Learnings from Technological Interventions in a Low Resource Language: A Case-Study on Gondi

Devansh Mehta [1*], Sebastin Santy[2*], Ramaravind K. Mothilal[2], Brij M.L. Srivastava[3], Alok Sharma[4], Anurag Shukla[5], Vishnu Prasad[1], Venkanna U.[5], Amit Sharma[2], Kalika Bali[2]
[1]Voicedeck Technologies, [2]Microsoft Research India, [3]INRIA France, [4]DN Developers, [5]IIIT Raipur

Abstract

The primary obstacle to developing technologies for low-resource languages is the lack of usable data. In this paper, we report the adaption and deployment of 4 technology-driven methods of data collection for Gondi, a low-resource vulnerable language spoken by 2.3 million tribal people in south and central India. In the process of data collection, we also help in its revival by expanding access to information in Gondi through the creation of linguistic resources that can be used by the community, such as a dictionary, children's stories, an app with Gondi content from multiple sources and an Interactive Voice Response (IVR) based mass awareness platform. At the end of these interventions, we collected a little less than 12,000 translated words and/or sentences and identified more than 650 community members whose help can be solicited for future translation efforts. The larger goal of the project is collecting enough data in Gondi to build a viable machine translation tool and speech to text interface that can help take the language onto the internet.

Keywords: Low-Resource Languages, Deployment, Applications

1. Introduction

Around 40% of all the languages in the world face the danger of extinction in the near future. Languages are not only a means of communication but also a carrier of tradition and cultures like verbal art, songs, narratives, rituals etc. When a language spoken in a particular community dies out, future generations lose a vital part of the culture that is necessary to completely understand it. This makes language a vulnerable aspect of cultural heritage and hence calls for their preservation. When it comes to saving such endangered languages, there are two aspects: Preservation and Revitalization (also referred to as 'revival linguistics') (Zuckermann, 2013). The former is concerned with how languages can be archived using different linguistic techniques so that it can serve as a lookup for future generations, while the latter focuses on ensuring that the language is resurrected into the daily fabric of people's lives. The biggest success story of a language getting revitalized is Hebrew (Fellman, 1973), which went from few native speakers to several million.

Initiatives like SOAS' Endangered Languages Documentation Programme (ELDP) [1] and The Language Conservancy (TLC) project [2] contribute mostly towards the documentation of endangered languages. However, language evolves with culture, and focusing solely on archival efforts misses out on how societies might have evolved differently had their language continued to be in use. In the present day and age of globalization and the integration of technology into almost every aspect of life, native speakers are turning to dominant languages at a faster rate than ever before to provide greater economic and social opportunities to future generations. Any revitalization effort undertaken today needs to include technological interventions that can comprehensively reverse this degradation. At a bare minimum, these languages need to be integrated with the Internet, which is becoming an ever dominant part of our lives, to ensure their survival and continued usage.

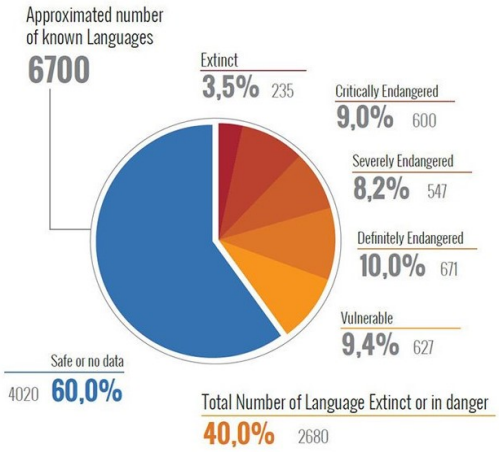

Figure 1: UNESCO 2017 World Atlas

We focus our efforts on Gondi, a South-Central Dravidian tribal language spoken by the Gond tribe in Central India. Gondi provides a unique outlook of how a language can be in danger even after having all the ingredients of a sustainable language like (1) long historical continuity (2) a population of 3 million people speaking it and, (3) is widely spoken in around 6 states of India with various dialects and forms. The complexities arise as Gondi is a predominantly spoken language with no single standard variety but a number of dialects, some mutually unintelligible. (Beine, 1994).

Deploying technology is a non-trivial task and there are legitimate concerns surrounding how language technologies should be implemented for low-resource languages (Joshi et al., 2019). It is difficult to simply transfer technologies prevalent in high resource language communities to minor-

* Equal Contribution
Contact email: kalikab@microsoft.com
[1]SOAS ELDP: https://www.eldp.net/
[2]TLC: https://www.languageconservancy.org/

ity communities for many reasons, the chief among them being the lack of data in low-resource languages. Our focus when working with the Gond community is thus centered more around devising novel approaches for data collection, unlike well-resourced languages where the focus is more on engineering. There also needs to be groundwork and identification of the real problems that can be solved by deployment of language technologies in minority communities. (Dearden and Tucker, 2015). Technical systems working in isolation from social contexts can be very dangerous to the ecosystem of these minority language communities and hence aren't ethically neutral exchanges of information. As far as possible, we have ensured that the technology interventions described in this study were led by the Gond community and non-profit organizations with a long history of working with the community. We have a deep appreciation of the fact that the problem we are trying to tackle is different from solving an engineering problem, since the expected outcomes have social rather than technical consequences.

In this paper, we deploy 4 technological interventions to help revitalize Gondi. The interventions are designed to achieve two objectives: (1) create a repository of linguistic resources in Gondi, with the intention of eventually using them to build language technologies like machine translation or speech to text systems that are essential for taking Gondi onto the internet; and (2) expand the information available to the Gond community in their language.

The first linguistic resource created is a Gondi dictionary that is accessible to the community as an Android app. The second is 230 children's books that were translated by the Gond community in a 10 day workshop. The third is an Android app that serves as a one-stop shop for Gondi content on the internet and also crowdsources translations from the community. The final intervention is a phone number that community members could call to gain awareness about a local election in their area, upon completion of which they earn mobile credits. These interventions respectively resulted in identifying 80-100 community members that participated in creating the 3,500 word dictionary; 20 community members that translated about 8000 sentences from Hindi to Gondi; 7 community members that translated 601 words; and 557 native speakers of Gondi that can be called for future workshops. The goal of the project is working with the Gond community members to collect enough data to build a machine translation tool and a speech to text interface that can help take Gondi onto the internet.

2. Context

According to the 2011 census (Chandramouli and General, 2011), the total population of the Gond tribe is approximately 11.3 million. However, the total Gondi speaking population is only around 2.7 million. That is, only about 25 percent of the entire tribe now speaks it as a first language. UNESCO's Atlas of the Worlds Languages in Danger (Moseley, 2010) lists Gondi as belonging to the vulnerable category. There is an added difficulty of creating resources for Gondi due to the linguistic heterogeneity within the Gonds. Spread over 7 states in India, Gondi is heavily influenced by the dominant language of each state to

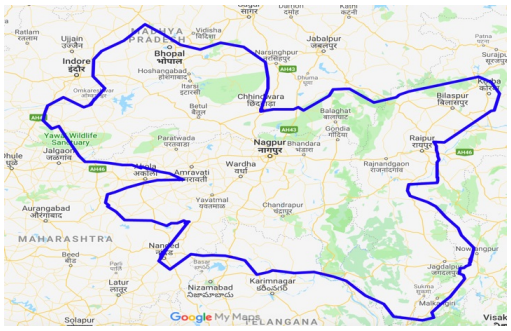

Figure 2: Gondi speaking areas in India

the point where a Gond Adivasi from Telangana(A Telugu speaking Southern state), finds it difficult to understand a Gond Adivasi from Chhattisgarh (a Central state with Hindi as the dominant language).

As a predominantly oral language, the proportion of Gondi speakers is expected to go down further as opportunities are shrinking to hear the language spoken outside of their everyday surroundings. For example, All India Radio, the only radio station in India allowed to broadcast news, does not have a single Gondi news bulletin in their broadcasts. There is no TV station or channel catering to Gondi speakers. There is also a severe dearth of online content in Gondi, resulting in members of the tribe having to learn a mainstream language in order to enjoy the benefits of internet connectivity. One of the few exceptions to this is The Wire [3], an Indian news outlet that has a news bulletin in Gondi that is published on Youtube. Views of their Gondi news broadcast range from 2,500 to 7,500, although it should be kept in mind that the Gondi spoken in the broadcast caters to the Gond Adivasis from the Central states of India and it is difficult for Gonds from the Southern states to understand the content.

Gondi is also not included in the 8th Schedule of the Indian Constitution, with the result that education and exams for government jobs cannot be administered in the language. The deleterious effects of this on their society are manifold. Gondi is considered the lingua franca of the local insurgents, who use the knowledge of the language and the perceived neglect by the government to recruit candidates from the tribe to join them (Kumar, 2019). Further, there are high dropout rates among children that speak Gondi as a first language. A 2008 UNESCO study found that children whose mother tongue is not the medium of instruction at primary school are more likely to fail in early grades or drop out, which in turn increases the chances of them joining the insurgency (Bühmann and Trudell, 2008). Working with the Gond tribes on reviving their language is thus important not just for cultural reasons, but may also serve as an instrument for bringing peace to their society.

3. Technological Intervention

There are relatively few cases that provide a holistic view of technology being used to help revive usage of an endangered or vulnerable language. The larger framework our interventions fall under is that language resource creation

[3]Gondi Bulletin: `https://youtu.be/M3q2ycJ_U7g`

feeds into building language technologies, which in turn enhances access to information in that language. To give an example of how this framework might operate, ensuring that Gondi is usable and accessible online greatly enhances the access to information. However, building the necessary language technology to take it online requires copious amounts of language data - by one estimate, a minimum of 100,000 translated sentences are required to build a machine translation tool (Koehn and Knowles, 2017), while at least 500 hours of transcribed and translated Gondi audio content is needed for a speech to text interface.(Huang et al., 2014)

Our efforts at collecting Gondi language resource creation and enhancing access to information in Gondi can be placed under broadly four buckets:

3.1. Gondi Dictionary Development

The Gondi speaking community is spread over 6 states in India. The dominant language in each state has crept into the dialect spoken by the Gond tribe residing there, with the result that it has become difficult for speakers from different parts to understand one another easily. In order to facilitate a democratized process of mutual intelligibility between the Gond groups, CGNet Swara organized seven workshops with Gondi speaking representatives from 6 states to develop a Gondi thesaurus containing all the different words used by speakers from various Gond regions. Some words, such as water, had as many as 8 different words for it. At the 8th workshop in 2018, which saw more than 80 people in attendance, the thesaurus was developed into a dictionary containing 3,500 words that were understandable by all groups, in essence 'purifying' the language of words that had entered their dialect due to the influence of the dominant state language. This dictionary not only enables translation of key words between Hindi and Gondi, but also understanding between the Gond tribes themselves. [4]

To enhance reach, the dictionary was made into an Android app, Gondi Manak Shabdkosh [5], that allows users to enter a Hindi or Gondi word and immediately hear or read its equivalent translation. This app is in many ways similar to the Ma! Iwaidja dictionary app [6], except that there is no wheel based interface for conjugation and sentence formation.

One of the use cases being explored with this app is in primary education, where tribal students and teachers are sometimes unable to communicate as the teacher does not know Gondi while the student does not know any other language. The hope is that the dictionary app will allow some basic communication and learning to take place.

3.2. Creating children's books in Gondi

One of the most effective ways of reviving a language is introducing it as a medium of instruction in schools, or at least as a subject. However, creating textbooks in that language is an important first step to achieving this goal.

Pratham Books is a non-profit publisher with a motto

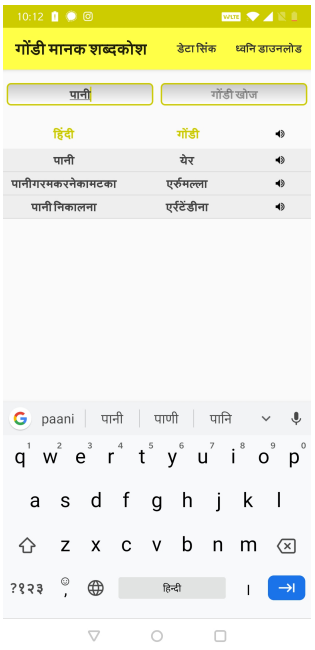

Figure 3: Gondi Manak Shabdkosh

"Book in every child's hand". Storyweaver[7] is an initative of Pratham Books that hosts more than 15,000 children's stories in various languages and dialects. A 10-day workshop was organized where 20 bilingual Gond tribals from three states of India came together and translated childrenâs books from Hindi to Gondi on the Storyweaver platform. These books were published on Storyweaver's website[8], resulting in the first ever online repository of childrenâs stories in Gondi. [9]

At the end of the workshop, 230 books and about 8000 sentences were translated from Hindi to Gondi. These stories, many of which introduced Gondi children to climate change for the first time, were printed out and distributed in primary schools in the Central Gondi speaking belt of Chhattisgarh. Efforts are now ongoing to convince the state government to include them as part of the school curriculum across the tribal belts of the state. [10]

3.3. Crowdsourcing Gondi translations

At the Pratham Books translation workshop, it was found that many participants wanted to continue the translation work from home, but there was no avenue for them do so. Taking inspiration from Aikuma(Bird et al., 2014), we developed Adivasi Radio, an Android application that presents users with Hindi words or sentences for which they need to provide the Gondi translation.

In the one month since its launch, there have been a total of 601 appearances translated through the app from 7 unique users. However, 5 of these users are either paid staff or volunteers hoping to become paid staff, indicating that soliciting translations from the community without monetary compensation may be a challenge. Moreover, the bulk of

[4] aka.ms/indian-express-gondi-dictionary
[5] aka.ms/gondi-dictionary
[6] ma-iwaidja-dictionary.soft112.com/
[7] storyweaver.org.in/
[8] aka.ms/storyweaver-gondi
[9] Article: aka.ms/hindu-gondi-nextgen
[10] Article: aka.ms/news18-climatechange

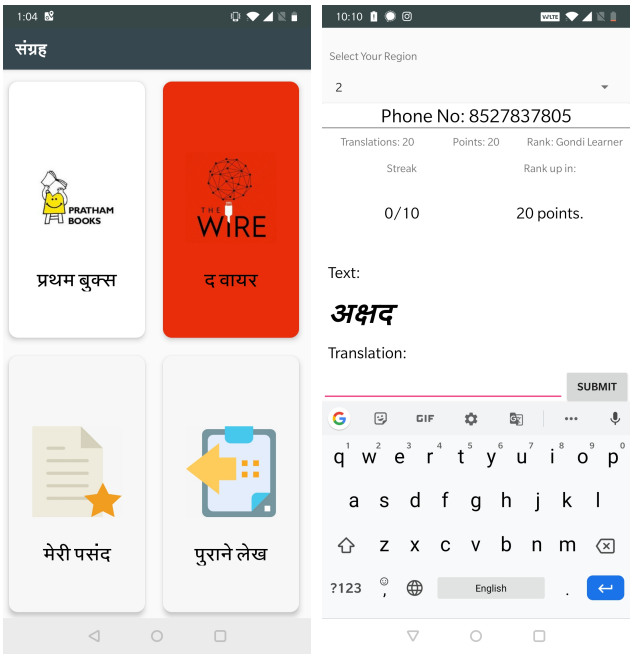

Figure 4: Adivasi Radio

the translations, 493, have come from one superuser. Future translation workshops will include a component on continuing the translation work through the app, with the hope that attendees continue translating even after they return to their village.

In addition to the translation role, Adivasi Radio also serves as a one-stop shop for Gondi content on the internet. The translated books from Pratham are accessible through the app, as are the Gondi stories from CGnet Swara. The information on The Wire's Gondi news bulletin are also available here.

To ensure these are comprehensible to populations that cannot read or write, a bootstrapped Devanagari text-to-speech system is integrated within the app to read out the written reports in Gondi. As more Gondi content proliferates online, we envisage Adivasi Radio as not only becoming the go-to place for native speakers to find outlets and sites publishing Gondi content, but also the primary medium for collecting the 100,000 sentences needed to build a machine translation tool between Gondi and other mainstream languages.

3.4. Disseminating Gondi content

Many recent studies on information dissemination in low-resource settings have relied on technologies, such as Interactive Voice Response (IVR), that could be used with internet-less mobile phones to connect to people and deliver information (Swaminathan et al., 2019; Chakraborty et al., 2019; Marathe et al., 2015; Moitra et al., 2016; Patel et al., 2010; Raza et al., 2018; Raza et al., 2019; Raza et al., 2013; Vashistha et al., 2015). As an example, Learn2Earn (Swaminathan et al., 2019) is an IVR based system that was used as an awareness campaign tool to spread farmers' land rights in rural India. Learn2Earn awards mobile talktime to users who call a toll-free number and listen to an awareness message, and answer all multiple-choice questions on the message correctly. Further, it has a peer-to-peer referral component where an additional recharge is provided for every new user successfully referred to the system. Learn2Earn was successful in spreading land rights awareness to 17,000 farmers in 45 days, starting from an initial set of just 17 farmers (Swaminathan et al., 2019).

Metrics	Value
Unique callers (total, during and after seeding)	(557, 480, 77)
Unique callers per day - during seeding (min, mean, median, max)	(13, 48, 49, 80)
Unique callers per day - after seeding (min, mean, median, max)	(0, 4, 3, 20)
Callers answering all questions correctly	313
Callers answering all questions correctly in their first call	104
Calls made by callers (min, mean, median, max)	(1, 3, 2, 64)

Table 1: **Summary statistics for our IVR-based deployment.**

In this intervention, we adapted the Learn2Earn technology to spread voter awareness among Gondi speakers in Dantewada (a rural district in the state of Chhattisgarh in India), during the time that a bypoll election was held. We chose to disseminate voter rights messages in Gondi, as the bypoll was crucial for Gondi speakers to establish representative and effective governance in their area. Further, prior work has shown that larger the contexts, identities and communicative functions associated with the language use, the more likely the language is expected to survive. (Walsh, 2005) We believe voter rights content in local language during elections has the potential to encourage conversations on topics of wider contexts and functions.

Learn2Earn was previously deployed in Hindi to spread awareness for voting awareness (Kommiya Mothilal et al., 2019). We extrapolated these experiments to Gondi to do that same in preparation for an upcoming bypoll in a Gondi speaking district. Similar to the original Learn2Earn implementation, the experiment started with a few 'seed' users and is intended to spread voluntarily through word of mouth or through peer-referrals thereon, both increasing the chances of language and content spread. Figure 5 shows the number of unique calls that were made to our system and the number of users who passed the awareness quiz over time. Further, the figure shows that a majority of quiz passers came to know about our system organically through word of mouth, though people knew that additional credits can be earned through referrals. Only a very few users continued to use the system even after the elections in September and after the seeding was stopped, probably to earn additional mobile talktime.

Table 1 describes important descriptive statistics of our experiment. We were able to spread voter awareness in Gondi to 557 speakers of which a majority (86%) were reached when we actively seeded. Nearly 60% of the users correctly answered all questions, either in their first attempt

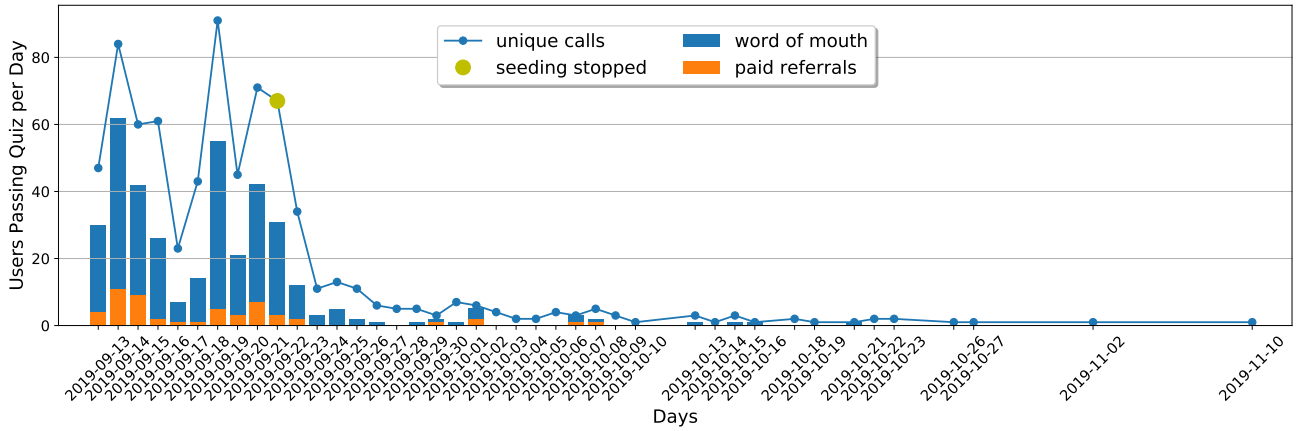

Figure 5: **Number of calls to the system (blue line) and the number of users who answered all questions correctly (bar plot). The blue and orange bars represent those users who came to know about the system through word of mouth and through referrals, respectively.**

or in successive attempts, indicating the effectiveness of the system for content dissemination. Further, many users called the system more than once, with one user calling 64 times (to refer someone, users had to place another call to the system). A follow-up survey revealed that 22 of 113 users surveyed could not vote as they did not have a voter ID card, potentially useful information for authorities to increase voter turnout. One disappointing finding was that only about 8.3 percent of users were women, indicating that a focus on outreach towards this community is needed.

We see three clear benefits from conducting Learn2Earn pilots in endangered or vulnerable languages. First, since it is entirely in the spoken form, only speakers of the endangered or vulnerable language can comprehend the content and earn the reward. Second, the phone numbers of the 557 users collected are an important dataset of speakers of that language and can be used for future translation workshops and related programs that help their economic and linguistic development. Finally, an oral language has a tendency to die out unless there is an opportunity for that language to be used outside of everyday surroundings, which periodic Learn2Earn campaigns on important issues help achieve.

4. Prior Work

4.1. Preservation Programs

The Endangered Languages Documentation Programme (ELDP) engages in documentation of endangered languages by creating a repository of resources for linguistics, the social sciences, and the language communities themselves. The Language Conservancy is a non-profit organization that strives towards revitalization of the world's endangered languages, restoring them to health and stability, and safeguarding them for future generations. They intend to implement this by introducing it as medium of instruction in schools by developing appropriate teaching methodologies and promoting and supporting the use of such languages beyond this setting - in homes and communities.

4.2. Language Revitalization Efforts and Successes

Hebrew is one of the languages which has shown the most promise in revitalization. Not being used since the second century, the usage of language in common conversations was kick-started by some Jewish communities in the 19^{th} century (Fellman, 1973). It has now risen to become the official language for the state of Israel and is spoken fluently by 7 million people.

However, such organic growth is almost impossible when the usage of language at this time is reinforced by many more factors. Ojibwe is an interesting example, where computer-based language learning technology was used and studies were conducted on how a particular multimedia tool might jumpstart communication in the Ojibwe language at home.(Hermes and King, 2013) Additionally, there were efforts where twitter was seeded with many Ojibwe tweets and hashtags to motivate individuals to converse in Ojibwe on twitter.

4.3. Deployed Applications and their Effects

Documentation and easy access to it is the first step in reviving a language. The Ma! Iwaidja is a mobile phone app which runs a dictionary of 1500 word and 450 English-Iwaidja translations along with audio for each. In addition, it provides a functionality to add words to the dictionary by the lay user. Due to their ubiquitous nature, crowdsourcing translations through apps are ideal due to their convenience and ease of use. Steven Bird's Aikuma app has leveraged its intuitive design to crowdsource 10 hours of audio or 100,000 words from indigenous communities in Brazil, Nepal and Papua New Guinea. (Bird et al., 2014)

5. Conclusion

While working with well-resourced languages, the main problem in designing language technologies is engineering. For low-resource languages, however, the main problem is one of designing methods for data collection upon which the language technology can be built.

To keep community members motivated through the process of collecting the large amounts of translations needed,

our team strived to achieve 2 simultaneous goals in each of the four technology interventions: the collection of data upon which language technologies such as speech to text or machine translation can be built, and expanding the access to information in that language, which the community members could point to as a demonstrable success. For example, the Learn2Earn pilot in Gondi not only provided Gond tribals with an opportunity to earn money for answering a quiz in their language and referring others to it, but it also provided a dataset of native Gondi speakers that can be called for future translation workshops. Similarly, translating children's stories and creating a standardized dictionary resulted in both data upon which machine translation tools can be built, and also tangible language resources that can be used by their community.

The larger goal of our research project is integrating Gondi with the internet, which requires at least 100,000 translated sentences. Our future efforts include an upcoming workshop to translate Wikipedia pages into Gondi and crowdsourcing more translations through the Adivasi Radio app. We plan to make all the data collected openly available and invite other researchers to participate in building language technologies that can benefit the Gond community.

6. Acknowledgements

We would like to thank Shubhranshu Choudhury from CGnet Swara, William Thies from Microsoft Research New England, Amna Singh from Pratham Books, and the director of IIIT Naya Raipur, Pradeep Kumar Sinha, for their invaluable assistance and advice. A word of thanks to all participants and workshop attendees from the Gond community, without whom this would not have been possible or meaningful.

7. Bibliographical References

Beine, D. K. (1994). *A sociolinguistic survey of the Gondi-speaking communities of Central India*. Ph.D. thesis, Univ. San Diego.

Bird, S., Hanke, F. R., Adams, O., and Lee, H. (2014). Aikuma: A mobile app for collaborative language documentation. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 1–5.

Bühmann, D. and Trudell, B. (2008). Mother tongue matters: Local language as a key to effective learning. *France: UNESCO*.

Chakraborty, D., Gupta, A., Team, G. V., and Seth, A. (2019). Experiences from a Mobile-based Behaviour Change Campaign on Maternal and Child Nutrition in Rural India. In *ICTD*.

Chandramouli, C. and General, R. (2011). Census of india 2011. *Provisional Population Totals. New Delhi: Government of India*.

Dearden, A. and Tucker, W. D. (2015). The ethical limits of bungee research in ictd. In *2015 IEEE International Symposium on Technology and Society (ISTAS)*, pages 1–6. IEEE.

Fellman, J. (1973). *The revival of a classical tongue: Eliezer Ben Yehuda and the modern Hebrew language*. Number 6. Walter de Gruyter.

Hermes, M. and King, K. A. (2013). Ojibwe language revitalization, multimedia technology, and family language learning. *Language Learning & Technology*, 17(1):125–144.

Huang, X., Baker, J., and Reddy, R. (2014). A historical perspective of speech recognition. *Commun. ACM*, 57(1):94–103.

Joshi, P., Barnes, C., Santy, S., Khanuja, S., Shah, Sanket, S. A., Bhattamishra, S., Sitaram, S., Choudhury, M., and Bali, K. (2019). Unsung challenges of building and deploying language technologies for low resource language communities. In *Proceedings of the 16th International Conference on Natural Language Processing (ICON-2019)*.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Kommiya Mothilal, R., Mehta, D., Sharma, A., Thies, W., and Sharma, A. (2019). Learnings from an ongoing deployment of an ivr-based platform for voter awareness. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, pages 257–261. ACM.

Kumar, S. R. (2019). Gondi language - identity, politics and struggle in india. *IOSR Journal*, 24(2):51–57.

Marathe, M., O'Neill, J., Pain, P., and Thies, W. (2015). Revisiting CGNet Swara and its Impact in Rural India. In *ICTD*.

Moitra, A., Das, V., Kumar, A., and Seth, A. (2016). Design Lessons from Creating a Mobile-based Community Media Platform in Rural India. In *ICTD*.

Moseley, C. (2010). *Atlas of the World's Languages in Danger*. Unesco.

Patel, N., Chittamuru, D., Jain, A., Dave, P., and Parikh, T. S. (2010). Avaaj otalo - a field study of an interactive voice forum for small farmers in rural india. In *CHI*.

Raza, A. A., Ul Haq, F., Tariq, Z., Pervaiz, M., Razaq, S., Saif, U., and Rosenfeld, R. (2013). Job opportunities through entertainment: Virally spread speech-based services for low-literate users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2803–2812. ACM.

Raza, A. A., Saleem, B., Randhawa, S., Tariq, Z., Athar, A., Saif, U., and Rosenfeld, R. (2018). Baang: a viral speech-based social platform for under-connected populations. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 643. ACM.

Raza, A. A., Tariq, Z., Randhawa, S., Saleem, B., Athar, A., Saif, U., and Rosenfeld, R. (2019). Voice-based quizzes for measuring knowledge retention in under-connected populations. In *CHI*.

Swaminathan, S., Medhi Thies, I., Mehta, D., Cutrell, E., Sharma, A., and Thies, W. (2019). Learn2earn: Using mobile airtime incentives to bolster public awareness campaigns. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):49.

Vashistha, A., Cutrell, E., Borriello, G., and Thies, W. (2015). Sangeet Swara: A Community-Moderated Voice Forum in Rural India. In *CHI*.

Walsh, M. (2005). Will indigenous languages survive? *Annu. Rev. Anthropol.*, 34:293–315.

Zuckermann, G. (2013). Historical and moral arguments for language reclamation. *History and Philosophy of the Language Sciences*, page 26.