Creating the KokBorok Language Resource at CoRSAL
Project Description

We propose to curate and archive materials on Kokborok (trp) a language spoken in Tripura State, India by about 101, 294 speakers (Census of India, 2011). Kokborok is one of three closely related members of the Bodo-Garo subgroup, which consists of Dimasa, Bodo, and Kokborok. It is the least documented of the three. We can use information from Dimasa and Boro to make better documentation and pedagogical materials of Kokborok quickly. Currently Chelliah is working with speakers of Dimasa and Boro to document Dimasa. Native speaking linguists in Assam are working with Scott DeLancey to write a grammar of Boro. The missing piece then, is KokBorok.

KokBorok is quickly being replaced by majority languages. In rural areas, KokBorok is still spoken, but in many speakers in urban areas, 20 and younger, speak Hindi and intergenerational transmission is not occurring. There is little government support for documentation that could lead to better pedagogical materials.

There will be two people working on this project. The first is Samir Debbarma, Assistant Professor in the Department of Kokborok, Tripura University and a native speaker of Kokborok. Samir has a PhD in Kokborok morpho-syntax from Assam University, Silchar. The second will be Shobhana Chelliah who is the Director of the Computational Resource of South Asian Language (CoRSAL) at the University of North Texas. The materials these two researchers hope to curate are from the following sources:

1. James Matisoff collection: Dr. James Matisoff provided a large selection of his papers to CoRSAL for archiving. These include materials from a field methods class at Berkeley on Kokborok conducted in the 1970s which include word lists, texts with morphological analysis, and papers on core areas of Kokborok grammar. Samir Debbarma knows the speaker consultant for this class. Many of the students in the class are still in Linguistics or related fields. We propose to scan these papers, which are mostly in IPA, so that they can be transferred to reusable data formats. We will contact the Field methods students and Kokborok consultant for digital rights permission so we can host the papers with appropriate metadata through the University of North Texas Digital Library CoRSAL archive.

2. The collection of Jonathan Evans: Jonathan Evans has a collection of 48 spoken Kokborok texts, in Roman and Bengali scripts, with lexical glosses in IPA. Twenty-five of the spoken sound files have been transcribed so far, with around 14 glossed. These spoken files include 20,017 tokens, 4,318 unique words, and 2,624 sentences. Thirty-seven of these are written texts, half of which have been glossed. These include 36,144 tokens, 10,043 unique words, and 3,848 sentences. Jonathan Evans has enthusiastically agreed to our curating these materials for this collection.

3. Additional materials that community members might want to contribute, especially audio and video of connected text and cultural events.

The curated Kokborok collection will help the community know what is already available and what further documentation efforts are needed for creation of pedagogical materials. To curate these materials we will have to:

1. Get digital rights permission from the contributors (such as the field methods students at Berkeley)
2. Use a high resolution camera to create digital files of all non-digital materials
3. Use OCR software to create reusable files.
4. Curate additional audio and video collections created by the community
5. Create CoRSAL metadata for each digital item
6. Upload to CoRSAL
7. Disseminate to the Kokborok community and CoRSAL community

For this archival project, we will use the following workflow:

1. Discuss with community leaders plans to create a collection
2. Secure required digital rights forms from Berkeley Fieldmethods students
3. Set up camera and table with clips and lighting.
4. Train Tripura students in use of high resolution camera
5. Name files according to CoRSAL guidelines, which is based on Dublin Core.
6. With native speaker students, create an oral or written transcription using SayMore
7. Create a landing page for the collection
8. Disseminate with a “language festival” type event with the community and announce via CoRSAL social media that this collection has been created.

This project is also about capacity building for creating collections from legacy materials in northeast India. For example, once the students in Tripura are trained in creating digital objects with the camera set up and in creating transcriptions with SayMore, a lot more can be done for the other languages in Tripura of which at least 6 are critically endangered. We are also working with students at the University of Guwahati, Assam, a state neighboring Tripura. At the University of Guwahati we will be providing training in digitization of legacy audio and text and in archiving existing materials on Boro and Dimasa in September 2020. As we have discovered over the past two years of conducting archiving workshops in India, there is very little infrastructure for long term preservation and access of documentary materials anywhere in India. This includes the large governmental organizations like the Central Institute of Indian Languages. Providing training and providing examples of how digitization and data management can lead to usable collections is a goal for this project and others that can follow.

Budget for Creating the KokBorok Language Resource at CoRSAL

The primary expense will be the camera, lights, and software. We would need funds to pay for the students from Tripura University who will be helping with the photography, sd cards, travel for Samir Debbarma to Kokborok communities from Agartala where Tripura University is located. We would also need funds for the UNT student who will scan the Matisoff material and enter the metadata into the UNT metadata editor. We would provide $150 to the UNT digital library for services. Chelliah has her own funds via Faculty Development Leave to travel to India.

Nikon D80016.1 effective megapixels and an EXPEED C2 image-processing engine	$450
Telephoto macro lens, the Nikon AF-S VR Micro-NIKKOR 105mm f/2.8G IF-ED	$800
OCR software: ABBYY FineReader 15 Standard	$200
Kaiser Repro Kid Copy Stand Kit with 23.25" Column, Grid Baseboard & Light Set with 2 Sockets	$314
Funds to pay for the students from Tripura University who will be assisting with photographing and creating metadata	$2000
SD cards	$40
Language festival for dissemination space and food	$50
Travel for Samir Debbarma to Kokborok speaking communities from Agartala	$300
UNT Digital Library	$150
UNT student for scanning at Digital Library	$300
	$4604