sounds (also known as phonemes). Traditionally rule-based,[39] AI-driven grapheme-to-phoneme (G2P) systems can now generalize more widely. By mapping written symbols to sounds or analyzing a grapheme's frequency,[40] these models can help expose or resolve inconsistencies in orthography proposals,[41] which is important as orthographic inconsistencies make it harder to create digital tools and usable data.[42] Newer G2P systems that use multilingual, multimodal, or metalinguistic data[43] can further support languages outside of those in the training dataset and significantly reduce phoneme errors that would otherwise degrade the quality of text-to-speech and other speech recognition systems further down the line.[44]

Morphological analyzers: Analyzing the structure and formation of words — also known as morphological analysis — is another foundational step for documenting and effectively encoding a language. Traditionally, these systems have been rule-based and are currently applied primarily to downstream tasks such as spell-checking, grammar correction, and search functions. Increasingly, however, linguists are recognizing the potential of AI-powered morphological analyzers as upstream tools that can support linguistic analysis itself during the early stages of the digitization process.[45] They have the potential to speed up the discovery of a language's morphological structure and enable the testing of hypotheses about the language.[46] Beyond research, such tools also carry pedagogical value: They can support literacy by teaching orthographic conventions during transcription, while still respecting community choices around linguistic variation and standardization.[47]

Optical character recognition (OCR) systems: While most Unicode proposals rely on human documentation, AI can support script encoding efforts indirectly by identifying, differentiating, and clustering glyphs (i.e., the specific graphical form used to visually represent a character) from scanned manuscripts. By drawing from methods used in library science[48] and paleography,[49] OCR systems that convert images into machine-readable text can aid the documentation and preservation of linguistic evidence which is needed for script encoding.

Image generation models: Although their use in this context is still nascent, image generation models hold potential to support typeface design and development once a script is encoded.[50] These models can extrapolate from a small set of reference characters[51] or transfer styles from higher-resource languages to generate complete digital fonts for the entire script.[52] This is particularly valuable for scripts with exceptionally large character sets, such as the Naxi Dongba script, which comprises over 1,000 characters.[53] Supporting even part of this process can save significant design time for an otherwise extremely labor-intensive task requiring specialized expertise. While font style transfer has been more common for high-resource languages like English[54] and Chinese,[55] it holds significant promise for complementing work with new, historic, or revived scripts.

Tools for unwritten languages: It is important to note that of those digitally disadvantaged, spoken-only languages whose speakers wish to pursue digitization, not all language communities may choose to develop a writing system for their language. As a result, technical approaches are emerging that enable the creation and use of digital tools for languages that do not have a script.

A key challenge in this situation involves data. How do you develop a machine translation model, a language identification system, or a voice assistant when you lack transcribed data for the language in question (because it is a language that cannot be transcribed)? Conventional speech recognition and synthesis pipelines will not work for unwritten languages. To address this, researchers are developing alternatives that bypass the need for orthography, such as speech recognition systems trained directly on speech audio or multimodal speech-to-image or image-to-speech models.[56] These innovations offer applications from accessibility tools to voice-driven online services.

3.3. AI Applications for Language Transcription

Once a language can effectively be rendered on devices through fonts, keyboards, and Unicode encodings, another set of AI tools can help accelerate work during the next stage, which involves actually bringing texts and audio in that language into the digital world. These language transcription and broader documentation and data collection processes are crucial not only for digital preservation purposes, but also for enabling and improving a wider variety of digital affordances for the language further down the line.

Language identification models (LID): Accurate language identification in text or audio is the sorting mechanism of any pipeline. They can filter vast amounts of language data crawled from the web or obtained from mixed-language archives and steer downstream tools and researchers to the target language inputs. While the field is mature for higher-resource languages, it faces a circular challenge when it comes to digitally disadvantaged ones: Models need training data to be able to identify the very languages for which data is not available.[57] To overcome this, researchers lean on multilingual[58] and self-supervised[59] approaches that learn from large unlabeled corpora. The open-source model GlotLID-M,[60] for instance, can detect 1,665 mostly low-resource languages in text, while GlotScript[61] can identify all 161 Unicode 15.0 scripts. These tools allow us to organize recordings and text scans by language before employing additional AI tools, thereby reducing potential model misfires on mixed or mis-tagged scripts or languages.

OCR systems (image-to-text): OCR models convert images — in this case low-resource scanned print and handwriting (e.g., palm-leaf manuscripts, early printed materials) — into searchable, editable text.[62] They have been shown to cut the time needed for manual transcription by at least half.[63] While accuracy on non-Latin scripts still lags,[64] recent advances on DDLs indicate that very low-data OCR is still viable: Fine-tuning open-source models (e.g., Google's Tesseract) on synthetic images or vision-language models (e.g., Meta's Llama or Alibaba's Qwen) on small amounts of labeled data greatly improves performance.[65] In addition to unlocking search and indexing functions, OCR can feed downstream machine translation systems with measurable success even when recognition is imperfect.[66]

Automatic Speech Recognition (ASR) systems (speech-to-text): Beyond computer vision, ASR systems unlock oral resources by turning speech into text, bypassing the transcription bottleneck where one minute of audio can take up to an hour to transcribe by hand.[67] Multilingual foundation models such as Open AI's Whisper[68] (covering approximately 100 languages) and Meta's Massively Multilingual Speech (MMS)[69] (covering 1,000-plus languages) provide ASR baselines for DDLs, even extending to languages absent from the training data.[70] To boost performance, researchers applied cross-lingual transfer (drawing from both related[71] or even unrelated[72] languages), leveraged

linguistic descriptions such as grammar books[73] to supplement scarce data, or developed privacy-preserving pipelines[74] for contexts where access to data is restricted as illustrated by work on Australian aboriginal languages such as Kunwinju and Muruwari. Once established, ASR can support community-facing uses — voice input, captions, and spoken interfaces for nonliterate users.

3.4. AI Applications for Supporting Language Tooling and Datasets

With speech and audio language data successfully digitized, the final stages of digitization involve creating and adapting a range of supporting digital language tools that ensure true digital inclusion. AI is increasingly the foundation of these tools, which enable speakers of a language to go beyond just writing their language digitally and be able to use and engage with their language in the digital sphere in accurate, safe, and accessible ways. In our schematization, we classify into basic digital affordances (e.g., machine translation systems, grammar- and spell-checkers, indexing, and search engines) and more advanced digital affordances (e.g., natural language processing, voice assistants, e-learning tools).

Machine translation (traditional, speech-to-speech, speech-to-text): Machine translation remains a cornerstone of language technology. Aside from serving as an important tool in its own right for speakers of the language to interact in the world and with other languages, it enables crucial access to tools and platforms otherwise limited to higher-resource ones. In the digitization pipeline, machine translation can help expand endangered-language datasets by translating materials from dominant languages. Importantly, translation is no longer confined to text: Emerging systems can now convert speech in oral-only languages into written translations (e.g., Language A audio → Language B text), aiding documentation where corpora consist mainly of recordings.[75] Others support direct speech-to-speech translation (S2ST), as in Meta's 2022 S2ST system for Hokkien, which outputs synthesized speech in English without requiring written content in Hokkien.[76]

Translation is no longer confined to text: Emerging systems can now convert speech in oral-only languages into written translations

Grammar- and spell-checking systems: Historically, basic tooling for low-resource languages has relied on rule-based systems — such as Hunspell for Sorani Kurdish,[77] Norway's Divvun project for Sámi languages[78] — which handcraft lexicons and rules to provide precise spell- and grammar-checking. Though highly precise, these systems require intensive linguistic expertise to build and maintain. By contrast, emerging AI-driven methods are more scalable, though more data-hungry. While some researchers treat spelling correction as a machine translation task, others are turning to LLMs instead.[79] Comparative studies show that neural MT-based models outperform both rule-based methods and LLMs, yet adoption of these methods individually (or in conjunction) still hinges on available resources.[80]

Text-to-speech (TTS) systems: AI systems that convert text into spoken audio are powerful tools that can be used for educational purposes (e.g., in classrooms, learning apps), accessibility purposes (e.g., screen readers, voice assistants), and more.

Historically limited to high-resource languages due to their reliance on vast amounts of training data, recent multilingual neural TTS models are dramatically lowering data requirements. Some frameworks generate intelligible speech for unseen languages using only text data, while others require as little as five minutes of audio.[81] Where a single fluent speaker is available, even small TTS systems can produce usable outputs, and models like ZMM-TTS now deliver speech output without any training data at all on the target language.[82] These advances in TTS systems can bootstrap access to voice assistants, audiobooks, inclusive digital interfaces, and other supportive tools for DDLs.

Forced alignment tools: Forced aligners, which map segments of audio to corresponding text, build directly on ASR and TTS technologies. Automatically synchronizing transcriptions with speech is helpful in cases where transcription is time-consuming and costly, and they may also leverage cross-lingual alignment when training data is scarce.[83] While this technique eases the transcription bottlenecks that plague under-documented languages, it has also been successful for developing learning resources for Indigenous languages.[84] For instance, forced aligners underpin interactive tools like ReadAlong Studio, a suite of tools that enable students to make sing-along and read-along audiobooks in their native language.[85]

Large language models (LLMs): LLMs are becoming the scaffolding for advanced digital affordances, and while they are famously data-hungry, researchers now combine data augmentation techniques[86] — such as distant supervision, rule-based labeling, or cross-lingual projection — with human-in-the-loop co-annotation to streamline and scale the process of creating labeled language corpora needed to use these models.[87] When used carefully and consistently with community-driven documentation, this can turn LLMs into powerful tools to accelerate the transition of endangered languages into the digital sphere.[88] A collaboration along these lines was successfully used by researchers to construct an expert-validated text corpora for Nüshu, a rare script that rose to prominence in the 19th century when it was used by Yao women in China.[89] Researchers at Dartmouth, with the support of expert annotators, were able to to train the GPT-4 Turbo LLM on such data and ultimately created an expanded, first-of-its-kind Nüshu-Chinese digital corpus.[90]

4. Recommendations for Responsibly Harnessing AI's Potential

While AI offers transformative opportunities for advancing digital inclusion, significant structural, informational, and procedural challenges must be addressed for its potential to be fully and responsibly realized for under-resourced and digitally disadvantaged languages.

4.1. Building Trust and Empowering Communities

Language digitization is not purely a technical exercise — it is fundamentally a social contract between researchers, developers, and the communities whose languages and cultural heritage are being digitized. Before digitization or tool development begins, it is essential to establish shared goals, processes, and outcomes that are grounded in trust and community empowerment. Trust is built through co-design, where communities are engaged not only in the later stages of language digitization (e.g., as language data sources) but also in the early stages as co-owners of the processes and decision-makers at critical digitization junctures.

How and when to use AI tools to support language digitization workflows should be carefully considered decisions made by language digitization researchers and practitioners in conjunction with the relevant language communities. Together, they may, for example, decide that certain aspects of language digitization should never involve AI, while others can benefit from supporting AI tools. They may also agree on data governance frameworks that ensure community ownership over their language data. These decisions should also account for the environmental footprint of AI systems, including energy and water use associated with model training and data centers, which in some cases are located on or near Indigenous lands — sparking conversations around environmental justice and the unevenly distributed costs of digital inclusion.[91]

Language digitization is not purely a technical exercise — it is fundamentally a social contract between researchers, developers, and the communities whose languages and cultural heritage are being digitized.

Community empowerment can take many shapes, such as participatory research designs that center the voices of language community members and the creation of dedicated platforms for continued consultation and feedback. These approaches can help align research goals with community priorities while also addressing structural imbalances in research agendas that are often optimized for high-resource languages.[92] Such biases risk overlooking the intersectional challenges faced by low-resource communities, including joint infrastructure, data, and funding constraints.[93]

Recommendations:

- *Foster community-engaged convenings and collaborations:* AI researchers and their institutions should prioritize convening research groups, workshops, and conferences that aim to identify AI tooling needs by centering the voices of individual community members. In doing so, they should invest in and collaborate with existing grassroots research networks like AmericasNLP[94] and Ghana NLP[95] by supporting joint meetings, hackathons, and regional symposia. These convenings not only build local capacity and foster innovation but can create trusted environments for AI co-design and ensure technologies are developed collaboratively and transparently while strengthening local ownership.

- *Build community-driven benchmarks and standards for basic and advanced digital language tools:* Language digitization researchers and practitioners should convene multidisciplinary expert panels, funders, and community representatives to continually reevaluate current digitization approaches, co-design comparative benchmarks for language tools, monitor the adoption of supporting AI tools, and guide funders toward the most effective, scalable, and inclusive strategies. This guarantees that community participation is embedded at every stage of development and deployment.

4.2. Laying the Groundwork: Strengthening Research Foundations

A rapidly growing body of research is developing and applying AI tools for digital inclusion efforts, including language digitization and preservation — we highlight many but by no means all of these efforts in this white paper. More research is urgently needed to continue mapping this landscape and exploring promising new AI techniques that could further digital inclusion workstreams and overcome data scarcity challenges. At the same time, there is little consensus on how widely existing AI techniques have been adopted in this field so far, how they have been incorporated into existing workstreams, and which are most effective in addressing bottlenecks. Additional research and consensus-building is needed to answer these questions and help determine which approaches should be prioritized by investors funding research in this space as well as organizations adopting such AI tools in their digital inclusion work.

As technical AI advances continue at lightning speed, it will also become even more crucial for linguists, language technologists, and other digital inclusion researchers and practitioners to work closely with AI researchers and developers. More interdisciplinary exchanges are critical to ensuring that AI developers are dedicating time and resources to creating tools that meaningfully address actual bottlenecks in language digitization rather than building tools optimized for well-resourced use cases. Moreover, many promising language technologies remain underrecognized because they originate in community-led contexts and fall outside traditional academic publication channels. Strengthening collaboration between AI researchers and practitioners working directly with DDLs would help surface this work, assess tool maturity, and see that research and investment flow toward field-tested solutions with genuine impact.

While AI systems, from OCR systems to machine translation models to LLMs, are useful tools to support language digitization work, they must be placed in the broader context of digital capacity-building and other digital inclusion work. Over-prioritizing AI could lead to sidelining equally critical foundational

research and tools that do not involve AI. Foundational research into the current status of DDLs, for example, still lacks widely agreed upon definitions, indicators, and figures. While a variety of organizations and initiatives (from UNESCO's World Atlas of Languages to Translation Commons and Unicode) have made significant progress on mapping which languages have functioning digital tools (such as keyboards, grammar checkers, or automatic speech recognition), there is as yet no stable, unified, and up-to-date reference that takes into account the status of community testing, integration, and uptake.

Over-prioritizing AI could lead to sidelining equally critical foundational research and tools that do not involve AI.

Recommendations:

- *Create reliable resources to track progress on DDLs:* Language digitization researchers and practitioners should develop a living global dashboard or index that tracks the status of language tool availability, quality, and adoption for DDLs. Various efforts to track and visualize progress on DDLs already exist, but they could benefit from consolidation and consensus-building.[96]

- *Invest in, expand, and evaluate effective AI tools for language digitization:* Language technologists, linguists, AI researchers, and funders should jointly assess the maturity, applicability, and real-world adoption of current AI tools for language digitization, including community-built tools that may fall outside academic visibility, and direct research and investment toward those with the greatest potential impact. This requires convening expert panels to evaluate technical approaches, identifying promising but underrecognized solutions, uncovering gaps, and ensuring that AI investments are paired with sustained funding for the "boring but vital" human labor and infrastructure that make these tools effective and scalable.

- *Create forums for interdisciplinary exchanges:* Language digitization researchers and practitioners should create mechanisms for collaborating with AI researchers, engineers, and ethics scholars, such as interdisciplinary workshops, AI sprints, or other platforms that connect AI experts with linguists/language technologists in need of supporting tools.

4.3. Improving Workflows: From Foundations to Practice

Historically, language digitization efforts have typically followed a strict step-by-step approach. In this white paper, we visualize the different AI tools that can support individual digitization workflows on a linear path (see Figure 2). However, this framing should only be understood as an aid rather than a prescriptive blueprint for how digital inclusion must unfold. In reality, language digitization is deeply non-linear, requiring practitioners to move back and forth between stages as new constraints, opportunities, and community priorities emerge.

What's more, strictly adhering to a sequential approach can unnecessarily lengthen timelines. While certain processes must remain sequential — for example, it is imperative to stabilize a language's

writing system before beginning work on script encoding — parallel work is possible for many individual workflows. Type designers, for instance, can begin work on typefaces before script encoding is complete. And voice data collection can happen in parallel to lengthier script development and digitization work. Parallel work can also mean deliberately establishing practices and infrastructure ahead of time for steps that will later need to be repeated across tools and iterations. With AI tools aiding many of these processes, parallel workflows are more possible than ever and can lead to significant leaps in the speed of digitization work.

Organizations and individuals might also harness a completely different set of tools to speed up their digital inclusion work for under-resourced languages: AI tools for internal productivity and communications. For organizations that are understaffed and underfunded, AI tools can automate routine tasks (e.g., note-taking, grant proposal writing, social media post generation) or optimize project management (e.g., task management, scheduling, resource allocation), freeing human capacity for community engagement and strategic planning. AI tools can also help guide effective work allocation between humans and AI systems involved in the language digitization tasks (such as data annotation), though it is crucial for humans to remain in the loop throughout.[97]

Recommendations:

- *Move to parallel workflows for language digitization work:* Language digitization practitioners should identify and formalize opportunities for parallel workstreams — taking into account the supportive role that AI tools can play — in order to accelerate time-to-capacity while maintaining a rigorous approach.
- *Leverage AI for organizational improvements:* Organizations and individuals leading digital inclusion should pilot AI-driven productivity tools that are tailored to the needs of small, resource-limited teams and enable human oversight.

4.4. Forming Coalitions: From Practice to Adoption

A significant barrier to digital inclusion is not simply the development of new technologies but their actual adoption by technology companies. In many cases, success hinges less on complex AI breakthroughs and more on overcoming logistical barriers: connecting the right people, platforms, and political will. Both formal and informal incentive structures often discourage or delay the uptake of proven solutions.

The case of Urdu is illustrative. The language remained digitally sidelined for decades due to the complexity of its preferred script, nastaʿlīq. It wasn't until the early 2010s that the first Urdu keyboard was developed for iOS.[98] Yet Apple continued to use the simpler Arabic naskh script until a viral open letter in 2014 helped push for the system-wide adoption of a nastaʿlīq font.[99] Meanwhile, calligraphers spent a decade handcrafting the first locally developed Urdu digital font, and individual researchers collaborated to create an open-source Urdu dataset that laid the foundation for AI-powered tools like autocorrect and predictive text.[100] Still, to this day, limitations in rendering, platform support, and system compatibility persist. This example underscores that without sustained support from major tech platforms, languages like Urdu risk being reshaped — or erased — by the very technologies designed to preserve them.

Māori is another important case study. Like many Indigenous languages, *te reo Māori* declined sharply

as a result of colonial assimilation policies. Even after the Polynesian language stabilized, mainstream digital platforms continued to privilege more dominant languages and offered little control over Māori language data.[101] In response, the Māori-led nonprofit Te Hiku Media pursued a different path: Rather than relying on Big Tech infrastructure, it built its own digital platforms and AI tools for speech recognition, prioritizing community consent and data sovereignty.[102] To produce the transcribed audio needed to train speech models to counter the language's decline, Te Hiku mobilized thousands of community members through coordinated recording campaigns. This produced one of the first high-performing automatic speech recognition systems for te reo Māori, alongside tools for transcription, pronunciation feedback, and language learning.[103] The case is evidence that meaningful progress is possible outside Silicon Valley, but sustaining digital inclusion requires long-term institutional support, enforceable data governance, and ongoing negotiation over control of a language's digital future.

Meaningful progress is possible outside Silicon Valley, but sustaining digital inclusion requires long-term institutional support.

Digital inclusion efforts cannot succeed without continuous attention and support from policymakers, funders, journalists, and the language communities themselves. The highly technical nature of digital inclusion work can often obscure its everyday importance. While AI policy and ethics discussions have started turning to issues related to low-resource languages, conversations today predominantly revolve around cutting-edge technologies like LLMs and their accessibility and accuracy for such languages. Scholars rightly point to training data scarcity issues that make the development of linguistically and culturally accurate LLMs challenging for DDLs.[104] Yet few delve into the core digital infrastructure that many languages still lack, without which the creation of new language data is difficult if not impossible. For billions of people, tools such as grammar checkers, spell checkers, diverse typefaces, and stable keyboards remain the priority — and without them, advanced AI applications are largely irrelevant.

Recommendations:

- *Explore and implement mechanisms to study and reform incentive structures surrounding language tool adoption:* Language digitization researchers and practitioners should work with industry leaders and policymakers to identify where market, reputational, or resource barriers discourage adoption, and create policies or partnership models that encourage integration of digital inclusion technologies.

- *Strengthen storytelling for general audiences:* Language digitization researchers and practitioners should develop clear, community-centered narratives that link digital inclusion to daily life, civic participation, economic activity, and emergency response.

4.5. Ensuring Cultural Sustainability: From Adoption to Impact

Digital inclusion work does not end once language tools are adopted. Ensuring the sustained and

culturally sensitive long-term adoption of language tools and development of more advanced technologies (such as LLMs for newly digitized languages) is just as important. In the midst of automation and acceleration, the continuous inclusion of language experts and members of the relevant language community is vital to ensuring genuine cultural representation and sustainable self-determination.[105] Excluding community perspectives risks alienating or even harming the very groups such tools are meant to serve and hurts long-term adoption once funding cycles end.

Digital inclusion work does not end once language tools are adopted. Ensuring the sustained and culturally sensitive long-term adoption of language tools and development of more advanced technologies (such as LLMs for newly digitized languages) is just as important.

An example of what happens when linguistic communities are left out of AI development is the steep cost, performance, and environmental penalties faced by non-English speakers in low-income countries when accessing LLMs.[106] Because LLM application programming interfaces (APIs) charge by the token, and many of these languages break into more tokens during processing than high-resource languages, speakers of languages like Bengali, Amharic, or Santali pay at least six times more for the same task.[107] They also generate a greater computational and emissions load. This penalty is not inherent to the languages themselves but to design choices made without including those communities at the table, a dynamic that reinforces existing power imbalances.[108] Resolving the issue requires involving diverse speakers throughout the AI development life cycle — from tokenization algorithms to corpus building, from model evaluation to pricing discussions — so that AI access becomes both fair and sustainable.

Researchers, including one of our co-authors, have outlined recommendations for future culturally aware language technologies, emphasizing the importance of scalable, diverse cultural knowledge and inclusive practices that must be embedded throughout the tool-development life cycle.[109] Just as communities and users should deliberate on whether to pursue digitization at all, they must also participate in the initial stages of problem selection and outcome definition. Such participation remains critical during data collection, as shown by Culture Cartography and CultureBank, a community-based cultural data collection methodology where native speakers work with LLMs to document cultural knowledge salient to themselves but currently unknown to LLMs via a mixed-initiative, human-AI collaboration manner.[110] This approach is an example of a scalable pathway for addressing the Western-centric biases in LLM training data and improving the capacity of LLMs to serve diverse global users.

Tool development itself must also directly engage intended users[111] through co-design and testing akin to Masakhane's participatory model — a grassroots research collective that advances NLP for African languages by and for African researchers and communities.[112]

Finally, continuous, rigorous, and context-aware evaluation and benchmarking is essential for evaluating success beyond performance metrics and for directing investment toward research that is not only scalable and effective but also inclusive of diverse cultural perspectives.[113]

Recommendations:

- *Empower culturally aware AI development:* Rather than treating communities merely as data sources, there should be greater investment in building technical capacity so community members can lead development of culturally aware LLMs that serve their needs and reflect their values. In other words, infrastructure and technical training need to be provided to community members to help them build systems grounded in their own language expertise and cultural knowledge alongside external collaborators.

- *Promote contextualized adoption and impact assessment:* Work directly with community members to identify both intended and unintended consequences of AI deployment for their context, drawing on their expertise about local norms and language practices to gain contextualized understandings and actionable mitigation plans.

Endnotes

1. Lindell Bromham et al., "Global Predictors of Language Endangerment and the Future of Linguistic Diversity," Nature Ecology & Evolution 6, No.2 (February 2022), 163–73, https://www.nature.com/articles/s41559-021-01604-y.
2. Declan Walsh, "The World Is Becoming More African," *New York Times*, October 28, 2023, https://www.nytimes.com/interactive/2023/10/28/world/africa/africa-youth-population.html.
3. Juan N. Pava et al., "Mind the (Language) Gap: Mapping the Challenges of LLM Development in Low-Resource Language Contexts," *Stanford Institute for Human-Centered Artificial Intelligence*, April 22, 2025, https://hai.stanford.edu/policy/mind-the-language-gap-mapping-the-challenges-of-llm-development-in-low-resource-language-contexts.
4. UNESCO Ad Hoc Expert Group on Endangered Languages, "Language Vitality and Endangerment," March 2003, https://ich.unesco.org/doc/src/00120-EN.pdf.
5. See, for example: Graham Lee Brewer, "Lakota Elders Helped a White Man Preserve Their Language. Then He Tried to Sell It Back to Them," *NBC News*, June 3, 2022, https://www.nbcnews.com/news/us-news/native-american-language-preservation-rcna31396; Marieke Glorieux-Stryckman, "AI Outrage: Error-Riddled Indigenous Language Guides Do Real Harm, Advocates Say," *The Gazette*, December 16, 2024, https://montrealgazette.com/news/ai-outrage-error-riddled-indigenous-language-guides-do-real-harm-advocates-say.
6. See, for example, Meta's Massively Multilingual Speech recognition model: Meta AI, "Introducing Speech-to-Text, Text-to-Speech, and More for 1,100+ Languages," May 22, 2023, https://ai.meta.com/blog/multilingual-model-speech-recognition/.
7. Charity Ferreira, "Stanford Initiative Seeks to Bring More Languages Online," *Stanford Report*, April 22, 2025, https://news.stanford.edu/stories/2025/04/Stanford-initiative-seeks-to-bring-more-languages-online.
8. Mark Davis, Greg Welch, and Steven Loomis of the Unicode Consortium are credited with coining the term "digitally disadvantaged language." See, for example: Isabelle A. Zaugg, "Digitally-Disadvantaged Languages," *Internet Policy Review* 11, no. 2 (2022): 1–11, https://policyreview.info/glossary/digitally-disadvantaged-languages; SILICON, "Inclusion for Digitally Disadvantaged Languages (Toral Cowieson, Maroua Bezzaoui, Samuel Benzecry)," September 1, 2025, https://youtu.be/drPys0zJTT4?si=VEB3rWYQRib95uZa&t=278.
9. Council of Europe, "European Charter for Regional or Minority Languages," May 11, 1992, https://rm.coe.int/1680695175; Bibi Halima and Keli Yerian, "Majoritized and Minoritized Languages," *Learning How to Learn Languages* (2024), https://opentext.uoregon.edu/languagelearningedition1/chapter/majoritized-and-minoritized-languages/.
10. UNESCO, "Language Vitality and Endangerment."
11. Department of Rural and Community Development and the Gaeltacht, "Minister Calleary Announces €4.9 Million in Funding for Irish Language Digital Projects in Dublin City University," November 21, 2025, https://www.gov.ie/en/department-of-rural-and-community-development-and-the-gaeltacht/press-releases/minister-calleary-announces-49m-in-funding-for-irish-language-digital-projects-in-dublin-city-university/.
12. Zaugg, "Digitally-Disadvantaged Languages."
13. Pava et al., "Mind the (Language) Gap"; Gabriel Nicholas and Aliya Bhatia, "Lost in Translation: Large Language Models in Non-English Content Analysis," *Center for Democracy & Technology,* May 23, 2023, https://cdt.org/insights/lost-in-translation-large-language-models-in-non-english-content-analysis/.
14. Yuecai Pan et al., "A New Dataset for Mongolian Online Handwritten Recognition," *Scientific Reports* 13, no. 1 (January 2023): 26, https://www.nature.com/articles/s41598-022-27267-8; Biligsaikhan Batjargal, "A Survey on Rendering Traditional Mongolian Script," *2010 International Conference on Asian Language Processing* (2010): 3-6, https://ieeexplore.ieee.org/document/5681554.
15. Donavyn Coffey, "Māori Are Trying to Save Their Language From Big Tech," *Wired*, April 28, 2021, https://www.wired.com/story/maori-language-tech/.
16. Welsh government, "Welsh Language Technology Action Plan," January 23, 2018, https://www.gov.wales/welsh-language-technology-action-plan-2018-2024.
17. Hellina Hailu Nigatu et al., "The Zeno's Paradox of 'Low-Resource' Languages" *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (November 2024): 17753–74, https://aclanthology.org/2024.emnlp-main.983.pdf.
18. This table is adapted from Ethnologue's Digital Language Support classification. See: David M. Eberhard, Gary F. Simons and Charles D. Fenning, eds.., *Ethnologue: Languages of the World*, 28th ed. (Dallas, Texas: SIL International, 2025), https://www.ethnologue.com/ethnoblog/welcome-28th-edition/; Gary F. Simons, Abbey L.L. Thomas, and Chad K.K. White, "Assessing Digital Language Support on a Global Scale," *Proceedings of the 29th International Conference on Computational Linguistics* (October 2022): 4299–4305, https://aclanthology.org/2022.coling-1.379/.
19. We use the term "potential" — rather than the original "digitally still" — to describe languages that currently lack digital support. This terminology aims to highlight the capacity and opportunity for these languages to be digitally integrated, while avoiding the normative assumption that non-digitization constitutes a deficiency.
20. Nina Markl and Stephen Joseph McNulty., "Language Technology Practitioners as Language Managers: Arbitrating Data Bias and Predictive Bias in ASR," preprint, arXiv, submitted February 25, 2022, https://arxiv.org/abs/2202.12603.
21. The Unicode Standard, https://www.unicode.org/standard/standard.html.
22. Kevin King et al., "Typotheque Indigenous North American Type," *Typotheque*, July 9, 2024, https://www.typotheque.com/research/north-american-research.
23. Unicode CLDR Project, https://cldr.unicode.org/.
24. See, for example: Mildred Europa Taylor, "Meet the Guinean Brothers Who Created the First-Ever Alphabet for Their Native Fulani Language," *Face2Face Africa*, April 1, 2020, https://face2faceafrica.com/article/meet-the-guinean-brothers-who-created-the-first-ever-alphabet-for-their-native-fulani-language.
25. Ramsha Saeed et al., "Detection of Offensive Language and ITS Severity for Low Resource Language," *ACM Transactions on Asian and Low-Resource Language Information Processing* 22, no. 6 (January 2023): 1–27, https://dl.acm.org/doi/full/10.1145/3580476.
26. Tanjim Mahmud, "Cyberbullying Detection for Low-Resource Languages and Dialects: Review of the State of the Art," *Information Processing & Management* 60, no. 5 (September 2023): 103454, https://www.sciencedirect.com/science/article/abs/pii/S0306457323001917.
27. Ahmed Aleroud et al., "An Examination of Susceptibility to Spear Phishing Cyber Attacks in Non-English Speaking Communities," *Journal of Information Security and Applications* 55 (December 2020): 102614, https://www.sciencedirect.com/science/article/abs/pii/S2214212620307791.
28. Asanilta Fahda and Ayu Purwarianti, "A Statistical and Rule-Based Spelling and Grammar Checker for Indonesian Text," *2017 International Conference on Data and Software Engineering* (2017): 1–6, https://ieeexplore.ieee.org/document/8285846.
29. The original graphic was co-developed by the Unicode Consortium and SILICON (Stanford Initiative on Language Inclusion and Conservation in Old and New Media), with involvement from Maroua Bezzaoui, Toral Cowieson, Mark Davis, Steven Loomis, and the CLDR-DDL subcommittee of the Unicode Consortium.
30. Bushra Ebadi, "Technology Alone Can't Preserve Endangered Languages," *Center for International Governance Innovation*, June 30, 2018, https://www.cigionline.org/articles/technology-alone-cant-preserve-endangered-languages/.
31. The Rosetta Project, https://rosettaproject.org/; Endangered Languages Project, https://www.endangeredlanguages.com/.
32. Antonios Anastasopoulos et al., "Endangered Languages Meet Modern NLP," *Proceedings of the 28th International Conference on Computational Linguistics: Tutorial Abstracts* (December 2020): 39–45, https://aclanthology.org/2020.coling-tutorials.7/; Sarah Lain, "Half the World's Languages Are Endangered—But AI Can Help Save Them," *Three Magazine*, July 22, 2024, https://www.threemagazine.com/language-headline/.
33. Shruti Rijhwani et al., "User-Centric Evaluation of OCR Systems for Kwak'wala," preprint, arXiv, submitted February 26, 2023, https://arxiv.org/abs/2302.13410.
34. Thomas Reitmaier et al., "Opportunities and Challenges of Automatic Speech Recognition Systems for Low-Resource Language Speakers," *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (April 2022): 1–17, https://dl.acm.org/doi/abs/10.1145/3491102.3517639.
35. See, for example: Claudio S. Pinhanez et al., "Balancing Social Impact, Opportunities, and Ethical Constraints of
36. A significant amount of work in this space is inevitably underrepresented, especially on the left side of the diagram. Much of it is grassroots-led, highly practical, and not published in academic venues, which tend to prioritize more polished, application-ready tools built on large existing datasets. This skews visibility toward later-stage technologies. Importantly, the relevance of these tools is also highly context-dependent: Some communities may find machine translation far less useful and better understood as an "advanced" affordance while placing greater priority on voice or digital assistants, functioning as more "basic" and immediately impactful tools. We thus echo calls for better and context-sensitive assessments of tool maturity so the field can identify which innovations meaningfully address the most urgent bottlenecks. Special thanks to the many scholars and practitioners whose work informed this visualization of the digital inclusion process: Anshuman Pandey, Andrew Glass, Mark Davis, Toral Cowieson, Debbie Anderson, Anushah Hossain, Steven Loomis, Conrad Nied, Manish Goregaokar, and Ben Joeng (Yang). All omissions or errors remain the sole responsibility of the authors.
37. Te Mana Raraunga Māori Data Sovereignty Network, https://www.temanararaunga.maori.nz/.
38. Advancing Indigenous Data Sovereignty, https://www.idsovandresearcher.com/about.
39. James Route et al., "Multimodal, Multilingual Grapheme-to-Phoneme Conversion for Low-Resource Languages," *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP* (November 2019): 192–201, https://aclanthology.org/D19-6121.pdf.
40. Alex Kasonde, "The Long March to Unicode: A Digital Approach to Variability in Icibemba Orthography," *Cogent Arts & Humanities* 12, no. 1 (March 2025), https://www.tandfonline.com/doi/full/10.1080/23311983.2025.2477347.
41. Pin-Jie Lin et al., "Modeling Orthographic Variation Improves NLP Performance for Nigerian Pidgin," *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation* (May 2024): 11510–22, https://aclanthology.org/2024.lrec-main.1006.pdf.

42. William D. Lewis, "Haitian Creole: How to Build and Ship an MT Engine from Scratch in 4 days, 17 hours, & 30 minutes," *Proceedings of the 14th Annual Conference of the European Association for Machine Translation (May 2010)*, https://aclanthology.org/2010.eamt-1.37/.
43. Xinjian Li, "Low-Resource Speech Recognition for Thousands of Languages" (doctoral dissertation, Carnegie Mellon University, July 2023), https://www.lti.cs.cmu.edu/people/alumni/alumni-thesis/li-xinjian-thesis.pdf.
44. Route et al., "Multimodal, Multilingual Grapheme-to-Phoneme Conversion for Low-Resource Languages."
45. Musica Supriya et al., "Developing a Hybrid Morphological Analyzer for Low-Resource Languages,"*Applied Sciences* 15, no. 10 (May 2025): 5682, https://www.mdpi.com/2076-3417/15/10/5682; Adam Wiemerslage et al., "Morphological Processing of Low-Resource Languages: Where We Are and What's Next," preprint, arXiv, submitted March 16, 2022, https://arxiv.org/abs/2203.08909.
46. Sarah Moeller et al., "IGT2P: From Interlinear Glossed Texts to Paradigms," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (November 2020): 5251–62, https://aclanthology.org/2020.emnlp-main.424.pdf.
47. Jiatong Shi, "Leveraging End-to-End ASR for Endangered Language Documentation: An Empirical Study on Yoloxóchitl Mixtec," *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics* (April 2021): 1134–45, https://aclanthology.org/2021.eacl-main.96.pdf; Zoey Liu et al., "*Not Always About You*: Prioritizing Community Needs When Developing Endangered Language Technology," *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* 1 (May 2022): 3933–44, https://aclanthology.org/2022.acl-long.272.pdf.
48. Eleonora Bernasconi and Stefano Ferilli, "Enhancing Symbol Recognition in Library Science via Advanced Technological Solutions," *Information* 16, no. 2 (February 2025): 119, https://www.mdpi.com/2078-2489/16/2/119.
49. Xiaolei Diao et al., "Ancient Script Image Recognition and Processing: A Review," preprint, arXiv, submitted June 24, 2025, https://arxiv.org/abs/2506.19208.
50. Zhiheng Wang and Jiarui Liu, "One-Shot Multilingual Font Generation Via ViT," preprint, arXiv, submitted December 15, 2024, https://arxiv.org/abs/2412.11342.
51. Samaneh Azadi et al., "Multi-Content GAN for Few-Shot Font Style Transfer," preprint, arXiv, submitted December 1, 2017, https://arxiv.org/abs/1712.00516.
52. Congwang Bao, Yuan Li, and En Lu, "Design and Implementation of Dongba Character Font Style Transfer Model Based on AFGAN," *Sensors* 24, no. 11 (May 2024): 3424, https://www.mdpi.com/1424-8220/24/11/3424.
53. Bao et al., "Design and Implementation of Dongba Character Font Style Transfer Model Based on AFGAN."
54. Shumeet Baluja, "Learning Typographic Style," preprint, arXiv, submitted March 13, 2016, https://arxiv.org/abs/1603.04000.
55. Baoyao Zhou, Weihong Wang, and Zhanghui Chen, "Easy Generation of Personal Chinese Handwritten Fonts," *2011 IEEE International Conference on Multimedia and Expo* (July 2011): 1–6, https://ieeexplore.ieee.org/document/6011892.
56. Aren Jansen et al., "A Summary of the 2012 JHU CLSP Workshop on Zero Resource Speech Technologies and Models of Early Language Acquisition," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (May 2013): 8111–15, https://ieeexplore.ieee.org/abstract/document/6639245; Odette Scharenborg et al., "Speech Technology for Unwritten Languages," *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (February 2020): 964–75, https://dl.acm.org/doi/10.1109/TASLP.2020.2973896; Xinsheng Wang et al., "Synthesizing Spoken Descriptions of Images," *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (October 2021): 3242–54, https://ieeexplore.ieee.org/document/9581052.
57. Tommi Jauhiainen et al., "Automatic Language Identification in Texts: A Survey," preprint, arXiv, last revised November 21, 2018, https://arxiv.org/abs/1804.08186.
58. Zhaodi Qi, Yong Ma, and Mingliang Gu, "A Study on Low-Resource Language Identification," *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference* (November 2019): 1897–1902, https://ieeexplore.ieee.org/document/9023075.
59. Gabriel Bernier-Colborne and Cyril Goutte, "Challenges in Neural Language Identification: NRC at VarDial 2020," *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects* (December 2020): 273–82, https://aclanthology.org/2020.vardial-1.26/.
60. Amir Hossein Kargaran et al., "GlotLID: Language Identification for Low-Resource Languages," preprint, arXiv, last revised July 2, 2024, https://arxiv.org/abs/2310.16248.
61. Amir Hossein Kargaran, François Yvon, and Hinrich Schütze, "GlotScript: A Resource and Tool for Low Resource Writing System Identification," preprint, arXiv, last revised March 27, 2024, https://arxiv.org/abs/2309.13320.
62. S. P. Sharan et al., "Palmira: A Deep Deformable Network for Instance Segmentation of Dense and Uneven Layouts in Handwritten Manuscripts," *Lecture Notes in Computer Science* (Springer Nature) 12822 (September 2021): 477–91, https://link.springer.com/chapter/10.1007/978-3-030-86331-9_31; Christian Reul, Uwe Springmann, and Frank Puppe, "LAREX: A Semi-Automatic Open-Source Tool for Layout Analysis and Region Extraction on Early Printed Books," preprint, arXiv, submitted January 20, 2017, https://arxiv.org/abs/1701.07396.
63. Rijhwani et al., "User-Centric Evaluation of OCR Systems for Kwak'wala."
64. Nevidu Jayatilleke and Nisansa de Silva, "Zero-Shot OCR Accuracy of Low-Resourced Languages: A Comparative Analysis on Sinhala and Tamil," preprint, arXiv, last revised August 25, 2025, https://arxiv.org/abs/2507.18264.
65. Yan Hon Michael Chung and Donghyeok Choi, "Finetuning Vision-Language Models as OCR Systems for Low-Resource Languages: A Case Study of Manchu," preprint, arXiv, July 9, 2025, https://arxiv.org/abs/2507.06761; Alik Sarkar et al., "Printed OCR for Extremely Low-Resource Indic Languages," Communications in Computer and Information Science (Springer Nature) 2474 (December 2024): 108–22, https://link.springer.com/chapter/10.1007/978-3-031-93691-3_9.
66. Oana Ignat et al., "OCR Improves Machine Translation for Low-Resource Languages," *Findings of the Association for Computational Linguistics* (May 2022): 1164–74, https://aclanthology.org/2022.findings-acl.92/.
67. Alexander Zahrer, Andrej Zgank, and Barbara Schuppler, "Towards Building an Automatic Transcription System for Language Documentation: Experiences from Muyu," *Proceedings of the Twelfth Language Resources and Evaluation Conference* (May 2020): 2893–2900, https://aclanthology.org/2020.lrec-1.353/.
68. Alec Radford et al., "Robust Speech Recognition via Large-Scale Weak Supervision," preprint, arXiv, submitted December 6, 2022, https://arxiv.org/abs/2212.04356.
69. Vineel Pratap et al., "Scaling Speech Technology to 1,000+ Languages," preprint, arXiv, submitted May 22, 2023, https://arxiv.org/abs/2305.13516.
70. Shao-Syuan Huang et al., "Enhancing Multilingual ASR for Unseen Languages via Language Embedding Modeling," preprint, arXiv, submitted December 21, 2024, https://arxiv.org/abs/2412.16474.
71. Jayadev Billa, "Leveraging Non-Target Language Resources to Improve ASR Performance in a Target Language," *Interspeech* (2021): 2581–85, https://www.isca-archive.org/interspeech_2021/billa21_interspeech.html.
72. Shreya Khare et al., "Low Resource ASR: The Surprising Effectiveness of High Resource Transliteration," *Interspeech* (August 2021): 1529–33, https://www.isca-archive.org/interspeech_2021/khare21_interspeech.html.
73. William Lane and Steven Bird, "Local Word Discovery for Interactive Transcription," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (November 2021): 2058–67, https://aclanthology.org/2021.emnlp-main.157/.
74. Nay San et al., "Automated Speech Tools for Helping Communities Process Restricted-access Corpora for Language Revival Efforts," *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages* (May 2022): 41–51, https://aclanthology.org/2022.computel-1.6/.
75. Scharenborg et al., "Speech Technology for Unwritten Languages."
76. Chen Zhang et al., "UWSpeech: Speech to Speech Translation for Unwritten Languages," *Proceedings of the AAAI Conference on Artificial Intelligence* 35, no. 16 (May 2021): 14319–27, https://ojs.aaai.org/index.php/AAAI/article/view/17684; "Meta's New AI-Powered Speech Translation System for Hokkien Pioneers a New Approach for an Unwritten Language" (blog), October 19, 2022, https://ai.meta.com/blog/ai-translation-hokkien/.
77. Sina Ahmadi, "Hunspell for Sorani Kurdish Spell Checking and Morphological Analysis," preprint, arXiv, submitted, September 14, 2021, https://arxiv.org/abs/2109.06374.
78. Divvun – Sámi language technology, https://divvun.no/en/.
79. Thierno Ibrahima Cissé and Fatiha Sadat, "Advancing Language Diversity and Inclusion: Towards a Neural Network-Based Spell Checker and Correction for Wolof," *Proceedings of the Fifth Workshop on Resources for African Indigenous Languages@ LREC-COLING* (May 2024): 140–151, https://aclanthology.org/2024.rail-1.16/; Agnes Luhtaru, Elizaveta Korotkova, and Mark Fishel, "No Error Left Behind: Multilingual Grammatical Error Correction with Pre-Trained Translation Models," *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics* 1 (March 2024): 1209–22, https://aclanthology.org/2024.eacl-long.73/; Frank Palma Gomez, Ali Rozovskaya, and Dan Roth, "A Low-Resource Approach to the Grammatical Error Correction of Ukrainian," *Proceedings of the Second Ukrainian Natural Language Processing Workshop* (May 2023): 114–120, https://aclanthology.org/2023.unlp-1.14/.
80. Mamadou K. Keita et al., "Grammatical Error Correction for Low-Resource Languages: The Case of Zarma," preprint, arXiv, last revised February 16, 2025, https://arxiv.org/abs/2410.15539; Linda Wiechetek et al., "Rules Ruling Neural Networks: Neural vs. Rule-Based Grammar Checking for a Low Resource Language," *Proceedings of the International Conference on Recent Advances in Natural Language Processing* (September 2021): 1526–35, https://aclanthology.org/2021.ranlp-1.171/.
81. Takaaki Saeki et al., "Learning to Speak from Text: Zero-Shot Multilingual Text-to-Speech with Unsupervised Text Pretraining," preprint, arXiv, last revised May 27, 2023, https://arxiv.org/abs/2301.12596; Florian Lux, Julia Koch, and Ngoc Thang Vu, "Low-Resource Multilingual and Zero-Shot Multispeaker TTS," preprint, arXiv, submitted October 21, 2022, https://arxiv.org/abs/2210.12223.
82. Gangular Singh Irengbam et al., "Text to Speech System for Meitei Mayek Script," preprint, arXiv, submitted August 9, 2025, https://arxiv.org/abs/2508.06870; Chen Gong et al., "ZMM-TTS: Zero-Shot Multilingual and Multispeaker Speech Synthesis Conditioned on Self-Supervised Discrete Speech Representations," preprint, arXiv, last revised August 27, 2024, https://arxiv.org/abs/2312.14398.

83. Graham Neubig et al., "A Summary of the First Workshop on Language Technology for Language Documentation and Revitalization," *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages and Collaboration and Computing for Under-Resourced Languages* (May 2020): 342–51, https://aclanthology.org/2020.sltu-1.48/; Juho Leinonen, Sami Virpioja, and Mikko Kurimo, "Grapheme-Based Cross-Language Forced Alignment: Results with Uralic Languages," *Proceedings of the 23rd Nordic Conference on Computational Linguistics* (May 2021): 345–50, https://aclanthology.org/2021.nodalida-main.36/.
84. Malgorzata Ćavar, Damir Ćavar, and Hilaria Cruz, "Endangered Language Documentation: Bootstrapping a Chatino Speech Corpus, Forced Aligner, ASR," *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (May 2016): 4004–11, https://aclanthology.org/L16-1632/.
85. ReadAlong Studio, https://readalong-studio.mothertongues.org/#/.
86. Michael A. Hedderich et al., "A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios," *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (June 2021): 2545–68, https://aclanthology.org/2021.naacl-main.201/.
87. Minzhi Li et al., "CoAnnotating: Uncertainty-Guided Work Allocation between Human and Large Language Models for Data Annotation," preprint, arXiv, submitted October 24, 2023, https://arxiv.org/abs/2310.15638; Pava et al., "Mind the (Language) Gap."
88. Daniel Oberhaus, "The Race to Save the World's Vanishing Languages," *The Atlantic*, December 5, 2024, https://www.theatlantic.com/sponsored/google/the-race-to-save-the-worlds-vanishing-languages/3954/.
89. Liming Zhao, *A Compendium of Chinese Nüshu* (Beijing: Tsinghua University Press, 1992), re-released as an ebook in 2019, https://books.google.com/books?id=0jC2zQEACAAJ.
90. Ivory Yang, Weicheng Ma, and Soroush Vosoughi, "NüshuRescue: Reviving the Endangered Nüshu Language with AI," *Proceedings of the 31st International Conference on Computational Linguistics* (January 2025): 7020–34, https://aclanthology.org/2025.coling-main.468/; Harini Barath, "Language Preservation Efforts Get an AI Boost," *Dartmouth News*, September 4, 2025, https://home.dartmouth.edu/news/2025/04/language-preservations-efforts-get-ai-boost.
91. Adam Zewe, "Explained: Generative AI's Environmental Impact," *MIT News*, January 17, 2025, https://news.mit.edu/2025/explained-generative-ai-environmental-impact-0117; Miacel Spotted Elk, "For Indigenous Communities, AI Brings Peril — And Promise," *Grist*, August 14, 2025, https://grist.org/indigenous/indigenous-peoples-examine-impact-of-ai-on-communities/.
92. Sebastian Ruder, Ivan Vulić, and Anders Søgaard, "Square One Bias in NLP: Towards a Multi-Dimensional Exploration of the Research Manifold," preprint, arXiv, submitted June 20, 2022, https://arxiv.org/abs/2206.09755.
93. Orevaoghene Ahia, Julia Kreutzer, and Sara Hooker, "The Low-Resource Double Bind: An Empirical Study of Pruning for Low-Resource Machine Translation," *Findings of the Association for Computational Linguistics: EMNLP 2021* (November 2021): 3316–33, https://aclanthology.org/2021.findings-emnlp.282/.
94. AmericasNLP, https://turing.iimas.unam.mx/americasnlp/.
95. Ghana NLP, https://ghananlp.github.io/.
96. See, for example: W3C, "Language Matrix: International Typography on the Web," https://w3c.github.io/typography/gap-analysis/language-matrix.html; Script Encoding Initiative, "Scripts to Encode," https://sei.berkeley.edu/scripts-to-encode/.
97. Li et al., "CoAnnotating"; Zijie J. Wang et al., "Putting Humans in the Natural Language Processing Loop: A Survey," preprint, arXiv, submitted March 6, 2021, https://arxiv.org/abs/2103.04044.
98. Alizeh Kohari, "How to Bring a Language to the Future," *Rest of World*, February 9, 2021, https://restofworld.org/2021/bringing-urdu-into-the-digital-age/.
99. Mudassir Azeemi, "An Open Letter to Tim Cook, and Jonathan Ive," *Medium*, October 9, 2014, https://azeemi.medium.com/an-open-letter-to-tim-cook-and-jonathan-ive-be38cf88d8ee.
100. Zeerak Ahmed, "Announcing Maḵẖzan," *Matnsaz*, November 3, 2019, https://matnsaz.net/blog/2019/11/announcing-makhzan.
101. Karen Hao, "A New Vision of Artificial Intelligence for the People," *MIT Technology Review*, April 22, 2022, https://www.technologyreview.com/2022/04/22/1050394/artificial-intelligence-for-the-people/.
102. Papa Reo, "Mātauranga – Research," https://papareo.nz/#matauranga.
103. Angie Lee, "Māori Speech AI Model Helps Preserve and Promote New Zealand Indigenous Language," Nvidia, January 16, 2024, https://resources.nvidia.com/en-us-global-public-sector/te-hiku-media-maori-speech-ai.
104. Pava et al., "Mind the (Language) Gap."
105. Abeba Birhane et al., "Frameworks and Challenges to Participatory AI," *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, (October 2022): 1–8, https://doi.org/10.1145/3551624.3555290.
106. Aivin V. Solatorio et al., "Double Jeopardy and Climate Impact in the Use of Large Language Models: Socio-economic Disparities and Reduced Utility for Non-English Speakers," preprint, arXiv, submitted October 14, 2024, https://arxiv.org/abs/2410.10665.
107. Orevaoghene Ahia et al., "Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models," preprint, arXiv, submitted May 23, 2023, https://arxiv.org/abs/2305.13707.
108. Pava et al., "Mind the (Language) Gap."
109. Wilhelmina Nekoto et al., "Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages," *Findings of the Association for Computational Linguistics: EMNLP 2020* (November 2020): 2144–2160, https://aclanthology.org/2020.findings-emnlp.195/.
110. Caleb Ziems et al., "Culture Cartography: Mapping the Landscape of Cultural Knowledge," preprint, arXiv, submitted October 31, 2025, https://arxiv.org/abs/2510.27672; Weiyan Shi et al., "CultureBank: An Online Community-Driven Knowledge Base Towards Culturally Aware Language Technologies," preprint, arXiv, submitted April 23, 2024, https://arxiv.org/abs/2404.15238.
111. James Landay, "'AI For Good' Isn't Good Enough: A Call for Human-Centered AI," *Stanford Institute for Human-Centered Intelligence*, February 13, 2024, https://hai.stanford.edu/events/ai-good-isnt-good-enough-call-human-centered-ai.
112. Masakhane, https://www.masakhane.io/.
113. Michael J. Ryan, William Held, and Diyi Yang, "Unintended Impacts of LLM Alignment on Global Representation," preprint, arXiv, last revised June 6, 2024, https://arxiv.org/abs/2402.15018.


Stanford University
Human-Centered
Artificial Intelligence


Stanford | SILICON