24

Language Documentation & Conservation Special Publication No. 15
Reflections on Language Documentation 20 Years after Himmelmann 1998
ed. by Bradley McDonnell, Andrea L. Berez-Kroeker & Gary Holton, pp. 248–255
http://nflrc.hawaii.edu/ldc/
http://hdl.handle.net/10125/24826

Reflections on language documentation in India

Shobhana Chelliah
University of North Texas

The last twenty years have seen efforts to support the study of minority and lesser-studied languages of India from varied stakeholders: these include the Indian government, international and Indian nonprofit organizations, indigenous and state-level cultural and language committees and institutes, and individuals with a passion to preserve and document their cultures and languages. Their efforts have led to mixed success due to conflicting ideologies, history, and resource availability (Annamalai 2003). Basing my observations on my research, personal experience and engagement with language documentation activities in the country, I provide an overview of the current state of language study and my hopes and efforts for future of language documentation and description in India.

1. A philosophy of language science and its consequences Woodbury defines language documentation as "the creation, annotation, preservation, and dissemination of transparent records of a language" (2011:159). While there is plenty of traditional data gathering toward language description in India (Abbi 2001), the other facets of language documentation mentioned in Woodbury's definition are still emergent.

A common viewpoint in linguistics departments in India is that extended and varied data collection, such as that needed for the creation of a documentary corpus, is time consuming and not useful for scientific publications. As result, much of the language data collected is at the level of the word or clause and is collected through responses to questionnaires. Language science, produced and sanctioned in this way, is reflected in the very successful society and related summer school, known as the Formal Studies in the Syntax and Semantics of Indian Languages (FOSSSIL). The stated goal of FOSSSIL is to:

> undertake schemes/activities leading to the development of formal linguistic descriptions and analyses of Indian languages and to provide facilities and act as a forum for exchange of information, ideas and experience in the

BY-NC-SA

ISBN: 978-0-9973295-3-7

> practices and techniques in formal linguistic descriptions and analyses of Indian languages.[1]

Many of the leading linguists in India are on the governing body of FOSSSIL and likewise, many of the programs around the country have this same focus. A consequence of this focus is the lack of encouragement towards resource creation, especially the lack of language data in the form of annotated corpora. Annotated corpora, however, are central to the goals of language documentation since such corpora provide the interpretive apparatus to audio and video records. When the corpus includes a variety of genres and events, even without audio and video, the corpus can illustrate and linguistic and cultural practices for future generations.

In the Indian context, annotated corpora could provide data sources to transform language science in the region. Looking at the descriptive theses and dissertations produced over the last 20 years, we see that linguists-in-training need more scaffolding in their attempts at language description. With only translation to guide data gathering, these researchers often miss less observable or reportable grammatical features like evidential systems (I discussed this in some detail in Chelliah 2001). Naturally occurring language samples could stimulate new avenues of grammatical investigation and thus prompt new linguistic discoveries. An easy way to do this would be to require Ph.D. theses that use minority or lesser-studied language data for typological, descriptive, or other analytic argumentation to include annotated texts in appendices with data cross-referenced in the body of the thesis. Another possibility would be to establish a new format for MA theses where these are not just descriptive sketches but are in the main annotated corpora, collected, analyzed, and annotated using the gold standard for this type of data analysis and including a grammatical sketch (see Woodbury 2014). This would serve several purposes: it would challenge authors to move beyond the fill-in-the-blanks method of grammar writing to one that attempts to account for patterns found in the corpora; provide training in transcription, annotation, analysis, and translation; and create useful documentary materials.

2. National infrastructure and linguistic culture India, through policy and practice, has privileged its larger languages, so it is unclear how efforts at documenting minority and lesser-studied languages will overlay with this linguistic culture in the long term. Indian states were reorganized based on language after independence (King 1997). National policy is in place to recognize languages and provide resources for their maintenance if the language: (1) Has an ancient literary tradition, in which case it falls under the "classical" language category; (2) Has speaker population of over a million and a writing system, which can support inclusion in the constitution's schedule of national languages (the 8th schedule). Twenty-two languages have this status and afford speakers rights to early childhood education and the ability to take national exams in that language; (3) Has a "pure" variety, which is often the higher register in a diglossic situation. The effort to preserve such purity is the ideology that led to the curation and preservation of ancient Vedic hymns in the first millennium and led to the need for the speaker of other varieties to have the interpretative grammatical insights offered by Panni (Scharf 2013, Schiffman 1996:152). In an environment increasingly hostile to the secular ideals of the Indian constitution, ideologies that use Hindi as a symbol for Hinduism have contributed to an imagined dialect continuum that has impeded accurate language identification. For

[1] `http://www.fosssil.in/index_fosssil.htm`

example, the 2001 Indian Census assigns membership of 47 varieties as dialects of Hindi and overall 1652 varieties into 122 languages (Abbi 2004: 2).

On the other hand, the Indian government both indirectly and directly supports language documentation of minority languages. An example of indirect support that affirms language documentation as a worthy activity was seen in 2013 when the Indian government conferred a high national honor, the Padma Shri, to Anvita Abbi for her work documenting the languages of the Andaman.

Examples of direct support for language documentation and dissemination are through infrastructure such as the creation of institutes and projects. The Sahitya Akademi, an organization dedicated to the promotion of the literature of major Indian languages, recently added a Centre for Oral and Tribal Literature. Under, the direction of Anvita Abbi in 2015–2017, this center released several publications on lesser-known languages, initiated a new series titled *Unwritten Languages*, and started a digital collection of oral literature. The Indira Gandhi National Centre for Arts Cultural Archive (IGNCA) located in New Delhi houses artifacts of anthropological interest including language materials such as microfiche of pre-20th century manuscripts from North East India. The IGNCA includes a growing digital repository.

The Office of the Registrar General and Census Commissioner under their Language Division oversees the Linguistic Survey of India (LSI). Initiated in 1981, this work is an ambitious updating of the pre-independence Linguistics Survey of India compiled, edited, and published by George Grierson between 1903–1928. The volumes produced by LSI are extensive in the coverage of demographic information, list of languages and numbers of speakers including a type of social network analysis laying out which languages are home languages, and which languages have wider function. Also included are maps and sociolinguistic information. The linguistic descriptions include useful overviews of nominal and verbal morphology including verb conjugation and simple sentence types. Descriptions are more detailed for larger languages like Oriya (42 pages), than for smaller languages like Relli (23 pages). The format for all are the same since they were collected using a standardized survey: a word list, sentence list, and a story in English for the connected text sample which appears to be the Prodigal Son from the book of Matthew in the New Testament Bible. This is one of the texts translated in the Grierson LSI.[2] The preface of one volume also notes that, "the whole of the interview was recorded us[ing] a tape recorde[r], the recording of which were transcribed in the field, using narrow phonetic transcription based on the International Phonetic [Alphabet] (Banthia 2002: viii)". Although some of the original Grierson LSI gramophone recordings are available online through the University of Chicago South Asia Digital Library, the modern LSI recordings are not publically accessible in either analog or digital format.

In 2007 Government of India also earmarked 280 crore (approximately $40 million USD in 2018 conversion rates) towards the documentation of endangered languages. The Central Institute of Indian Languages (CIIL) in Mysore was commissioned to oversee the training of hundreds of field linguists to document speech varieties village-by-village to gather information on language structure, script, and literature (Srivatsa 2012). This project is no longer in place as it was first conceived although work on endangered languages continues at CIIL. In 2013, the Ministry of Human Resource Development initiated the Scheme for Protection and Preservation of Endangered Languages (SPPEL) with the immediate goal of providing a grammar, dictionary, and ethnolinguistic sketch

[2]PDFs of these descriptions and the preface information are viewable Linguistic Survey of India website: `www.censusindia.gov.in/2011-documents/lsi/ling_dnh.html`

for 117 languages of 10,000 or fewer speakers with a long-term goal of covering 500 languages.[3] CIIL, in collaboration with academic and cultural institutions, is leading the implementation of this scheme across India. While the SPPEL website includes ethnolinguistic notes; sample audio clips from word lists with transcription; and a bibliography for some of the listed languages, progress towards the goal of grammars and dictionaries appears to be slow. SPPEL has also undertaken a massive effort to train native speakers and younger linguists in data elicitation, audio recording, and acoustic analysis. A related annual conference, The International Conference on Endangered and Lesser Known Languages, features invited lectures by national and international documentation and archiving experts.

The University Grants Commission in the 2000s set up centers to support endangered languages: for example, the Centre for Endangered Languages of Northeast in Tezpur University in Assam, and the Centre for Endangered Languages & Mother Tongue Studies at the University of Hyderabad in Telangana. In addition to documentation and preservation, the vision for the centers is to instill in speakers an appreciation for their languages and to support language transmission through creation of pedagogical material. This linking of social responsibility and linguistic work is relatively new for India and is articulated with increasing frequency. An example is the purpose statement of the new conference, Approaches and Methodologies for the Study of Indigenous and Endangered Languages, which states:

> Language is an integral part of the social identity and ethnicity of a group. In order to preserve a social group's or tribe's cultural identity, it is essential to understand the urgency and draw a roadmap for preserving their cultural, social and linguistic heritage and identity.[4]

To these newer dedicated centers, we can add established programs that have been offering field-based linguistics courses: Jawaharlal Nehru University and Delhi University in Delhi; the Central Institute of Indian Languages in Mysore; the University of Guwahati and Tezpur University in the Assam; and Chandigarh University in Punjab.[5]

Language documentation in India is taking place where two contesting ideologies exist: one that articulates the value of linguistic diversity and the other that supports the large and religiously relevant. Speakers of minority languages contest dominant ideologies through grassroots movements valorizing their languages and affirming cultural and group identity through language. At the same time, some of the same groups support the dominant ideology by, for example, lobbying for and gaining acceptance to the 8th Schedule of languages (e.g., Meiteiron (Manipuri) with approximately1.2 million speakers was included in the 8th schedule in 2005). The fluid alignment and misalignment with state policy is reflected in the ways we affirm identity in India, from code switching, code selection, conformity and resistance.[6]

[3] `www.sppel.org/soligadoc.aspx`

[4] `https://linguistlist.org/issues/29/29-414.html`

[5] The existing literature on the ethical conduct in language documentation still focuses on the "white researcher and native community". The Indian context needs consideration of a much more complex mix of "researchers", including missionaries, literacy experts, national surveyors, national and international academics, speaker-academics, and citizen-documenters.

[6] I discuss this in reference to name choice in Chelliah (2005).

3. Institutional incubators and grassroots movements In parallel with formal linguistics programs in India, language documentation and preservation activities are numerous.

At the Northeast Indian Linguistics Conference which meets in Assam and nearby states every two years, I learned that many communities are eagerly pursuing orthography development and dictionary creation, some with the help of missionary groups interested in Bible translation. I have written about individuals who have worked to document their cultural events and practices with minimal training (Chelliah 2016). In the Lamkang community in Manipur state, Mr. Beshot Khular has produced three books and a set of audio and video DVD on proverbs, traditional narratives, and traditional song and dance. Reverend Daniel Tholung recorded elders telling traditional stories and created film on dances during major festivals to capture specific outfits, headdresses, and other ornamentation used during those dances. He paid close attention to the vocabulary used during these events and to the specific uses of objects. Another remarkable documenter is Mr. Somi Roy, the founder and managing trustee of the Imasi Foundation. The mission of the IMASI foundation is to promote Manipuri culture through documentation, preservation, and dissemination of literature. Recently, Mr. Roy, an accomplished translator of literary works and screenplays from Manipuri to English, has added verbal art to his list of interests. For example, he is creating online resources on a cycle of songs sung in classical Manipuri Sankirtan style by an all-women's choir called the Jalakeli performance. His work fits well with Woodbury's definition of documentation: videos of performance of the songs of the Jalakeli; interviews with the singers of the Jalakeli; and documents about the Jalakeli tradition. At Guwahati University, we find under the direction of Professor Jyotiprakash Tamuli, a growing cadre of students working with both Indian and non-Indian linguists to produce substantial documentation such as Krishna Boro's descriptive grammar of Hakhun Tangsa and Prafulla Basumatary work on Boro grammar, both with accompanying audio and video documentation. There are many others inspired by the atmosphere and training provided at Guwahati University, who work with their communities to document their language.

Two high-profile ventures to "document" language and culture in India have brought India's linguistic diversity to the public's attention. Ganesh Devy, once a professor of English and now self-taught linguist, is the conceptual and practical lead for the People's Linguistic Survey of India (PLSI). Devy, organizing under a nonprofit called the Bhasha Research and Publication Centre, utilized the effort of 3500 volunteers including native speakers or speakers of related languages, linguists, and historians to gather language information state-by-state. The group has identified 780 languages and plans to publish collections of descriptions to cover all the languages they have identified. Since there is no accompanying audio and video recording along with the publications, the activities for PLSI do not fit the definition of language documentation laid out in Himmelmann's seminal article to which this volume is dedicated. However, we can recognize PSLI for starting a popular conversation in India on linguistic diversity, prompting the former Indian Prime Minister Manmohan Singh, to state in a public lecture that India is ignoring its linguistic diversity and that more should be done to bring minority languages to the digital age (Banerjee 2017).

An International nongovernmental organization, the Living Tongues, has also been active, creating dictionaries with audio accompaniment on Munda and other languages

of Arunachal Pradesh.[7] Living Tongues provides workshops to train speakers on how to add to existing lists of words so that both audio and transcription can be crowd sourced. The long-term effects of this training are yet to be seen - it is hopeful that these interventions will create local documentary linguists and stimulate language activism that will result in long lasting useful language documentation. I should also mention that there are many individuals from universities worldwide that work on documentation projects in India. The major traffic for PhD documentation work is from Singapore (Nanyang Technological Institute); Thailand (Payap); Australia (Melbourne and Sydney); the UK (the School of African and Oriental Studies); Switzerland (University of Zurich); Germany (Cologne and Max Planck, Nijmegen); and many universities in the United States. Stephen Morey (Melbourne) and Mark Post (Sydney) have provided training for local linguists on a consistent basis for several years and the quality of curation practices has greatly improved due to their efforts. Missionary groups from the United States, especially those related Summer Institute of Linguistics are in many parts of India doing literacy work which often includes dictionary creation and orthography development.

4. Building on strong beginnings We can see that there are enormous resources available to engender lasting and valuable language documentation in India. In my view, all that is needed now to spring forward from traditional data gathering for language description to data gathering for language documentation is a culture that rewards archiving of primary source materials. We need to shift the focus from training that results in quick-fix grammatical sketches to training that results in rich documentation that we can later harvest for description for pedagogy, revitalization, or science. That is, we need to focus on audio, and video data collection, metadata creation, data management, and data curation. A resource for training, discussion, and sharing on language documentation, such as the US Institute on Collaborative Language Research[8] or Summer School in Language Documentation and Linguistic Diversity[9] specially tailored for India would be hugely supportive of this growth.

Another need are local repositories and national or at least regional archives for digital language data. I am inspired by the many native speaker-linguists and language activists who work, in spite of all manner of obstacles, to write, read, publish, teach, proclaim, and celebrate their languages and cultures. At the University of North Texas (UNT), I am working with colleagues to create a repository for annotated corpora of South Asian languages which we are calling the Computational Resource for South Asian Languages (CoRSAL). With CoRSAL, my hope is to establish the value of annotated corpora (with source audio and video) for multiple fields of study and for multiple stakeholders and to raise the prestige of creating such corpora. We plan for the CoRSAL repository to house data in formats that can be easily accessed and used by indigenous groups, linguists, and other researchers for a range of purposes, including language science, computational linguistics, language reclamation and revitalization, language teaching, and investigations into diverse cultures and histories. We banking on the idea that "If we build it, they will come"! The CoRSAL concept has already spun off into interesting subprojects led by my UNT colleagues: metadata improvement for language archiving by Oksana Zavalina; user-centered design of language archives by Christina Wasson;

[7] https://livingtongues.org/india2015/

[8] https://en.wikipedia.org/wiki/CoLang

[9] http://www.bu.edu/applied-linguistics/2014/01/23/summer-schools-international-summer-school-in-language-documentation-and-linguistic-diversity-stockholm-sweden/

information seeking behaviors of indigenous populations by Mary Burke; and shared data formats for cross corpora comparison by Alexis Palmer, Manish Srivastava (from the International Institute of Information Technology – Hyderabad), and me. In addition, many of my colleagues working in Northeast India, senior and junior, are partnering in the project by contributing existing corpora from languages spanning from Arunachal Pradesh to Tripura. We have partners in Assam with whom we are working to create complimentary archives for audio and video and accompanying annotated materials. We hope to find funding not only for the CoRSAL concept but for grants to train and support documentation.

True collaborative language documentation is still a rarity in India. But my belief is that progress in language science and documentary linguistics is going to come from communities seeking support from linguists to create resources for revitalization. These agents of change will take us out of the doldrums and blow us into a storm of linguistic discovery. I hope to be right in the middle of that storm.

References

Abbi, Anvita. 2004. Vanishing diversity and submerging identities: An Indian case. Paper presented at the *conference of Dialogue on Language Diversity, Sustainability and Peace*, 20–23, May 2004. Barcelona, Spain.

Abbi, Anvita. 2001. *A manual of linguistic fieldwork and Indian language structures*. Munich: Lincom Europa.

Annamalai, E. 2003. The opportunity and challenge of language documentation in India. In Peter K. Austin (ed.), *Language Documentation and Description Volume 1*, 159-167. London: School of Oriental African Studies.

Banerjee, Rumu. 2017. Manmohan Singh stresses on the need to tap potential of India's linguistic diversity. *The Times of India*, August 4 2017. (`http://timesofindia.indiatimes.com`)

Banthia, Jayant Kumar. 2002. Foreword. In S. P. Datta (ed.), *Linguistic survey of India special studies Orissa, p. 5.* Kolkata: Language Division, Office of the Registrar General.

Chelliah, Shobhana. 2005. Asserting nationhood through personal name choice: The case of the Meithei of Northeast India. *Anthropological Linguistics*, 47(2). 169–216.

Chelliah, Shobhana. 2001. The role of text collection and elicitation in linguistic fieldwork. In Paul Newman & Martha Ratliff (eds.), *Linguistic fieldwork*, 152–165. Cambridge: Cambridge University Press.

Chelliah, Shobhana. 2016. Responsive methodology: Perspectives on data gathering and language documentation in India. *Journal of South Asian Languages and Linguistics* 3(2). 176–196.

Grierson, George Abraham (ed.). 1903–1928. *Linguistic survey of India.* Calcutta: Office of the Superintendent of Government Printing.

King, Robert D. 1997. *Nehru and the language politics of India.* Delhi: Oxford University Press.

Scharf, Peter M. 2013. Linguistics in India. In Keith Allan (ed.), *The Oxford handbook of the history of linguistics*, 227–258. Oxford: Oxford University Press.

Schiffman, Harold. 1996. *Linguistic culture and language policy.* London: Routledge.

Srivatsa, Sharath S. 2012. New Linguistic Survey of India to begin in April next year. *The Hindu*, Updated: March 22, 2012 10:27. (`https://www.thehindu.com/`)

Woodbury, Anthony. 2011. Language documentation. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge handbook of endangered languages*, 159–186. Cambridge: Cambridge University Press.

Woodbury, Anthony C. 2014. Archives and audiences: Toward making endangered language documentations people can read, use, understand, and admire. In David Nathan & Peter K. Austin (eds.), *Language documentation and description, Volume 12: Special Issue on Language Documentation and Archiving*, 9–36. London: School of Oriental African Studies.

Shobhana Chelliah
shobhana.chelliah@unt.edu