Routledge
Taylor & Francis Group

Learning, Media and Technology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/cjem20

Language, artificial education, and future-making in indigenous language education

Uma Pradhan & Joyeeta Dey

To cite this article: Uma Pradhan & Joyeeta Dey (05 Nov 2023): Language, artificial education, and future-making in indigenous language education, Learning, Media and Technology, DOI: 10.1080/17439884.2023.2278111

To link to this article: https://doi.org/10.1080/17439884.2023.2278111

Published online: 05 Nov 2023.

LEARNING, MEDIA AND TECHNOLOGY
https://doi.org/10.1080/17439884.2023.2278111

Routledge
Taylor & Francis Group

Language, artificial education, and future-making in indigenous language education

Uma Pradhan [a] and Joyeeta Dey[b]

[a]Institute of Education, University College London, London, UK; [b]NIAS, National Institute of Advanced Studies, Bangalore, India

ABSTRACT

This paper examines how language-based artificial intelligence is envisaged to imagine new futures for indigenous languages. It draws on the visions, programmes, and plans of six language initiatives that are developing language technology for often-marginalised indigenous, tribal, and minority (ITM) languages, such as Gondi, Maithili, Rajasthani and Mundari, in India. We note three distinct discourses: (1) technological optimism in utilising these new opportunities by claiming space for otherwise-marginalised languages, (2) the imperative for collaborative and collective work in order to address sparse datasets, and (3) the need to negotiate the contested nature of imagining a new collective future. This paper argues that indigenous language technology is not just a technical project but a contested process of subverting linguistic hierarchy through the 'active presencing' of these languages. Overall, the paper emphasizes the need for a nuanced approach that recognizes the interplay between technology, language education, and broader social and political factors.

ARTICLE HISTORY
Received 9 May 2023
Accepted 27 October 2023

KEYWORDS
Artificial intelligence; language education; indigenous language; marginalised language; India

Introduction

There has been a surge of new initiatives that use artificial intelligence (AI) to advance language education, including often-marginalised Indigenous[1], tribal, and minority (ITM) languages. The use of self-learning predictive models, machine-learning systems, and automated data processing through Natural Language Processing in these language technologies are expected to create innovative spaces for Indigenous, tribal, and minority (ITM) language education (Bali et al. 2019; NITI Aayog 2018). As ITM languages usually lack institutional spaces for learning like schools, these language technologies are anticipated to open up new learning opportunities, through its revitalization and increased functionality. This paper examines the discourse surrounding the new possibilities that arise from human-technology interactions and how these interactions may facilitate the (re)imagination of new futures for Indigenous languages, in contrast to their predicted disappearance and endangerment in the coming years (UNESCO 2010). Although there is a growing body of literature that recognizes the crucial role of technology in facilitating language learning, its focus tends to be on high-resource languages such as English (Ranathunga et al. 2023). These studies often neglect the fact that minoritized and low-resource languages may not have access to a large corpus or standardised language resources that are available to high-resource languages.

CONTACT Uma Pradhan u.pradhan@ucl.ac.uk 20 Bedford Way, London WC1H 0AL, United Kingdom

Even when the same technologies are used for both high-resource languages and ITM languages, the latter is likely to be impeded by an array of additional obstacles. This paper explores how language technology is expected to address the challenges that are faced by low-resourced and minoritized languages that are not mainstreamed in formal education or other institutions.

Emerging scholarship on human-technology interaction urges that we must 'take seriously the entanglement of the material and the ideational' (Littau 2016, 83). It emphasizes that technology should not be viewed simply as a new instrument for old tasks, but rather as a source of new creative possibilities that can generate innovative actions and meanings. In this paper, we explore whether and how this new technology is envisaged to offer unique perspectives on Indigenous languages and their futures. For this purpose, this paper draws on an analysis of six initiatives that use language technology for education in India, with a specific focus on the minoritized languages such as Gondi, Maithili, Rajasthani and Mundari languages. We specifically focus on the rationale of these initiatives, the activities that they engage in, and the constraints faced by them, in order to discuss how technologies are utilised to create conditions for language learning despite the broader structural constraints that hinder their learning.

While we acknowledge that Ed-tech in general still tends to represent digital neocolonialism (Adam 2019), there is always more complex tension that exists between the 'global' and 'local' in technology (Gallagher and Knox 2019). This paper explores the discourses beyond the understanding of technology as a hegemonic, surveillant and controlling force, which is used for the subjugation and erasure of Indigenous languages, to the contours of possibility offered by artificial intelligence to advance education for low-resourced languages. Firstly, the paper discusses the possibility of automation, the ability to work virtually, and the logic of world-making in these technologies are seen to potentially open up spaces for distinct visions of future-making that may not otherwise be accessible. Secondly, given the limited support available to the low-resource languages, we discuss that creative and collaborative work is articulated as an important way in which ITM language technologies have ensured the 'presence over absence' of minoritized knowledges (Vizenor 2008, 1). Collaborative work is seen as a way to overcome an additional burden of negotiating social and linguistic hierarchies, biases in existing technology, and tensions between various social actors. This paper argues that the use of language technology for Indigenous language education is, therefore, not just a technical project. Rather it is a contested and ongoing attempt to subvert linguistic hierarchy and imagine more equitable technological futures that are open to plural knowledge systems. Overall, the paper highlights the importance of adopting a nuanced approach that acknowledges the power-laden relationship between technology, language education, and broader social and political factors.

Artificial education, language, and future-making

With the rapid development of artificial intelligence (AI), this new technological space is increasingly being presented as a promising avenue for new futures in global education systems (Nemorin et al. 2023), including for language learning. Though the exact definition of the term AI is quite contested, it is often used to refer to self-learning predictive models, such as a machine-learning system, and more generally automated data processing by computational artefacts. This ability to learn independently, process big data, and automate work is seen to be transformative in expanding possibilities for the future, going beyond what appears to be possible in the now (Hautala and Ahlqvist 2022). These technological developments are expected not only to optimise education governance and improve learning outcomes (Selwyn and Gašević 2020) but also to make education more equitable and bridge the learning gaps (UNESCO 2022). It is in this vein that the UNESCO document 'Reimagining Our Futures Together', outlines a new vision for the digital in educational futures with a focus on inclusion, reiterates the need to create a 'more supple digital environment' that is not limited to 'specific and dominant strain of knowledge, unique to the post-Renaissance West' but more responsive to multiple knowledge systems (UNESCO 2022, 36).

Despite these grand visions to harness digital technologies, the ongoing development of artificial intelligence tends to overwhelmingly focus on mainstream and high-resource languages such as English. Applications such as Google Translate, Duolingo, Wooloroo etc. are increasingly becoming popular and responsive to the language learning needs of the language communities. However, these applications still require a large language corpus for Natural Language Processing i.e., the ability of computers to process and analyse human language, usually available only to high-resource languages. Given these social and political contexts, which do not necessarily create conditions that support minoritized languages, the development of these technologies for Indigenous, tribal, and minority (ITM) languages is faced with additional challenges of adequate resources, biases in existing technology, and linguistic hierarchies. While the use of AI in language education enables the reconfiguration of space, time and responsibility with the potential of language pedagogy integrated into a variety of ways, their access, availability, and utility may be different for different communities (Selwyn and Facer 2014). Thus, imaginations of the potential futures remain severely limited, especially in contexts where Indigenous languages are required to adapt to the existing AI infrastructures.

Given the unequal foundations, this future-making project is fundamentally a task to envision the rearrangement of power relations. While the possibility of automation and logic of world-making offer new arenas for language development, imagining Indigenous futures also requires an acknowledgement of language hierarchies that tend to overpower technical developments and find ways to indigenise the technology. Smith et al. (2016) identify this as a 'gnawing sense of mayhem' as different knowledge systems attempt to come together, within a broader context of unequal relations. These underlying power relations often materialise in the form of inadequate language data, lack of language recognition and absence of or contested standardisation, effective pedagogic instruments, and incompatibility with existing 'foundation models' of machine learning, as well as the disconnect between goals of the technology applications to the needs of the community (Mager and Katzenbach 2021). These constraints may also arise from the positionality and embodiment of those tasked with imagining, especially given the 'whiteness' of artificial intelligence spaces (Cave and Dihal 2020). Especially, given the uncomfortable history of technology-Indigenous interactions, where technology has been mainly used for the subjugation and erasure of different knowledges, and the further mainstreaming of dominant languages, imagining Indigenous futures through technology poses both conceptual and practical challenges.

Bringing together Indigenous languages and AI technology is, therefore, a very challenging task in imagining 'Indigenous futurity' (Tuck and Gaztambide-Fernández 2013). Scholars contend that this new future is fundamentally different from 'settler futurity' which requires the erasure of Indigenous people. Instead, the imagination of these new futures opens spaces for more plural epistemology and does not necessarily erase other knowledges, including settler epistemology (Tuck and Gaztambide-Fernández 2013; Yang 2000). This potential for engagement between western and non-western knowledge generates what Battiste and Henderson (2021) refer to as Trans-Systemic Knowledge Systems which expand spaces across various knowledge systems. The future, here, is not as a thing but as a continuous and iterative process of ongoing action (Bryant and Knight 2019). Such creative and political reimagining of the future is distinct from more popular 'predictive', and 'anticipatory' engagements which are increasingly overtaking educational discourse (Facer 2011). These may come alive through interventions in the present, through a particular reframing of past events, or in the planning of future visions. Within these relational spaces, Indigenous ways of 'knowing, being, and doing' find expression (Martin and Mirraboopa 2003), allowing us to envisage the possibilities of the time yet to come (Naidoo 2016).

Given that the most predicted future for the ITM languages is of death and disappearance (Moore 2006; UNESCO 2010), the construction of new future/s also includes an act of refusal of 'Indigenous absence and erasure' through 'active presencing' of multiplicity of knowledges (Vizenor 2008, 1). Vizenor (2008) explores how activism utilizes public art and graffiti murals to inhabit spaces, highlighting their capacity for world-making and future-making for minoritised

communities. And as some studies show, AI has been deployed by different communities in creating digital art to explore various possibilities of imagining a new world by (i) representing community memories as well as (ii) imagined futures, while enabling collaboration and digital visibility (Christen 2004). In overcoming both the ideational and practical gap between technology and Indigenous knowledges, recent studies show that Indigenous media futures employ the twin approach of creativity and politics – looking at how technology can be used to build these futures, and what opportunities for connection and rearrangement of power relations they afford (Cartee 2003).

It is, therefore, important to understand the narratives, ideas, and discourses that shape these imaginations of the new futures for ITM languages. Narratives are not merely semantics but are crucial in constructing 'disparate facts in our own worlds and weave them together cognitively in order to make sense of our reality' (Patterson and Monroe 1998, 315). This paper explores these discourses as spaces that hold the potential to instigate social reconfigurations by constructing new knowledge, new representations and new subjects (Somers and Gibson 1994). Building on these insights, we examine whether and how these narratives perceive new technology as offering fresh perspectives on minoritised languages and their futures. In doing so, while recognizing AI for ITM languages as a project to develop technology for diverse languages, we also critically recognize AI as a social, cultural, political, and relational practice that cannot be separated from the existing social hierarchies (Crawford 2021).

Research context and methodology

This paper explores the technology initiatives that use artificial intelligence (AI) to advance minoritized languages in India. The use of technology in language education in India is embedded within the complex history of policy on minority or Indigenous languages. Legal protections for linguistic minorities in India are anchored in Article 19 of the Indian Constitution (Reddy 2019). Given that India has 1369 different languages, many of which have disappeared, Constitutional protection is viewed as necessary for the goal of maintaining cultural diversity, especially since there are entrenched legal and educational obstacles to language equality such as Article 351 which directs the promotion of Hindi as a link language. However, this has resulted in Hindi having an 'imperialising effect' on linguistic minorities. Thus, the digital revitalisation of minority languages effectively serves as an anti-imperialisation move.

In order to address this language inequality, the Supreme Court of India ruled that a linguistic minority is any group that has at least one spoken language, regardless of whether or not it has a written script (Tyagi 2003). This was followed by The National Commission for Minorities Act 1992 (a commission to safeguard the rights of linguistic minorities), the National Framework for Action on Minorities 2005 (advocating the use of minority languages in education) and the National Policy on Education 2003 (supporting the use minority languages in the media). These struggles for the protection of language rights, including the legal recognition of the rights of linguistic minorities, show the inherently political nature of language within tribal activism in India, and its resonance with the shared threat of language erasure for Indigenous communities around the world (Shulist 2018; Tyagi 2003).

In recent years, the Indian government has turned to technology in its project of revitalizing minority and Indigenous languages (IndiaAI 2021; Indian Express 2020). The initiatives range from digitizing scripts of minority languages in standardized forms that allow for use across platforms, creating digital resources for learning and teaching these languages including e-learning platforms, the software, information processing tools, human-machine interfaces created by Department of Electronics and Information Technology (DeitY) as part of the Language Technology Promotion Scheme and the use of AI by the National Translation Mission whose focus on regional languages includes minority languages (Reddy 2019; Tyagi 2003). The National Strategy for Artificial Intelligence (NITI Ayog 2018) also articulates a clear vision of AI's potential to transform the education sector. With national initiatives such as AI4 Bharat and recent policy

development in National Education Policy, the scope and scale for the developments and use of AI in education are expanding. In this paper, we draw on the analysis of work by six initiatives on minoritised language technology, led and managed by Aripana Foundation, Microsoft, Indian Institute of Technology Kharagpur, Rajasthani Bhasha Academy, BR Ambedkar University, and Pratham Digital (see Table 1). These projects cover languages spoken in various states of India (See Figure 1).

We purposely selected these initiatives as they are working on different minoritised languages, specifically, Gondi, Maithili, Rajasthani, and Mundari. The initiatives are of varying scales – ranging from Microsoft, a tech behemoth, to Aripana Foundation, a small non-profit working out of rural Bihar. One was a university project housed out of BR Ambedkar University, involved in creating tech for multiple low-resource languages as part of a university project. Rajasthani Bhasha Academy, a large web portal for language learning, provided resources exclusively in Urdu language. They hosted pre-recorded animated lectures, gamified vocabulary acquisition exercises as well as automated assessments in the form of quizzes. Microsoft Research Lab is working on Project ELLORA – Enabling Low Resource Languages, with a team of scientists and engineers developing Machine Learning and AI, collaborating with non-state, often philanthropic, organisations to 'empower marginalised populations' using technology. The project on building the Mundari App is funded by Microsoft and created by the Indian Institute of Technology, Kharagpur. They are also collating data on this low-resource language by recording, transcribing, translating and creating a digital script of Mundari Bani. This app allows Mundari speakers access to the digital world. Aripana Foundation is a grassroots organisation based out of rural Bihar. As a part of their work on reading and literacy, they translate children's books into Maithili. They have also collaborated with Google to create a digital script for Maithili and collated data for automated translation. Pratham Digital, a wing of Pratham Education Foundation, works on AI and ML on specific technologies including producing voice-based question answering and automated grading for their large-scale assessment ASER.

For the initial identification of the projects, we searched on the internet for examples of organisations or projects on technology-led language teaching. We deliberately selected initiatives that use technologies for low-resourced languages that are not mainstreamed in institutional spaces such as schools. This was followed by systematically gathering information available on these initiatives that are available on their websites, reports and published news. We then identified

Table 1. Key features of technology initiatives.

Technology Initiatives	Language	UNESCO Status[2]	Developed With	Key Features of the programme	Regions spoken in India
Aripana Foundation	Maithili	Potentially vulnerable	Adult Maithili speakers	Language database, typing software	Bihar, Jharkhand
Microsoft Research	Gondi	Potentially vulnerable/ endangered	Adult Gondi speakers	Language database	Chhattisgarh, Madhya Pradesh, Andhra Pradesh, Telangana, Maharashtra, Odisha
Mundari App by IIT Kharagpur	Mundari	Potentially vulnerable	Adult Mundari speakers of West Bengal	Web-based app	Jharkhand, Chhattisgarh, Odisha, West Bengal, Assam
Rajasthani Bhasha Academy	Rajasthani	Not listed[3]	Adult Rajasthani diaspora	Video lectures	Rajasthan, Haryana, Gujarat, Uttar Pradesh, Madhya Pradesh
BR Ambedkar University	Multiple low resource regional languages	Endangered Languages	With adult populations	Language database typing software	Across the country
Pratham Digital	Non-hindi Regional languages	Endangered Languages	With children across the country	Voice Recognition software	Across the country

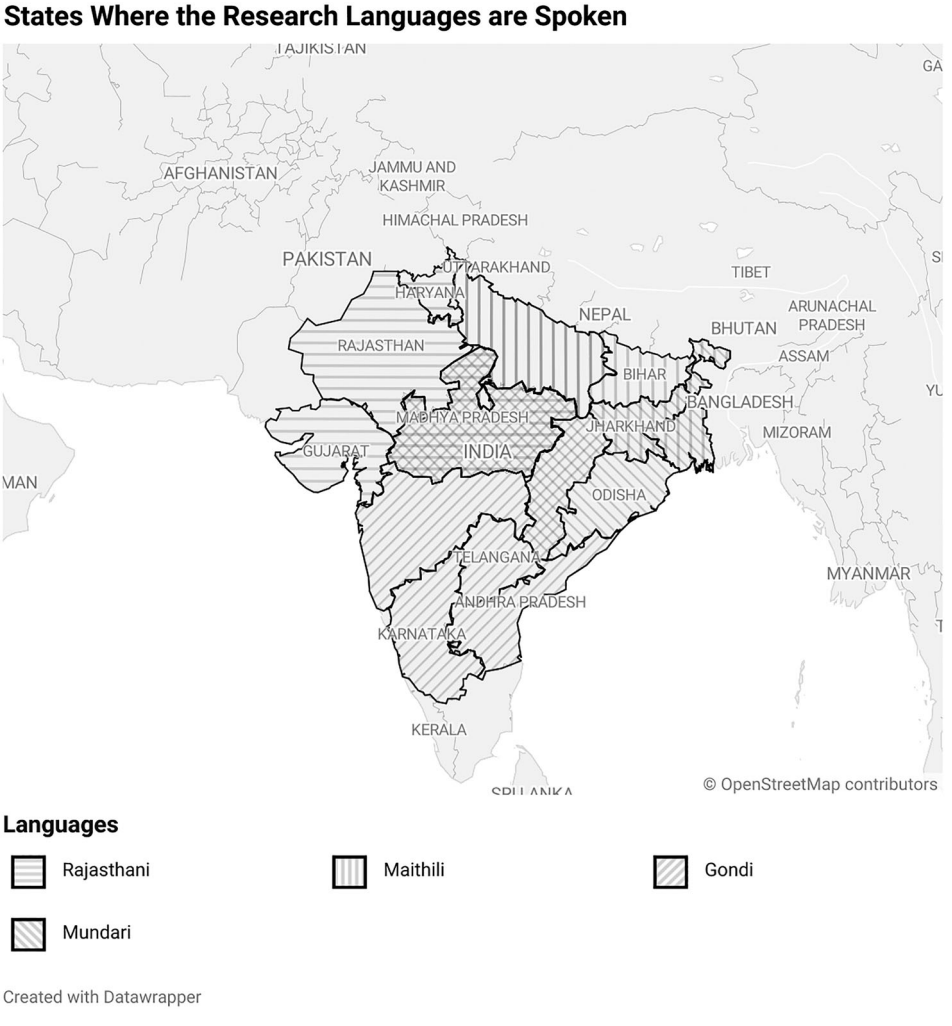

Figure 1. Map.

interesting textual points around three distinct themes. First, we analysed the rationale for using technology and artificial intelligence for language education. Second, we also looked at their different programmes and activities. We have also analysed the design of the technology initiatives looking at the degree of participation that was accorded to Indigenous people in shaping such design (McLoughlin 2000). And third, we explored the constraints faced by the initiatives, and the strategies used to overcome them. We used an iterative process of coding, layering, and identifying meaningful themes so that these unstructured texts could be analysed for more meaningful thematic analysis. We then followed five stages of thematic analysis: Familiarity with the data; Generating initial codes; Looking for themes; Reviewing themes; and Naming and defining themes (Braun and Clarke 2006). This process of analysis also enabled us to identify the themes that show the tensions between the dominant and minority knowledge paradigms. The literature we have referred to which frames our analysis is careful to also draw on the work of Indigenous scholars (Macgilchrist, Potter, and Williamson 2022). The following sections will discuss these themes, particularly how the role of technology has been conceptualised and implemented so as to navigate the broader structural constraints that hinder Indigenous language learning.

Technological optimism and 'active presencing' of indigenous languages

The language initiatives were overwhelmingly guided by the idea that view language-based AI is a powerful tool to support ITM language education, especially to facilitate language learning, preserve endangered languages, and promote cultural diversity. This technological optimism was demonstrated by all initiatives, as they sought to improve access to services like banks and libraries, as well as promote language education and revitalization. While distinct in their approaches, all six initiatives conceived the language learning platforms and digital infrastructures as a way to use, preserve or keep the language alive. This 'active presencing' of language in technological spaces was seen to be particularly important since these languages were not mainstreamed in formal education, one of the concerns was to address the absence of these languages in the spaces of learning. Microsoft Research Lab explains Project ELLORA'a purpose as 'a step to preserving a language for posterity' and ensuring the 'longevity of their language' (Shiyas 2023). IIT Kharagpur engaged in a study to find what the 'community needs to keep the language alive'. Aripana Foundation launched its Project Bhasini to 'transcend the language barriers' and build high-quality language datasets to train state-of-the-art AI models. The foundation aims to create these language platforms by 'leveraging the power of artificial intelligence' Similarly, Rajasthani Bhasa Academy aims to address the 'absence of Rajasthani from educational institutions' (RBA 2023).

The possibility of a new future, distinct from the current reality in which minority languages are disappearing every day, was envisioned as achievable through 'active presencing' (Vizenor 2008) of ITM languages in public domains at multiple levels. Firstly, it is by creating spaces for non-mainstream languages that these organisations attempted to reimagine technology beyond its mediation through a small number of dominant languages. Secondly, to realise the new technological future, one of the most challenging tasks was to accurately build models that could incorporate Indigenous sounds and orthography. Since both sounds and texts are the primary media through which AI operates, these human-technology interactions create the engagement space that demands changes in both technology and languages. By archiving varied language data, on the one hand, these interactions open the space to envision the new technological futures that normalise the presence of Indigenous languages in technological spaces. And on the other hand, they also transform these languages dramatically by standardising spoken words in the data archives.

We particularly noted that developing language technology for ITM languages also meant addressing the differing levels of marginality. Some languages such as Maithili and Rajasthani work with an already-developed script and literary tradition. While other languages such as Gondi and Mundari are primarily oral language traditions. These levels of marginality pose different challenges to the integration of technology as well as to the preparation of the language for education. The work of creating language technology for minoritized languages, therefore, had several layers. The first most expensive and labour-intensive layer involves collating voice and text data on the languages, then human transcribing and labelling this data, which requires linguistic expertise. For low-resource languages, this is the least available. Organisations such as Microsoft are collecting data at scale from communities, in voice-to-text format for those languages without a script, and text-to-speech and vice versa for the rest (Abraham et al. 2020). Of course, even for Microsoft, this process was not without its challenges.

These new technological interventions allowed people to access and occupy spaces that were otherwise not available to minoritised and low-resource languages. In order to build linguistic data, small but locally rooted organisations like Aripana Foundation collaborated with Google to create language input tools, in their specific case the Tirhuta keyboard for Maithili. These input tools have been developed with the explicit motivation of supporting futures where messaging services, content creation on the internet, and access to resources on the internet will be available to Maithili speakers and writers (Aripana Foundation Website). While these services need human-collected and transcribed data, services like text prediction and grammar correction can be done by training machines on pre-existing data. Organisations like Rajasthani Bhasha Academy are invested

in providing these digital services to speakers and writers of Rajasthani languages (Rajasthani Bhasha Academy Website). But translating these languages to the digital medium is both producing a requirement for and leading to inadvertent problems of standardization. As there are multiple versions of Rajasthani, choosing one form, and thus tacitly standardizing it in the digital domain.

The subsequent layer, which is built on the pre-existing layers, is in creating products for the user end. These products can range from Aripana using the digital tools to translate English children's books on Pratham's Storyweaver app to Maithili to be used for their education programme, to the IIT-KGP team creating a user interface (UI) in the form of an app with a chatbot for teaching, learning and communication in Mundari (Bali et al. 2019). These teams are working on textual mapping for language technology, with specialised teams working on developing apps for languages like Mundari. Similarly, the work going on at BR Ambedkar University is attempting to build user-end assistive technology for teaching Indigenous languages in formal settings like schools to cover the deficit in the availability of Indigenous language teachers.

Despite technological optimism, the task of envisioning technological futures for Indigenous languages involved an immense translation effort or act of 'brokerage' (Lachney 2017) to bridge different logic and knowledge systems and make them accessible to each other. Translation, at times, requires expressing 'what others say and want, why they act in the way they do and how they associate with each other' in one's own language (Callon 1984, 223). To facilitate machine learning, the organizations and institutions mentioned earlier employed various technological tools to incorporate language data. However, South Asian languages possess unique typographical characteristics, such as font, and morphological characteristics, such as word structure, tone, and emphasis. This diversity is often threatened by the introduction of 'foundation models' such as GPT3, BERT, RoBERTa, BART, and T5, which serve as the language-agnostic foundation for most natural language processing (NLP) systems and are known to homogenize languages (Bommasani et al. 2021). Non-scheduled languages, which lack official recognition and frequently have no written texts, are not included in the data, and even among those, there is a dearth of dialectical diversity representation (GIZ 2020). Consequently, all six organizations faced numerous challenges in adapting these minoritised languages to fit into the existing AI infrastructure and modifying AI to create space for these languages.

The visions underpinning these ventures are dominantly motivated by ideas of preservation and 'revitalisation' of 'endangered' languages (Eisenlohr 2004), by making them more present in technological spaces. This is expected to serve the purpose of not just meeting community needs but also transforming the digital landscape towards diversity and decolonization, thus connecting those who fall outside of the traditional sphere of the digital (Meighan 2022). Here, Tuck and Yang's (2012) reminder that decolonisation is not a metaphor is quite instructive, as these initiatives bring to the forefront Indigenous futurity through the tangible transformations of technology and the honouring of different epistemologies. In this context, making minoritized languages more visible through 'active presencing' is seen as one of the important gains claimed by AI-enabled Indigenous language technologies.

The imperative for collaborative and collective action

The need to overcome technical challenges, sparse language data, insufficient resources, and inadequate institutional support meant that these language technologies could not be built only by technical teams. This predicament immensely shaped these organizations' approaches to building collaborative and collective action, bringing together a diverse range of expertise. The team pages of all six initiatives show a multi-disciplinary team of technical experts, linguists, language activists, and teachers. It included a range of other actors – community liaisons, children, citizen volunteers, and user communities. Given the inadequate data and language corpus, these initiatives worked closely with user communities as language experts and not only linguists. Aripana Foundation discusses working in collaboration with academic and research institutions but also

recruited Maithili Annotators, Maithili Experts and Project Coordinators (Aripana Foundation Website).

The collaboration was not limited to technical teams that created specific programmes and platforms. The teams, therefore, were mainly led by technical experts but were composed of individuals with varied interests coming together with their loosely articulated goals for minoritised languages, which materialised in the form of language platforms. Pratham Digital mentions that they are 'actively collaborating with youth-driven start-ups, content creators and interest groups' to create audio data of children reading out text and numbers (Pratham Website). This has enabled Pratham to create datasets to train and automate machine learning on children's voices. Here, children are seen not only as end users of the technology but also as actors creating important data sets for the technology. They also worked with teachers to create Knowledge Graphs, which would streamline any subject topic into a set of interconnected keywords. Similarly, while Microsoft has an expert in Natural Language Processing and Data and applied scientist, they work closely with local communities to create the base datasets. For this purpose, Microsoft partnered with Karya (a crowdsourcing platform) to work with communities in rural and semi-urban areas, so that they could build high-quality language speech datasets in various Indian languages. For highly marginalised languages such as Mundari, Microsoft worked closely with an anthropologist, who is a member of the Munda community, and translated sentences to build a language corpus.

This collaborative work was seen to be imperative, especially since mainstream spaces have remained elusive to minoritized languages (Rajpurohit and Kothari 2021). Unlike mainstream and high-resource languages such as English, languages such as Gondi, Mundari, Rajasthani and Maithili required more fundamental tasks such as creating datasets and language standardisation. In this case, an important task for the team was to build the agency of the programme through machine learning. Especially with the possibility of automation and machine learning, there is a new space of interactive and intuitive learning that is being imagined in Indigenous language learning, that is self-paced and not constrained by institutional limitations. Rajasthani Bhasha Academy shows that with machine learning abilities such as natural language processing, AI tools can be created and enriched to provide better online translations. Similarly, Microsoft has been creating a Hindi-Gondi machine translation tool, which will enable people to access all the information that is available in Hindi.

The process of collaborative and collective action took on many different forms in these contexts. One such expression was the emerging partnerships for the creation of systems that recognized the specificities of different languages. This required people with different skill sets. The collaboration between Aripana Foundation and Pratham through their platform, Storyweaver, allowed them to bring together different expertise, to start diversifying children's literature and further literacy through children's literature (Aripana Foundation 2020). For this purpose, Aripana Foundation first digitised the Maithili language. However, they were faced with difficulty due to the absence of certain Maithili sounds in Hindi script (i.e., devanagari). Thus, the Aripana team converted these Maithili sounds into written symbols (*matras*) that could be incorporated into the Hindi script. Aripana Foundation and Pratham also created a new typing tool that eased this process and made it possible to type in Maithili language. Through these modifications, the organisations were able to use Google Translate app to create a corpus of stories in Maithili language. These stories enabled educators to start the conversation on Maithili stories, delve deeper into a topic, and build alternative perspectives that open up children's worlds and shape their aspirations.

The conceptualisation of collaborative and collective action in this context is not necessarily determined by a well-organised set of actions taken by already-existing groups of individuals. In this process, a variety of actors continuously engage with both opportunities and constraints while forging the relationships that make sense for this loosely put-together collective effort (Melucci 2013). The emergent AI then becomes an embodiment of this teamwork that materialised a particular vision of Indigenous future-making, by centering the sounds, the words and the stories that honour different communities' connection with their people, traditions, and knowledges.

Challenges to language futures

This future-making project, while taking place in apparently inert technological spaces, was invariably entangled with the existing relations of power and language hierarchies. The language initiatives overwhelmingly articulated the need to navigate multiple and often contrasting priorities at different levels. Some language initiatives seem to prioritise the 'unifying' different dialects. Rajasthani Bhasa Academy, for example, is working towards a recognition of an 'umbrella category' Rajasthani that would bring together approximately 22 different languages spoken in the region of Rajasthan, as they feel it is easier to claim space for one language rather than demand recognition of 22 different languages. They are building on 'a sense of linguistic oneness in Rajasthan' and 'the broad inter-intelligibility of the various forms' (Rajpurohit and Kothari 2021). Thus, Rajasthani Bhasha Academy seeks to facilitate language learning through this unified language form of Rajasthani language. These alternate platforms also enable organisations such as Rajasthani Bhasha Academy to establish spaces to diversify language education in spaces outside the official and institutionalised spaces. Technological innovation has, thus, opened up the potential for the production of new linguistic knowledge.

One of the starkest challenges in the development of Indigenous language technology is the tensions in response to the needs of the community. While the linguists and technical team might create the programme with the intention of language conservation or revitalisation, the community might want something as simple as access to resources. The creation of technology that is responsive to the user demands a more open-ended vision from the creators. Many of the existing language technologies are built with academic visions for linguistic analysis, thus creating a disconnect between the motivations of different actors. For example, Microsoft noted that the younger generation does not adhere strictly to the pure form of the Mundari language in their everyday language use. Instead, they use a hybrid form that combines elements from Bengali, Mundari, and Oriya (Bali et al. 2019). This is compounded by the fact that there is minimal digital content in the Mundari script, thereby reducing the incentives to use Mundari and its script (Shiyas 2023). This has meant that language workers have had to continue recording the nuances of the language and its use (Mitra 2019) and find ways to respond to them in the technology design (Bali et al. 2019).

These organisations are also challenged by the biases against the marginalised populations who do not constitute a relevant 'market' for most technology or have no presence and access to digital spaces. When Pratham created language resources with children, their main obstacle was the lack of adequate data on children's voices (Goenka 2023). Even though the requisite infrastructure for these operations was also being built through libraries of digitalised voice and textual data on Indian languages like INLTK, Indic NLP Library, Stanford NLP, the existing voice recognition did not work properly for children and thus making it inadequate for the success of the programme. To tackle this challenge, Pratham had to start by developing a language corpus from voice notes recorded by the children themselves, ensuring that the data was free from background noise or other sounds that could disrupt machine learning. The corpus was then transcribed and annotated by volunteers, including citizen volunteers, local community members, and teachers who administered the test. As a result of this effort, Pratham has been able to automate the language recognition and learning process, allowing children to use their voices to translate and seek assistance from apps like Alexa.

Most prevalent challenge among all six initiatives was the lack of comprehensive and sufficient linguistic data for these languages, an issue that frequently stems from inadequate investments in these languages. For instance, Aripana Foundation found that the existing technology failed to recognize the languages, dialects, and faces they were working with, resulting in a lack of annotated corpus for meaningful use and scaling. Such data-related constraints continue to reinforce language hierarchies, marginalizing certain languages and excluding specific communities from natural language processing (NLP) applications. This problem is even more exacerbated and difficult to

detect, given that Indian Language NLP currently has no 'benchmarks', or a set of tasks, by which the functionality of a software is evaluated (Sambasivan et al. 2021). Additionally, the lack of an open-source language database further compounds the problem by increasing costs and limiting access for smaller organizations or individuals (GIZ 2020).

While community engagement and participation have been integral to developing and implementing language-based AI, their involvement has remained sporadic and limited in scope. Here, the powerful actors – like governments and Big Tech- perform the act of agenda-setting, though increasing vocabularies of 'participation' of the affected group is becoming more well accepted, and while still led by technology companies, efforts are being made to seek out their voices in providing a direction for what resources are to be created. The Microsoft project – while collaborating with local partners – most explicitly embodies the tensions of unequal access to resources. For the Mundari language, Microsoft researchers collaborated with IIT Kharagpur in 2018, who worked closely with Mundari communities on the translation of Hindi sentences into Mundari (Dhapola 2023). The roles assigned to different groups of actors, and their representation in the language-based AI show how technology could perpetuate existing power structures. For the technology to be accessible to the wider community, it often needs to be mediated by commercial companies that have the resources to scale and reach the target population. This process requires a significant investment of time, effort, expertise, and funding, which is a severe limitation for low-resource languages. Consequently, inequalities become an inherent part of ongoing technological change. Despite opening up new possibilities, the technology remains constrained by the inability to support all languages equally.

Conclusion

AI-based language technology is often seen as a tool to imagine a new world. This paper examines the discourses on new possibilities that these human-technology interactions are expected to unlock and how technologies are envisaged to enable the imagination of new futures for minoritized languages. In this paper, we examined six language initiatives, across big and small actors in this space, that are working towards utilising these technologies to preserve and promote Indigenous, tribal and minority (ITM) languages in India. The focus on these ITM languages allows us to understand how technology is expected to navigate broader structural constraints that hinder the teaching/learning of minoritized languages, regardless of whether this is supported by big actors such as Microsoft or small organisations such as Aripana foundation. We analysed how technology and artificial intelligence is conceived, implemented, and innovated in these initiatives. We particularly paid attention to the ideas, narratives, and discourses that shape the actualisation of this cultural, political and economic project of imagining a more equitable future.

The overarching visions of these initiatives show three distinct discourses. Firstly, they were shaped by technological optimism in utilising the new technological opportunity for ITM languages, claiming these spaces for languages that have been historically excluded from public spaces and technology, and while doing so charting out new technological futures of minoritized languages. This 'active presencing' of non-mainstream languages in technological spaces was also seen as a way to go beyond deficit thinking and reshape this technology by utilising a range of knowledge systems. Secondly, the possibility of using a virtual environment, automation, and machine learning for ITM language learning, while offering an opportunity, was also a huge project that demanded collaborative and collective work. These collaborations occur from a place of both creativity and necessity, since there are currently no organisations with the combination of technical expertise and social reach into disenfranchised communities that could achieve this on their own. Thirdly, the process of Indigenous future-making was also articulated to address hierarchies and power relations of the past that persist into these futures and creatively subvert them. It is also interesting to note how this small but emerging space has drawn together a large range of actors with diverse, sometimes conflicting, aims and how these different visions have come together in a joint project.

This paper shows that while there is technical optimism about the possibilities offered by artificial intelligence, there is also an emphasis on the need to build strategies to overcome social and linguistic hierarchies, address biases in existing technology, and tackle tensions between various social actors. This paper argues that the use of language technology for minoritized language education is understood not just as a technical project but as a contested and ongoing attempt to subvert linguistic hierarchy through the 'active presencing' of these languages. Overall, the paper underscores the importance of adopting a nuanced perspective that acknowledges the intricate interplay among technology, language education, and wider social and political considerations. By exploring how language initiatives utilise technology to promote language learning, this paper provides insights into the challenges and opportunities associated with this emerging field of AI and language learning.

Notes

1. In India the use of the word 'Indigenous' to refer to communities (or their languages) is not used officially, as it is assumed that this definition implies that the rest of the people in the country are not 'native'. However, many low-resourced native languages in India are subject to marginalisation or erasure due to the dominance of official languages (like Hindi and English at the national level and other state languages at the regional level) and the work we explore here falls under the framework of a global movement led by the United Nations for protection and revitalisation of minority languages. This is also in line with the Indigenous movement around the world, where the capital I makes reference to a reclamation of identity with an aim to address the shared threat of language erasure for Indigenous communities.
2. UNESCO World Atlas of Languages identified different levels of vulnerability and endangerment (Available at https://en.wal.unesco.org/discover/, accessed on 2 August 2023).
3. Rajasthani language is not listed in UNESCO World Atlas of Languages because it is a language that is still seeking formal recognition. Rajasthani Bhasa Academy is seeking to 'unify' 22 different dialects spoken in the region of Rajasthan under an 'umbrella category' of Rajasthani, in order to make a stronger demand for one language rather than 22 different languages.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by UCL Research Culture Award, 2022 [grant number: 570091].

ORCID

Uma Pradhan http://orcid.org/0000-0002-0540-4085

References

Abraham, B., D. Goel, D. Siddarth, K. Bali, M. Chopra, M. Choudhury, P. Joshi, P. Jyoti, S. Sitaram, and V. Seshadri. 2020. "Crowdsourcing Speech Data for Low-Resource Languages from Low-Income Workers." In *Proceedings of the 12th Language Resources and Evaluation Conference*, 2819–2826.

Adam, T. 2019. "Digital Neocolonialism and Massive Open Online Courses (MOOCs): Colonial Pasts and Neoliberal Futures." *Learning, Media and Technology* 44 (3): 365–380. https://doi.org/10.1080/17439884.2019.1640740.

Aripana Foundation. 2020. "Helping Build Input Tools in Maithili." Aripana Stories. Accessed Januray 11, 2023. https://www.aripanafoundation.org/2020/05/22/maithili-input-tools/.

Aripana Foundation Website. Accessed January 11, 2023. https://www.aripanafoundation.org/about/#tab-14642.

Bali, K., M. Choudhury, S. Sitaram, and V. Seshadri. 2019. "Ellora: Enabling Low Resource Languages with Technology." In *Proceedings of the 1st International Conference on Language Technologies for All*, 160–163.

Battiste, M., and S. K. J. Henderson. 2021. "Indigenous and Trans-Systemic Knowledge Systems." *Engaged Scholar Journal* 7 (1): i–xix.

Bommasani, R., D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, et al. 2021. *On the Opportunities and Risks of Foundation Models*. Stanford: Center for Research on Foundation Models (CRFM).
Braun, V., and V. Clarke. 2006. "Using Thematic Analysis in Psychology." *Qualitative Research in Psychology* 3 (2): 77–101. https://doi.org/10.1191/1478088706qp063oa.
Bryant, R., and D. M. Knight. 2019. *The Anthropology of the Future*. Cambridge: Cambridge University Press.
Callon, M. 1984. "Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fishermen of St Brieuc Bay." *The Sociological Review* 32 (1 Suppl): 196–233. https://doi.org/10.1111/j.1467-954X.1984.tb00113.x.
Cartee, A. 2003. "Protecting the Indigenous Past While Securing the Digital Future: The FTAA and the Protection of Expressions of Folklore." *International Law Review* 1:203.
Cave, S., and K. Dihal. 2020. "The Whiteness of AI." *Philosophy & Technology* 33 (4): 685–703. https://doi.org/10.1007/s13347-020-00415-6.
Christen, K. 2004. "Properly Warumungu: Indigenous Future-Making in a Remote Australian Town." PhD diss., Santa Cruz: University of California.
Crawford, K. 2021. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven, CT: Yale University Press.
Dhapola, S. 2023. "How Microsoft's Project ELLORA is helping Small Languages Like Gondi, and Mundari become Eloquent for the Digital World." Accessed April 10, 2023. https://indianexpress.com/article/technology/how-microsofts-project-ellora-is-helping-small-languages-like-gondi-mundari-become-eloquent-for-the-digital-world-8413587/.
Eisenlohr, P. 2004. "Language Revitalization and New Technologies: Cultures of Electronic Mediation and the Refiguring of Communities." *Annual Review of Anthropology* 33:21–45. https://doi.org/10.1146/annurev.anthro.33.070203.143900.
Facer, K. 2011. *Learning Futures: Education, Technology and Social Change*. London: Routledge.
Gallagher, M., and J. Knox. 2019. "Global Technologies, Local Practices." *Learning, Media and Technology* 44 (3): 225–234. https://doi.org/10.1080/17439884.2019.1640741.
GIZ. 2020. *A Study on Open Voice-data in Indian Languages*. New Delhi: Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH.
Goenka, K. 2023. "Great Expectations: Mainstreaming AI and ML in Education." *Pratham Open School Blog*. Accessed April 11, 2023. https://prathamopenschool.org/Blog/GreatExpectations.
Hautala, J., and T. Ahlqvist. 2022. "Integrating Futures Imaginaries, Expectations and Anticipatory Practices: Practitioners of Artificial Intelligence between now and Future." *Technology Analysis & Strategic Management* 34: 1–13.
IndiaAI. 2021. "IIT Madras is Building Language Technology for Indian Languages." https://indiaai.gov.in/case-study/iit-madras-is-building-language-technology-for-indian-languages.
Indian Express. 2020. "Kerala: E-learning, in Tribal Dialects, Gets Underway in Wayanad." https://indianexpress.com/article/education/kerala-e-learning-in-tribal-dialects-gets-underway-in-wayanad-6502183/.
Lachney, M. 2017. "Culturally Responsive Computing as Brokerage: Toward Asset Building with Education-Based Social Movements." *Learning, Media and Technology* 42 (4): 420–439. https://doi.org/10.1080/17439884.2016.1211679.
Littau, K. 2016. "Translation and the Materialities of Communication." *Translation Studies* 9 (1): 82–89. https://doi.org/10.1080/14781700.2015.1063449.
Macgilchrist, F., J. Potter, and B. Williamson. 2022. "Reading Internationally: If Citing is a Political Practice, who are we Reading and who are we Citing?" *Learning, Media and Technology* 47 (4): 407–412. https://doi.org/10.1080/17439884.2022.2140673.
Mager, A., and C. Katzenbach. 2021. "Future Imaginaries in the Making and Governing of Digital Technology: Multiple, Contested, Commodified." *New Media & Society* 23 (2): 223–236. https://doi.org/10.1177/1461444820929321.
Martin, K., and B. Mirraboopa. 2003. "Ways of Knowing, Being and Doing: A Theoretical Framework and Methods for Indigenous and Indigenist Re-Search." *Journal of Australian Studies* 27 (76): 203–214. https://doi.org/10.1080/14443050309387838.
McLoughlin, C. 2000. "Cultural Maintenance, Ownership, and Multiple Perspectives: Features of Web-Based Delivery to Promote Equity." *Journal of Educational Media* 25 (3): 229–241.
Meighan, P. J. 2022. "Indigenous Language Revitalization Using *TEK-Nology*: How Can Traditional Ecological Knowledge (TEK) and Technology Support Intergenerational Language Transmission?" *Journal of Multilingual and Multicultural Development* 1–19. https://doi.org/10.1080/01434632.2022.2084548.
Melucci, A. 2013. "The Process of Collective Identity." In *Social Movements and Culture*, edited by H. Johnston, 41–63. London: Routledge.
Mitra, D. 2019. "IIT set to Launch an app in Mundari to Keep Indigenous Language Relevant." *Times of India*, February 2019.
Moore, R. E. 2006. "Disappearing, Inc.: Glimpsing the Sublime in the Politics of Access to Endangered Languages." *Language & Communication* 26 (3-4): 296–315. https://doi.org/10.1016/j.langcom.2006.02.009.

Naidoo, L. 2016. "Hallucinations." Ruth First Lecture, Wits University, Johannesburg, 17 August. http://witsvuvuzela.com/wp-content/uploads/2016/08/Hallucinations_RUTHFIRST_August2016_FINAL.pdf.
Nemorin, S., A. Vlachidis, H. M. Ayerakwa, and P. Andriotis. 2023. "AI Hyped? A Horizon Scan of Discourse on Artificial Intelligence in Education (AIED) and Development." *Learning, Media and Technology* 48 (1): 38–51. https://doi.org/10.1080/17439884.2022.2095568.
NITI Aayog. 2018. *National Strategy for Artificial Intelligence*. Delhi: NITI Aayog.
Patterson, M., and K. R. Monroe. 1998. "Narrative in Political Science." *Annual Review of Political Science* 1 (1): 315–331. https://doi.org/10.1146/annurev.polisci.1.1.315.
Pratham Website. Accessed January 11, 2023. https://www.pratham.org/2021/06/01/research-scientist-speech-and-audio-pratham-digital/.
Rajpurohit, D. S., and V. Kothari. 2021. "Constitutional Recognition for the Rajasthani Language Continues to Remain Elusive." Scroll India Feb 28, 2021. Accessed April 11, 2023. https://scroll.in/article/988086/is-rajasthani-a-single-language-or-a-spectrum-of-many-related-but-distinct-tongues.
Ranathunga, S., E. S. A. Lee, M. Prifti Skenduli, R. Shekhar, M. Alam, and R. Kaur. 2023. "Neural Machine Translation for low-Resource Languages: A Survey." *ACM Computing Surveys* 55 (11): 1–37. https://doi.org/10.1145/3567592.
RBA (Rajasthan Bhasa Academy). 2023. "Rajasthan Bhasa Academy Website." Accessed April 11, 2023. https://www.rajasthanibhashaacademy.org/.
Reddy, P. A. 2019. "Linguistic Minorities in India: Entrenched Legal and Educational Obstacles to Equality." LSE Blog. Accessed August 20, 2023. https://blogs.lse.ac.uk/southasia/2019/02/21/linguistic-minorities-in-india-the-entrenched-legal-and-educational-obstacles-they-face/.
Sambasivan, N., E. Arnesen, B. Hutchinson, T. Doshi, and V. Prabhakaran. 2021. "Re-Imagining Algorithmic Fairness in India and Beyond." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 315–328. https://doi.org/10.1145/3442188.3445896.
Selwyn, N., and K. Facer. 2014. "The Sociology of Education and Digital Technology: Past, Present and Future." *Oxford Review of Education* 40 (4): 482–496. https://doi.org/10.1080/03054985.2014.933005.
Selwyn, N., and D. Gašević. 2020. "The Datafication of Higher Education: Discussing the Promises and Problems." *Teaching in Higher Education* 25 (4): 527–540. https://doi.org/10.1080/13562517.2019.1689388.
Shiyas, A. 2023. "Microsoft Research Project Helps Languages Survive and Thrive." Microsoft News 30 January, 2023. Accessed April 10, 2023. https://news.microsoft.com/en-in/features/microsoft-research-project-helps-languages-survive-and-thrive/.
Shulist, S. 2018. *Transforming Indigeneity: Urbanization and Language Revitalization in the Brazilian Amazon*. United Kingdom: University of Toronto Press.
Smith, L. T., T. K. Maxwell, H. Puke, and P. Temara. 2016. Indigenous Knowledge, Methodology and Mayhem: What is the Role of Methodology in Producing Indigenous Insights? A discussion from Mātauranga Māori.
Somers, M. R., and G. D. Gibson. 1994. "Reclaiming the Epistemological Other: Narrative and the Social Constitution of Identity." In *Social Theory and the Politics of Identity*, edited by C. Calhoun, 35–99. Oxford, UK: Blackwell.
Tuck, E., and R. A. Gaztambide-Fernández. 2013. "Curriculum, Replacement, and Settler Futurity." *Journal of Curriculum Theorizing* 29 (1): 72–89.
Tuck, E., and K. W. Yang. 2012. "Decolonization is not a Metaphor." *Decolonization: Indigeneity, Education & Society* 1 (1): 1–40.
Tyagi, Y. 2003. "Some Legal Aspects of Minority Languages in India." *Social Scientist* 31 (5/6): 5–28. https://doi.org/10.2307/3518031.
UNESCO. 2010. *Atlas of the World Languages in Danger*. Paris: UNESCO.
UNESCO. 2022. *World Atlas of Languages*, Paris: UNESCO. https://en.wal.unesco.org/.
Vizenor, G. 2008. *Survivance: Narratives of Native Presence*. Lincoln, NE: University of Nebraska Press.
Yang, K. S. 2000. "Monocultural and Cross-Cultural Indigenous Approaches: The Royal Road to the Development of a Balanced Global Psychology." *Asian Journal of Social Psychology* 3 (3): 241–263. https://doi.org/10.1111/1467-839X.00067.