Archiving Endangered Mundā Languages in a Digital Library

Satyabrata Acharya, Debarshi Kumar Sanyal, Jayeeta Mazumdar, Partha Pratim Das

National Digital Library of India

Indian Institute of Technology Kharagpur, Kharagpur – 721302, India

satya.linguistics@gmail.com, debarshisanyal@gmail.com, jamaz82@gmail.com, ppd@cse.iitkgp.ac.in

Abstract

In the age of globalization and cultural assimilation, the number of speakers in many indigenous languages is fast dwindling. According to the UNESCO Atlas of the World's Languages in Danger, several languages of the Mundā family spoken by indigenous people predominantly in the eastern part of India are under the threat of extinction. In this paper, we present a study of the linguistic features of the endangered Mundā languages. We then propose the idea of a digital archive to collect and preserve textual, audio, and video documentation of these languages. We also explore the role of advanced technologies like artificial intelligence in the design of the archive. We believe our efforts will lead to the preservation and revitalization of the endangered Mundā languages.

Keywords

Endangered Language, Language Documentation, Digital Library, Digital Archive, Mundā Language Family, Linguistic Diversity

Introduction

A lost language echoes a lost culture, and it reflects an invaluable knowledge lost. Once a language is likely to become extinct in the near future, is obviously in danger. Of the estimated 7,111 known living languages in the world today (Simons 2017), nearly half are in danger of extinction and are likely to disappear in this century (Wilford 2007). Significant numbers of endangered languages disappear instantaneously at the moment of death of the sole extant speakers. Indeed, dozens of distinctive languages currently have only one native speaker still living, and that person's death will mean the extinction of the particular language: It will no longer be spoken, or known, by anyone. Others are lost slowly in the bilingual cultures since native languages are overwhelmed by the leading languages (Woodbury 1993). According to the Atlas of the World's Languages in Danger, at least 43% of the total languages spoken in the world are endangered (Moseley, 2010). When no one speaks the language, it is said to have died or become extinct.

In the document ‘Language Vitality and Endangerment', United Nations Educational, Scientific and Cultural Organization (UNESCO) has specified six degrees of endangerment that ‘may be distinguished with regard to intergenerational transmission'. Followed by UNESCO's Atlas of the World's Languages in Danger (http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CLT/pdf/aboutEndangeredLanguages-WV-EN-1.pdf), languages have been classified under seven heads based on intergenerational language transmission: (1) safe, i.e., the language is spoken by all generations; (2) vulnerable, i.e., most children can

speak the language, but they speak it only in some places, (3) definitely endangered, i.e., children no longer learn it as a mother tongue at home), (4) severely endangered, i.e., parents do not converse among themselves or with their children in this language, but grandparents and older generations speak it), (5) critically endangered, i.e., grandparents and older people speak it partially and infrequently; younger generations do not speak it at all, and (6) extinct, i.e., no one speaks it.

A language dies when its speakers disappear or switch to other languages. Globalization and cultural assimilation have accelerated language deaths in recent times. In 2003, Daniel Abrams and Steven Strogatz introduced an insightful mathematical model to capture the dynamics of dying languages (Abrams & Strogatz 2003). They mathematically modeled the competition between languages and explained why languages die. Daniel Abrams and Steven Strogatz considered two languages and assumed that a language's attractiveness depends on its current numbers of speakers and its perceived status. The perceived status encodes the social or economic opportunities of its speakers. The model predicts that two languages cannot exist together in a stable way; one will push the other to extinction. The perceived status is found to be an important indicator of the fate of a language; if the status degrades, the chances of its extinction quickly increase. Abrams and Steven Strogatz observed wherever bilingual or multilingual societies coexist; there has been little mixing among the linguistic populations. Therefore, one way to protect a language from extinction is to raise its perceived status.

A large number of indigenous languages spoken in tribal-dominated regions of India belongs to Astro-Asiatic Mundā language family. UNESCO has identified a total of 12 Mundā languages as endangered (UNESCO 2011) including Mundari, Birhor, Kharia, Turi, Korwa, Koda, Korku, Juang, Gadaba, Sora, and Bonda. Despite a few of the Mundā languages, bilingualism is widespread. At the present break-neck speed of assimilation, most Mundā languages are going to be extinct to the end of this century (Driem 2007). A significant number of Mundā language communities are now under a massive demographic and socioeconomic encumbrance to assimilate linguistically to the local Indo-Aryan majority languages, e.g., Bangla, Hindi, and Oriya. Till date, many Mundā communities throughout India and Bangladesh are virtually forced to cope with a different language and culture losing their own origin and identities in order to survive.

In general, a three-step response strategy has been recommended to save an endangered language (Austin & Sallabank 2014): (1) language documentation, i.e., producing textual and audio-visual documents of the syntax, semantics, and oral traditions of the language; (2) language revitalization, i.e., increasing the number of active speakers in the language, and (3) language maintenance, i.e., providing support to the language so that it is protected from those who might reduce its speaker count.

Motivation

This paper explores how endangered Mundā languages can be archived in a digital library. It supports the tasks of *language documentation, language description and language revitalization* of an endangered language. The core objectives of this paper are the following:

I. Creating a digital repository for preserving multimedia collections of endangered Mundā languages and the culture of the communities.
II. Making a safe and long-term repository for the language documentation collections.
III. Making the collections available to researchers, communities, and the public through the digital library setup.
IV. Supporting users in discovering and accessing the documents and recordings by means of a single point of accesses.
V. Enabling the users to browse and access the collections through the online catalog of a digital library

Contribution

This paper presents a detailed review of the characteristics of the Mundā languages. It sketches the design of an archive for endangered Mundā languages. It also proposes to augment the archive with cyber applications based on advanced technologies like artificial intelligence. It is hoped that with these initiatives, it is possible to protect the languages with the following contributions:

- Revitalization and Maintenance
- Preserving information on cultural resources and language diversity for upcoming generations and researchers
- Introducing accountability of archiving an endangered language in a digital library

Roadmap: The next section gives a brief report of the degree of endangerment and linguistic descriptions of the languages. The related works of the present study are described in the third section. It is followed by a section describing the methodology to build the archive. The fifth section gives a broad picture of the elements and architecture of the archive. The sixth section explores how artificial intelligence-based techniques can contribute to the initiative of language documentation and revitalization. The seventh section concludes the paper.

Endangered Mundā Languages

Demographic Classification and Degree of Endangerment

Mundā languages belong to the Austroasiatic family, and these are largely distributed into southern and northern branches. It has been classified into various subdivisions as shown in Figure 1.

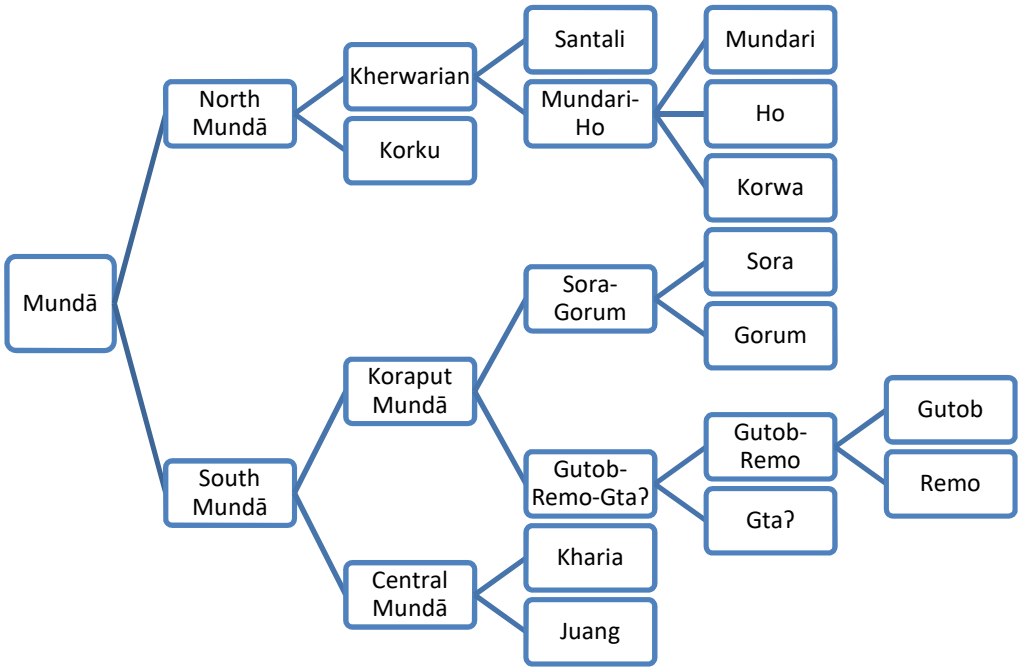

Figure 1: Classification Mundā Languages (Source: Diffloth, Gérard. 1974)

This paper is an attempt to find a systematic exposition of indigenous and endangered Mundā languages in India by crafting digital documentation of their linguistic description and cultural orientation. As per the UNESCO Atlas of the World's Languages in Danger (The Guardian 2011), eleven Mundā languages viz. ; Mundari[unr], Birhor [biy], Kharia [khr], Turi [trd], Korwa [kfp], Koda/Kora [cdz], Korku [kfq], Juang [jun], Gadaba/ Bodo Gadaba [gbj], Sora [srb], and Bonda/Remo [bfw] were considered as endangered languages subsequently with parametric degrees of endangerment. An estimated 5,000 speakers of the definitely endangered Turi language live in West Bengal, Madhya Pradesh and Odisha whereas an estimated 25,000 speakers of vulnerable Kodā language live in West Bengal, Madhya Pradesh and Bihar. Khāriā and Mundāri are both 'vulnerable' in their class of endangerment. An estimated 7,50,000 Mundari speakers live in Bihar, Odisha and West Bengal and 1,77,000 speakers of Khāriā live in Bihar, Odisha, West Bengal, and Madhya Pradesh.

The South Mundā language Juang covers almost 1,7,000 speakers in the Kyonjhar and Dhekānāl districts of Orissa, whereas Khaṛia dialects have over 1,90,000 speakers largely in Choṭā Nāgpur, Rāncī, and Orissa. The language Sora has approximately 2,50,000 speakers in Orissa and Andhra Pradesh. Remo language has only 2,500 speakers in the Jayapur hills of Korāpuṭ. The language known as Geta has approximately 3,000 speakers in Koraput, Malkangiri, Kudumulgumma, Chitrakonda, Khairput and on either side of the Sileru River in the East Godāvarī district. The Mundā language Korku has almost 200,000 speakers in southwestern Madhya Pradesh and adjoining parts of Maharashtra, particularly in the Satpuḍā range and Mahādev hills. Muṇḍari has approximately 7,50,000 speakers in Siṃhabhum, Manbhum, Hazaribag and Palamu districts. Estimated 1,50,000 Bhumij speakers of Mundari language still survive in Bihar, Orissa and Madhya Pradesh. The seminomadic Birhoṛ language is waning with below two thousand speakers in Siṃhabhum, Southern Palāmu, Southern Hazārībāg, and Northern and Northeastern Ranchi. Koḍa language is spoken by approximately 2,5,000 people in Choṭa Nagpur. Turi is spoken by an estimated 2000 people in West Bengal, Palāmu, Ranchi, Siṃhabhum, Raygaḍh, and Chattisgaḍh. Endangered Mundā languages and their demographic variations have been mentioned in Table 1 to depict an estimated scenario of the languages and their status of endangerment.

Table 1: A list of endangered Mundā languages and demographic variations

Languages with Language Codes	Number of Speakers	Degree of Endangerment	Location	Available Writing System	Glottolog
Mundari [unr]	750000	Vulnerable	Bihar, Odisha, West Bengal, Bangladesh, Nepal	Mundari Bani, Devanagari, Bengali–Assamese script, Oriya script	mund1320
Birhor [biy]	2000	Critically Endangered	Chhattisgarh, Odisha, West Bengal, and Maharashtra	It does not have a script and is performed orally	birh1242
Kharia [khr]	200000	Vulnerable	India (Jharkhand, Chhattisgarh, Odisha, West Bengal, Assam, Tripura, Andaman and Nicobar Islands), Nepal	Devanagari, Bengali script, Oriya scrip	khar1287
Turi [trd]	2000	Critically	Jharkhand, Madhya	It does not	turi1246

		Endangered	Pradesh, Odisha	have a script and is performed orally	
Korwa [kfp]	35000	Vulnerable	Madhya Pradesh, Bihar, Chhattisgarh (Surguja, Jashpur, parts of Raigarh district)	It does not have a script, and is performed orally	koda1256
Koda/Kora [cdz]	25000	Vulnerable	West Bengal (Bankura and Bardhaman districts), Odisha, Bihar, Bangladesh (Rajshahi Division)	Bengali (Bangla) script	koda1236
Korku [kfq]	200000	Vulnerable	Betul district, Hoshangabad and East Nimar in Madhya Pradesh, and Akola, Amravati, Buldana districts in Maharashtra.	Devanagari script	kork1243
Juang [jun]	17000	Vulnerable	North Angul, east Dhenkanal, south Keonjhar districts in Odisha	Oriya (Odia) script	juan1238
Gadaba/ Bodo Gadaba [gbj]	26262	Vulnerable	Telengana (Andhra Pradesh): Visakhapatnam district; Odisha: Koraput district, Lamtaput sub-district, 40 villages; Malkangiri district, Khoirput sub-district	Oriya (Odia) script	bodo1267
Sora [srb]	250000	Vulnerable	Andhra Pradesh, Odisha, Bihar, Madhya Pradesh, Tamil Nadu, and West Bengal	Oriya (Odia) scrip, Sora Sompeng script [Sora]. Telugu script [Telu]	sora1254
Bonda/Remo [bfw]	2500	Critically Endangered	Malkangiri, Khoirput, and Bondo Hills in Odisha	Oriya (Odia) script	bond1245
Geta [gaq]	3000	Severely endangered	East Godavari district in Andhra Pradesh; Koraput and Malkangiri districts in Odisha	It does not have a script and is performed orally	gata1239

In terms of the present degree of endangerment of such Mundā languages, it is necessary to undertake an organized work for proper documentation of these languages. Among the above mentioned existing endangered Mundā languages Birhor, Turi and Remo/Bonda have been identified as Critically Endangered. As per the Expanded Graded Intergenerational Disruption Scale (EGIDS) measured by the 'Ethnologue: Languages of the world' [Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). 2019.], Turi language has been shown in the language cloud in Figure 2 (Source: https://www.ethnologue.com/cloud/trd, Ethnologue 2019) as a critically endangered dying language and marked red.

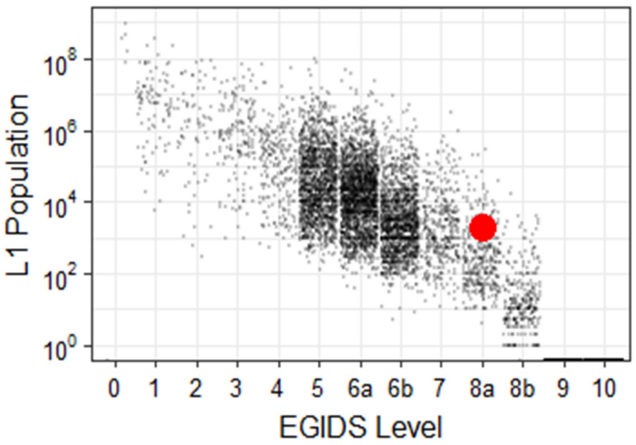

Figure 2: Turi in Language Cloud (Ethnologue 2019)

Shared Linguistic Features of Mundā languages

I. Phonology

1. Vowels

Mundā vowel systems are generally simpler than other Austroasiatic languages. It is typically a triangular system of five vowels, like the Indo-Aryan and Dravidian languages. In Table 2, we find a striking exception in Sora, whose three central vowels look very Southeast Asian.

Table 2: A set of shared Mundā cardinal vowels

High	i	ɨ	u
Mid-tense	e		o
Mid-lax	ɛ	ə	ɔ
Low		ɑ	

It is essential to reconstruct three central vowels for each Mundā subgroup: Sora-Gorum, Gutob-Remo , Kharia-Juang, Khewarian, and Korku-Kherwarian . In 1989 Diffloth gave evidence of creaky-voiced

vowels in Proto-Mon-Khemer. Vowel registers are in South Asia; if the correspondence can be resolved, this would be another Mon-Khemer-like feature of Mundā (Donegan & Stampe 2002).

2. Consonants

Followed by the place and manner of phonetic articulation, total 28 consonants, and 23 phonemes are commonly available in Mundā language family. As an example, we are listing the Mundari consonants in Table 3.

Table 3: A list of Mundari consonants

		Labial	**Dental**	**Retroflex**	**Palatal**	**Velar**	**Glottal**
Stop	**voiceless**	p	t̪	ʈ	t͡ɕ	k	ʔ
	aspirated	(pʰ)	(t̪ʰ)	(ʈʰ)	(t͡ɕʰ)	(kʰ)	
	voiced	b	d̪	ɖ	d͡ʑ	g	
Fricative			s̪				h
Nasal		m	n̪	ɳ	ɲ	ŋ	
Approximant		w	l	ɽ	j		
Trill			r				

Dravidian languages have influenced Mundā phonology in case of the acquisition of some retroflex consonants. In contrast to Indo-Aryan and Dravidian languages, Mundā languages typically have unreleased final consonants. Initial and final consonant clusters are not permitted in Mundā languages, and the occurrence of preglottalized consonants is quite distinctive. Final stops before vocalic suffixes in Mundā languages alternate with their voiced equivalents (Stampe Patrica 2002).

3. Tonality

Mundā languages are generally non-tonal, even though we find Korku syllables with a difference of tonality between high and low tone.

II. Morphology

1. Mundā morphology is much more complex and multifaceted than that of an average Austroasiatic language. It is fundamentally agglutinating. Furthermore, it employs reduplication and a variety of affixes (prefixes, infixes, and suffixes) to formulate nominal and verbal derivatives.

2. The most important characteristic feature of the agglutinating Mundā languages is the case marker, which is added after the object.

3. There are two gender classes, animate and inanimate in Mundā language; the first is divided into human and non-human. The grammatical numbers of Mundā have been distributed into singular, dual, and plural. It is striking the existence of inclusive/exclusive forms of the first person-plural-pronoun, i.e., there are two kinds of 'we', one includes the speaker, the other excludes him.

4. Verbs decide person, gender, and number with the subject by incorporating affixes or by adding them to the word that immediately precedes the verb.

5. A variety of suffixes indicates tense, aspects, and modality. As well for suffixes, structures with auxiliary verbs may be active to express tense. As like many other languages, the tense and aspect features are closely related, but their relative importance is different in the northern and southern languages: in the first one's aspect is prevalent, in the second ones tense.

5. There are different voices in Mundā: middle, passive, reflexive, reciprocal, and causative.

III. Syntax

In terms of syntactic patterns, Mundā syntax is quite distinctive from other Austroasiatic languages. Instead of Subject-Verb-Object (SVO), Mundā languages have a Subject-Object-Verb (SOV) rudimentary word-order. In this context, they are closer to Dravidian languages of India, though in contrast with them their order is quite strict.

IV. Lexicon and Vocabulary

Mundā lexicon has been inclined by adjacent Indo-Aryan languages which have had, however, little impact at the structural level. The opposite can be said of Dravidian languages. The unique linguistic unity of Mundā and Mon-Khmer has been refreshed, and it still breaks, mainly on lexical cognates (Bhattacharya 2000). The degree of similarity between Mundā languages is exposed in their shared lexicon.

Endangered Mundā Scripts and Writing System

There are only three scripts available which have been created specifically for writing Mundā languages; Sora Sompeng for the Sora language, Ol' Chiki for the Santali language, and Varang Kshiti for the Ho language. As per the degree of endangerment, the language Sora is now Vulnerable with 2,50,000 speakers (The Guardian, 2011) covering the states Andhra Pradesh, Assam state, Odisha, Bihar, Madhya Pradesh, Tamil Nadu, and West Bengal. The Sora Sompeng (Sorang Sompeng) script was created by Mangei Gomango in 1936 and was used in religious contexts (Everson Michael 2009). The Sora Sompeng (Sorang Sompeng) script shown in Figure 3 is quite distinctive with the following distinguishing characters:

1. The Sora language is written in an IPA-based script developed by Christian missionaries, and in the Telugu and Oriya scripts. There are twenty-four letters in the Sora Sompeng syllabary, named for the twenty-four deities in the Sora pantheon (Stephanie Holloway 2010). The eighteen consonant letters convey an inherent [ə] vowel ([ɔ] may or may not be written post-consonantally. Therefore, the characteristic vowel could be said to merge [ə] and [ɔ]). Unlike many of the South Asian syllabaries, there are no vowel diacritics. Vowels except the [ə] are written both initially and postconsonantally using six self-governing vowel characters.

2. Sora follows the Mundā pattern of using dental [t] and retroflex [ɖ], but not retroflex [ʈ] or dental [d] (which fill out the Brahmic pattern). Retroflex loan sounds (including [ʈ], [ʂ] and [ɳ]) are indicated by writing the one Sora Sompeng diacritic to the left of the closest equivalent letter. Dental [d] is not differentiated from retroflex [ɖ] in writing. Retroflex sound [ɽ] is also native to the Sora language.

3. Aspirate stops are also challenging for Sora Sompeng writing. Aspiration is not distinctive in native Sora, so is omitted in writing Sora words, but needs to be represented in writing a number of loan words from adjacent languages in which it is distinctive. The letter h cannot be used to indicate aspiration; it is used for representing a glottal stop. Nouns in Sora must have two syllables, and a glottal stop is often inserted halfway through the vowel in a mono-syllabic noun to split it into two syllables. Therefore, where aspiration needs to be written, it is written with the closest non-aspirate letter followed by the letter j.

4. It is thought that vowel length is generally not written. The exception to this is in cases where a long [a:] at the start of a word conveys some kind of grammatical information about the word, or in cases where it changes the stress pattern of the word. In these cases, the letter a is written twice.

5. Vowel-nasalization is quite unique in spoken Sora, but it is not clear whether this is represented in a written form.

6. Sora Sompeng has no script-specific punctuation. The Latin full stop, comma, semicolon, exclamation mark, mathematical symbols, and parentheses are used.

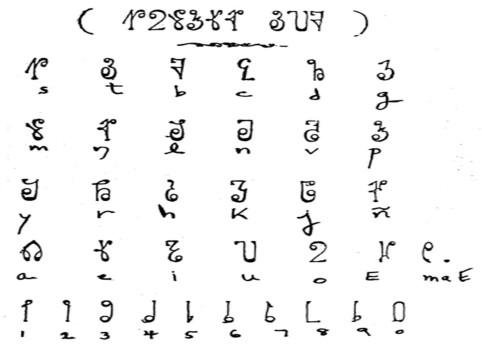

Figure 3: The Sora Sompeng Script (Source: Mahapatra 1978)

Related Works

1. DELMAN

The Digital Endangered Languages and Music Archives Network (DELMAN) presently works for documenting and archiving endangered languages and cultures worldwide (DELMAN 2003). It is an international network of archiving data on the linguistic and cultural diversity of endangered languages. DELAMAN is projected as an open organization for connecting any other organizations in the archiving and preservation of endangered languages and music.

2. ELAR

The Endangered Languages Archive (ELAR) of SOAS University of London is a digital repository preserving and publishing multimedia collections of endangered languages (https://www.soas.ac.uk/elar/). ELAR archive holds collections all through the entire world with regional grip in Africa, the Middle East, Asia, Australia, and Latin America. Till date, recordings encircling more than 450 distinctive languages can be found in ELAR. Collections in ELAR mainly contain audio and video recordings of language, verbal art, songs, narratives and performances of rituals.

3. Pangloss Collection

The Pangloss Collection, developed by the LACITO centre of CNRS in Paris, is a widely spread archive for endangered languages. As a member of the OLAC network of archival repositories, Pangloss Collection aims to store and enable access to audio recordings of endangered languages all over the world. The Pangloss Collection provides free online access to documents of unprompted spontaneous speech (Pangloss 2016). The Pangloss collection provides access to original recordings simultaneously along with transcriptions and translations (CoCoON 2017).

4. DOBES Portal

The DOBES Archive covers language documentation data of the languages in danger becoming extinct around the world. This portal gives access to the materials of the archive and provides information about the DOBES Endangered Languages Documentation programme (http://dobes.mpi.nl/dobesprogramme).

5. Living Tongues Initiatives

Living Tongues Institute began a multi-year project in 2005 to widely document the lexica and grammars of the modern Mundā language family (Jennifer 2008). The major output of the project was creating a set of talking dictionaries and online grammar for Ho, Remo, Gtaʔ and, Sora (Kari 2009). The initiative was led by linguist Dr. Gregory D. S. Anderson.

The works for Mundā languages initiated by Living Tongues Institute focused specifically the tasks of language documentation in a large-scale. Till date, the Living Tongues research team has not taken any initiative for digital archiving of the endangered Mundā languages as a single point of access.

Data Collection Methods

We propose to organize data collection under the focused areas of fundamental or basic research, field study, content analysis, and laboratory experiments. Here the fundamental or basic research will be concerned with a theoretical framework and underlying rules of linguistics and cultural study for a grammatical description of entire languages and anthropological survey of the communities.

A very large variety of data should be collected through field study. Based on the degree of endangerment, the informants of the survey may be divided into two categories: (1) grandparents and the older generation, (2) the parent generation and the younger generation.

Interviews should be designed in three ways: informal, structured, and focused. There will be a goal-oriented set of questions for structured and focused interviews. As supporting tools, the project will use a voice recorder for audio data collection and video recorder to capture the ambiance, field data, and overall expression. The field data will be transcribed and translated during the survey. Content analysis denotes the document analysis based on the abovementioned field study. It aims to fulfill the aims of the proposed research by extracting and analyzing the relevant grammatical and anthropological data from the proposed field study.

In laboratory experiments, several experiments can be carried out for linguistic and cultural documentation. Recorded pronunciation and phonetic notations may be documented through graphical representations to examine the numbers of phonemes, place/s of articulation, manner/s of articulation, and tonal qualities, among others. On the other side, the root words of the selected Mundā languages can be grouped into different sets. All members of any such set will exhibit the same morpheme-alterations during suffixation. Generalization and classification of root words into classes or paradigm sets will help in identifying the morphophonemic rules for each class of words. It is hoped that this will help to generate a comprehensive lexical and grammatical database of these endangered Mundā languages.

The data collected in a field-survey will be described with a set of standardized metadata. Moreover, the archive aims to take care of the long-term perseverance of digital materials.

The language documentation depositories in the archive will contain the following types of components:

- Audio and video recordings with different depths of annotation
- Transcriptions and translations together with morphosyntactic glossing (Simons 1998).
- Photographs and drawings bundled into groups of photos documenting processes
- Videos and music recordings of cultural activities, rituals, and social performances
- Documents on the genealogical affiliation of an endangered language accompanied by its socio-linguistic contexts, grammatical and phonetic features

Elements of the Archive

We envision a comprehensive archive of the selected Mundā languages. It might be designed as part of a more general digital educational library like the National Digital Library of India (https://www.ndl.gov.in/). It will have the following components.

I. Repository of textual and non-textual documents:

It will include the following:

(a) Written documentation: A documentation of written forms of oral Mundā languages and their culture will be archived. It includes different texts produced by the community as well as research documents that enlighten and inform others of their language, lifestyle, and culture. It is important to develop fonts and virtual keywords for these languages so that users can type seamlessly and produce written documents easily.

(b) Audio Documentation: It will be an aid in research for annotation, transcription, and translation of speech corpus. Moreover, it will help preserve the oral forms that encode folklores, songs, recitations, music, daily utterances, and rare verbal expressions of the Mundā community.

(c) Visual Documentation: Photographic documentation and video documentation of diverse aspects of languages and culture of these communities will be prepared and preserved. It will provide a graphic window into ways in which the languages are used by the communities.

II. Metadata elements

While preserving endangered languages in digital format Metadata takes a front-seat role to disseminate the records. The broad range of language documentation in both text and audio-visual forms needs to be properly annotated for the recall in digital space. There are various metadata standards and we have opted for Dublin Core metadata element set (Dublin Core Metadata Initiative, August 2007) to describe the digital resources, as presented in table 4.

Table 4: Metadata elements on archiving endangered languages

Metadata Registry Name	Value
dc.contributor	Any person or institution or agency is responsible for creating the work.
dc.creator.researcher	The researcher is responsible for the creation of the work.

dc.language.iso	Language in which the resource is written for text and the language in which the interview or video is shoot.
dc.coverage.temporal	Temporal period, period label, date, or date range
dc.description	Description of the item/work.
dc.subject.ddc	Classification of the item.
dc.title	Title of the work.
dc.title.alternative	Transliteration of the wok in regional language.
dc.publisher	Institution/Agency/Person who makes the work available in the public domain/market.
dc.publisher.date	Date of the work published.
dc.format.mimetype	The digital format of the work. www.e.g-pdf/epub/rdf/odc/html

III. Online talking dictionary/dictionaries of endangered languages

Bilingual and multilingual dictionaries of endangered and vulnerable languages help non-native speakers, and other people understand the meaning of words in these languages. Recently the Odisha Government has published bilingual dictionaries of several endangered languages spoken in the remote tribal areas of Odisha (Satyasundar, 2018). A talking dictionary of a language is an online interactive tool that allows users to listen to high-quality audio recordings of words and phrases in that language, and also to enrich the database with new uploads. Typically, a talking dictionary also contains meanings of the words in a mainstream language like English as well as descriptive images so that users can easily understand them. A talking dictionary of the Kera-Mundāri language is accessible at http://talkingdictionary.swarthmore.edu/kera_mundari/; it is developed by the Living Tongues Institute of Endangered Languages. We also plan to develop talking dictionaries of the endangered Mundā languages.

IV. Opening of socialization of traditional Mundā culture and languages

A common website will be developed and launched for public access to all the above digital content and associated applications. The website can serve as a one-stop access point to the digital archive of the languages. For example, the textual and multimedia documents and the dictionaries can be made accessible through the website. Additionally, users may be allowed to upload new content on the languages and cultures of these communities through the website into the archive. The interface should be very user-friendly to promote greater use of language resources.

We believe, along with the above digital tools, there must be initiatives to hold periodical workshops, seminars and training camps where the attendees will communicate in these languages and discuss the problems faced by the native communities. There is indeed a strong need for capacity building of the languages and communities in terms of culture and major communicative existence. These steps can help revitalize and maintain indigenous languages.

Artificial Intelligence in Language Archiving

Artificial intelligence (AI) has a very high potential in contributing to efforts in language preservation. We enumerate below some of the ways AI can help in archiving and revitalization of Mundā languages. The first of the three ways mentioned below is useful to the public directly while the remaining two ways are relevant to the engineers designing various tools for the archive.

I. Conversational chatbots: Due to the low number of speakers in indigenous languages, the scope of conversation in these languages is reduced, which further decreases the influence of these languages. One way to counter this is to construct AI-based chatbots that can talk to people in these languages. The advances in natural language processing have enabled machines to engage in meaningful dialog with humans in many mainstream languages like English. It is technically possible to train these chatbots in endangered indigenous languages and then have them talk with humans in these languages. But it needs a large corpus of conversation texts in the corresponding languages because training modern deep learning-based models require enormous data. These corpora can be collected in ways already discussed. The chatbots can use either a text-based conversation or voice-based conversation. In the latter case, voice recordings from indigenous speakers are necessary for machines to learn the accents and phonetic details. Following are two instances where successful AI-powered robots have been constructed to revitalize endangered languages (Constantin, 2019). Scientists at the ARC Centre of Excellence for the Dynamics of Language (CoEDL) based in Australia have built a social robot called Opie (http://www.opierobots.com/) that uses Google's AI platform to teach children heritage languages through games, stories, and lessons. The robot displays human-like responses (such as facial expressions) to children's reactions. It is trained with 40,000 hours worth of spoken material in six indigenous languages spoken in Australia and, five languages are spoken in Asia-Pacific. Reobot is an AI-enabled chatbot that understands te reo Māori, an indigenous language of New Zealand. It can reply to messages in both English and Māori.

II. Automated annotation, transcription, and translation: Language documentation projects typically require ethnolinguists and language experts to annotate texts (i.e., associate labels with text spans) and transcribe voice recordings collected from speakers of endangered languages. It also entails translating them to more mainstream languages, so they are understandable to a larger audience across the world. These tasks require enormous human effort spanning days and months of laborious work. Thanks to advances in AI, this task can be speeded up with the latest machine learning tools. Automatic speech recognition systems have been built for indigenous languages (Besacier et al., 2014). These techniques do not always generate very accurate results, but they do provide significant assistance in the documentation of endangered languages. Technology companies like Microsoft and Google in collaboration with universities and research centers, have produced translators for endangered languages. For example, Microsoft Translator (https://translator.microsoft.com/) supports Yucatec Maya and Querétaro Otomi, which are spoken only by a few thousand people in parts of Mexico (Charney, 2015). Similar translators should be designed for Mundā languages. Google and CoEDL have designed a pipeline to simplify the development of automatic speech recognition systems for languages that have a very low speaker base; this, in turn, aids the task of language documentation (Foley, 2018). Note, however, that most of these systems need a corpus of text or speech where the annotation/transcription/translation (as the case may be) has been manually done. This corpus is needed to train the machine learning algorithms. The trained model then works on new data.

III. Automatic data augmentation: Machine learning-based solutions to the revitalization of endangered languages often suffer from the lack of adequate labeled corpus, whether it is speech recognition or language translation or conversation or the like. These languages are sometimes called acutely low-resource languages as it is difficult to find high quality transcribed and labeled audio data and labeled text data for them. Deep learning-based approaches that have produced very high-quality speech recognizers and language translators for mainstream languages require extremely large training datasets. Therefore, it is unrealistic to expect these state-of-the-art techniques to be directly applicable to indigenous languages. Hence researchers have been motivated to devise various methods to augment the sparse datasets available

for indigenous languages. Methods like noise addition, pitch augmentation, and speed augmentation have been used to augment speech data and improve the accuracy of speech recognition systems for Seneca, an endangered indigenous language of North America (Jimerson, 2018a). Dictionaries and grammar rules have also been leveraged to generate new texts for enhancing the corpus (Jimerson, & Prud'hommeaux, 2018b). Another approach to increase available data is to use a generative adversarial network (GAN) which is a powerful class of machine learning systems (Kontzer, 2019). Given a training set, a GAN learns to generate new data with the same statistics as the training set. For example, a GAN that is trained on a given set of images can produce new images that are similar to the ones in the training set and might look authentic to human eyes. Similarly, given a collection of recorded speech, a GAN can generate new but similar data containing characteristics in the recorded version. These synthetic datasets can help develop more accurate AI-based tools for the preservation of threatened languages.

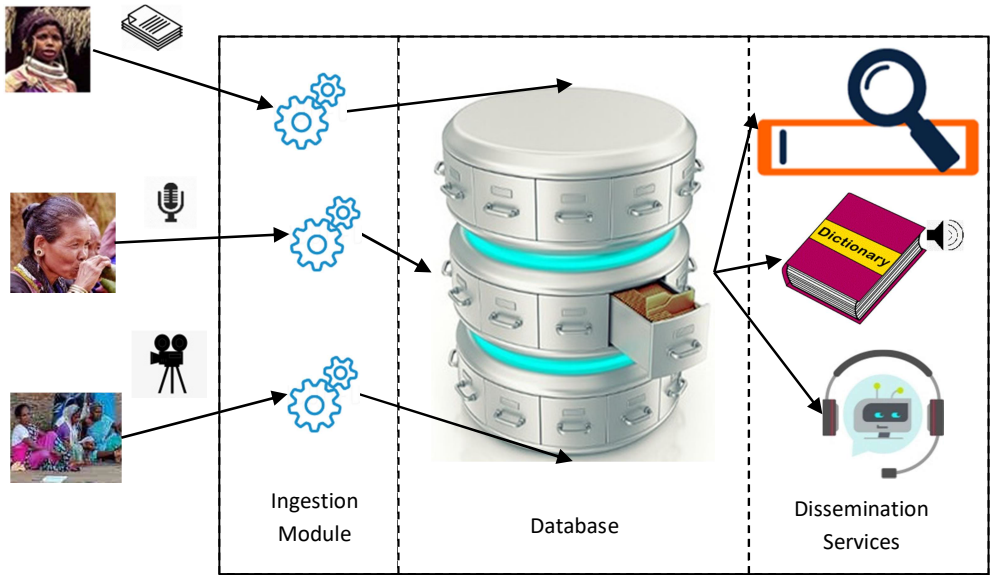

Figure 4: Architecture of the Mundā language archive

Figure 4 shows the architecture of the proposed language archive. Textual, audio, and video documentation collected through fieldwork and other methods are ingested using a variety of software tools. These records are stored in appropriate databases. It may be recommended to anonymize the data by removing personally identifiable information to protect user privacy. Each language resource is curated with appropriate metadata. Different dissemination services may be implemented on top of the database so users can easily discover and use the resources. The most important among them is a search engine that allows faceted search and browses the database. The search engine should support different filters to limit the displayed results and thus, cater more precisely to the user's information needs. Another interesting service could be a talking dictionary that, given a word in a Mundā language, can speak it out (thus, familiarizing the user with its pronunciation) and show the meaning and example uses of the word. Users may be interested to learn the indigenous languages by engaging in conversation with speakers of these languages. However, due to the low number of Mundā speakers, it may not be possible. A possible alternative is to design chatbots that can

understand and reply in a Mundā language. The chatbots may also support language translation, i.e., translate a text input in a mainstream language like English to a Mundā language, and vice-versa.

Conclusion

We have looked at the alarming state of some of the indigenous languages of the Eastern part of India. In order to protect them from extinction, a concerted effort is needed from social scientists, engineers as well as policymakers. We have proposed the idea of an archive where the nuances of these endangered languages will be captured, preserved, and available for others to study or better still, practice. Artificial intelligence tools will play a pivotal role in keeping them alive. We expect linguistic documentation, engaging virtual reality, and artificial intelligence-based tools, and a continuous capacity building exercise will revitalize the endangered Mundā languages and the centuries-old cultures that speak through them.

References

1. Abney, S., & Bird, S. (2010). The human language project: building a Universal Corpus of the world's languages. In Proceedings of the 48th annual meeting of the association for computational linguistics (pp. 88-97).
2. Daniel M. Abrams, Steven H. Strogatz. (2003). Modeling the dynamics of language death. https://www.math.uh.edu/~zpkilpat/teaching/math4309/project/nature03_abrams.pdf
3. Austin, P. K., & Sallabank, J. (Eds.). (2011). The Cambridge handbook of endangered languages. Cambridge University Press.
4. Barik, Satyasundar. (2017). Tribal communities in Odisha are speaking up to save their dialects. https://www.thehindu.com/news/national/other-states/tribal-communities-in-odisha-are-speaking-up-to-save-their-dialects/article18713925.ece
5. Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. Speech Communication (PP. 85-100).
6. Bhattacharya, Sudhibhushan. (2000). Kinship Terms in the Munda Languages. https://www.jstor.org/stable/40457389
7. Matthias, Brenzinger. Language Diversity Endangered. https://www.degruyter.com/viewbooktoc/product/36447
8. Charney, S. (2015, February 23). For language, technology is both the medium and the message. (Retrieved August 18, 2019). https://blogs.microsoft.com/blog/2015/02/23/language-technology-medium-message/.
9. Constantin, S. (2019, January 16). Can AI help save endangered languages? (Retrieved August 18, 2019). http://www.alignthoughts.com/can-ai-help-save-endangered-languages
10. Diffloth, Gérard. (1974). Austro-Asiatic Languages. Encyclopædia Britannica (pp. 480–484).
11. Diffloth, Gérard. (1989). Proto-Austroasiatic creaky voice. Mon-Khmer Studies (PP. 139-154).
12. Digital Endangered Languages and Musics Archives Network. http://www.delaman.org
13. DOBES Archive. http://dobes.mpi.nl/dobesprogramme
14. Drude, S. (2003). Language vitality and endangerment. International Expert Meeting on UNESCO Programme Safeguarding of Endangered Languages. (Retrieved August 18, 2019).

http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CLT/pdf/Language_vitality_and_endangerment_EN.pdf
15. Evans, Nicholas & Toshki Osada. (2005). Mundari and argumentation in word-class analysis (pp. 442–457).
16. Evans, Nicholas & Toshki Osada. (2005). Mundari: the myth of a language without word classes (pp.351–390).
17. Everson, Michael. (2009). Proposal for encoding the Sora Sompeng script in the UCS (pdf). http://std.dkuug.dk/jtc1/sc2/wg2/docs/n3647.pdf
18. Foley, B., Arnold, J. T., Coto-Solano, R., Durantin, G., Ellison, T. M., van Esch, D. & Olsson, O. (2018). Building Speech Recognition Systems for Language Documentation: The CoEDL Endangered Language Pipeline and Inference System (ELPIS). In Proceedings of the 6th International Workshop on Spoken language Technologies for Under-Resourced Languages (SLTU 2018) (pp.205-209).
http://www.unesco.org/culture/en/endangeredlanguages/atlas.
https://www.researchgate.net/publication/265199702_Endangered_Languages_of_South_Asia
19. Hughes, Jennifer V. (2008). 'Racing to Capture Vanishing Languages'. The New York Times. (Retrieved February 22, 2009).
20. Jane, Patrica & Stampe, David.South-East Asian Features in the Munda Languages: Evidence for the Analytic-to-Synthetic Drift of Munda.
http://www.ling.hawaii.edu/austroasiatic/AA/bls2002.pdf
21. Jimerson, R., & Prud'hommeaux, E. (2018). ASR for documenting acutely under-resourced indigenous languages. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018) (pp.4161-4166).
22. Jimerson, R., Simha, K., Ptucha, R. W., & Prudhommeaux, E. (2018). Improving ASR Output for Endangered Language Documentation. Proceedings of the 6th International Workshop on Spoken language Technologies for Under-Resourced Languages (SLTU 2018) (pp.187-191).
23. Kontzer, T. (2019, January 2). The oral of this story? AI can help keep rare language alive. (Retrieved August 18, 2019).
https://blogs.nvidia.com/blog/2019/01/02/deep-learning-preserves-seneca-language/
24. Living Tongues Project. https://livingtongues.org/
25. Lyderson, Kari. (2009). Preserving Languages Is About More Than Words. Washington Post. (Retrieved March 16, 2009).
26. Mahapatra, Kh. 1978-79. SoraN SompenN: A Sora Script.
http://sealang.net/sala/archives/pdf8/zide1999three.pdf
27. Moseley, Christopher (ed.). (2010). Atlas of the World's Languages in Danger (3rd ed.). Paris: UNESCO Publishing. (Retrieved August 18, 2019).
28. Mundari language. Wikipedia. https://en.wikipedia.org/wiki/Mundari_language
29. Newberry, J. (2000). North Munda dialects: Mundari, Santali, Bhumia. Victoria, B.C.: J. Newberry.
30. Pangloss Collection. https://lacito.vjf.cnrs.fr/pangloss/index_en.html
31. Patricia, Donegan & Stampe, David. (2002). Proceedings of the Twenty-Eighth Annual Meeting of the Berkeley Linguistics Society: Special Session on Tibeto-Burman and Southeast Asian Linguistics. (pp.111-120).
32. Peter K. Austin, Julia Sallabank. (2014). Endangered Languages: Beliefs and Ideologies in Language Documentation and Revitalisation; 1 edition (November 25, 2014), British Academy.

33. Peterson, John. (2015). Introduction – advances in the study of Munda languages. DOI: https://doi.org/10.1515/jsall-2015-0008
34. Satyasundar B. (2018, Nov. 24). Odisha now has a lexicon for rare tribal languages. The Hindu. (Retrieved August 18, 2019). https://www.thehindu.com/news/national/odisha-now-has-a-lexicon-for-rare-tribal-languages/article25588109.ece
35. Simons, Gary F. (1998). The nature of linguistic data and the requirements of a computing environment for linguistic research. In "Using Computers in Linguistics: a practical guide", John M. Lawler and Helen Aristar Dry (eds.). London and New York: Routledge, (pp.10-25).
36. The Dublin Core Metadata Initiative. https://www.dublincore.org/
37. The Endangered Languages Archive (ELAR). https://www.soas.ac.uk/elar/
38. The Guardian, Guardian News and Media Limited. Endangered languages: the full list. 2011. https://www.theguardian.com/news/datablog/2011/apr/15/language-extinct-endangered#data
39. The Language Gulper: An insatiable appetite for ancient and modern tongues. http://www.languagesgulper.com/eng/Munda.html
40. Turi in the Language Cloud. Languages of the World. (2019). https://www.ethnologue.com/cloud/trd
41. UNESCO Atlas of the World's Languages in Danger, United Nations Educational, Scientific and Cultural Organization. (2011). http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CLT/pdf/aboutEndangeredLanguages-WV-EN-1.pdf
42. Van Driem, George. (May 2007). Chapter 14: Endangered Languages of South Asia. Handbook of Endangered Languages, Mouton de Gruyter, Editors: Matthias Brenzinger. (pp..303-341).
43. Wilford, J. (September 19, 2007). Languages die: But not their last words. New York Times. www.nytimes.com/2007/09/19/science/19language.html?ex=1347854400&en=03c91ba69ddbdb61&ei=5090&partner=rssuserland&emc=rss
44. Woodbury, Anthony C. (1993). A defense of the proposition, "When a language dies, a culture dies". Proceedings of the First Annual Symposium about language and society—Austin (SALSA).
45. Zide, Norman. (1996) (Eds.). Scripts for Munda languages. Oxford: Oxford University Press.