Recovering knowledge commons for the global south

Ghosh, Arjun
Indian Institute of Technology Delhi, New Delhi
arjunghosh@hss.iitd.ac.in

Abstract

The colonial encounter instituted the hegemony of documentary practices over oral, performative and manuscript practices. Only knowledge validated through the process of print publication could stand the test of legal scrutiny. On the one hand the Western epistemological quest glossed over ideas that existed through ephemera, on the other hand that knowledge which Western print practices imbibed from non-European traditions were henceforth, locked behind intellectual property regimes and restrictive archival practices. This tremendously skewed access to knowledge between the North and the South. Tools for digital transformation, particularly those that are based on artificial intelligence (AI) and machine learning (ML), can be culture specific eg. training data for Optical Character Recognition (OCR)/ Hand-written Text Recognition (HTR); datasets for natural language processing. In the absence of culture specific tools in underdeveloped societies Anglo-American interpretive categories and assumptions become the default. Further, Anglo-American institutions work to use their advantage in the balance of knowledge distribution to maintain their hegemonic position. In order to protect the South's access to human heritage and knowledge we need to develop technologies that leverage the potentials of digital communication for increased conversations among languages of the South.

Keywords: digital transformation, low resource languages, global south, natural language processing, typography

1 Learning from the onset of modernity

Writing in 1962 about the coming of the electronic age Marshall McLuhan compared the period to a similar moment of epochal shift in communication technology - "We are today as far into the electric age as the Elizabethans had advanced into the typographical and mechanical age" (McLuhan et al. 2011, 1). McLuhan was indicating the significant shifts in human perception and social and political reorganisation that were witnessed with the coming of modernity. As print technology, the principal driver of the typographic age evolved in western Europe, it was this region where the initial symptoms of modernity can be traced to. However, with the spread of colonialism the political, cultural and social dimensions of the typographic age affected the colonies at different times through the subsequent centuries. The influx of modernity under conditions of colonial contact created epistemic and structural distortions that continue to affect life in the former colonies even decades after political independence. I argue that much of these distortions were results of an imposition of epistemic categories that developed organically in western Europe but were imposed, often violently in the colonies. In this paper I seek to interrogate possible repetitions of the same distortions in the current epoch where we are seeing a movement from a predominantly documentary universe to an increasingly digitized world. Western compuutational models are often uncritically adopted in solutions for diverse fields like governance, pedagogy, research, entertainment and commerce. These should not be viewed as politically neutral. The paper will also engage in some loud thinking regarding ways to avoid the recolonial effects of digital transformation.

2 Typefaces and linguistic identities

In his work *Imagined communities: reflections on the origin and spread of nationalism* (1991), Benedict Anderson outlined the far reaching consequences of print in shaping homogenous national identities across a wide geographical extent (Anderson, 1991). The mechanization of the reproduction of knowledge created a dependency on the production process of the typeface. Once a type-foundry produces a typeface, the design gets used at multiple printing presses. Once a printing press produces a typeset and a printed page from it, identical copies of the printed page are distributed through the bounds

of the political entity thus producing and strengthening a community based on the common language that lay fixed in the pages printed from a common typecast. While the uniformities of the typeface and the printed page are replicated multifold creating national identities in Europe, in the colonies the national aspirations generated by print are thwarted by the realities of the colonial rule. While in the metropole the linguistic nationality coincides with the political entity that rides it, in the colony the political command simultaneously produces linguistic identities (albeit ironically against the administrative design) and mutilates the avenues for its realization. In India, the British colonial state instituted printing presses in various Indian languages for administrative and missionary motives. This work involved the creation of typefaces in these languages. The earliest typefaces were produced in Europe. The officers at the Fort William College in Calcutta, who were tasked with the training of British administrators in the colony, also took upon themselves to set the grammar of some of these languages. The result was a privileging of some language groups over others and undermining linguistic diversity – giving rise to 'standard' Bangla, Hindi or Marathi etc. (Darnton, 2002, 2001; Ghosh, 2004; Roy, 1995). The British administrators did have a more sinister design in the use of linguistic realignment for identity formation in South Asia – but more about that later.

3 QWERTY Keyboard to Touch Screens

Printing technology moved a long way since the days of Gutenberg and Caxton till the 1870s when the typing machine and the QWERTY keyboard were created. The ubiquity of the QWERTY keyboard in contemporary times is another fallout of British and American global domination in the 20th century. The QWERTY keyboard made the Roman alphabet the default language of documentation across the world, with names of people, places and communities having to be spelled in the Roman alphabet from the birth certificate to the death certificate. The documentary transformation of the world's governance, trade and cultural expressions by default became 'Romanised'.

When the early computers came along in the mid-20th century the QWERTY keyboard became the predominant input device. Even those languages that may have largely escaped "Romanisation" in the phase of documentary transformation faced the threat of being run over by the Roman alphabet or even by the English language altogether as English became the lingua franca of the world. Though the technology for printing did exist for various script systems, the relative access and diversity for typewriting systems and fonts were far more limited. When the earliest mobile phones made it to the market with their push button equivalent of the QWERTY keyboard, users took to it to avail text messaging services – causing further resort to "Romanisation".

However, with the advent of smart phones and touch screen systems the dependence on the Roman alphabet is challenged for the first time since the invention of printing. Touch screen interfaces make it technologically possible for users to switch between scripts and users can choose to type their messages and content in a language of their choice. In fact, most device manufacturers as well as application developers do offer the users a wide range of languages as the default system language. The development and availability of transliterator engines further make text content creation in a script and language of one's choice a real possibility for the first time in many centuries.

If the advent of print was simultaneous with the advent of modernity and therefore, the inception of the hegemony of the North over the South, can the versatility of the touch screen input devices and other input systems that circumvent the dependence on a single script, be the moment where the hegemony of the North is challenged? This question cannot be answered by technology alone but through making certain choices in the research and development of digital technologies.

4 Cultural impact of colonial documentation

Colonial contact brought Europeans to administer lands which till then largely depended on oral and customary means of governance. The advent of British colonialism in India saw an elaborate and institutionalized effort to extract performative, oral and customary knowledge into a documentary order through records and surveys. The British set up institutions like the Survey of India for conducting topographic surveys and producing detailed maps, Anthropological Survey India for documenting the population, the Cadastral Survey for determining land ownership and revenue assessment, the Geological Survey of India for enumerating the mineral resources. Through such institutions, the British administration sought to quantify and control the land, its people, its resources and its diversity.

In his essay "Colonialism and its forms of knowledge" Bernard S. Cohn recorded that the British used land records as a mechanism for manipulating ownership patterns and taxation in order to protect and consolidate their own interests (Cohn, 2021). Thomas Metcalf noted that the British used land records and surveys as mechanisms to order the landscape of Indian provinces and keep them under their control (Metcalf, 2002). According to David Ludden such efforts by the British colonial administration to bring land tenure systems and revenue records under documentary control disrupted traditional communities and support systems and generated new economic disparities (Ludden, 1985). In the Malabar region of present day Kerala the British intervention, the demand for land records and transfer of ownership through male heredity, resulted in the disruption of the matrilineal *tharavad* [1]. This in turn led to the fragmentation of the large joint family systems towards nuclear families causing loss of resources. One of the effects of these changes was the decline of traditional performance forms like Theyyam that were hitherto sustained by the resources generated by the *tharavad* (Robin, 1976; Saradamoni, 1999). Beyond disruption of existing social and cultural practices, these impositions of documentary order produced reactions among the people of South Asia. Ranajit Guha exposed how the British policies of land records and revenue collection periodically led to peasant rebellions (Guha, 1983).

5 Cleaving Urdu and Hindi

The only occasions where the British accentuated South Asian diversity were when diversity could be used to generate conflict. As noted earlier, in order to train British administrative staff the Fort William College was established in Calcutta. John Gilchrist was appointed as Supervisor at the Hindustani Department of the college. As scholars have recorded, Hindustani or Urdu was the lingua franca of most parts of northern, central and western India at the time when the British consolidated their control over most parts of the subcontinent [2]. The language, or more correctly, the set of languages identified as Hindustani or Urdu is said to have originated in army camps and *bazaars* of the Mughal period. It was there that local languages interacted with the Perso-Arabic-Turkic languages of commanders and soldiers of the Mughal army. It wasn't until the 18th century that Hindustani or Urdu gained a position of esteem in the Mughal court. Till then it remained a language of lay usage and largely oral circulation. When it did emerge as a language of poetry and literary expression it could be written in the Nastaliq-Persian script or the Nagri-Sanskrit script depending on the comfort of writer and the readers.

Emerging from European documentary consciousness where national identities were seen as culturally homogeneous – Gilchrist sought to find a difference between Hindi and Urdu. The British regarded caste and religion as the determining factor of Indian polity and saw a fundamental difference between Hindus and Muslims. British orientalist historians sought to divide Indian history into phases – a glorious Hindu past disrupted by a tyrannical Muslim rule. Having presented India's past in terms of civilizational conflict, the British presented themselves as the saviours of the Hindus. The syncretic culture of Hindustani and Urdu in the pre-colonial Hindustan did not fit with this

narrative. Gilchrist, while commissioning language textbooks, ordered that textbooks meant for Hindus would be written in the Nagri script, and those for Muslims would use Nastaliq. The former ultimately, through a purging of words of Perso-Arabic origin, became Hindi. The later, through a symmetrical purging of words of Sanskrit origin, became Urdu (Orsini, 2004; Rai, 2001; Stark, 2009).

As history has noted, the British pursued the engineering of differences between Hindus and Muslims more aggressively after the rebellion of 1857. This culminated in a bloodied partition of the subcontinent in 1947. Even today religious bigotry and rivalry continues to plague the politics and life of India and Pakistan. The seeds of documentary transformation undertaken by colonial rule continuing to bear ominous fruits. Similar instances of inter-ethnic violence, that owe their inception to colonial administrative interventions, continue to rage at various places across the global South.

6 Affordances of digital transformation

6.1 Virtual keypads

We have already discussed the possibility of text content generation in multiple scripts and languages through virtual keyboards on touch screen devices. It is possible today to imagine that a user can move directly from literacy in a non-European language to the use of the keypad in that language without having to negotiate the Roman script. While increasingly large numbers of people generate content daily in non-Roman characters we should still note that their usage is largely restricted to languages that are dominant in the print domain specific to that script. For instance, a Bangla user can choose from among two Bangla keyboards for installation on their device – Bangla (India) or Bangla (Bangladesh). Other languages like Sylheti or Hajong, which historically use the Bangla-Assamese script family do not appear as an option on most devices and platforms. Most of these virtual keyboards used predictive text and autocompletion for effective use. An option to use wider variety of languages aligned to particular scripts is absent. This is an area for further research and development. Such research and deployment can potentially reverse the documentary ecosystem where dominant languages hold sway over knowledge creation, distribution and access.

6.2 Digitization of documentary records

One of the driving forces of research and exploration in the digital humanities has been the digitization of documentary resources and records – both manuscript and print. There are two steps to the digitization of documentary heritage – first, is the scanning of the records to produce digital images; second, is the conversion of the images of these records to machine readable text. Digitization of documentary records can serve two important functions. First, making the records available in image formats helps in the preservation of these records which may undergo further damages and disintegration with continued handling by researchers. Second, since images can be accessed across the internet it can allow greater access to researchers without the need to travel to the physical archive. As scholars in formerly colonized parts of the world would vouchsafe, in order to study the history and heritage of their own cultures and people they have always had to travel to archives located in the former metropole. A visit to the India Office Records room of the British Library becomes almost imperative for a scholar who wants to study colonial history of South Asia. However, this exchange has never been one of equals. Due to prevailing currency exchange rates this interaction remains unequal. For scholars in the global South access to archives in the former metropole is more difficult to come by. In such situations scholars in the former colonies would rely on their ability to access archives nearer home where possible.

Most places in the global South face substantial challenges at both these steps. Digital scanning of documents is a resource intensive process. Most archives and libraries in the global South are not able to muster up enough resources to be able to undertake largescale digitization. In Indiacommendable efforts in digital scanning of records have been undertaken by institutions like

the National Archives, the Indira Gandhi National Centre for the Arts, the National Library in Kolkata, the National Manuscripts Mission and the Rekhta Foundation to name a few. Yet, there is still a general lack of support for digitization efforts. Hence, being the custodian of fast perishing records, escalating costs and lack of funding, many archives in the global South turn to institutions in the global North for funding. While such funding does help prolong the life of the physical records, it seldom increases the accessibility for scholars in the global South. While the archive in India that partnered in the digitization of the archival collection receives a copy of the digitized images, the nonavailability of a unified platform for accessing these resources means that scholars in the home country still have to travel to the archive to access these images. Hence, while the access for the scholars of the global North increases, the corresponding increase in access for the scholars from the global South is severely limited behind pay wall and the need for dollar grants. This is akin to what Roopika Risam terms as the "digital afterlives of colonialism" where the archival resources of the former metropole continues to set the terms of academic debates (Risam, 2018).

6.3 Machines reading text

The second step of digitization involves producing machine readable text from the digitized image records. There is a dearth of accurate, reliable and user friendly platforms for optical character recognition and handwriting text recognition in South Asian scripts and languages. The availability of machine readable text can be important from the point of view of scholarly inquiry for several reasons. First, machine readable text makes lesser demands on computer memory space as the files storing text are smaller than that of images. Hence, machine readable texts make the storage, distribution and access to resources more affordable. The absence of machine readable text for records in South Asian languages increases the costs for the archives housing resources in these languages. Second, machine readable texts allow deep search of the contents of the archive. In the absence of deep search, bibliographic search has to rely solely on metadata. The absence of deep search for non-Roman scripts can create a bias towards archival records that are available as machine readable text in Roman scripts.

Third, machine readable text can be subjected to computational methods of inquiry. The field of computational literary studies has made giant strides in areas like topic modeling and spatial analysis revealing patterns within large text corpora which were undiscovered by the human eye. Such techniques require trained natural language processing (NLP) datasets specific to each language being analysed. NLP datasets can be rules based, or modeled on labeled data. NLP tools can also be developed from large unlabeled data. Further, machine readable texts, whether sourced from digitized documentary records or from born-digital content, form the corpora that are used to train machine learning (ML) models that are used for predictive tools. In a paper that surveys the state of linguistic diversity and inclusion in NLP tools among the various languages of the world, Joshi et al. report that the gap between low resource languages and high resource languages can keep widening as NLP tools become more sophisticated. Low resource languages suffer at both ends. There being a dearth of resources like documentary data, OCR tools, virtual or phonetic keyboards, it is difficult to obtained labeled data on these languages. And since there is very little content on these languages ML approaches of working with unlabeled data are also not successful (Joshi et al., 2021).

The state and corporate push towards the use of ML-based tools for governance has been growing in recent years. Intervention ranges from early crime detection to generation of product demand. In the absence of adequate ML models that are trained on the languages and cultures of operation, these governance models make use of models that have been trained in other languages and other contexts (Ramesh et al., 2023). Hence, their predictive accuracy is culture specific. Since, most of these tools originate from the global North, the predictions based on these tools potentially seek to replicate the governance

structures and assumptions of the global North in the global South. In fine, there is a real and present danger of the imposition of Western knowledge systems and administrative priorities on the global South in the process of digital transformation – a phenomena that would replicate the documentary transformation under colonial rule. A case in point, is the dearth of labeled datasets for detection of hate speech, cyber bullying, fake news in South Asian languages. This has presented a major threat to life and liberty in South Asia. The absence of software tools to detect and act upon such instances lends credence to the exponential occurrences of internet shutdowns in India.

7 Strategies for the global south

While documentary recording and distribution of knowledge – whether through manuscripts or print – have led to a democratisation of access to knowledge, it also resulted in the creation of dominant languages. The technology of printing and the incidence of colonialism led to creation of a hierarchy of languages with a handful of print languages and scripts being dominant over thousands of spoken languages. Among these the English language and its Anglo-American base assumed the role of the lingua franca of the world. This hierarchisation of languages and undermining of diversity, though making global knowledge exchange possible also privileges the global North as the net holder and validator of knowledge. The hold of the global North over the flow of world's knowledge is further strengthened by intellectual property regimes that keep knowledge archives behind pay walls and control the development of new knowledge tools through placing source codes of applications behind copyright (Ghosh, 2014).

As we have seen, the imposition of European epistemic framework on largely oral and customary knowledge systems led to significant and long term disruptions in South Asia. The South Asian people, and indeed the people of the global South, are still contending with these disruptions brought about by a documentary transformation that was not allowed to organically develop from their societies and cultures. The people of the global South need to act concertedly to prevent this epistemic imposition being repeated in the phase of digital transformation.

1. Our governments and institutions have to step up the research to develop tools that will facilitate the digital transformation in languages of the global South. In India, the Bhashini project [3], supported by the Ministry of Electronics and Information Technology (MEITY), presents as it goals the development of OCR, NLP tools, machine translation and text to speech in the Indian languages to "enable all Indians easy access to the Internet and digital services in their own languages". While such efforts are commendable and the need of the hour, one must bear in mind that when we speak of "Indian" languages it should genuinely seek to empower a larger set of languages and not the ones that dominated print circulation. It is essential that we remember that, though cultures in the global North went through the trajectory of moving from the oral to the handwritten and then to the typographic before arriving at the digital, the global South has many instances where linguistic communities would be moving directly from the oral to the digital. Hence, an increased impetus must be given to generate digital tools for low resource languages. It is only when every linguistic community feels included in the circulation and revitalization of knowledge that we redevelop diversity. Such diversity can question the polarities of the contemporary world where certain cultures are the sum creators of knowledge and use that to perpetuate their advantage over other cultures.

2. Governance models powered by machine learning and artificial intelligence in the global South should be trained and be informed by the bodies of knowledge emerging for their respective cultures. This must include knowledge available in non-Roman scripts and low resource languages.

3. The copyright regime that was instituted as a response to the growth of printing, has served as an important basis for maintaining the continued hegemony of the global North over knowledge systems. In the digital epoch the need

for synergies between hardware and software systems mean that closed source software prevents the use of hardware. Proprietary control over software systems enable rent seeking that drive up costs for knowledge operators in the global South. Hence, the strategy of the global South should be to escape the rent trap. The development of tools of digital transformation, analysis and prediction should be undertaken in a culture of sharing. Knowhow about the research and development and source code of all developed applications should be shared under open licenses. This would enable rapid and development. Similarly, content generation, research and software development should be a people driven movement using crowdsourcing techniques. Educational planners should encourage children to understand and develop code at an early age so that they can learn to code for themselves and the society around them and not learn to be dependent on readymade, one-size-fits-all proprietorial applications. Moreover, it is particularly important for scholars of the humanities and social sciences to be able to understand code so that they can bring their knowledge of social and cultural systems to critique code.

4. Finally, peoples of the global South should develop avenues of exchange enabling South-South knowledge sharing. We need to move from a predominantly South-North-South trajectory of knowledge flow to South-South, North-South or South-North trajectories. The architecture of digital transformation for the global South should be designed to make possible a direct interaction between languages of the global South. ie. the digital transformation in the global South should be marked by choice and not compulsion.

Notes

[1] The ancestral home for the joint families among the Nair and Nambuthiri communities in Kerala.

[2] Rahul Sankrityana noted "Hindi incorporates all the languages which emerged after the eighth century A.D. in 'Suba Hindustan' - the region that is bounded by the Himalayas, and by all the regions associated with the Punjabi, Sindhi, Gujarati, Marathi, Telegu, Oriya and Bangla languages. Its older form is called Magahi, Maithili, Braj Bhasha, etc. Its modern form may be considered under two aspects: a widely disseminated form called Khari Boli ... and the various local languages which are spoken in different places: Magahi, Maithili, Bhojpuri, Banarasi, Avadhi, Kannauji, Brajmandali, etc…." (Quoted in Rai, 2001, p. 12)

[3] [https://www.bhashini.gov.in/en]

References

Anderson, B.R.O., 1991, Imagined communities: reflections on the origin and spread of nationalism. Verso.

Cohn, B.S., 2021, Colonialism and Its Forms of Knowledge: The British in India. Princeton University Press.

Darnton, R., 2002, Book Production in British India, 1850-1900. Book Hist. 5, 239–262. https://doi.org/10.1353/bh.2002.0005

Darnton, R., 2001, Literary Surveillance in the British Raj: The Contradictions of Liberal Imperialism. Book Hist. 4, 133–176. https://doi.org/10.1353/bh.2001.0007

Ghosh, A., 2014, Freedom from Profit: Eschewing Copyright in resistance Art. Indian Institute of Advanced Study, Shimla.

Ghosh, A., 2004, Cheap Books, 'Bad' Books: Contesting Print Cultures in Colonial Bengal, in: Gupta, A., Chakraborty, S. (Eds.), Print Areas: Book History in India. Permanent Black, Delhi, pp. 189–196.

Guha, R., 1983, Elementary Aspects of Peasant Insurgency in Colonial India. Oxford University Press.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., Choudhury, M., 2021, The State and Fate of Linguistic Diversity and Inclusion in the NLP World. https://doi.org/10.48550/arXiv.2004.09095

Ludden, D., 1985, Peasant System in South India. Princeton University Press.

Metcalf, T.R., 2002, An Imperial Vision: Indian Architecture and Britain's Raj. University of California Press.

Orsini, F., 2004, 'Pandits, Printers and Others: Publishing in Nineteenth Century Benares,' in: Gupta, A., Chakraborty, S. (Eds.), Print Areas: Book History in India. Permanent Black, Delhi, pp. 103–138.

Rai, A., 2001, Hindi nationalism. Orient Blackswan.

Ramesh, K., Sitaram, S., Choudhury, M., 2023, Fairness in Language Models Beyond English: Gaps and Challenges. https://doi.org/10.48550/arXiv.2302.12578

Risam, R., 2018, New Digital Worlds: Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy. Northwestern University Press, Evanston.

Robin, J., 1976, The Decline of Nayar Dominance: Society and Politics in Travancore, 1847 - 1908. University of Sussex Press.

Roy, T., 1995, "Disciplining the Printed Text: Colonial and Nationalist Surveillance of Bengali Literature," in: Chatterjee, P. (Ed.), Texts of Power: Emerging Disciplines in Colonial Bengal. University of Minnesota Press, London and Minneapolis, pp. 30–62.

Saradamoni, K., 1999, Matriliny Transformed: Family, law and ideology in twentieth century Travancore. Sage Publications, New Delhi.

Stark, U., 2009, An Empire of Books: The Naval Kishore Press and the Diffusion of the Printed Word in Colonial India. Permanent Black, New Delhi.