Shifting Paradigms of Multilingual Publishing and Scholarship in India

REEMA CHOWDHARY

Abstract: The shifting paradigms of multilingual publishing and scholarship in India refer to the evolving trends and approaches in academic publishing and scholarly activities within a multilingual and diverse linguistic context. A growing body of work in digital humanities in last decade has addressed the problems of lack of geo-linguistic diversity (Galina, 2014) and language sensitivity (Spence and Brandao, 2021) while drawing our attention towards larger geopolitical factors that inform the development of global knowledge infrastructures (Fiormonte 2012). The research paper traces the historiography and offers a comprehensive analysis of the evolving landscape of multilingual publishing and scholarship in India. As a linguistically diverse nation, India has a rich tapestry of languages, each with its unique cultural and historical significance. This diversity has significantly influenced the paradigms of publishing and scholarship in the country. The study begins by examining the linguistic diversity of India and its impact on publishing. It explores the challenges and opportunities presented by a multitude of languages and scripts, highlighting the need for inclusive and accessible publishing practices. It investigates the role of technology and the internet in reshaping publishing and scholarship. It delves into the accessibility and reach of digital content in various languages, considering how these technological advancements have democratized knowledge dissemination. The study critically examines how the key facets like open access initiatives, digital publishing, community engagement, multilingual scholarship, and cultural preservation integrate the value of multilingualism into a more equitable and epistemically scholarly communication and publishing in India, hence, fostering inclusivity and promoting diverse knowledge systems. It then discusses a series of case studies with a special focus on Sindhi script OCR to investigate the need and importance of such mechanism in place for language inclusivity within a language-heterogeneous country.

Keywords: Multilingual Publishing, Scholarship in India, Historiography, Language

https://doi.org/10.3998/jep.5592

Multilingualism is a defining characteristic of India's diverse cultural landscape, profoundly shaping its publishing and scholarly endeavors. India's linguistic richness and diversity are evident in the dissemination and preservation of its numerous languages and dialects. The linguistic census of 2011, released in 2018, recorded more than 19,500 languages and dialects as mother tongues in India. Among these, mother tongues spoken by 10,000 or more speakers have been classified and grouped under appropriate languages at the national level, resulting in a total of 121 languages (ORGI 2022, vii). The historical significance of these languages in the realm of publishing and scholarship is monumental, shaping the transmission of knowledge, literature, and cultural expressions. However, the fact of "English as hyper central language" (de Swaan 2001) disrupts the existing linguistic dynamics and impedes language diversity. The Eighth Schedule of the Indian Constitution recognizes 22 official languages: Assamese, Bengali, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Malayalam, Manipuri, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Sindhi, Tamil, Telugu, Urdu, Bodo, Santhali, Maithili, and Dogri. Six languages in India are listed as "Classical" in the Eighth Schedule of the Constitution: Tamil (declared in 2004), Sanskrit (2005), Kannada (2008), Telugu (2008), Malayalam (2013), and Odia (2014). India has by and large incorporated most of its languages and dialects as official state languages, yet a large proportion of digital tools and many publishing houses function with a monolingual mindset. This diversity also presents challenges, hindering accessibility and equity in the publishing domain. This article aims to dissect this duality of language heterogeneity and explore the diverse landscape of multilingual publishing and scholarly initiatives in India. The article explores the dual nature of language heterogeneity, considering it both as an opportunity and a challenge in integrating and implementing multilingualism within a more equitable and epistemically just publishing system. A case study of Sindhi script optical character recognition (OCR) introduces the reader to multilingual digital access initiatives and investigates ways in which language heterogeneity in India poses a challenge and an opportunity to establish an equitable publishing system.

Historical Roots of Multilingual Publishing and Scholarship in India

The historical roots of India's linguistic plurality can be traced back to ancient civilizations. The Vedic texts, composed in Sanskrit, form the cornerstone of India's literary heritage. Sanskrit's influence transcended regional boundaries, serving as a lingua franca for scholars and intellectuals across the Indian subcontinent. The medieval period saw the rise of regional languages as a medium for scholarly pursuits and literary expression. Tamil, Telugu, Kannada, Bengali, Urdu, and other vernacular languages flourished, giving rise to a diverse array of literary genres, from poetry and epics to philosophical treatises and scientific texts. The Mughal era further contributed to multilingual publishing in India. Persian, the court language, served as a medium for historical chronicles, administrative records, and scholarly discourses. Persian works on literature, science, and philosophy coexisted alongside indigenous languages, thus creating a multicultural and multilingual scholarly environment.

The colonial era, marked by the British Raj, introduced English as a language of administration and education. While English became a crucial link language, it also had a transformative impact on Indian print publishing and scholarship. The introduction of printing presses facilitated the dissemination of knowledge in various languages, albeit with an emphasis on English. Serampore Mission Press, established in 1800 by William Carey, Joshua Marshman, and William Ward in Serampore, West Bengal, emerged as a pioneering institution in this field (Mondal 2014). They recognized the importance of printing materials in various Indian languages to spread knowledge and literacy. The press's establishment marked a significant milestone in multilingual printing in India as it printed works in Bengali, Sanskrit, Tamil, Telegu, Marathi, and Assamese, among others. The presses, operated by missionaries such as the Calcutta School Book Society Press (1817) and the American Mission Marathi Press (1813), significantly contributed to printing materials in regional languages. India's independence in 1947 brought renewed interest for regional languages and indigenous knowledge systems. The Indian Constitution recognized multiple languages as official, advancing a multilingual approach to governance, education, and cultural preservation. Vernacular newspapers, journals, and books became crucial tools of communication and furthering regional identities. The promotion and preservation of regional languages gained impetus, fostering a resurgence in vernacular literature, academic research, and cultural exchange. This created a non-homogeneous publishing market structured according to the region and language. The publishing sector ecosystem, therefore, continued to be made up of large, medium, and small publishing houses.

Paradigm of Multilingualism in Indian Publishing and Scholarship

The late 19th and early 20th centuries witnessed a renaissance of vernacular literature, with prominent authors such as Rabindranath Tagore, Bankim Chandra Chattopadhyay, and Subramania Bharati contributing immensely to their respective regional languages. Their works were pivotal in revitalizing indigenous languages and fostering cultural pride. The post-independence era saw the emergence of language movements advocating for the recognition and preservation of linguistic identities. The reorganization of states along linguistic lines in 1956 and the subsequent adoption of the Eighth Schedule of the Constitution, recognizing 22 official languages, underscored the importance of linguistic diversity in India's nation-building process. Translation played a crucial role in facilitating cross-cultural exchange and promoting multilingualism in India. Translators such as A. K. Ramanujan, G. A. Grierson, and U. R. Ananthamurthy bridged linguistic divides by rendering literary classics and scholarly texts into multiple languages, thus enriching the literary landscape.

However, so-called modern nations are often perceived of as monolingual entities, or bilingual at most, and the multilingual nation may be seen as one that has still not sorted out what its main, common, or national language is (Kothari 2018). Repeatedly, the imperative of Western modernity and nation-building has spurred a specific technocratic imagination that seeks the efficiency of a common language and perceives multiple languages as a developmental burden. Due to the complex history of social and cultural formations in India, the political boundaries of linguistic states often do not align with their cultural boundaries. The debates around multilingual translation, publishing, and scholarship are many and ongoing, but the intricacies of producing a work in a nation with a linguistic continuum are challenging. Perhaps responding to this, Rita Kothari (2018) says, "Hegemonic languages shrink the multilinguality, while minor ones stretch themselves sometimes to the point of relinquishing the places they come from. Complicating these criss-crossing relations that are both temporally and spatially, horizontally and vertically determined, is the presence of English, a subject that demands its own treatment, and yet, never, as the case tends to be, without its relationship with other Indian languages" (8–9). The multilingual publishing landscape in India is indeed a complex and multifaceted phenomenon that encompasses linguistic diversity, sociocultural dynamics, and technological advancements. The linguistic diversity in India, characterized by a multitude of languages and dialects, presents both challenges and opportunities for multilingual publishing (Huckle 2021). There is growing emphasis on the need for a paradigm shift towards multilingual education, transcending monolingual ideologies and embracing dynamic multilingualism in pedagogy. The role of multilingual users has been recognized as crucial after the COVID-19 pandemic, emphasizing the significance of understanding their role in information diffusion and adjusting publishing strategies accordingly (Chen et al. 2021).

Multilingual print publishing in India embodies a prominent space, but epistemic work is needed to develop shared infrastructure and allow diverse fields to engage more meaningfully with digital, multilingual publishing and scholarship. While English remains a significant player, the robust presence of regional language publishing needs a greater multilingual focus in what Alan Liu (2020) calls "the techne of diversity," meaning that technical innovation drives the understanding of diversity and the understanding of diversity drives technical innovation. Digitization has undoubtedly affected transcultural/trans linguistic dynamics, but there are also concerns that digital inclusions are not multilingual enough and do not offer equitable support across all the languages. The minority languages are greatly understudied and under-resourced owing to inconsistent funding and infrastructural support. Paul Spence (2021), in his report *Disrupting Digital Monolingualism*, states that "more dialogue is needed with digital practitioners and researchers to co-create search-driven tools and methods" (25) to capture multiple linguistic resources and codes. By referring to a series of examples and a case study on Sindhi OCR, this article introduces the reader to the dynamics, challenges, trends, and opportunities in multilingual publishing and scholarship in India.

Multilingual Publishing Initiatives

Several significant government and institutional initiatives seek to meet the requirements of multilingual users, authors, translators, and publishers, thereby fostering the establishment and promotion of multilingualism within a specified context. The following initiatives outline the possible solutions, dynamics, and challenges to raising cultural diversity, equity, and inclusion through promising alternatives for multilingual publishing.

Sahitya Akademi, India's National Academy of Letters, was founded on March 12 1954, and is an initiative independent of the Indian government that promotes literature in 24 Indian languages. The National Book Trust, the Publications Division (formerly under the Ministry of Human Resource Development, now under the Ministry of Information and Broadcasting), the Centre of Translation in Bengaluru, National Translation Mission, and Indian Literature Abroad are some significant projects and organizations that work in the field of translation. Academic publishing extends beyond English with journals and books in languages such as Hindi, Tamil, Telugu, and others. The National Council of Educational Research and Training (NCERT) publishes educational material in multiple languages. Institutions such as Sahitya Akademi, the Indian Council of Social Science Research (ICSSR), and various universities publish scholarly journals. Magazines such as *Katha* and *Indian Literature* promote multilingual literary content, featuring poems, stories, and essays in diverse Indian languages.[1]

Such initiatives are the digital pathways and road to canonicity that create awareness for research and knowledge production in Indian languages. These platforms play a pivotal role in elevating awareness and accessibility to scholarly resources, cultural content, and educational materials across various linguistic landscapes in India. They act as repositories for indigenous knowledge, enabling a comprehensive understanding of regional cultures, histories, and traditions. By creating a digital space for these languages, these initiatives not only preserve linguistic diversity but also empower communities

[1] See the appendix for additional details and examples of multilingual publishing initiatives in India.

by offering them a space to express, share, and perpetuate their rich heritage through literature, arts, and academia.

Multilingualism and Digital Inclusion in Scholarly Publishing

Gunnar Sivertsen (2018) argues that local language use in scholarship is needed to foster engagement with stakeholders and the public. Various languages and communication platforms influence diverse audiences (Hicks 2004). Diego Chavarro, Puay Tang, and Ismael Ràfols (2017) demonstrate that non-English journals fulfill distinct communication roles compared to dominant English publications. They provide avenues for researchers to enter the publishing sphere and share their work on topics that are less often addressed in mainstream channels.

Publishing practices vary in Indian languages due to their multivariant histories. The initiatives surveyed above signify the concerted efforts made by various organizations, government bodies, and NGOs to support and propagate multilingualism across India. They contribute significantly to language preservation, literacy, and the promotion of diverse cultures and knowledge systems through multiple languages. These regional language publications encompass a wide spectrum, including novels, poetry, short stories, essays, folk tales, and academic texts. There's a growing emphasis on translated literature, which enables cross-pollination of ideas and promotes cultural exchange. Translation initiatives have become instrumental in bridging linguistic gaps and making literature accessible to diverse audiences across the country. Digitalization has revolutionized the publishing industry and offers new avenues for authors, publishers, and readers. E-books, audiobooks, and online platforms have augmented traditional print publishing by enhancing accessibility and outreach, especially in regional languages. Social media and digital platforms have emerged as spaces for literary engagement, enabling authors to connect with readers and literary communities.

The scope of multilingualism in publishing can be sketched by valuing the goal of "equivalence" in publishing by multilingual scholars (Curry and Lillis 2014) and varied government and private initiatives. Networked activities as well as advanced language technology dynamics can incorporate low-resourced/minority languages such as Sindhi (a non-Latin right-to-left script). The accessibility of such languages in the digital world helps to facilitate a digital space in which multiple voices are valued. The concept of language accessibility encompasses not only the promotion of diversity but also the critical aspects of inclusion and equity. It is ethically imperative to challenge and improve the existing status quo by establishing platforms that bolster the digital multilingual publishing ecosystem. These platforms are essential for enhancing the visibility and participation of diverse linguistic communities, and creating a more inclusive digital environment. The current scope of digital inclusion is quite limited, necessitating a more thoughtful and engaging approach to knowledge production. It is crucial to develop strategies that not only establish digital spaces but also revive and amplify the existing work undertaken in minority languages across India. Additionally, preserving linguistic diversity amid globalization remains a concern, as the dominance of English could overshadow indigenous languages. Government initiatives and literary organizations have been instrumental in promoting and supporting regional language publishing. Grants, awards, and fellowships incentivise authors and publishers in regional languages, driving creativity and sustaining multilingual publishing.

In essence, the contemporary publishing industry in India is a dynamic amalgamation of languages and cultures. While such efforts acknowledge the colonial origins of English, they also decenter the dominance of standard English(es) and decolonize knowledge production (Bhambra, Gebrial and Nişancıoğlu 2018; Santos 2017). The dual heterogeneity in India, about linguistic diversity and socioeconomic disparities, not only produces an opportunity but also poses a significant challenge to equitable publishing. The historiography of multilingual publishing in India reflects a dynamic narrative, showcasing the resilience of regional languages and their evolution through various historical epochs. The multilingual landscape, while rich and diverse, faces hurdles in ensuring fair and inclusive access to publishing platforms and resources across all languages. These challenges extend to issues such as (1) access and representation, (2) resource allocation, (3) economic disparities, (4) policy and infrastructural gaps, and (5) technological challenges.

First, in regard to access and representation, the exact number of languages and dialects spoken in India is difficult to determine precisely due to various factors including regional variations, dialect continua, and the inclusion of minority languages. The Eighth Schedule of the Indian Constitution recognizes the aforementioned 22 languages as scheduled languages, which are given official status at the national level. However, numerous dialects and sub-dialects are spoken within each language group. Certain languages and dialects often have limited representation in mainstream publishing, leading to the marginalization of diverse voices and cultures. Lack of access to publishing opportunities for smaller linguistic groups perpetuates inequalities. Minority languages such as Gondi, Dogri, Konkani, Tulu, Meitei, Santali, and Sindhi face challenges such as a lack of institutional support, declining speaker populations, or limited access to education and resources.

Vernacular literature has gained prominence, with publishing houses dedicated to various languages, but still there exist profound nuances and modalities of publishing from language to language across India. With the widespread dominance of English and state regional languages, “the state of scholarly publishing infrastructure in India is precarious and access to scholarly articles in Indian languages is difficult”

(T, Arora, and Menon 2017). States such as Maharashtra, Tamil Nadu, West Bengal, and Karnataka have thriving newspapers in Marathi, Tamil, Bengali, and Kannada, respectively. For instance, *Ananda Bazar Patrika* (Bengali) and *Dinamalar* (Tamil) are widely circulated regional newspapers. The translation ecosystem of India existed long before the advent of the printing press, and it ranges from multilingual usage (using Hindi and English as link languages) to using language in the aid of nationalism and nation-building (both pre-and post-independence). Literary translation across several languages in which writers were composing was considered a "uniting" factor in the culturally diverse Indian nation-state. However, the state-owned languages subscribe to a monolingual paradigm with the claim that each state should be free to use its state language for official purposes. Moreover, the discourse on translation in an Indian context does not treat translation as functioning in a multilingual public sphere or among Indians who simultaneously function in multiple languages (Israel 2021).

India is a multilingual country, but there are insufficient efforts toward multilingualism through the official use of multiple languages or translation. The role of language in the nation and identity building have thus sustained (but also remained limited through) varied state initiatives and projects over 75 years. The participatory platforms that design and manage open access infrastructure and open communication are limited in their embrace of India's vast multilingual, multidisciplinary, and multi-stakeholder content, which would allow more outreach within and beyond the research community. The varied multilingual projects aimed at preserving, promoting, and disseminating cultural and linguistic diversity are numerous and comprehensive (see appendix). Thus, the challenges of access and representation faced by certain languages and dialects in India have profound implications for linguistic diversity, cultural representation, and societal inclusion. Addressing these challenges is crucial for equitable access to publishing platforms and resources across all languages, thereby promoting inclusivity and preserving the richness of linguistic and cultural heritage in the country.

Second, unequal distribution of resources, including funding, infrastructure, and technology, creates barriers for publishing in various languages. Linguistic hierarchies, dominance of certain languages, lack of institutional support for minority languages, and challenges in educational contexts lead to unequal treatment of languages in India. This disparity affects the quality and reach of publications, hindering equitable access to knowledge. According to the Census of India 2001 (Pandya and Gohil 2023), while India has 22 officially recognized languages, 30 languages are spoken by the majority of the population, along with 440 dialect languages. Among the widely spoken and published languages in India, Hindi holds a prominent position. With 422 million speakers, constituting 41% of the population, Hindi is numerically and proportionally the largest indigenous language community in India (Pandita 2014). Hindi publications account for 37.5% of periodicals in the country (Pandita 2014). While Hindi is a dominant language, other languages such as English, Urdu, Marathi, and Telugu also play crucial roles in the linguistic fabric of India. English, in particular, has become a key language for research communication, with Indian management journals predominantly published in English (Bajwa and König 2017). Marathi and Telugu, spoken by millions of people, contribute to linguistic diversity but do not hold an equal space with Hindi and English. The dominance of languages such as English, Hindi, and a few other major languages creates a linguistic hierarchy that influences the visibility and representation of other languages. This hierarchy impacts the access and opportunities available to languages beyond the dominant ones, leading to disparities in recognition and usage. Moreover, the highly multilingual nature of India poses challenges in educational settings, where the language of instruction may not align with the mother tongue of students. This discrepancy disadvantages learners and further contributes to the unequal response towards languages in the country.

A third challenge is that economic constraints impact the publication and distribution of content in regional languages, leading to a concentration of resources and opportunities in dominant languages and hindering equitable representation. In India, economic constraints often limit the investment and infrastructure available for publishing and distributing content in regional languages. Publishing houses prioritize languages with larger markets and higher profitability, such as Hindi or English, over smaller regional languages. This preference for dominant languages results in a concentration of resources, including funding, marketing, and distribution networks, further marginalizing content in regional languages. Inadequate policies and infrastructural support for multilingual publishing pose a fourth challenge by hindering the growth and sustainability of publications in minority languages. Infrastructure for distributing publications is concentrated in the urban centers or primarily caters to languages with larger markets, such as Hindi or English. As a result, publications in minority languages such as Tulu, Gondi, Dogri, Konkani, Meitei, and Santali do not always reach readers in remote or rural areas where the language is spoken, limiting their accessibility and visibility. Moreover, the limited access to the internet, digital publishing platforms, and e-commerce channels further restricts the reach of these publications to a wider audience, particularly readers who consume the content online.

Lastly, language is a primary means for communicating information and knowledge. The ability to access content on the internet in a language that one can use is a key determinant for the extent to which one can participate in knowledge societies (UNESCO 2023). Multilingualism in cyberspace is important for a pluralistic, equitable publishing environment, but most of the world's languages are not present in cyberspace and its resources are being increasingly marginalized. Technologies that facilitate publishing, like OCR, often support dominant languages only, making it difficult to digitize and publish content in less widely spoken languages and further limiting their accessibility. OCR presents both an opportunity and a challenge in accessing multilingual scholarship. In some ways, OCR offers new possibilities for studying and incorporating translingual and transcultural dynamics. The advent of artificial intelligence (AI), machine learning, and natural language processing (NLP) presents a shifting paradigm of publishing from the limits of print publishing to digital publishing with its increased accessibility to linguistic diversity or intercultural communication. AI technologies have the potential to revolutionize multilingual publishing in India by enabling publishers to overcome language barriers, enhance content creation and distribution, and provide personalized experiences for diverse linguistic audiences. Embracing technology for multilingual digitisation, allocating resources to support diverse publications, and implementing inclusive policies can lead to more equitable publishing practices in India. Integrating and practicing the value of multilingualism within scholarly and publishing systems involves nurturing equity and epistemic justice across languages.

One of the ways to address the challenges listed above is to develop a multilingually enabled digital knowledge infrastructure for non-Latin script languages in particular. For example, Sindhi, an official language of India, faces marginalization due to limited accessibility in terms of both speakers and written resources. The language lacks substantial digital infrastructure, hindering its widespread use and accessibility across India. The following section discusses the challenges associated with developing and introducing language technology to a low-resourced language such as Sindhi by presenting a case study on developing a Sindhi OCR.

The Current Multilingual Publishing Landscape in India: A Case Study on Sindhi OCR

India's diverse linguistic landscape has nurtured a rich tradition of multilingual publishing, showcasing a multitude of languages and cultures. In recent years, advancements in technology have significantly influenced the country's publishing landscape, particularly by enabling the digital representation of various languages. The remainder of this article will focus on the role of Sindhi OCR technology in multilingual publishing in India. This study provides Sindhi OCR as one example to understand the technological advancements and challenges of working on low-resourced languages. The attempt is to map the current status of publishing and scholarship in a multilingual setting.

Digital advancements have revolutionized publishing, enabling the representation of lesser-known languages. OCR technology is a prime example, facilitating the digitization of texts in various scripts. Despite progress, challenges persist, especially in developing OCR systems for low-resourced languages, hindering their representation in the digital realm. OCR for low-resourced languages is challenging and has not attained optimal accuracy for print and handwritten texts. *Low-resourced languages* in India refer to languages that have limited linguistic resources, digital text data, and technological support compared to major languages. While the exact list may vary, some examples of low-resourced languages in India include Bodo, Dogri, Kashmiri, Konkani, Maithili, Manipuri (Meitei), Nepali, Sanskrit, Santali, Sindhi, and Tulu. These languages may have fewer speakers, limited written resources, and less availability of linguistic tools and technologies compared to widely spoken languages, such as Hindi, English, Bengali, Telugu, or Tamil. Developing language technology resources and tools for these low-resourced languages is essential to preserve linguistic diversity, promote cultural heritage, and enable digital inclusion for speakers of these languages.

OCR in South Asian languages such as Sindhi continues to pose a formidable challenge in achieving comprehensive digitization of literary works. Sindhi as a second and third language of India loses its literary presence due to historical influences, sociolinguistic factors, educational policies, and cultural shifts. The digitization of cultural heritage plays a pivotal role in preserving linguistic diversity and historical narratives. Sindhi exists with its unique scripts and rich literary tradition within the multilingual fabric of the country. OCR has been proven effective and precise when applied to text recognition in English language; however, its proficiency for South Asian languages, which include scripts that are both right to left (RTL) and left to right (LTR), remains limited and less accurate. Sindhi, in particular, presents an additional layer of complexity due to its utilization of two distinct scripts (Perso-Arabic and Devanagari). This complexity is further compounded by variations in script styles and the prevalence of handwritten manuscripts. These challenges underscore the unique features that distinguish Sindhi from other languages, necessitating specialized approaches in linguistic analysis and computational processing. The following case study provides a multilingual setting for a third language of India while it also highlights the importance of adapting an OCR for a unique script like Sindhi script. It attempts to offer tools and best practices for creating searchable digital formats and visualizations, thereby improving the accessibility of these valuable cultural treasures.

Significance of Sindhi OCR

Sindhi has a long history going back to the 10th century CE. Sindhi is one of the Indo-Aryan languages deriving its script from Persian and Arabic. It is spoken by roughly 30 million people in Pakistan, where it holds a status of an official language. It is also spoken in India where it is a scheduled language, without any state-level official status. The main writing system is Perso-Arabic script, which accounts for the majority of Sindhi literature. In India, both the Perso-Arabic and Devanagari script are used. Sindhi script follows the rules of bidirectional writing of the Arabic script. It is written

RTL, but for numbers the writing is LTR. Sindhi script has a sum total of 52 characters in its alphabet as compared to Arabic with 28, Persian with 32, and Urdu with 39 characters. Thus, Sindhi constitutes the largest extension of the original Arabic script.

OCR in South Asian languages, notably Sindhi, faces challenges hindering the comprehensive digitization of literary works. A substantial amount of work is available in computerizing the Sindhi vocabulary, but the work towards the development and implementation of OCR technology is still at a foundational level. Sindhi faces greater challenges due its cursive script and non-Latin orientation. Character recognition is much easier for Latin scripts than for non-Latin scripts. OCR for Sindhi involves complexities with the orientation of dots, differentiating characters, and more variety of placements (see Table 1). The two dots, vertical and diagonal, and three dots with two directions, pointing upward and pointing downward, pose a challenge to recognizing the characters in its right form.

Like Arabic script, most of the characters in Sindhi have more than one shape when forming a word, ligature, or a compound word, as shown in the following isolated characters:

ٺ , ٿ , ج , ح , ک , ڪ

Additionally, the use of a single base shape with different placement and orientation of dots make machine learning complicated. Forming a word or ligature in cursive writing involves connecting a character to its preceding, succeeding, or both adjacent characters. This characteristic is known as the cursive nature of the language. A ligature can consist of isolated characters or a combination of one, two, or more characters (see Table 2). This aspect poses greater challenges when integrating grammar, searching, and sorting functionalities into Sindhi OCR.

The single character shape represents multiple characters with the only difference in the number of dots, position, and orientation of dots. D. N Hakro et al. (2016), Shafique et al. (2017), and other scholars working on Sindhi language translation,

Table 1: Dots in Sindhi characters

Number of dots	=	Characters
With single dot	12	ن, ف, غ. ظ, ض, ز, ذ, ڊ, خ, ڄھ, ج. ب
With two dots	11	ي, ڱ, ڳ, ق, ڍ, ڌ, ڄ, ڃ, ٺ, ت ,
With three dots	06	ش, ڎ, چ, ث, ٿ, پ
With four dots	05	ڦ, ڙ, ڇ, ٿ, ڀ
With small (ط)	01	ڻ
		ء ,ه,و ,م ,ل ,ڳھ ,گ ,ک,ڪ ,ع ,ط ,ص ,س
Without dot	17	ر, د, ح, ا,
Total number of characters	**52**	

Table 2: Different forms of ligature in cursive

1	More than one ligature forms a word	موتمار صحتمند
2	Ligature and isolated characters form a word	بهادر
3	One ligature forms a word	ٿلهو پلو
4	Only isolated characters	رازو ادارو
5	More than one ligature and isolated character	هردلعزيز

transliteration, and OCR clearly discuss the issues and challenges with Sindhi script and propose varied solutions for its development and implementation. Yet scholars in the field are not able to attain more than 84% accuracy to convert the text images into an editable readable format.

This article does not specifically address the technical impediments in the Sindhi OCR but rather underscores the lack of cognizance to establish a robust digital platform that can increase the opportunities for the multilingual scripts. An OCR for a low-resourced language such as Sindhi can offer, as Masoud Ghorbaninejad, Nathan P. Gibson, and David Wrisley (2023) say, scholarship that includes "not only the study of ancient or historical languages in the centuries-long tradition of Orientalist scholarship but also the modern, often multilingual societies that themselves require multiscript, and multidirectional digital environments" (48). Translation integration, machine translation, and crowdsourcing are some of the possibilities that need to be integrated for a dynamic and robust multilingual ecosystem. The technological advancements are in place, but their Anglocentrism is part of the bias towards LTR scripts, as English and many other languages of the post-industrial world are written from left to right (Fiormonte 2016; Galina Russell 2014; Mahony; Meza 2019). The Sindhi script needs to be romanized in order to make it technologically accessible, which requires efforts to move away from the orientalist bias and LTR orientation. Most of the tools and platforms that function well for LTR languages function poorly for RTL or bidirectional text.

Language accessibility is not solely a matter of diversity, where numerous voices are valued, but also a question of inclusion and equity. In this context, there exists an ethical imperative to enhance the current state of affairs, ensuring that global colleagues not only have a seat at the table but are also empowered to actively contribute to and drive the work forward. The current multilingual publishing landscape in India reflects a blend of tradition and technology. While major languages dominate, initiatives such as Sindhi OCR exemplify the potential to digitally represent and preserve less-resourced languages. As of January 2023, a search yielded over 200,000 issues and six million code commits related to "RTL," with over 38,000 of these issues remaining unresolved.[2]

[2] Code commits are changes submitted to the code repository. See the commits and issues at https://github.com/search?q=%22RTL%22&type=commits; open issues are listed at https://github.com/search?q=%22RTL%22+state%3Aopen&type=

Efforts to address technological limitations and promote inclusivity in the digital sphere will be crucial in nurturing India's linguistic diversity in the digital age. Sindhi training datasets are not yet sufficiently available for research, and the availability of resources is minimal. The available datasets are also very limited due to the low-resource nature of the Sindhi language. There are two ways for the natural selection of datasets for the Perso-Arabic script as both a Unicode character set and an Extensible Markup Language (XML) file. Sindhi uses Unicode for the storage. Multiple file formats used for Sindhi text are ".txt" or ".doc," ".xcls," ".spss," among others. Most social technologies, such as text messages, WhatsApp, Facebook, online websites, and translators, use the romanized Sindhi text instead of Sindhi script. The vowel placement is similar to the English language. The romanized Sindhi text provides greater accessibility for social outreach, but issues with the Sindhi script remain a pertinent challenge for scholars of the Sindhi language. The diverse characteristics of Sindhi script, including its unique writing style, incorporation of Persian and Arabic scripts, extensive use of additional characters and ligatures, variations in dot orientation, prevalence of characters with single base shapes, and larger character set compared to other Arabic adopting languages, combined with issues such as skew and noise, present significant challenges in the development of an OCR system for Sindhi script.

In a diverse setting where the Sindhi population is approximately 3.8 million among the 1.42 billion people in India, there exists a linguistic minority due to lack of state patronage. This bias is characterized by a persistent inclination towards tools and platforms that perform effectively for LTR languages while often proving inadequate for RTL or bidirectional text. The issue of inclusion of RTL digital culture is a pressing need within the larger global community. An increasing number of institutes, such as the American University of Beirut, New York University Abu Dhabi, University of British Columbia, and Islamicate Digital Humanities Network, work in and between the RTL and LTR societies.[3] Sindhi OCR is just one example to demonstrate the impact of a "monolingual fallacy"—the implicit assumption that everywhere in the world people read and write in one language, one script, and one reading direction which eventually ends up conditioning the development of software for specific languages only. The result is a transformative shift in multilingual publishing, characterized by the initial focus on widely spoken languages and the subsequent integration of low-resourced languages into scholarly discourse.

Issues&ref=advsearch&l=&l=. Of course, not all of these have to do with support for RTL scripts or layout, but a quick perusal shows that a great many of them do.

[3] For the latter, see Islamicate Digital Humanities Network: The Next Generation, https://idhn.org.

Conclusion

The current multilingual publishing landscape in India reflects a blend of tradition and technology. While major languages dominate, initiatives such as Sindhi OCR exemplify the potential to digitally represent and preserve less-resourced languages. Efforts to address technological limitations and to promote inclusivity in the digital sphere will be crucial in nurturing India's linguistic diversity in the digital age. Scholarly publication in languages other than English is difficult, but platforms such as those listed above and in the appendix offer an ambitious example of publishing scholarly works in regional languages and serve as a collaborative space for Indian literature scholarship. Contemporary paradigms of multilingualism in Indian publishing encapsulate dynamic challenges and evolving trends within a multi-layered ecosystem. There exists a pressing need to foster equity and epistemic justice across languages. The promotion of language inclusivity in scholarly publishing anticipates policy development, encouraging submissions in diverse languages, providing translation services for abstracts, summaries, or key findings, and embracing more open access models to disseminate research across languages without financial barriers. Investing in multilingual digital repositories, OCR technologies, NLP, and AI for low-resourced languages presents a challenge as well as an opportunity to bridge the digital divide. Practicing the value of multilingualism within scholarly communication and publishing systems will ensure research in less represented languages is not marginalized but rather integrated into the broader scholarly discourse.

This article has revealed a wide disconnect between the aspirations of multilingual scholarly publishing and a more equitable democratization of the knowledge system which can strengthen the multilingual publishing ecosystem in India. India's linguistic diversity, historically intertwined with its publishing and scholarly traditions, remains a source of strength and vitality. The plethora of languages continues to present a vibrant cultural milieu and serves as a testament to the inclusivity and resilience of India's societal fabric. There exists a demand for trusted digital infrastructure to generate the new forms of knowledge production, circulation, and access to many avenues of multilingual publishing in India.

Based on the historiography of multilingual publishing and scholarship in India, there is a pressing need to provide community engagement and collaboration, resources, training, and funding to support researchers working in diverse linguistic contexts; establish evaluation and recognition criteria acknowledging contributions in multiple languages within academic assessments and promotions; encourage multilingual journals or platforms that embrace diverse languages, promoting cross-cultural dialogue and understanding; ensure cultural sensitivity and ethical considerations in translating and representing diverse languages and knowledge systems; and develop mechanisms for peer review in multiple languages to maintain academic rigor across linguistic boundaries.

Appendix

Multilingual Initiative Examples

Sahitya Akademi organizes literary festivals promoting multilingual literature and showcasing works from diverse Indian languages. It holds India's largest multilingual library as well as the Centre for Oral and Tribal Literature, with a purpose to create a "House of Voices" to archive the original oral texts available in different languages in audio and audio-video formats. These oral texts are accompanied by written translations in scheduled languages and English for wider distribution. The Akademi is a national organization which works actively for the development of Indian letters and to foster activities in all of the Indian languages. It is the only institution that takes up literary activities in all 24 languages, including English. The Akademi gives awards to literary works and translations from and into the languages of India. Sahitya Akademi publishes and serves as a huge scholarly platform in a linguistically diverse country like India.

The Bharatavani project, launched in 2016, is an initiative by the Ministry of Human Resource Development, which offers multilingual online dictionaries, terminology, and educational resources in several Indian languages. According to the 2011 census, there are 121 languages presented in two parts: (1) 22 languages included in the Eighth Schedule to the Constitution of India and (2) 99 non-scheduled languages. Of the 270 identifiable mother tongues, 123 are grouped under scheduled languages, and 147 are grouped under non-scheduled languages. Moreover, there are several languages/mother tongues (grouped as "others" in the Constitution) that are spoken by fewer people in India, thus limiting languages to their oral form (Choudhary et al. 2023). In the face of globalization and digitization many regional and indigenous languages have been marginalized, endangering the cultural heritage they embody. The Bharatavani project is an initiative which focuses on recording sociocultural and linguistic information about 121 Indian languages and making it accessible to larger audience. It is an online knowledge repository in and about all languages in India in a multimedia format through an online portal. The formal education in India is closely tied to dominant majority languages such as English. The preference of English-medium education marginalizes the major regional and constitutional languages, hence weakening them across various sectors of Indian society. As a result, it is crucial to ensure that content development in these marginalized languages, including their cultural components (textual, auditory, and visual), is easily accessible through an online presence. The project aims to bridge the digital divide and promote equal access to knowledge and information. The Bharatvani project is planned and executed by the Central Institute of Indian Languages (CIIL), Mysore, in the form of a mobile application and a web portal where registered users can access the encyclopedia, language learning material, dictionaries, glossaries, textbooks, and more in 121 languages for free. It is the first of its kind in the world with over 200 operational dictionaries in multiple languages and subject combinations. The Bharatvani project is the world's single hub of important indigenous content in India (Choudhary et al. 2023). The goal is to make an indigenous knowledge repository of India's respective languages available through a robust digital platform. Bharatavani constitutes an integral facet of Digital India, striving to construct a knowledge ecosystem that bridges the gap created by the digital divide.

Through its resolution on January 18, 1968, the Government of India placed significant emphasis on the country's educational and cultural progress. It projected the need to implement the Three-Language Formula, a joint effort between the Union Government and State Governments. This policy aimed to include an Indian language, preferably a south-Indian language, alongside Hindi and English in Hindi-speaking regions. In non-Hindi-speaking areas, the formula proposed Hindi along with a regional language. The imperative to preserve and propagate knowledge across Indian languages underscored the establishment of Regional Language Centres under CIIL. CIIL functions under the Department of Higher Education, Ministry of Education. The institute is the largest catalytic force responsible to assist, advise, contribute, protect, and promote language, language resources, and corpus. CIIL develops language resources, corpora, and technologies for various Indian languages, aiding in multilingual research and development. CIIL was established to coordinate the development of Indian languages through scientific studies, promote interdisciplinary research, and contribute to mutual enrichment of languages. The Bharatavani project, Linguistic Data Consortium for Indian Languages (LDC-IL), National Translation Mission (NTM), National Testing Service–India, and Scheme for Protection and Preservation of Endangered Language (SPPEL) are some of the schemes and projects under CIIL. Such enterprise is performed inter alia, in order to stimulate and support research and development from various linguistic streams to a common head. While creating and advancing a digital environment, CIIL builds resources for electronic publishing in varied languages of India.

Such examples present the bilingual and multilingual institutional practices that create a niche platform for progressive writing, promoting, and publishing of Indian languages. It underscores the historical patterns that highlight and analyze the necessity for multilingualism, countering the colonial dominance of English as the primary medium for institutional and cultural practices. Other examples include the Bhasha Research and Publication Centre founded in 1996, an organization that functions with a vision to "voice" the Adivasi (Tribal) community of India. The Bhasha Research and Publication Centre holds an extensive collection of freely available online audio-video documentation of indigenous communities, promoting multilingualism. Digital

Empowerment Foundation/Digital Literacy and Inclusion runs programs and campaigns promoting digital literacy and inclusion in various Indian languages, enabling multilingual access to technology. Project E-VIDYA is an initiative by the Government of India to provide multilingual e-content for school education across states in various languages. NTM is a translation initiative that works towards translating knowledge texts between Indian languages, fostering multilingual scholarship and access to diverse knowledge systems. Technology Development for Indian Languages (TDIL) focuses on technology development for Indian languages, facilitating multilingualism in digital platforms and tools. Duolingo and Rosetta Stone are some of the language learning apps and platforms that offer courses in Indian languages alongside global languages, promoting multilingualism. Language dictionaries and learning portals Shabdkosh, Rekhta, Bhashini, and others are online platforms offering dictionaries, language learning tools, and resources in multiple Indian languages, fostering multilingual proficiency.

KSHIP, or Knowledge Sharing in Publishing, is an independent publishing center established and managed by the Indian Institute of Technology Indore (IITI). The center aims to offer access to scholarly publications in languages other than English. KSHIP is involved in the development of multilingual open access scholarly publishing focused primarily on the humanities and social sciences. As a multilingual publishing house, KSHIP focuses on scholarly monographs and translations in Indian languages and invites scholars to host journals in Indian languages. The idea is to cater to a much larger humanities and literature research ecosystem by making literature research accessible in Indian languages. KSHIP offers an ambitious example of publishing scholarly works in regional languages, thus promoting the conversation on multilingualism as an opportunity and on the democratization of knowledge in public spaces.

References

Awan, Shafique Ahmed, Zahid Hussain Abro, Akhtar Hussain Jalbani, Dil Nawaz Hakro, and Maryam Hameed. 2018. "Handwritten Sindhi Character Recognition Using Neural Networks." *Mehrun University Research Journal of Engineering and Technology* 37, no. 1: 6. https://doi.org/10.22581/muet1982.1801.17.

Bajwa, Nida ul Habib, and Cornelius J. König. 2017. "On the Lacking Visibility of Management Research from Non-Western Countries: The Influence of Indian Researchers' Social Identity on Their Publication Strategy." *Management Research Review* 40, no. 5: 538–55. https://doi.org/10.1108/mrr-02-2016-0036.

Bhambra, Gurminder K., Dalia Gebrial, and Kerem Nişancıoğlu, eds. 2018. *Decolonising the University*. London: Pluto Press.

Chavarro, Diego, Puay Tang, and Ismael Ràfols. 2017. "Why Researchers Publish in Non-mainstream Journals: Training, Knowledge Bridging, and Gap Filling." *Research Policy* 46, no. 9: 1666–80. https://doi.org/10.1016/j.respol.2017.08.002.

Chen, Ninghan, Xihui Chen, Zhiqiang Zhong, and Jun Pang. 2021. "The Burden of Being a Bridge: Understanding the Role of Multilingual Users during the COVID-19 Pandemic." arXiv, submitted April 9, 2021. https://doi.org/10.48550/arxiv.2104.04331.

Choudhary, Narayan, L. R. Premkumar, Chandan Singh, Shubhana Mondal, Shivangi Priya, Beluru Sudarshan, P. Perumal Samy, and Shailandra Mohan. 2023. "Bharatavani Project—Reviving Linguistic Diversity and Cultural Heritage in India: A Case Study." University of North Texas Libraries, UNT Digital Library, July 3, 2023. https://digital.library.unt.edu/ark:/67531/metadc2114300/.

Curry, Mary Jane, and Theresa M. Lillis. 2014. "Strategies and Tactics in Academic Knowledge Production by Multilingual Scholars." *Education Policy Analysis Archives* 22. https://doi.org/10.14507/epaa.v22n32.2014.

Fiormonte, Domenico. 2016. "Toward a Cultural Critique of Digital Humanities." In *Debates in the Digital Humanities 2016*, edited by Matthew K. Gold and Lauren F. Klein, 438–58. Minneapolis: University of Minnesota Press. https://doi.org/10.5749/9781452963761.

Galina Russell, Isabel. 2014. "Geographical and Linguistic Diversity in the Digital Humanities." *Literary and Linguistic Computing* 29, no. 3: 307–16. https://doi.org/10.1093/llc/fqu005.

Ghorbaninejad, Masoud, Nathan P. Gibson, and David Joseph Wrisley. 2023. "Right-to-Left (RTL) Text: Digital Humanists Plus Half a Billion Users." In *Debates in the Digital Humanities 2023*, edited by Matthew K. Gold and Lauren F. Klein, 234–48. Minneapolis: University of Minnesota Press. https://doi.org/10.5749/9781452969565.

Hakro, D. N., I. A. Ismaili, A. Z. Talib, Z. Bhatti, and G. N. Mojai. 2014. "Issues and Challenges in Sindhi OCR." *Sindh University Research Journal (Science Series)* 46, no. 2: 143–52. https://sujo.usindh.edu.pk/index.php/SURJ/article/view/5337/3618.

Hakro, D. N., M. Memon, S. A. Awan, Z. A. Bhutto, and M. Hameed. 2016. "Isolated Optical Character Recognition." *Sindh University Research Journal (Science Series)* 48, no. 4: 839–44. https://sujo.usindh.edu.pk/index.php/SURJ/article/view/4705.

Hakro, Dil Nawaz, and Abdullah Zawawi Talib. 2016. "Printed Text Image Database for Sindhi OCR." *ACM Transactions on Asian Low-Resource Language Information Processing* 15, no. 4: 1–18. https://doi.org/10.1145/2846093.

Hicks, Diana. 2004. "The Four Literatures of Social Science." In *Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems*, edited by Henk F. Moed, Wolfgang Glänzel, and Ulrich Schmoch, 473–96. New York: Kluwer Academic.

Huckle, Jacob. 2021. "Multilingualism and the International Baccalaureate Diploma Programme (IBDP): An Analysis of Language Learning in the IBDP in Light of the 'Multilingual Turn.'" *Journal of Research in International Education* 20, no. 3: 263–80. https://doi.org/10.1177/14752409211059267.

Israel, Hephzibah. 2021. "Translation in India: Multilingual Practices and Cultural Histories of Texts." *Translation Studies* 14, no. 2: 125–32. https://doi.org/10.1080/14781700.2021.1936149.

Kothari, Rita, ed. 2018. *A Multilingual Nation: Translation and Language Dynamic in India*. New Delhi: Oxford University Press.

Kulczycki, Emanuel, Raf Guns, Janne Pölönen, Tim C. E. Engels, Ewa A. Rozkosz, Alesia A. Zuccala, Kasper Bruun, et al. 2020. "Multilingual Publishing in the Social Sciences and Humanities: A Seven-Country European Study." *Journal of the Association for Information Science and Technology* 71, no. 11: 1371–85. https://doi.org/10.1002/asi.24336.

Liu, Alan. 2018. "Digital Humanities Diversity as Technical Problem." *Alan Liu* (blog), January 15, 2018. https://doi. org/doi:10.21972/G21T07.

Mahony, Simon. 2018. "Cultural Diversity and the Digital Humanities." *Fudan Journal of Humanities and Social Sciences* 11:371–88. https://doi.org/10.1007/s40647-018-0216-0.

Masoud Ghorbaninejad, Nathan P. Gibson, and David Joseph Wrisley. 2023. “Right-to-Left (RTL) Text: Digital Humanists Plus Half a Billion Users.” In *Debates in Digital Humanities 2023*, edited by Matthew K. Gold and Lauren F. Klein, 47–73. Minneapolis: University of Minnesota Press. https://doi.org/10.5749/9781452969565

Meza, Aurelio. 2019. “Decolonizing International Research Groups: Prototyping a Digital Audio Repository from South to North.” *Digital Studies/Le champ numérique* 9, no. 1: 7. https://doi.org/10.16995/dscn.303.

Mondal, Amrita. 2014. “Educational Intervention and Negotiation: A Case Study of Serampore Mission and New Education.” *Exploring History* 5/6, nos. 2/1 (December 2013–June 2014): 75–89.

ORGI (Office of the Registrar General and Census Commissioner, India). 2022. *Language Atlas of India 2011*. New Delhi: Ministry of Home Affairs, Government of India. https://censusindia.gov.in/nada/index.php/catalog/42561.

Pandita, Ramesh. 2014. “Growth and Distribution of Hindi, English, and Urdu Periodicals in India: An Analysis (1941–2013).” *DESIDOC Journal of Library & Information Technology* 34, no. 4: 309–16. https://doi.org/10.14429/djlit.33.5920.

Pandya, Chintan, and Jasmine Gohil. 2023. “Between Quandary and Squander: An Analysis of Preservation Practices of Vernacular Literature at University Libraries.” *Collection and Curation* 42, no. 2: 46–52. https://doi.org/10.1108/cc-04-2022-0014.

Santos, Boaventura de Sousa. 2017. *Decolonising the University: The Challenge of Deep Cognitive Justice*. Newcastle upon Tyne: Cambridge Scholars Publishing.

Sivertsen, Gunnar. 2018. “Balanced Multilingualism in Science.” *BiD* 40. https://doi.org/10.1344/BiD2018.40.25.

Spence, Paul. 2021. *Disrupting Digital Monolingualism: A Report on Multilingualism in Digital Theory and Practice*. Language Acts and Worldmaking Project. https://doi.org/10.5281/zenodo.5743283.

Spence, Paul Joseph, and Renata Brandao. 2021. “Towards Language Sensitivity and Diversity in the Digital Humanities.” *Digital Studies/Le champ numérique* 11, no. 1. https://doi.org/10.16995/dscn.8098.

T, Shanmugapriya, Shaifali Arora, and Nirmala Menon. 2017. “Developing Database for Scholarship in Indian Languages and Literatures.” *Asian Quarterly: An International Journal of Contemporary Issues (AQ)* 15, no. 4: 85–99.

de Swaan, Abram. 2001. *Words of the World: The Global Language System*. Cambridge: Polity.

UNESCO. 2023. “Multilingualism.” Information for All Programme, April 20, 2023. https://www.unesco.org/en/ifap/multilingualism.