Using the Wikimedia sphere for the revitalization of small and underrepresented languages in India

This report explores opportunities within the Wikimedia movement and projects to help revitalise small and underrepresented languages in India and provide recommendations to CIS's Access to Knowledge team in furthering this effort. The report is mainly based on a roundtable conversation on Digital Access in Bhubaneswar with a diverse range of backgrounds and professions, including independent researchers, representatives from non-profit organizations, retired government officials, Wikimedia contributors (both Odia and Santali), ecological activists, directors of research institutes, consultants, and journalists. This was organized by the Access to Knowledge team of CIS in collaboration with Vasundhara, Bhubaneswar.

This is a report by Subodh Kulkarni with editorial oversight and support by Tanveer Hasan and Soni Wadhwa.

This strategic note discusses a broad program idea of offering barrier-free open access to resources in various underrepresented languages in India.

Languages spoken in the Republic of India belong to several language families, the major ones being the Indo-Aryan languages spoken by 78.05% of Indians and the Dravidian languages spoken by 19.64% of Indians. Languages spoken by the remaining 2.31% of the population belong to the Austroasiatic, Sino–Tibetan, Tai–Kadai, and a few other minor language families and isolates. According to the People's Linguistic Survey of India, India has the second highest number of languages (780), after Papua New Guinea (840). Ethnologue lists a lower number of 456.

The UNESCO endangerment classification is as follows:

1. *Vulnerable*: most children speak the language, but it may be restricted to certain domains (e.g., home)
2. *Definitely endangered*: children no longer learn the language as a 'mother tongue' in the home
3. *Severely endangered*: language is spoken by grandparents and older generations; while the parent generation may understand it, they do not speak it to children or among themselves
4. *Critically endangered*: the youngest speakers are grandparents and older, and they speak the language partially and infrequently
5. *Extinct*: there are no speakers left

North-East India is home to more than 200 languages, out of which 82 are listed as *Vulnerable*, 63 as *Definitely Endangered*, 6 as *Severely Endangered*, 46 as *Critically Endangered* and 6 as *Extinct* (The Guardian Dataset). Arunachal Pradesh is the state with the highest number of languages, with as many as 66 languages spoken there, while West Bengal has the highest number of scripts, nine, and around 38 languages. The state of Odisha has 62 Scheduled Tribes who speak as many as 74 dialects. Their ethos, ideology, worldview, value- orientations and cultural heritage are rich and varied. Odisha has the unique distinction of having 93 different Scheduled Caste communities spread over 30 districts and 314 blocks of the state having different dialects. Apart from the languages of the North-East and the state of Odisha, there are several other languages all over India that deserve better representation on the Internet. While a handful of these languages enjoy status and visibility as official languages of the states and thereby hold some currency as widely spoken languages in their linguistic territories, there are many more languages that do not have speakers counting beyond a few hundred. Examples include the Bellari language (Spoken in Karnataka by 1000 speakers), the Toda language (Spoken in Tamil Nadu by 1600 speakers) and the Naiki language (Spoken in Maharashtra by 1500 speakers). What these languages do share in common with the languages of the North-East mentioned earlier is that they all lack free and open source knowledge and data.

Some of these languages are the official languages of the states and are widely spoken in this region. On the other hand, some of the languages have a few hundred native speakers. However, irrespective of the size of the native population or official status of the language, they all lack free and open source knowledge & data.

These languages show a range of marked cross-linguitic features which pose several interesting questions to Linguistic theories and speech processing research. Moreover, the close geographical proximity of these languages makes them vulnerable to changes in multiple linguistic levels, making these languages an excellent resource to study language change. Despite this, these languages severely lack digital preservation. One of the major reasons that contribute to the lack of resources is the difficulty in human access to some of the areas in these regions. Moreover, with English and Hindi being used as a lingua franca in these regions, the actual number of speakers proficient in their native language is much fewer than the number shown in the census reports. This makes it more important than ever to initiate a preservation process which does not primarily depend on fieldwork while also increasing the presence of the language in the digital sphere.

As language technologies advance and more sophisticated tools are built using Artificial Intelligence, the divide between low resource languages and others is likely to get even larger as a common prerequisite of these advanced systems is the existence of a large amount of digital data. Low resource languages are at a risk of being left behind.

Research on these languages by researchers are mostly conducted by collecting data personally, which causes a huge hindrance to the research process, as most of it remains as a private collection or published in closed journals. Moreover, data collection through fieldwork is particularly challenging in this region due to the restricted access to most of the disturbed areas.

The goal of this program is to facilitate the study of these languages by making existing resources discoverable and building open-source structured datasets and tools using the Wikimedia sphere to enrich the language research landscape of small and underrepresented Indian languages.

Role of CIS-A2K

- To design and commission relevant research studies in collaboration with language communities to define the premises of the program. The plan is to work with languages which are being written in single or multiple scripts in the pilot phase.
- To develop strategies regarding the integration of language datasets with Wikimedia projects
- Skill building of volunteers and community leaders in Wikimedia projects
- Structure of local knowledge to be compiled for contribution
- To identify the specific Wiki projects such as Wikipedia, Wikimedia Commons, Wikidata, Lingua Libre etc to build the archives of these languages
- Designing outreach and knowledge dissemination processes
- To develop partnerships with other academic, social, cultural and research institutions in the language sector for the sustainability of the project
- Material support - Sound recorders, microphones, hard discs, laptop, scanner, internet hardware
- Financial support - Remuneration of intern/fellow, internet data recharge

Specific objectives

a. Empowering the communities by enhancing digital literacy and connecting them with the world of knowledge and people outside.
b. Revitalizing/enriching the languages by increasing their use, coverage and depth using technological interventions.
c. Creating an ecosystem for developing language learning resources and tools; particularly, in the context of the New Education Policy.
d. Enabling scholars and researchers to overcome the challenge of finding appropriate data and expanding the knowledge on these languages.
e. By using the Wikimedia sphere, the infrastructural and technological support is secured, so that these languages are able to function in the digital world.

It is important to realise that these objectives can introduce new dynamics into other spheres of activity, such as education and the development of language.

Methodology

Our target languages broadly belong to two sets:
1) Languages which are primarily spoken in various states of India and have some or no digital presence on the internet.
2) Endangered languages which have extremely limited or no digital presence.

Survey of ongoing work

Several individuals and institutions are working on languages across the globe. There are significant initiatives in India also to revitalise the small languages in the digital sphere. Some of these are listed in the reference section at the end. An exhaustive survey of all such efforts will be done to map the present status as well as a listing of stakeholders. The target languages for A2K's future work and the potential collaborators will also be identified through these exercises.

Digital Dictionary Making

A dictionary is a vital resource for any language learning. The idea of collaborative dictionaries using platforms like Wiktionary or Wikidata Lexemes eliminates the need for expert lexicographers and terminologists and rather follows the method in which the users enter data as new entries, definitions, and so on, and the same is reviewed by editors, once published. An offline e-dictionary application using this dataset could be developed to overcome the problem of sparse internet connectivity where the user is only expected to download & install the application once and use the dictionary offline at any moment.

Data Acquisition Strategies

1. Leveraging Crowdsourcing using LinguaLibre for the creation of Speech CorporaGiven the scarcity of text and speech corpora for these low-resource languages, the main potential source for dataset creation is by crowdsourcing.

2. Using Optical Character Recognition techniques -
   The digitisation of texts in the public domain would be done and made available freely by uploading them on Wikimedia projects. The digital copy will be made machine-readable using Optical Character Recognition (OCR).

Processing the acquired data

a. Preprocess
b. Processing Speech Corpora
c. Processing Bilingual Parallel text Corpora

Housing datasets

a. Wiki Commons for media files
b. Wikidata for Lexemes
c. Wikisource for texts

Capacity Building workshops

- Promoting the language among the young speakers of the community, since they are the future of the language and if it survives, it will belong to them.
- Help language speakers possess up-to-date digital competencies and feel confident about them to actively participate in the digital world and increase content in their own native language.
- Promoting contributions on platforms like Storyweaver, Pratham Books, Eklavya etc.

Promote the upskilling of native speakers and other disseminators

- Facilitate knowledge exchange through participatory mechanisms both virtually and face-to-face.
- The potential communities would be introduced to Incubator for building new Wikimedia projects

Educational development

- Applying Open access philosophy to advance language pedagogy.
- Develop language learning resources and tools, particularly, in the context of the New Education Policy.

References

1. Wikipedia articles
2. SCSTRTI, Odisha - https://www.scstrti.in/index.php/resources/mle-initiative/bilingual-dictionaries
3. Most populous languages of Odisha - https://commons.wikimedia.org/wiki/File:Languages_of_Odisha.svg
4. People's Linguistic Survey of India - https://www.peopleslinguisticsurvey.org/
5. The state and fate of linguistic diversity and inclusion in the NLP world - https://aclanthology.org/2020.acl-main.560/
6. Bhasha India - https://www.microsoft.com/en-in/bhashaindia
7. Omniglot - https://www.omniglot.com/index.htm
8. Bharatavani - https://bharatavani.in/
9. Storyweaver - https://storyweaver.org.in/
10. Dimasa Thairili - https://www.dimasathairili.com/
11. SIL International - https://www.sil.org/
12. Ethnologue - https://www.ethnologue.com/
13. Global Recordings Network - https://globalrecordings.net/en/
14. Glottolog - https://glottolog.org/
15. Endangered Languages Project - https://endangeredlanguages.com/