Building a Common Voice Corpus for Laiholh (Hakha Chin)

Kelly Berkson[†], Samson Lotven[†], Peng Hlei Thang[†], Thomas Thawngza[†], Zai Sung[†], James C. Wamsley[†], Francis Tyers[†], Kenneth Van Bik[‡], Sandra Kübler[†], Donald Williamson[†], Matthew Anderson[†]

[†] Indiana University, [‡] California State University Fullerton

[†] {kberkson,slotven,phthang,tzthang,zhsung,jwamsley,ftyers,skuebler,williads,andersmw}@indiana.edu,

[‡] kvanbik@exchange.fullerton.edu

Abstract

In this paper, we discuss our efforts to build a corpus for Laiholh, also called Hakha Chin. Laiholh is spoken in Chin State in Western Myanmar, in parts of India and Bangladesh, and in several Burmese refugee communities in the US. Indiana, for example, is home to about 25,000 Burmese refugees. The ultimate goal of our team is to contribute to the development of speech translation technology that will be of benefit, both in general and in the local community in Indianapolis. Translation tools would be of great use in local emergency rooms, schools, and businesses. In pursuing our (admittedly lofty) goals, we are building a growing community of speakers, field linguists, computational linguists, and computer scientists. As a team, we have worked to share our different skill sets and mobilize the wider community around collecting data via Mozilla's Common Voice platform. We present here a reflection on the project thus far, the kind of description we wish had existed when we were first building this collaboration and determining preliminary project goals. We hope that other communities and language activists who are thinking about developing speech technology may benefit from hearing about our motivations, concerns, experiences, and successes.

1 Introduction

One of the largest incoming refugee groups in the US is from Myanmar (Burma). Since 2008, more than 109,000 Burmese refugees have settled in the United States (Centers for Disease Control and Prevention, 2016), and at present admission rates for Burmese refugees are second only to the Democratic Republic of the Congo (DRC). From October 2017 through September 2018, 3,555 people from Myanmar and 7,878 from the DRC have been admitted to the US. (Bureau of Population and Migration, 2018).

Many Burmese refugees in the US are members of the Chin ethnic group, and the number of Chin churches, businesses, and community organizations is on the rise. Indiana is home to organizations such as the Burmese American Community Institute (BACI) and the Chin Community of Indiana (CCI), which focus on community support, integration, and advocacy. The BACI has a specific goal of preparing Burmese American students for higher education, e.g. with summer college prep and research programs, and their efforts have had growing success. Many graduates of their summer research program—including three of the authors of this paper—have enrolled at Indiana University, Bloomington as undergraduates. These circumstances have led to many opportunities for interaction and collaboration between linguists and students, a situation which is certainly echoed at many other schools around the world.

Members of the Chin communities in the US speak 30 or more under- and un-documented languages from the Kuki-Chin branch of the Tibeto-Burman language family. One of these languages is Laiholh, also sometimes called Hakha Lai, Lai Chin, or Hakha Chin (Bedell, 2001; Matisoff, 2003; Peterson, 2016; Van Bik, 2006). Laiholh is used as a vehicular language in Chin State and is spoken by as many as 10,000 people in Indiana—including four of the authors of this paper—as either a first or second language (Executive Director of the Burmese American Community Institute, 2018). Many other languages (e.g. Falam, Lautu/Lutuv, Mara, Matu, Zophei) are also spoken, albeit often by fewer people. Laiholh is not endangered, and we believe that will remain true despite the disruptive nature of many of the factors contributing to the formation of the diaspora communities in Indiana and beyond. Our hope, however, is that we can make a methodological contribution to communities working with smaller and endangered languages by sharing details about the

Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 5–10,
Honolulu, Hawai'i, USA, February 26–27, 2019.

participatory and collaborative nature of our work. The work we describe here focuses on Laiholh, for purely strategic reasons: given the current composition of the community in Indianapolis, our hope is that Laiholh resources will have the greatest effect. We also hope to replicate our efforts with other Chin languages in Indiana, including endangered languages, in future work.

In this paper we describe one specific collaboration, but our central concern is not unique to us: we seek to build tools that will enhance communicative options. As noted by the United Nations High Commissioner for Refugees, the world refugee population is higher than it has ever been: someone is displaced every two seconds. Situations may develop quickly, and refugee communities around the world face communicative challenges. Existing literature focuses on communication challenges in, e.g., medical settings (Carroll et al., 2007; Morris et al., 2009) and primary and secondary schools (MacNevin; Naidoo, 2011). Language learning, even when proceeding well, takes time, and developing automatic solutions is even more time consuming. It is sometimes difficult to determine how we can best be of assistance in the face of pressing needs because the path to usable products is long. Herein, we outline one way in which student speakers, community members, linguists, and computer scientists can work together to begin to do so.

2 Community Needs

In developing our community of collaboration, one explicitly stated goal was to actively seek consensus on the projects we take on. We want to be of use, and the discovery and articulation of community needs has required reflection on the part of the native speakers and active listening on the part of the rest of the team. Discussions with team members, members of the wider Burmese community, and those who interact with them (e.g interpreters, speech pathologists) have allowed us to compile a list of the varied, real, and current needs of the local Burmese refugee community. Many of these needs revolve around language–a reality echoed in many other refugee communities worldwide.

Many challenges can be mediated with the help of a human interpreter, and the native speaker authors of this paper are often asked to interpret for family and community members. As such, they have firsthand knowledge of situations where translation is needed. Examples include:

- at the hospital and dentist (during check-in/check-out, before translators arrive)
- when paying bills (e.g., utilities) or interacting with insurance agents (e.g., car accidents)
- at state/government offices like the Bureau of Motor Vehicles (address changes, license plate renewal, ID card creation) or post office (for address changes, sending/reading mail)
- at car dealerships or local businesses (negotiation, sales)
- in interactions with the police (e.g. when pulled over), in court (pre-trial hearings), in jail (paying bail, calling a lawyer)
- at work (understanding/negotiating contracts/policies, talking to HR, requesting time off through FMLA, training)
- at school (parent-teacher interactions, administrative messages, meetings with speech pathologists)
- learning local regulations, e.g. Dept. of Natural Resources (hunting/fishing laws, licenses)
- with banks and credit card companies (understanding policies, paying bills)
- with the city government (e.g., to request building permits, pay parking fines)
- with US Citizenship and Immigration Services (citizenship paperwork, in-person interactions)
- for voting, voter registration, candidate info.

Professional interpreters can be employed to help in many situations—events such as court appearances require the human interpretation skills of a trained professional, for instance—but there are many situations where calling an interpreter is either not practical or not possible. Our student co-authors and others of their generation often interpret for community members, but they also attend school and work part-time so their availability is limited. If a community member cannot be accompanied by a bilingual interpreter to pursue a change of address at the Bureau of Motor Vehicles, a note saying "I need an address change form" may suffice. Such a solution fails as soon as a follow up question is posed, however, and in these mundane but crucial situations, technology could be of use.

There are also time-sensitive situations where technology could make a critical difference in the lives of the Burmese American community– in emergency rooms, for example, there is generally a lag between when patients check in and when interpreters arrive. In Indianapolis, experience suggests that often a Burmese translator (instead of a Laiholh translator) arrives, which is a problem because many Laiholh-speakers do not speak Burmese. Thus, while in some situations summoning the wrong translator might constitute only an annoyance, playing guessing games with interpreter language in an ER can be deadly. Simply put, the scale of the need (25,000 refugees in Indiana alone) requires what is currently an unreasonable amount of work for humans. As such, speech translation would be a boon—for many reasons, to many people, in our community and in others worldwide. It is against the backdrop of these realities that we, the authors, came together to form a collaborative team of speech community members, linguists, and computer scientists.

3 A Developing Collaboration

In December of 2017, a subset of the authors had a series of meetings to discuss collaboration concerning developing machine translation and automatic speech recognition (ASR) capabilities for under-resourced languages. We chose the development of a Common Voice system for Laiholh (see below) as the first step towards the larger aim. Since that time, our circle of collaborators has grown. The native speakers involved in this project went on to work as language assistants in a Field Methods class on Laiholh and on the Common Voice project described here. Several field methods students have continued to work on Laiholh. Input from computational linguists and computer scientists has informed the way field linguists collect, organize, and prepare data. Dialogue between all parties has been ongoing, and many of us work closely with one another on a near-daily basis.

Soon after our initial meetings, word of our interest in Chin languages spread and community members (both those in the Chin community and those who interact with them) began to contact us to describe challenges they encounter. Speech language pathologists from a local school district expressed a need both for basic materials on the languages spoken by their nearly 5,000 Burmese students and for help determining which language(s) children are acquiring at home. We talked to doctors who had little way to interact with Chin patients and difficulty providing them with written materials, and heard stories suggesting occasional patient discomfort when translators were involved in sensitive medical conversations. We met with the members of the Myanmar Students' Association on campus and learned that dozens of students were interested in working to develop language materials. As our list of needs and resources grew, it became clear that we needed a larger structure for data collection and a platform designed for inclusion, so that we could involve many eager parties. We needed to organize a group of individuals with different skill sets and different backgrounds around a common project, one with the potential to push us towards our growing list of goals.

We now provide a brief overview of Common Voice, noting why we believe its corpus-building structure and potential to lead to the development of voice recognition technology will help move us toward our larger goals.

4 Common Voice

Our long-term goal is to develop automatic speech translation for Laiholh, and Mozilla's Common Voice platform[1] offers two necessary outcomes that bring us closer to that goal. It facilitates both the creation of a public domain spoken corpus and the development of speech-to-text software. Speech data is collected via a phone or browser app from any native speaker willing to donate their voice to the corpus. Once the corpus is large enough, Mozilla will use machine learning software to develop speech-recognition technology This moves us closer to our larger goal of speech translation because once a written Laiholh sentence can be generated from spoken language, that written language can be used as input for text-based (Laiholh–English or English–Laiholh) machine translation technologies.

There are four specific ways that using Common Voice facilitated our work: (1) it provided us with a clear project structure and delineated, attainable goals; (2) it gave us an existing interface for data collection so we did not have to create one from scratch; (3) it stores the data collected, so we do not have to secure storage space for hundreds of hours of audio; and (4) it offers access to machine learning technology in the creation of a speech recognition system, which is a prerequisite for machine translation.

[1] https://voice.mozilla.org/

Before speech data collection can begin with Common Voice, two projects have to be completed. First, the Common Voice interface must be translated into the target language. For Laiholh, we completed this in Summer 2018. Next, 5,000 written sentences (in the target language) have to be collected. These written sentences are presented for users to read aloud, meaning that literacy is an important component of interacting with Common Voice. As such it may not be viable when working with languages that do not have widely-used orthographies.

Construction of a written corpus for Laiholh consisting of 5,200 sentences was completed in October 2018. The bulk of the work was completed by our native speaker undergraduate co-authors, who spent many hours a week in Summer 2018 thinking up sentences. As a larger group, we also sat together and thought through scenarios: "What do we need to say during parent-teacher conferences? What questions might a doctor ask at a check-up? What do we say when we're texting with friends?" Many sentences were also gleaned from an online Laiholh dictionary created by one of our co-authors, Kenneth Van Bik, and Laiholh author Joel Ling gave us permission to borrow sentences from some of his books. Finally, we also asked for input from other community members when translating some of the terminology in the interface to ensure that the decisions we were making on a day-to-day basis would result in an app that is user-friendly for everyone.

Common Voice offers a simple user interface where speech community members can provide two types of data: (1) Recording, where users are presented with a series of sentences which they are instructed to record; or (2) Validation, where users see a sentence, hear a recording from another user, and assess whether what they saw matches what they heard. Over time, as people submit and validate recordings, a large data set is collected and housed by Mozilla. The data set remains public domain and can be downloaded at any time for our own use, or by others. The data is structured as written sentences paired with multiple audio files and judgments of which audio files are accurate renderings of the provided sentences.

To ensure the development of robust technology, particularly given that Laiholh is spoken by a multilingual diaspora community, our goal is to record highly varied data that includes many accents, dialects, and voice qualities. This will ensure that we can train a robust speech recognition system. The only way to procure such a set of learning data is through widespread community use of the Common Voice app. This, in turn, means that we need to find ways to disseminate information about the project to as many people in the wider community, those beyond our smaller community of collaboration, as possible.

4.1 Preliminary Data

We began the speech data collection stage of the project in mid-November 2018. During the first three months of data collection—from November 14, 2018 to January 14, 2019—260 users have contributed by donating their voices and 310 users have contributed by validating sound clips. Altogether, 4,500 audio clips totalling 5 hours and 53 minutes of audio data have been submitted and 2 hours and 44 minutes have been validated. The average number of clips contributed per user is 17.4, and the top user contributed 243 clips.

To the best of our knowledge, this dataset constitutes the largest spoken corpus of Laiholh in existence. The Chin Cable Network Channel (CCN) made a video tutorial in Laiholh on how to use the app which was shared on CCN's Facebook page. We have also been invited to share brief presentations about the app during events at local churches.

While the total amount of data collected continues to grow day by day, growth of the corpus is clearly most robust when we are actively working to publicize it. During the first week that it was live, for example, more than 2 hours' worth of data were recorded. Community buy-in is hugely important in this work, and one request we received during the early weeks of data collection had to do with the style of the sentences included in the corpus. In particular, we were asked to add additional sentences that represented more informal domains such as texting and online chatting. In response to this request, we have pulled back from publicizing for the time being in order to devote time to increasing/diversifying the sentences in the corpus.

5 Developing Trust

In building a community based on different skill sets and different understandings, we have found two parts of the process where we have had to place trust outside of ourselves in order to pursue the project with Common Voice. First, we needed to trust that the Common Voice platform would function as advertised and that, if we dedicated time and resources to it, we would eventually get the desired output. Second, we needed

to trust that the wider Laiholh-speaking community would be able and willing to access Common Voice and record sentences. We turn now to describing why our developing community decided to place its trust in this project and why we are hopeful that we are on the right path.

We decided to pursue Common Voice because we were able to read online about other communities who were involved in the project. We read blog posts about various language groups hosting Common Voice "sprints" focused on collecting a lot of data in a short period of time, and we could envision doing that with our community. We read about the ethics and principles espoused by Common Voice.[2] We liked the emphasis placed on open access and public domain resources. One of our co-authors had inside knowledge of Common Voice, and talked with the team extensively—answering questions, and sharing information. Everything we heard was to our liking, and we felt comfortable moving forward.

With regards to the Laiholh speaking community and our questions about whether people would be interested in donating their time and voices to the project, we were able to put our concerns to rest very early on due to the enthusiasm we encountered. From college students to community leaders, we were met with excitement and interest everywhere we went. Organizations like the Chin Cable Network Channel and the Chin Youth Network of North America have helped us by creating video tutorials and advertising videos. A representative of Chin Baptist Churches USA (CBC-USA) offered to advertise the project to all of its 110 churches across the country with its 30,000 or so members. These positive reactions, in addition to the response when Common Voice in Laiholh went live, reassured us that with continued effort on our part we will continue to see robust community participation in data collection.

6 Conclusion

Seeking to develop speech recognition and machine translation technology is a sizeable goal that involves many steps. To accomplish this goal, we will need to use all of our diverse strengths and skills, and we will need to develop a common "language" that will allow researchers from computational linguistics, computer science, linguistics, and from the language community to develop our capacity to collaborate and to trust in one another. One challenge that we continue to work to confront has to do with communicating complex facts to non-experts: the language experts on our team have knowledge about Laiholh that can be difficult to convey but crucial for the computer scientists to understand. Similarly, the long and complicated path that we hope will result in working speech technology for Laiholh—and the way in which specific components of the project like building the Common Voice corpus are related to that larger goal—is clearer to the computer scientists and computational linguists than to the other members of our team, or indeed to other members of the larger community. To maintain energy and enthusiasm, however, it is crucial for the technical experts to find ways to make the steps more transparent. Working to ensure that all members of the team are engaged and empowered is an ongoing goal, one that has been well-served by coming together around the shared Common Voice project.

Our goals are lofty, but the payoff if we succeed will also be very high as it will dramatically improve the lives of local, national, and international community members. The hope, too, is that successful collaboration will increase our knowledge about how speech recognition and machine translation can work for other languages with few computational resources but with a strong community buy-in.

[2]For example, see the following blog post: https://medium.com/mozilla-open-innovation/more-common-voices-24a80c879944

Acknowledgments

This project was supported in part by the Indiana University College of Arts and Sciences Ostrom Grants Program and by the IU Department of Linguistics.

References

George Bedell. 2001. The syntax of deixis in Lai. *Linguistics of the Tibeto-Burman Area*, 24(2):157–171.

Elaisa Vahnie. Executive Director of the Burmese American Community Institute. 2018. Private Communication.

Jennifer Carroll, Ronald Epstein, Kevin Fiscella, Teresa Gipson, Ellen Volpe, and Pascal Jean-Pierre. 2007. Caring for Somali women: Implications for clinician–patient communication. *Patient Education and Counseling*, 66(3):337–345.

Centers for Disease Control and Prevention. 2016. Burmese refugee health profile. Technical report, U.S. Department of Health and Human Services.

Joanne MacNevin. Feeling our way in the dark: Educational directions for students from refugee backgrounds. Master's thesis, University of Prince Edward Island.

James A Matisoff. 2003. *Handbook of Proto-Tibeto-Burman: System and Philosophy of Sino-Tibetan Reconstruction*. University of California Press.

Meghan D Morris, Steve T Popper, Timothy C Rodwell, Stephanie K Brodine, and Kimberly C Brouwer. 2009. Healthcare barriers of refugees post-resettlement. *Journal of Community Health*, 34(6):529.

Loshini Naidoo. 2011. What works? A program of best practice for supporting the literacy needs of refugee high school students. *Literacy Learning: The Middle Years*, 19(1):29–38.

David A Peterson. 2016. Hakha Lai. *The Sino-Tibetan Languages*, page 258.

Refugees Bureau of Population and Migration. 2018. Refugee arrivals by placement state and nationality. Technical report, U.S. Department of State.

United Nations High Commissioner for Refugees. 2018. Refugee statistics. Technical report, United Nations Refugee Agency.

Kenneth Van Bik. 2006. *Proto-Kuki-Chin*. Ph.D. thesis, University of California, Berkeley.