Interdisciplinary Journal of Linguistics
Volume [15] 2022, pp. 139-150

THE DIGITIZATION OF AN ENDANGERED LANGUAGE USING OPEN-ACCESS RESOURCES: THE CHOLANAIKKAN CASE STUDY

Priyanka Pradeep Kumar*

Abstract

The paper introduces the working stages of the first-ever initiative in multimedia format for Cholanaikkan /IPA: cʰoːl̪ə̆n̪aɪkʰən/. The aim is to promote preliminary multilingual education based on the mother tongue for the language speakers of this under-resourced minority language. The paper will bring into perspective the language policies and socio-linguistic issues accelerated by the global pandemic in case of this endangered language spoken in a linguistically diverse country. The paper will further discuss in subsequent sections- the motivations, the processes, and the challenges involved in developing a digital dictionary for endangered languages using the methodological underpinnings of lexicography and theoretical understanding of sociolinguistics. Furthermore, the project hopes to explore the scope of language maintenance by expanding the domains of language use and contribute to the possibility of language revitalization efforts among the community members. In conclusion, drawing on the experiences from this dictionary compilation project, the paper will account for knowledge dissemination with regard to the key role played by open-access resources while documenting an endangered language during COVID pandemic.

Keywords: Cholanaikkan, Endangered Language, Indian Tribal Language, Digital Dictionary, Socio-linguistics, Lexicography

1. Introduction and Discussion

Around the world, the rate at which languages are dying or are forced into hibernation-"sleeping beauties[1]" (Zuckermann, G., 2020) is alarming. Though throughout our human history language loss and language shift has been viewed as an inevitable process, over the recent years there has been an increased recognition and growing awareness that more attention needs to be paid to this area. The latest report by Ethnologue[2] posits that as of the year 2022, 3,045 languages are endangered, which accounts for 42.5% of all known languages spoken across the globe. According to the UNESCO report by Moseley (2010), the number of languages identified to be endangered is 2,471. This includes 197 languages

* English and Foreign Languages, Hyderabad, India.

spoken in India which makes it the country with the greatest number of endangered languages of the world. While the numbers keep shifting over the years and though linguists have questioned the relevance of GIDS scale (Fishman, 1991), EGIDS scale (Lewis & Simon, 2010), UNESCO's parameters and Ethnologue's parameters of endangered languages in Asian context wherein the notion of oral languages co-existing with written languages is not taken into serious account, the fact remains that world's language diversity is at threat. As per the study by Bromham, L. et al. (2022), around 1500 known languages will not be spoken by the end of this century and the ongoing trend of language loss could triple within the next 40 years. These overwhelming statistics surrounding endangered languages and the lag in implementing appropriate measures, point towards the eventual loss of cultural identity, traditions and wisdom that a language speaker has the moral right to keep to survive in a community. "Certain concepts, beliefs, sensations can only be expressed in the language which they are coined, shaped, experienced; there are very significant meanings that are forever untranslatable" (Pattanayak, 2014, pp. 293). For academia and linguists especially, it would mean loss of data that could push and challenge theoretical advancements. Considering the above scenario, it can be asserted that language families are likely to vanish in the coming decades. Hence, endangered languages must receive the highest priority for directing preservation efforts and research through digital resources in the 21st century for the purpose of extending linguistic investigations as well as language maintenance. The present paper reports on the issues, reasons and stages of developing a mutilingual digital dictionary of an endangered language with the example of Cholanaikkan. The project involved the creation of a mobile application of Cholanaikkan dictionary with audio by using open access resources."Dictionaries of endangered languages are no longer merely static records that preserve language but are now being developed as dynamic, multi-functional tools for language maintenance and revitalization" (Ogilvie, S.2011. pp389).

1.1 Scheduled, Non-scheduled and Endangered Languages in India

India is a country with immense linguistic diversity while also being a distinct polity that has prevalent linguistic hierarchy with respect to: a). The varying socio-economic status of the speakers spread across different geographical regions, b). Equivocal degree of prestige factor attached to these languages by administration and native speakers and, c). The availability of linguistic resources (newspapers, textbooks, literature, etc.) and domains of language use (formal vs. informal). Based on the census report (2011)

released by the government, presently 1,369 mother tongues are spoken across the country. The census report does not account for all the languages spoken in the country owing to the decision post 1971 census by the government of India that languages with less than 10,000 speakers need not be included in the official language list of 8th schedule of the Indian constitution. As a consequence, the linguistic hierarchy is now more apparent than ever with the introduction of the broad classification of 22 languages (Assamese, Bengali, Bodo, Dogri, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Maithili, Malayalam, Manipuri, Marathi, Nepali, Odia, Punjabi, Sanskrit, Santali, Sindhi, Tamil, Telugu and Urdu) as the scheduled languages and rest of the languages and mother tongues spoken across the country as non-scheduled languages. Later, the census of India (2011) apart from mentioning 99 non-scheduled languages, reported an "Others" category combining the mother tongues of 1,875,542 speakers without disclosing the names of these languages or speaker-language ratio and thereby, further complicating the linguistic hierarchy.

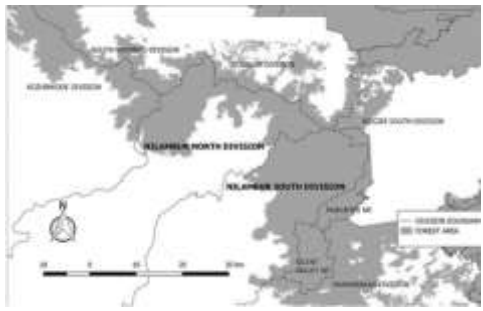
The scheduled languages are notably promoted through language policies in the education system and receive financial support from the administration. They are associated with high prestige in the country and learning them is perceived to bring better job opportunities and financial securities to its speakers. On the other hand, the non- scheduled languages are low-resourced languages due to lack of funding and the comparatively lesser prestige factor associated with them. Besides this, the speakers of non-scheduled mother tongue languages are mostly marginalized linguistic minorities (including indigenous/tribal communities), further pushing this disparity between scheduled and non-scheduled languages.

According to the survey by Kumar & Lahiri (2022), the case of endangered languages in India is not better than non-scheduled languages despite substantial discourse regarding language endangerment, having access to some governmental funding and setting up of national projects like Scheme for Preservation and Protection of Endangered Languages (SPPEL) and Schemes of University Grants Commission. The reason highlighted for this is the lack of availability of language resources and the study additionally encourages researchers to develop dictionaries and digital tools for language-learning to seize the pace of language endangerment situation in India.

1.2 Sociolinguistic Profile of Cholanaikkar

Cholanaikkar/Cholanaika forms an indigenous community residing deep inside the Karulai and Chungathara forest ranges of Western

Ghats. This is in the Nilambur region of the Malappuram district in the southern state of Kerala in the Indian subcontinent. The community prefers to stay isolated and away from both agrarian and urban lifestyle as is commonly adopted by the Indian citizens. They are among those unique tribes that still depend largely on forest produce for their survival. They were added into the list of Scheduled tribes of Kerala after amending the Scheduled Castes and Scheduled Tribes Orders Act (2002). This particular community is listed under PVTG[3] (Particularly Vulnerable Tribal Group) by the Indian administration. Also known as *"Caveman of Kerala"* (Bhanu, 1989), even to the present date, they use natural caves as a safe dwelling space in the forest. Cholanaikkar has been identified as a numerically smaller community with a population of less than 400 members[4]. The community speaks a distinct Dravidian language[5] called Cholanaikkan and as per the UNESCO's endangerment criteria, this lesser-studied language is identified by Sreelakshmi (2021) as "Definitively endangered". Further, she observes that lack of script, declining speakers and restricted domains of language use-Household conversations and in marketplace (as code language) has accelerated the threat of language loss for the Cholanaikkan.



Cholanaikkan settlement area

	Literacy (Overall) in %	Literacy ST in %
India	74	59
MALAPPURAM		
Kerala	94	67

Cholanaikkan is learnt as mother tongue by the younger generation of the community but the inter-generational transmission of this mother tongue is interrupted at the age of five when they are shifted to the Indira Gandhi Model Memorial residential school (IGMMRS), Nilambur. There they are not allowed to speak their mother tongue and have to comply with the monolingual education with the dominant regional language. The only language of

instruction used in this school is the vernacular as well as also a scheduled language- Malayalam. This education scenario has led to high school drop-out rate and low school enrolment among the Cholanaikkan students. The community has the lowest literacy rate when compared to other scheduled tribe communities of Kerala and also when compared to the overall literacy rate of the state. The proportion of children not attending schools is comparatively the highest in Cholanaikkan community with 36.84 percent[6]. A major reason for the dropout is that not even primary education is provided in the mother tongue and the students find it difficult to cope with language of instruction as the entire schooling period from primary to higher secondary is in Malayalam. Another reason is that the teachers are neither trained to teach in Cholanaikkan nor are there any textbooks or literature available in this language.

School drop-out rate	
2010-11	54 students
2011-12	110 students
2012-13	172 students
2013-14	113 students
2014-15	55 students
2015-16	39 students
2016-17	15 students
2017-18	86 students
2018-19	91 students
2019-20	71 students

Source: RTI Act (2021)

Indira Gandhi Model Memorial Residential School, Nilambur

2. Motivation for the Cholanaikkan Dictionary Application

The current language documentation project of Cholanaikkan began with an aim to address certain critical issues around this endangered language. Firstly, the language lacks an in-depth linguistic analysis; secondly, it is not yet featured in international linguistic information resources like Ethnologue/ Glottolog or even in the national linguistic resource like SPPEL and thirdly, there is no classification available for this language for historical comparative research in future. Finally, the community has always preferred living close to nature; thereby possessing a trove of traditional wisdom that they will express only in their language and loss of this language without proper documentation could mean that knowledge is lost forever. But as the project was progressing, the COVID pandemic set in and, the author gathered the data through RTI (Right to Information) to reach the realization from this reliable resource that the children of this community are suffering from pandemic-induced accelerated education gap.While the entire world shifted to online education; due to the particular geographical location of Cholanaikkan community and poor reception of telephone network and internet; these children were not able to attend classes through mobile phones. Neither could they watch educational videos through TV because their homes do not have electricity. A possibility for language documentation through fieldwork was not even imaginable for the author within the purview of the health-based vulnerability of this numerically smaller tribe during pandemic and the government's strict restrictions for outsiders visiting them. During this period, an earlier published open-access repository on this language was discovered and the author used this legacy work to draw on oral and written vocabulary to create a multilingual mobile application of Cholanaikkan dictionary. The dictionary contains audio files of Cholanaikkan vocabulary to facilitate easy understanding for speakers, educators and administration to easily understand and communicate. Furthermore, it also aids in developing learning tools and health instructions. And, most importantly unlike most of the mobile applications, this particular application developed by the author does not require internet connection to function (i.e) it can be used offline. The application was developed using open-access tools to keep it cost-effective, risk-free and at the

same time, to also make it possible for easy-corrections upon feedback from the community members.

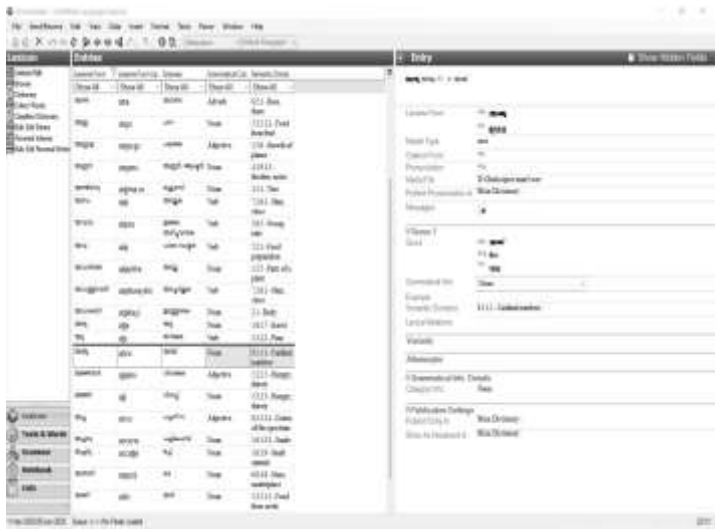
3.Methodology

Developing a digital dictionary application for endangered languages requires one to address certain unique set of challenges before launching the first version. The first step involved in this project was to determine the target users. The insights to make this choice were drawn from interviews with the teacher of IGMMRS School, Nilambur. In the case of Cholanaikkan mobile dictionary, it was decided that it should benefit the language community by fostering language use among the youth. Likewise, the academic community too must find it helpful to get a basic understanding of the language in order to promote literacy and bridge the educational gap in the community. Moreover, digital dictionaries can inculcate a sense of pride in the cultural identity within the members of Cholanaikkar community and can aid policy-makers to create educational resources and health bulletins in their mother tongue. Hence, it was an obvious choice to develop a digital dictionary that would benefit all three (speakers, educators and administrators) to ultimately address the issue of Cholanaikkan language endangerment.

The second step after determining the target users was to collect the basic vocabulary for creating the dictionary application. For this purpose, the author relied mainly on text-corpus method. In this case, the Cholanaikkan language repository available on DOBES[7] archive was utilized. The repository mostly consisted of continuous speech files of broad variety of genres elicited by an elder male member of Cholanaikkan community and ELAN[8] files created by the depositor for the speech files. The individual lexicons required for the Cholanaikkan Dictionary Application were extracted from ELAN files using the software Audacity[9]- a free and open source audioeditor.

The third step entailed the compilation of the dictionary using FLEx[10] software- an open-access tool developed by SIL. Each lexicon in this dictionary consists of two parts: the Headword and the Commentary to Headword. The Headword further included: a). The vernacular language- Malayalam. b). The oral language with a phonetic gloss. c). The audio file extracted from the repository. And, the Commentary to Headwordconsists: a) Trilingual gloss (Malayalam, English and Hindi). b). Grammatical information, and c). Semantic domains. A total of 250 words have been compiled in the current stage of Cholanaikkan dictionary.

Image of Cholanaikkan FLEx data



Headword and Commentary to headword

The final step to the project was to publish the dictionary in the format of a mobile application. Dictionary App Builder[11] (DAB) software, another open access tool developed by SIL was used for this purpose. It supports direct export of XHTML files from FLEx. It enables one to create a dictionary application that does not need internet to work and the dictionary application can be distributed offline through Bluetooth or SD card.

Screenshots of Cholanaikkan Dictionary App

4. Scope and Limitation

The UNESCO GEM Report (2016) states that about 40% of the global population does not receive education in their mother tongue. This impacts the learning capacities and academic performance of children especially those from linguistic minorities leading to high school drop-out rate as was observed in the case of Cholanaikkan students. Moreover, inclusive frameworks of language policy, multilingual education involving mother tongue instruction combined with digital tools for language learning is advocated as a critical way for revitalizing endangered languages. The present project of Cholanaikkan Dictionary Application posits a similar agenda to encourage mother tongue based multilingual education for the community amid high school drop-out rate and low enrolment rate. From being a major document of endangered languages, these days digital dictionaries also play a key role in language learning.

The limitation of this project is that this dictionary application does not account for lexical change in the language that could have happened through language contact with dominant languages over the years after the language was documented for the repository that forms a critical resource for the current version of Cholanaikkan Dictionary Application and also feedback from community members could not be included due to COVID related restrictions in the settlement area.

Acknowledgements

I am thankful to Mr. Harish (school teacher, IGMMRS) for giving the telephonic interview and patiently answering my questions. I am also thankful to research assistant Anjaly, who helped me with vocabulary compilation. I am also grateful for the continuous support of my supervisor for motivating me to find ways to implement the language documentation strategies around the unusual situation of the COVID pandemic.

Endnotes

1. Term coined by Prof. Zuckermann for languages with no native speakers
2. See https://www.ethnologue.com/
3. PVTG is a government scheme for the most vulnerable of scheduled tribes with declining population.
4. Inconsistent number of population: 124 (2011 census) and 363 (KIRTADS census, 2011).
5. According to Muralidharan, R., (1988). *A Descriptive study of Cholanaikka*. Ph.D. dissertation. Annamalai: Annamalai University
6. See Report on Socio economic status https://www.stdd.kerala.gov.in
7. See https://dobes.mpi.nl/
8. See https://archive.mpi.nl/tla/elan
9. See https://www.audacityteam.org/
10. See https://software.sil.org/fieldworks/

    See https://software.sil.org/dictionaryappbuilder/

Works Cited

Bromham, Lindell, et al. "Global Predictors of Language Endangerment and the Future of Linguistic Diversity." *Nature Ecology and Evolution*, vol. 6, no. 2, Dec. 2021, pp. 163-73, https://doi.org/10.1038/s41559-021-01604-y.

Census of India. Primary Census Abstracts, Registrar General of India, Ministry of Home Affairs. Government of India, 2011.

Fishman, Joshua A. *Reversing Language Shift: Theoretical and Empirical Foundations of Assistance to Threatened Language*. Multilingual Matters, 1991.

Grenoble, Lenore A. "Atlas of the World's Languages in Danger." *Journal of Anthropological Research*, vol. 67, no. 2, July 2011, pp. 294-295, https://doi.org/10.1086/jar.67.2.41303301.

GIDS. Revue Roumaine de Linguistique/Romanian Review of Linguistics.

United Nations Educational, Scientific and Cultural Organization (UNESCO). "Global Education Monitoring Report Summary: Education for People and Planet: Creating Sustainable Futures for All.", 2016.

Kumar, Ritesh, and Bornini Lahiri. "Language Resources and Technologies for Non-Scheduled and Endangered Indian Languages." *Conference on Sanskrit and Indian Languages: Technology, arXiv preprint arXiv:2204.02822, 2022.*

Lewis, Melvyn Paul, and Gary F. Simons. "Assessing Endangerment: Expanding Fishman's GIDS." *Revue Roumaine de Linguistique*, vol. 55, 2010.

Pattanayak, D. P. "Language and Cultural Diversity: The Writings of Debi Prasanna Pattanayak." *Indira Gandhi National Centre for the Arts and Orient Blackswan*, 2014, p. 293.

Ogilvie, S. "Linguistics, Lexicography, and the Revitalization of Endangered Languages." *International Journal of Lexicography*, vol. 24, no. 4, Aug. 2011, pp. 389-404, https://doi.org/10.1093/ijl/ecr019.

K.M, Sreelakshmi. "Assessing Language Vitality: A Case Study of Cholanaikkar Tribe in Kerala, India." *Language in India* vol. 21, 2021, http://www.languageinindia.com.

Zuckermann, Ghil'ad. *Revivalistics: From the Genesis of Israeli to Language Reclamation in Australia and Beyond*. Oxford UP, 2020.