Exploratory User Research for CoRSAL

7 DECEMBER 2016

REPORT PREPARED FOR
Shobhana Chelliah, Director of CoRSAL and Professor of Linguistics, University of North Texas

BY STUDENTS OF ANTH 4701/5110 DESIGN ANTHROPOLOGY
Duha Al Smadi, Sebastian Barnes, Molly Blair, Miyoung Chong, Robin Cole-Jett, Aaron Davis, Samantha Hardisty, Jenny Hooker, Corderon Jackson, Tori Kennedy, Janette Klein, Brittany LeMay, Melanie Medina, Kenneth Saintonge, Anh Vu

PROFESSOR
Christina Wasson, Professor of Anthropology, University of North Texas

Table of Contents

1. The Project

Christina Wasson

Introduction

This project, *Exploratory User Research for CoRSAL*, was an exploratory ethnographic study to generate a foundational understanding of how different user groups might use a planned language archive for South Asian languages. The language archive is being developed under the guidance of Shobhana Chelliah, Professor of Linguistics at UNT. Chelliah was therefore the client for this project. The language archive is called the Computational Resource for South Asian Languages, or CoRSAL for short. Our research project will be used by the CoRSAL team to help plan the design of CoRSAL's infrastructure, and laid the groundwork for further studies that will take a deeper look at issues surrounding the design and use of the planned language archive.

The overarching research question for this project was:

- What are the needs of each major user group with regard to this future language archive?

Within this overarching question, we investigated a number of subquestions:

- What is the relationship of these users to South Asian languages?
- What are their current cultural practices of depositing, accessing, using and sharing information about these languages, as relevant?
- What kinds of content would be most important to these users for a future language archive?
- What are their technological constraints and preferences? (Internet access, mobile app vs computer app, etc.)
- To the extent that these users currently use language archives, how well do the language archives meet their needs?
    - What problems do they encounter, and how do they work around those problems?
    - What would they like language archives to do that they currently don't do?

Collaborative Contexts

This project was embedded in three different collaborations.

CoRSAL Team

In August 2016, Shobhana Chelliah put together a team to work on the development of CoRSAL. One of the primary initial tasks of the team was to apply for funding. The team has applied for funding from the PIRE program of the National Science Foundation. The application is still pending. There are three rounds of competition. The first round was successful. As of December 2016, the team was waiting to hear whether they had succeeded in round two. Even if the PIRE application is not successful, the team will apply for funding from other NSF programs. Team members are:

- Shobhana Chelliah, Principal Investigator
- Rodney Nielsen, Co-Principal Investigator
- Alexis Palmer, Co-Principal Investigator
- Christina Wasson, Co-Principal Investigator

So at a broader level, the whole CoRSAL team may be regarded as a client of our project.

Collaboration with Chelliah's Language Documentation Class

Shobhana Chelliah and Christina Wasson discovered that they were both teaching classes that related to CoRSAL on Wednesday evenings in fall 2016. They decided to take advantage of the situation, and organized three joint class meetings over the course of the semester.

The topic of Chelliah's class was linguistic data management and tools. Students worked on Lamkang language materials, preparing them for future deposit in CoRSAL. They learned to take linguistic data from recording to analysis to preparation of a packet that can be uploaded to a digital repository.

Wasson and Chelliah thought it would be interesting for students in each class to learn about each others' work. The interactions between students could promote interdisciplinarity and help each class gain a broader perspective on their activities.

Collaboration with Design Class at Illinois Institute of Technology

This class will collaborate with a design class taught by Santosh Basapur in spring 2017 at the Institute of Design, Illinois Institute of Technology. Students in the design class will use findings from this project to further develop design ideas for CoRSAL. Students from the present class have expressed readiness to communicate with Basapur's project. We have not finalized the exact process, but Skype and email are both likely communication tools.

Study Participants

As Wasson and Chelliah were planning this project, Chelliah identified four main user groups for CoRSAL:

- Language communities
- Computational linguists
- Other linguists who want to use CoRSAL as a source of data for research
- Linguists who are depositors to the archive and/or archive managers

We conducted research with 16 study participants, drawn from all four of these user groups. In most cases, the study participants were identified by Shobhana Chelliah. The table below summarizes information about the interviewees. They are described in more detail in Chapters 2 and 3.

User Group	*Abbreviation*	*Location*	*Number*
Language communities	LC	India	3
Computational linguists	CL	United States	4
Other linguists	OL	United States, Australia, Netherlands	5
Linguist depositors + archive managers	DM	United States	4

For the language community, students interviewed members of the Lamkang tribe in Northeast India. This is the language community that Chelliah is working with most intensively at present. For the depositors and archive managers, students interviewed members of Chelliah's research team, who are currently preparing Lamkang data for deposit in CoRSAL.

Please note that the phrase "study participant" is used in this report to refer to interviewees. The term "subject" is generally avoided by anthropologists today because it conjures up images of experiments. And "informant" has had negative connotations since Watergate.

Research Methods

This project was conducted by students in a class on design anthropology, and therefore follows the methods characteristic of this field.

> "Design anthropology" describes the practices of anthropologists who collaborate with designers and team members from other disciplines in order to develop new product ideas (Wasson 2000). The primary contribution of the anthropologists lies in the ethnographic research they conduct with users, or potential users, of the product being envisioned, in order to learn about the everyday practices, symbolic meanings, and forms of sociality with which a successful new product would need to articulate. Designers and other members of product development teams draw on findings from such research to develop design ideas that fit the lived experience of intended users… Generally speaking, design anthropologists work at the "fuzzy front end" of the product development cycle. This is where exploratory research takes place that may lead to the conceptualization of new products; it precedes the actual product development process (Wasson and Squires 2012:26) (Wasson 2016).

Design anthropology is a common approach to conducting research for the design of technologies, as part of an interdisciplinary process that is variously termed user-centered design, human-centered design, user experience, or human-computer interaction, depending on the context in which the process takes place. The findings of this project will be used by the CoRSAL team to help plan the design of CoRSAL's infrastructure, as well by the design class at the Illinois Institute of Technology described above.

The research design for this project was approved by UNT's Institutional Review Board, as required of all UNT research that may be published or presented at conferences.

Data Collection

Students conducted in-depth interviews with all study participants. The interviews were semi-structured, meaning that students started with an interview guide containing a list of topics to ask about, but also asked extensive follow-up questions that were not listed on the guide. This allowed the student researchers to pursue conversational topics that were illuminating to the project but that had not been predicted beforehand.

The linguists were interviewed at their place of work, so they would be in front of their computers. The interviews include walk-throughs of data and software, where the study participants showed the student researchers what they do with language data and any language archives they used.

The interviews typically lasted 1-1 ½ hours. They were all recorded. They were generally conducted by two-person teams of students.

For the Lamkang community members in Northeast India, interviews were conducted by phone, due to their limited Internet access. These interviews were audio recorded. Interviews with all other study participants were video recorded. Some were conducted face-to-face, and some took place via Skype. For the latter interviews, software was used to record the Skype screen.

Data Files and Storage

After each interview, the pair of students who conducted the interview wrote detailed field notes, with time code inserted periodically to correlate the field notes with the recordings. A few students went above and beyond by fully transcribing the recordings.

Project data files therefore consisted of two types:
- Recordings (mainly video, some audio), adding up to a total of 20 hours
- Field notes

All data files were uploaded to a Google Drive created for the class project. This made all data available to all class members. The password was only shared with members of the research team.

Analysis

Sharing Fieldwork Experiences and Collaboratively Brainstorming Patterns

Initial analysis of the data occurred in class meetings, by all students working together as a group. As student conducted their interviews during weeks 4-7, they would report on their experiences during the next class meeting. Their classmates would discuss and compare insights from interviewees, with a focus on pattern identification. Students would describe a pattern they saw, supporting it with one or more "instances" of that pattern from the conversation with their study participant. These instances and patterns were noted in a brainstorming document. Once one student had identified a pattern, other students usually volunteered "instances" from their own fieldwork, and the class would discuss the pattern and perhaps modify or add nuance to it.

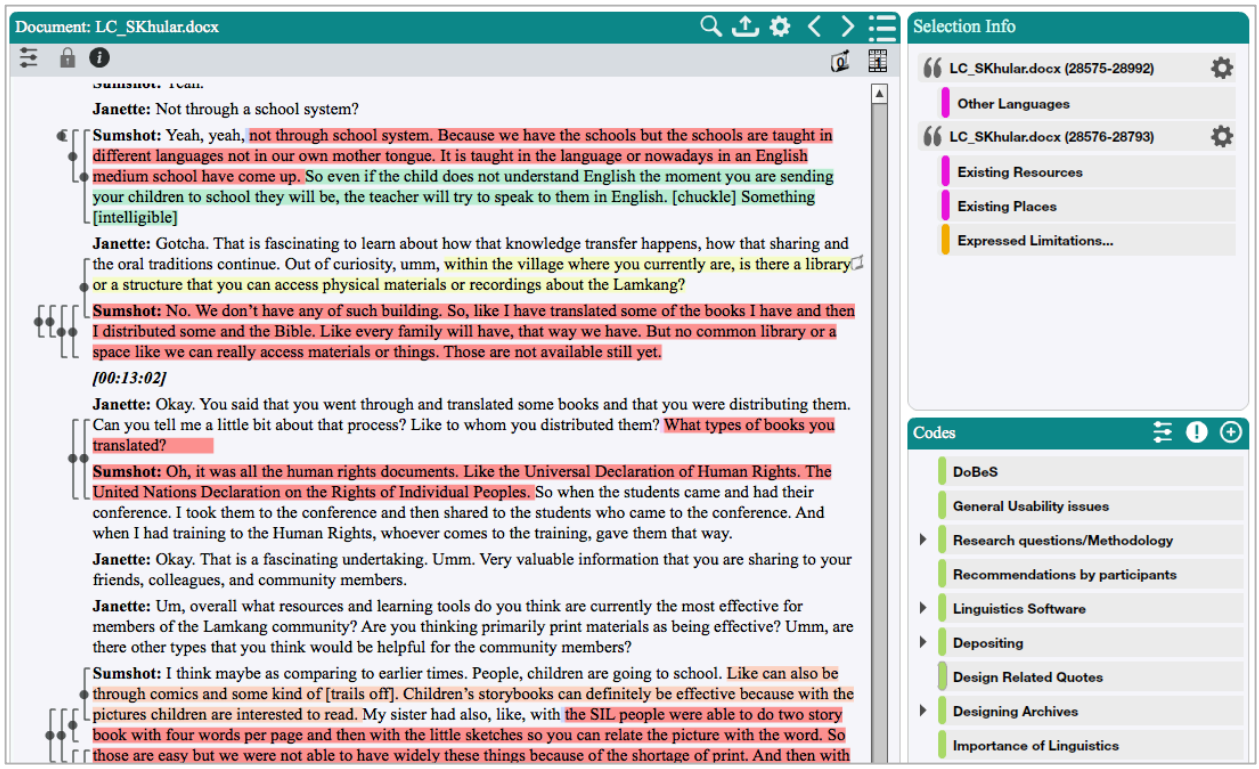
In week 8, the class presented their fieldwork experiences to Shobhana Chelliah, Rodney Nielsen, and Chelliah's class, with the aid of slides and video clips. Further collaborative analysis took place during that meeting. The collaborative brainstorming analysis continued during weeks 9 and 10.

By week 9, the class had developed a clear idea of what the main patterns in the data were. These patterns then became the chapter topics for the final report. Students divided up the chapter topics among themselves.

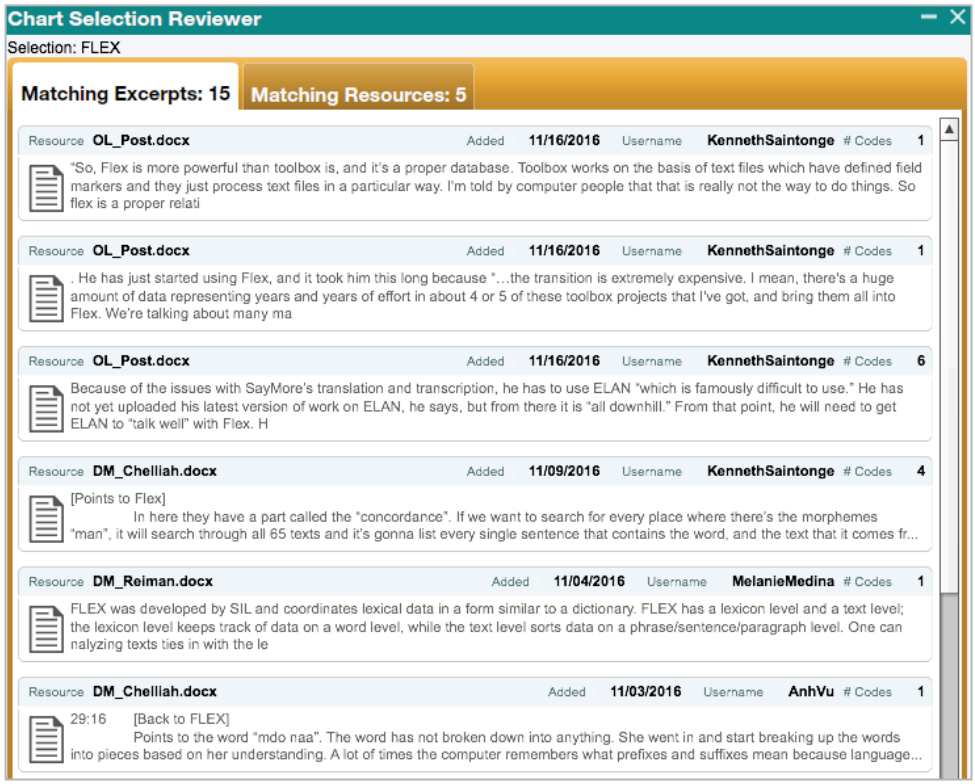
Rigorous Qualitative Analysis Using Dedoose

In order to subject their chapter topics to a rigorous and thorough analysis, students used a qualitative data analysis program called Dedoose. Dedoose is a browser-based program and as such was easy to work with and collaborate on when not in class. With Dedoose, students were able to code every "instance" in the field notes and associate it with the appropriate pattern. Codes are like tags; they are used to mark pieces of text. Codes are defined by the person who creates them; they can be specific or broad, depending on what kind of information the researcher wants to mark in the data. What was useful for the class was that Dedoose could generate reports of pieces of text marked with a particular code. This helped them closely examine patterns across all 16 field notes. Students created 96 codes in Dedoose, and applied the codes to 1156 field notes excerpts.

Below is a screen shot showing part of a field notes document in Dedoose. This is what students looked at when they were coding. The field notes in this case are Sumshot Khular's interview. The colored pieces of text are excerpts that have been coded; each code receives a different color. The codes for the orange excerpt at the top are listed in the top right corner.



All excerpts that have been assigned a particular code can then be viewed together as a list. The screen shot below shows the beginning of the list of excerpts for the code "FLEX."



During weeks 9-12, students analyzed their chapter topics with the assistance of Dedoose. Chapter drafts were completed in week 12. Christina Wasson led the editing process, and the report was finalized by week 14.

Overview of Research Findings and Design Implications

One of the things that has been lacking in the design of many previous language archives is a serious consideration of who the intended user groups are, and how the archive would need to be designed in order to meet the needs of each of these groups. A focus on the particularities of different user groups is a foundational contribution that design anthropology and user-centered design make to the field of language archives.

Furthermore, CoRSAL seeks to target an unusually complex and challenging combination of user groups. Up until now, no language archives have been specifically designed to accommodate the needs of *both* language communities and computational linguists, in addition to typological linguists.

Through our research with CoRSAL's targeted user groups, we have concluded that it would be helpful if CoRSAL could create three main portals. These portals would be customized to the specific needs of:

- Language community members
- Researchers
- Depositors

Within these main portals, there could be further levels of customization, for instance for:

- Different language communities, and multiple uses within those communities
- Different types of researchers

Chapters 2-9 describe the results of our study in detail. Each chapter contains two halves. First, the *research findings* are presented. Then, *design implications* are provided based on those findings.

Chapters 2 and 3 describe the user groups with whom we conducted research and the design implications of their cultural contexts and practices. Chapters 4-7 present common challenges faced by users of language archives. Chapters 8-10 offer concluding reflections about CoRSAL. Chapter 10 brings the design implications from all chapters together in a single, summarized list.

Note on Quotes

A note concerning the quotes in this report: the quotes are in most cases taken from the *field notes* of the interviews, meaning that they are paraphrases of the interviews rather than direct quotes. However, in some cases direct quotes are included.

Quotes are labeled with the user group abbreviation (LC, CL, OL, DM – see table on page 2), and the last name of the interviewee.

USER GROUPS

2. The Lamkang Community

Janette Klein, Tori Kennedy, and Samantha Hardisty

Photo by Sunshot Khular. Children dancing at Thamlapokpi Village,1 January 2017.

Background on the Lamkang Community

The Lamkang live in the state of Manipur in India. Members of the Lamkang are mainly concentrated in the southern part of Manipur (Lewis et al. 2016). Traditionally, the main religion in Manipur was an animistic religion called Sanamahism which involved ancestor and spirit worship. People had also adopted some Hindu practices, and for some castes there was a strong identity with Hinduism. The transition to Christianity in the last century has created a developing hybrid of old and new cultural elements. This switch has has impacted their culture significantly.

As of 1991, the Lamkang tribal population in Manipur was 4,031. By the 2001 census, the Lamkang population in Manipur was 4,524 and per the 2011 census, the Lamkang population in Manipur was 7,770 based upon the table 1.24 demographic status of scheduled tribe population and its distribution (Ministry of Tribal Affairs 2013, 153. There may have been a change in the census data collection process between 2001 and 2011. Since 1951, the government of India has recognized Lamkang as a scheduled tribe. The Lamkang language belongs to the Tibeto-Burman group of languages and the tribe's history can be traced back to the 1st century A.D. and the Manipuri kings chronicles (Sankhil 2012). While the use of the Lamkang language is vigorous, it is still categorized as a developing language while traditional stories and folk traditions are not being transferred on to ensuing generations (Lewis et al. 2016, Sankhil 2012).

Here is a list of Lamkang villages, given to us by Sumshot Khular:

1. Angbrasu	14. Sektaikarong	27. Diiringkhuu
2. Betuk Sengkreng	15. Damloonkhupii/Thamlapokpi	28. Charangching Khullen
3. Paraolon	16. Damzol	29. Charangching Khunou
4. Lunkharlon	17. Thamlakhuren	30. Charangching Khunkha
5. Charlong	18. Dulksenlon/Leingangching	31. Phaidam
6. Kotel Khutun	19. Nungpanlon/Nungkangching	32. Kurnuloon
7. Kongpe	20. Keithelmanbi	33. Komsen
8. Laiktla/Lamkang Khunjai	21. Angkhel Chayang	34. P. Raalringkhuu
9. Ksen Khupii	22. New Chayang	35. Nyongkong
10. Leipungtampak/Rindamkhuu	23. Lamrinkhuw	36. Kana Charlong
11. Old Lamkang Khunthak	24. Aibuldam	37. Chingkhir
12. New Lamkang Khunthak	25. Daampi	38. Ringkhuu
13. Lamkang Khunthak	26. Lamkang Colony	39. Seljool

Interviewees

In order to develop a more comprehensive understanding of the Lamkang community, the Lamkang culture, and community member engagement with the Lamkang language, we interviewed three community members, all of whom are native Lamkang speakers.

Reverend Daniel Tholung

Reverend Daniel Tholung was the first individual interviewed. Rev. Tholung identifies himself as a minister and language preservation activist. He graduated from university 15 years ago. He is currently an ordained Christian minister serving the Lamkang community in South Manipur, where he lives with his wife and children. His younger sister and their parents also live with him.

Rev. Tholung indicated that he attends many conferences and workshops regarding Bible translation and Lamkang language preservation. He is actively participating in Bible Translation Consultant Development Workshops (TCDW) for the Asia-Pacific region. He stated that he often applies what he learns about language preservation to assist other communities in their own preservation endeavors. Rev. Tholung also continues to advance his skills and training in linguistics by furthering his education in language based courses at a university. In 2010, Rev. Tholung visited UNT as a visiting scholar and participated in a Field Methods class. Additionally, in 2013, Rev. Tholung participated in the Orthography Workshop held in Guwahati, Assam. During this workshop, Rev. Tholung continued to assist in the development of a Lamkang language orthography under the direction of Shobhana Chelliah from the University of North Texas, David Peterson from Dartmouth University, and Thangi Chhangte. While working on Lamkang language transcription, Rev. Tholung uses Transcriber 1.5 or 1.6 when Internet access allows.

Sumshot Khular

Sumshot Khular was the second individual interviewed from the Lamkang community. Khular is a human rights and peace activist. She holds an M.A. in Theory and Practice of Human Rights from Essex University, UK and an M.A. in Linguistics from Manipur University. Additionally, Khular was awarded a Fellowship in Oral Literature in 2016 from the Firebird Foundation for Anthropological Research for her project *Documentation of the Lamkang Language*. She worked with Central Institute of Indian Languages (CIIL), Mysore, on the Endangered Language Project. She is the Executive Director of Community Action and Research for Development and has been actively involved in various grassroots organizations promoting education, human rights, gender, development, and peace. Khular is also the Vice-President of the NAGA Women's Union and has worked for the Centre for Social Development (CSD), Imphal, the Indian Social Action Forum (INSAF), New Delhi, and the Foundation for Social Transformation (FST), Guwahati. She has translated international human rights documents into the Lamkang language such as the Universal Declaration of Human Rights, the United Nations Declaration on the Rights of Indigenous Peoples, and the Convention on the Elimination of All Forms of Discrimination Against Women. She has also written two story books for children in Lamkang. Khular is currently living in the Thamlakhuren village located in the Chandel district with her family.

Swamy Ksen Tholung

Swamy Ksen Tholung was the third Lamkang community member individual interviewed. Swamy Tholung categorizes himself as an elder of the community, who is dedicated to preserving the language. Swamy Tholung is a pastor at a Baptist church in the Lamkang community, which holds the largest local religious institution membership. In addition to working with Bible International from 1987-2001, Swamy Tholung translated the New Testament to Lamkang and is currently working with a literature society on additional translation initiatives. Swamy Tholung is an English-Lamkang translator, writer, and presenter at language workshops who is fluent in six languages.

Research Findings

The Importance and Meaning of the Lamkang Language

Throughout the course of the interviews, the importance of the Lamkang language to the community members shone through repeatedly. The Lamkang language is more than just a language spoken within the home and the community. Based upon the interviews, it was learned that the Lamkang language functions as a source of identity and cultural pride for members of the Lamkang community. It is a compliment to the cultural ceremonies and traditional dress that are visible hallmarks of the Lamkang heritage. The Lamkang language serves as an auditory symbol that connects community members to each other and to their place in history. Selected comments on the role and importance of the Lamkang language included:

> "The Lamkang Language is my mother tongue. My heart language. I speak, I live, and this is my life… I live with it, yeah. I use it! In my everyday life." (LC D. Tholung)

> [On the Lamkang language] "... I think language is one of the basic forms of communication and it is also an entity for us. Because the language itself shows our identity and the richness of community is all expressed by language. Whether it is a folk song, folk tales or in riddles or proverbs or whatever that we use it all rests through the medium of language. So, it is an important thing. And without the language we are nothing. That is how I take it." (LC Khular)

> "So that, at least, my dream, and my prayer, and my desires are that somehow my language can be preserved." (LC S.K. Tholung)

The Lamkang language is an embedded oral tradition within the Lamkang tribe and community. In discussing with the interviewees the transference of the Lamkang language and culture from generation to generation, interviewees noted that the Lamkang traditions are passed on by folk song, traditional ceremonies, folktales, and collective learning with the elders imparting knowledge through practice and hands-on-application. Learning the Lamkang language is not part of the educational curriculum within the region. Rather, Lamkang is spoken within the home and the immediate community. As such, the role of family and the community were stressed as being critical aspects for the dissemination of the Lamkang language.

> "As a community. I think, like, as you are born, even when you are not able to speak, parents would speak or even the other siblings would speak to you in your own language. That is how we learn, I think. Because we don't really have any other forms of learning. We are having

> the Bible, the hymn book, and now we are trying to develop some books for the children and all. But as I remember when we grow up. We don't really also read or write as such but then with conversation in the family, talking in the community that is how one learns. Listening stories, or yeah, talking to anyone in the community, that is the only that way we learn" (LC Khular)
>
> "Do not have many [educational or learning] materials or institutions. All learning and usage are oral; it is used in the community. The children learn from being around the language since birth." (LC S.K. Tholung)
>
> "To be very honest, we don't have any materials. We don't have any institutions to develop the Lamkang language. Because, our tribe, we don't have any Lamkang schools. All schools are government schools. So this is one of the things we are under privileged. So the way that we learn the language is from birth. In the community, Lamkang, and the children learn it from birth. They are born in a family that speaks Lamkang, but we do not use any other means or materials to teach the language." (LC S.K. Tholung)
>
> "We learn in the community, in the home. In that way, it depends on the festivities and it also depends on the farming. Whatever the activities that we do. It's like, in a way it is collective learning and in the family you are taught… It is like, orally transmitted. You are taught and you are taught by observing and practicing. Whatever items as you learn them." (LC Khular)

Khular also made note of the pervasiveness of Lamkang as a spoken language within the community.

> "Anyone in the Lamkang community we speak with our own language only…we speak with the small children, with the elders, or whoever is in the community in Lamkang only." (LC Khular)

Globalization and Language Assimilation

The impact of globalization was brought forth by all three interviewees during the course of the discussions on the Lamkang language. The educational structure within the region has shifted over the past years resulting in children being sent to boarding school where the state language or English are the mandated spoken and written languages. As such, even though Lamkang may have been learned as the language spoken in the home while young, once the child is sent to school the daily practice of the Lamkang language ceases, thus weakening/fading its retention within the child's long-term memory.

> "[Lamkang is learned] not through the school system. Because we have the schools but the schools are taught in different languages, not in our own mother tongue. It is taught in the language or nowadays in an English medium school have come up. So even if the child does not understand English the moment you are sending your children to school they will be, the teacher will try to speak to them in English." (LC Khular)

Similarly, as much of the programming on the radio, television, or through the Internet is transmitted in either the state language or in English. The constant exposure to languages other than Lamkang further weakens the ability of the language learner to develop skills in the Lamkang language. Examples of these concerns are outlined in the following statements made by the interviewees.

"You may learn those languages [state language or English] but as we didn't have any of the programs in any of our languages [Lamkang], in the long run, maybe, we may not able to learn our language but we will easily be learning others language. But may find difficult to learn our language or if the children, for example, nowadays many send their children when they are young to a boarding school. Once they send them out then they lose the chance of communicating every day with the children in your own mother tongue. So they speak definitely in the state language or in English. And then they will find difficulty when they come back home to communicate in their own mother tongue…With all this…yeah, invasion of all these big languages, or different languages because we are not just confined to our community but as we move out, as we interact more, that may be a possibility" (LC Khular)

"Any the youngsters, when we ask them to write up a story or a book, they will just write it using many other languages." (LC S.K. Tholung)

"One would be like we are exposed to all this media like the TV, with the coming of TVs we have different languages coming in, not just English movies but [?], the Manipuri, or the Hindi, which are all different languages and people keep watching them." (LC Khular)

As Lamkang is a tribal language, its use is not prevalent outside the villages where it is the mother tongue. This too contributes to the dilution of the impact of the Lamkang language and its assimilation as more and more frequently, individuals are traveling outside the borders of their towns and villages to interact with surrounding communities. Language assimilation is a very real concern as expressed in the following statements.

"This is a problem because they cannot normally and properly use their language in public and are compelled to use other languages. Only amongst themselves do they have the liberty to use their language." (LC S.K. Tholung)

"Our Lamkang language is very distinct. Some tribes do not have the culture we have, but it is now almost impracticable and people sometimes are not practicing it. Also, we are living together with some bigger tribes, and so that is the dominating tribe. So sometimes we are not free to do so. So some hindrances are there, in the same way, that our language is." (LC S.K. Tholung)

"There is no restriction. The Lamkang constitution has helped us to do this, but there is one barrier that our population is a very small tribe." (LC S.K. Tholung)

"I'm at home… [with] the family, so with my sister we speak, with my aunt, with my - everyone in the village we speak the language so it's like we use it every day. Only when we go to the city or when we go to a different town where we cannot use our language and we have to use the state language to communicate with others then we use the state language. But as long as it is in the community, in the house, in the village we use our own language to communicate." – (LC Khular)

"Mostly the Lamkang people, well, among the educated Lamkang people they can speak English, sometimes Manipuri, sometimes Hindi, and Lamkang. Lamkang people mostly they understand at least 3 languages. In our community, we speak mostly Lamkang." (LC S.K. Tholung)

"I personally feel very insecure, because whenever a Lamkang person delivers a speech or message, if it is a write up almost 60-70% of both languages are mingled with Lamkang. People seldom use it, they feel it is odd to use it. (LC S.K. Tholung)

> "[We are] a small tribe and because in this place where we are having multi-tribes or different tribes living together it is also important to preserve and promote the language… at some point there may a time when the language can be also very much getting a lot of assimilations or borrowings. Change may evolve but like, how best can we preserve the language or promote it?" (LC Khular)

Learning and Resources in the Lamkang Community

Limitations with Existing Resources

Promoting awareness and understanding of the Lamkang language is a vital component to ensuring its revitalization. As Lamkang is an oral language, there is a need for steps to be taken in the development of a written language that is taught to all community members, according to our Lamkang community member interviews. A few resources are already available in the Lamkang language, including a New Testament and a Children's Bible translated by Swamy Ksen Tholung, a hymnal, and assorted Human Rights tracts including the *Universal Declaration of Human Rights* and the *United Nations Declaration on the Right of Individual People* as translated by Sumshot Khular. Yet concern remains that these are too advanced for most community members to understand without further training in foundational Lamkang language skills, which could be obtained through the development of basic grammatical aids.

> "Our expectation and our hope, when this book [the translation he created of the Children's Bible] comes, in addition to our bible and our hymn book, that will help the Lamkang people to understand more of the language, and there will be more use to use the language." (LC S.K. Tholung)

> "We are halfway in a way, to finalizing the orthography what to be used, because we have two Bibles but people find it difficult to read. And we have no base, like the alphabet are not there. So without the alphabet we have two huge books that are too difficult for children to read or any adult even to read" (LC Khular)

As demonstrated by the above statement from Sumshot Khular, the development of foundational materials is critical, as it is the perception that without these tools, attempts to impart literacy and revitalize the Lamkang language will only be partially successful. Already, steps have been taken to begin development of a mutually agreed-upon orthography, a basic grammar in conjunction with Shobhana Chelliah, and a basic picture dictionary in conjunction with the Summer Institute of Linguistics (SIL) based in Dallas, TX. However, these resources have not yet been completed or disseminated to the community members. Additionally, concern was expressed by study participant Ross that development of an official orthography faces additional challenges, as shown in the following quote:

> "The problem of devising from unwritten language is very serious one because political questions are coming up. Due to the difficulty and socio-political problem, taking care of language and finding a good sample of language that is not slanted in one way is very hard" (OL Ross)

Desired Resources

Discussions with the interviewees produced several unique insights into the resources that are desired for continuing to promote the Lamkang language. Due to the presence of strong familial ties within the Lamkang community, a desire was expressed by the interviewees for the children and the younger generation to receive priority in becoming fluent in Lamkang. This desire led to

Image retrieved from: http://neipeople.blogspot.com/2014/10/lamkang.html

several recommendations from the interviewees for language learning materials specifically targeted to younger school-aged children, including a documented alphabet, storybooks, comics, and animations.

> "Because a lot of the children are watching the cartoons and things like that. So, if instead of watching the English, if those voices could be in Lamkang that can be also, I think, helpful." (LC Khular)

> "Some pictorial kind of dictionary that can be used and by which we can also learn the language for children. And we also working on the small children's stories." (LC Khular)

> "[Could] also be through comics and some kind of [trails off]. Children's storybooks can definitely be effective because with the pictures children are interested to read. Having something like a comic book or booklet kind, I think that can be also useful or some kind of short animation, DVDs and things that is shared people can still watch them in their home TVs." (LC Khular)

As the documentation of the language continues, additional resources in the form of the completed grammar currently being developed by Shobhana Chelliah and her team at UNT, and the picture dictionary currently being developed in collaboration with SIL will be especially beneficial to adults with no or limited linguistic training. But beyond the development of the materials themselves, an important consideration is the production of materials that can be dissemination to all of the community members in addition to information literacy programs.

> "The SIL people were able to do two story book with four words per page and then with the little sketches so you can relate the picture with the word. So those are easy but we were not able to have widely these things because of the shortage of print." (LC Khular)

> "We had an Australian lady, we had Mongolian, we had one [Norwegian?] and now we have one of the Russian girls who is trying to help in literacy development. So that's the time that we have workshops and they try to come and assist in the workshop." (LC Khular)

> "We had a workshop, that which she collected some of the legendary Lamkang stories. Four of us went, and then there she asked us the grammatical question. Like in this portion; what is the noun, what is the verb, what is the adjective, and all of the other things. For all of these things we were interviewed, and we have given our best." (LC S.K. Tholung)

> "But it is, it is, like, quite different from how Shobhana's working because hers is more academic and advanced than what we are doing here. Like you have something of a picture, then the word in Lamkang and one in English. So you know what it is in English." (LC Khular on the development of the Lamkang picture dictionary)

Collaborative Language Material Development

The importance of the development of language learning materials being a collaborative endeavor was reinforced several times by different interviewees.

> "We involve everyone because we have the elders coming in and we have young people who are able to read and write. We ask them, the elders to tell their story, they try to write them down. We also went to collect the word. Everyone. Elders are involved, young people are involved. A collective kind of effort." (LC Khular)

> "Everybody also gives their time and effort. Everybody feels the ownership. They are being part of the process. Which is also a good one." (LC Khular)

> "Because in the beginning people thought "oh, we cannot do it. Only the experts can do it." But then we all said you know the word, and you can tell the story, so you are our resource. And the young person who don't really know the story but he can write down so he contributes his time helping to put into written form. So the elder [unintelligible]. And the young one being included in the process that makes everyone happy that all is part of the process" (LC Khular)

Reinforcing this collaborative mindset, Shobhana Chelliah offered some additional insight during her interview into her perception of the language documentation project with members of the Lamkang community.

> "The Lamkang project was built from a different perspective, which was untraditional. It began with the community reaching out to the team and sending in materials. So the initial materials were random and messy. There were different writing systems, very little phonetic transcription. The word for word translation was provided by the community, but hardly any global understanding of the text, or even sentence by sentence translation because it was difficult to do." (DM Chelliah)

The community members have collaborated with various organizations to preserve the language. Often this has resulted in little benefit for the Lamkang. Swamy Ksen Tholung described an instance when he and a group of community members traveled very far to record one audio story. Based on his interview it appears that this recording is not in the possession of the Lamkang. The perception on the part of the community is that the community members have shown eagerness to do whatever needed to preserve their language and traditions, but their dependency on outside resources, who often appear to be unreliable, is hindering their process.

> "The reason is, I should say, we do not have many privileges. We cannot afford the audio recordings, video recordings. So that is our disadvantage and we do not have that possibility. … One organization helped us just to produce one audio cassette. The audio recording was done...about 600 km away from our state. We went there and our voice was recorded. And it was just the story of Jesus and the story of his disciples. And we do not have like alphabetical, literature, audio or the video." (LC S.K. Tholung)

The community feeling at this time is that little to no progress is being made in documenting their language. In fact, a 3,000-root lexicon of the Lamkang Language has been developed by the Lamkang research team. Tyler Utt continues to add texts to the database from which lexemes are culled and added to the lexicon. The result is to be an online dictionary and grammatical sketch with pedagogical notes to assist the community with creating language teaching books. However, as of this time these materials and knowledge have yet to be shared with the Lamkang community. Because these developments were not shared it is generating a potential friction between the community members and external agencies, especially expressed through our interviewees, who are involved in the preservation process.

As previously indicated, the interviewed Lamkang members share the view that the language materials development is a collaborative process. Their hope is that the collaboration takes place among members of the community, including the elders and youth of the community, and also external organizations and researchers who are vested in the preservation of endangered language materials. It is our hope and recommendation that as the research and archiving of language materials progresses, the scope and strength of the collaboration between researchers and language communities will also continue to increase.

Lamkang Language and Collaborative Preservation

Due to concerns over the aging population within the Lamkang community, a sense of urgency for preservation of the Lamkang language was conveyed by the interviewees. Preservation of the heritage and cultural knowledge held by the community elders is a critical aspect to ensuring the longevity of the Lamkang community as a whole. The concerns and efforts extended by the community members to preserve the folktales, stories, ceremonies, and traditions are expressed within the following statements from the interviewees.

> "So, our resources are these people, the old people. I am very much concerned that if all of them have expired or died then the younger generation might not the original traditions that might have been handed down. This is the concern that we have." (LC D. Tholung)

> "The chief of the Lamkang village (Beshot Khullar) had been writing about songs, folktales and other traditional things. He needed more help because the tradition and language were dying out, and the young kids were not learning the language." (DM Chelliah)

> "The type of people within the community that currently possess the archival materials are the elders around 70-80 years old. The concern is that these people are the only resources to collect information about the community and they are passing away due to old age. Three elders in this position died this year. My concern is that if all of them have died then the younger generations might never learn the original traditions that could have been handed down." (LC D. Tholung)

> "We are working with SIL, a lady came here to India to make an alphabet book, but that is just the initial [phase]. The people, well, we are almost late [referring to it being almost too late to preserve the language]. This is why when these people of UNT are trying to help, I am very grateful." (LC S.K. Tholung)

Similarly, Swamy Ksen Tholung indicated that the shifting focus to Christian traditions has impacted the preservation of Lamkang ceremonies and traditions as expressed below:

> "In the church they speak Lamkang, and for marriage...in past practices when someone was sick and in past offerings to gods. There is no scholar or documenter of the past or present story and culture of Lamkang, so nothing is preserved. Marriages and burials are not recorded. They use the Christian traditions now. They are different from the practices of the pre-Christian Lamkang, which is about 20-30 years back." (LC S.K. Tholung)

The imminent loss of historical language traditions and culture is further compounded by globalization and language assimilation through the changing educational structure within Manipur and its villages.

The Digital Divide, Information Literacy and Access

Geographic Constraints

Geographic considerations must be taken into account when developing CoRSAL. All three interviewees expressed concerns over the remote nature of the villages and the surrounding region and its lack of technological infrastructure. To contextualize the remote nature of the area, Rev. Daniel Tholung states the following:

> "Our town is at the end of the way and after our place the hill country starts. So, we are at the end of the village…Mine is a small village surrounded by seven hills. It is a small village and the Hindus, who make up 90% of the state of Manipur, are integrating. His [Rev. Daniel Tholung's] village is in the foothills to remain away from the non-Christians. The Christian

villages are all close together within these foothills and it requires walking up and down the hills to reach anyone. They have to go up pretty high to reach their town." (LC D. Tholung)

Telecommunication and Technical Infrastructure

An important consideration in developing a language archive that is intended for use by the language community members is to determine what the telecommunication and technical infrastructure in the area will support. A surprising finding that emerged throughout the course of the interviews was the lack of a supporting regional telecommunication infrastructure in Manipur and more specifically within the villages where the Lamkang community members reside. Although, according to our interviewees, many individuals have smartphones, it is our perception that the degree of connectivity and functionality of the applications and their connectivity will need to be carefully assessed prior to assuming that these devices may complement or supplement standard Internet access via a desktop or laptop computer. Statements from the interviewees, such as Swamy Ksen Tholung, indicated that a major concern for this project is the lack of Internet access. Additional statements from the interviewees illustrating these concerns and supports the finding that the Internet at this time is not a viable resource for the community include the following:

> "Internet access is a major problem and that some villages do not even have electricity. There is limited internet access in the nearby towns but very few people within his community use it." (LC D. Tholung)
>
> "With the ground situation, where we are still having all these, people are using mobiles but they cannot really access Internet or connectivity with Internet is bit of uncertain, still even now. So, yeah. It will be a bit difficult for community people." (LC Khular)
>
> "The best place internet connection would be at the university nearby the village for Internet access." (LC D. Tholung)
>
> "Because we say we access Internet, but in some places it is okay, in some places like we have to go to certain mountain and then on to speak the phone in some villages. In some villages unless there is power we cannot call or receive calls. There won't be any connectivity. So having the university archive to be accessed by the Lamkang community would be very minimal. We cannot be able to access as much because only those who are curious or those who are able to access Internet who are away from home in the community who are in towns and cities may be able to do that. But for people in the communities, in villages to access them would be, yeah, still a difficulty." (LC Khular)
>
> "There is one Internet shop in the entire Chandel district, they must pay to use it. Some people have broadband." (LC S.K. Tholung)
>
> "The whole day I was trying to write a mail and open and it was like off and on, and it was not really possible, unless maybe I go to city in the capital. That can be possible but I am in the village right now as my cousin is unwell and I have to be at home. So I cannot leave her and go. I cannot express it is really difficult." (LC Khular)

It was repeated numerous times that the local university currently offered the best Internet connection the Lamkang people could have access to. Yet, the university itself does not have many resources either; as Sumshot Kulhar put it, "the linguists department only had three or four computers". Although limited, it is still a great resource but it is far enough away from the village that it is not possible to visit daily if a community member desired. It is also not visited by everyone within the community. Our three interviewees mentioned that they did indeed visit

whenever they could, but it should be emphasized that our interviewees are part of the educated class within the Lamkang community.

An example of this is provided by the background of Sumshot Khular. Khular was the first in her family to be given a fellowship to study abroad. Yet as she furthered her education, she felt she needed to really do something for her community, to give back what she learned. So, that is why she returned to her village to develop human rights training and documentation in her own native language of Lamkang. She wanted to be able to develop her language and culture but found it was impossible to do from a distance. At the same time, being in her village and trying to contact people in other villages or those from outside the region on a state, national, or international level who could help her in this project proved futile, as they were unreachable due to the technology and connectivity issues.

Literacy

The struggles inherent in creating an orthography for the Lamkang language also produce challenges in promoting literacy in the Lamkang community. Given that an orthography for Lamkang is still under development, few writings currently exist. This means that the opportunity for community members to develop literacy skills is limited. On top of the limits of the existing materials for the language, as of yet, a formal educational structure within the community does not exist.

> "Because even for adults. You can also do adult literacy program. But then not possible to do it. There are many people who wanted to know how to read and write but the facilities to teach them is not there." (LC Khular)

> "But no common library or a space like we can really access materials or things. Those are not available still yet." (LC S.K. Tholung)

In the past, the structure of education was informal; it relied on oral teachings, and was taught at home and throughout the community without a designated building that would traditionally be construed as a school.

The complications of deciphering the spelling systems of the Lamkang language seem to be persistent and strong. One instance of this was Swamy Ksen Tholung's claim that the community language should actually be spelled "Lamkaang". This exemplifies the struggles the community is enduring in regard to spelling.

CoRSAL and the Language Community's Archival Needs

Emerging from the interviews was the concern that CoRSAL as it has been presented to them will not be a functional archive for the community members. This fear developed from the aforementioned obstacles including the lack of internet connectivity, low literacy rates, technology access in general, the educational structure, and lack of resources.

> "[Rev. Daniel Tholung] believes that as an Internet archive, our CoRSAL project would not be able to be used except by college students in the cities at the university where Internet access is available." (LC D. Tholung)

> "I think it will be just hardly 5 or 10 persons in the community might be using the [language archive] computers, whatever, for people who work. But then, not all." (LC Khular)

Several ideas were presented by the interviewees as options for archiving Lamkang language and traditional materials including the establishment of a physical archive, improved software and devices for collecting video and audio recordings and training on archiving processes and

procedures. Development of a physical archive as a repository for archiving Lamkang language and traditional materials was the most desired idea.

> "The museum or resource library could not be located within [his small] village because they do not have electricity and they also have communication problem. These things can be decided later on for the convenience of location but, for now, he thinks that being located in the villages nearby the towns may be best." (LC D. Tholung)

> "If we can have a place, or a room, especially designed to archive all of this for the collections. It could be any kind of traditional artifacts. Maybe a phone, maybe a musical instrument. So far nobody knows how to use the traditional kind of music instruments. It is diminishing, day by day. So, we need some place or at least something. Some computers, not to be taken by anybody but to keep in one proper place so that we can do all this work… We can construct a decent house, like one to three rooms. It does not have to be very expensive. If we can work in that place, it could help it the project of archiving…Especially when people come together. The materials are scattered. We have to tell them to bring all of what they have. We will ask the individuals to come who are keeping the materials, and we can organize. Otherwise, we do not have a place for where we can collect these things. There is no place." (LC D. Tholung)

> "What I intended most is that the language has an archive; to preserve language, how to archive language. But to my understanding, the language in a written form, then some history and some stories if we are privileged to write and keep that in the archive; it will include all of the traditions and custo Like, we have customary pattern ceremonies and then traditions. All of these things, unless we could get this into a written form. And, if we could develop some art; like develop some particular type of arts, crafts for the archive. I think that would be very beneficial to the Lamkang people." (LC S.K. Tholung)

The second collection of ideas for archiving the Lamkang language and traditional materials included acquisition of improved software and devices for collecting video and audio recordings and training on archiving processes and procedures.

> "They need a machine to better record audio recordings as well as expertise in the area of archiving to help them along in the field. Currently they are using a simple recording device that is not up to date for archiving these songs and stories. This device does not function properly as it is not clear and allows for excessive background noise. This problem transfers over to their transcription process as it becomes too difficult to transcribe from audio they cannot hear. They are experiencing a similar problem with their video recordings as they require proper guidance in how to conduct a video recording for an archive." (LC D. Tholung)

> "He has a software called Transcriber 1.5 or 1.6 version to help him write down the language. He is having problems with the software because of the internet connection and because of corruption. He hopes to have better software access and to learn more on these progra" (LC D. Tholung)

Design Implications

Design can influence how a user behaves and in what ways the artifact and/or materials they are using affects them. This section discusses opportunities to fulfill the purpose of creating a language archive that preserves the Lamkang language, as well as the other languages that will be integrated into the CoRSAL archive. It focuses on the design implications that are targeted to serve the Lamkang community despite the acknowledged issues of the existing technological

infrastructure and lack of resources that we, at this stage, cannot remedy as they are outside the scope of CoRSAL.

Include Linguistic and Cultural Materials of Interest to Language Communities

The depository should be not only for linguistic materials, but also for art, comics, customs (which they hope to document in writing), legends and stories both oral and written, history, and possibly government and religious documents. Through the development of literacy within the community, the range of genres deposited is expected to grow. For instance, based on the issues that interviewees identified as being important to them, such genres might include materials on women's issues, and traditional cultural arts such as music and dance.

Empower Community Members to Develop Materials

Addressing the desire to create arts, comics, and engaging learning material: we recommend empowering community members to develop these materials, in collaboration with the CoRSAL team. Developing these materials within the community increases the probability of use, but more importantly, it would encourage community collaboration and active engagement as participants in the archiving and language preservation process. This would mitigate the "systematic disenfranchisement" of indigenous groups with respect to archives, discussed by Shilton and Srinivasan (2007, 89), which is important to members of the Lamkang community. Collaboration in this manner would also facilitate a bridging of the generation gap many interviewees fear will continue to develop and widen.

However, to do this, there is a need for material goods, such as scanners and printers. Funding and/or contributions from external parties may be needed. The simple addition of technology of this type would allow for the production and distribution materials freely available for use by all members of the community. Participation in this type of initiative would increase production of materials for the archive and transform it beyond a static source for linguists into something that effectively meets the archival needs of the language community.

Even as the Lamkang recognized their own lack of clarity about an orthography for their native language, we expect that within other language communities the extent of writing systems will vary. Looking to the future after the creation of the initial orthography, the preservation of Lamkang as a living language will rely on growth of its written documents as a means to promote the continued use of the language. To facilitate continued growth and use, basic language learning materials may be needed and the development of such should be tiered based on user's language level. For example, simple picture and story books, comics, and animations would be beneficial to those at the beginning levels of language learning and can also be used to engage children. Intermediate language learning materials would include grammars, worksheets, audio/video resources. Scholars, educators, and researchers would necessitate more advanced linguistic learning materials.

Engage Both Older and Younger Generations

We recommend a participatory process of data gathering that engages both older and younger generations in a cross-generational manner. Instructions should be clear and simple, with a natural language design structure. Also, we suggest starting the participatory data collection as soon as possible, to address the community concerns of urgency surrounding the decline of community elders who are the current repositories of cultural traditions, heritage, and language.

Provide Hard Copies

Our thoughts at this point are to create hard copies of information that may be stored and maintained by the Lamkang community members as a permanent fixed part of the CoRSAL archive at a designated satellite location within the Lamkang community. Hard copies would only be made public after they had been approved by local community members. Understanding the issues that Lamkang community members are facing with technology constraints, we believe hard/paper copies are the best route for what we can provide and implement as a physical representation of the archive.

The CoRSAL platform should make it easy to print the materials: distributing the materials in a physical form would make it easier to share with a wide range of community members. This would encourage them to use the materials. At the same time, it would be advantageous to store the electronic version online so that it could not be destroyed or lost. An option on the CoRSAL site for users who lack access to printers would be to request CoRSAL to mail the printed materials.

Partner on Seeking Funding for Local Space and Tools

The desire by community members, as expressed by the interviewees, is for a physical location that would house their artifacts and provide a work space for the language preservation. They also desire resources to benefit the archiving process, such as more or better quality audio and video equipment. Similarly, to be able to use an online language archive they would need at least a permanent desktop computer and Wi-Fi modem. Unfortunately, these are resources we cannot provide as part of the current project plan for CoRSAL. However, we would like to strongly encourage the solicitation of other funding sources that could be involved in the provision of the physical resources the community members need to offset the lack of technical infrastructure.

Provide Tutorials on Language Preservation and Use of CoRSAL

The development and presence of tutorials on the landing page of the CoRSAL platform is recommended. Tutorials could provide examples of the approaches and methods other communities are using to preserve their culture and traditional artifacts, further encouraging the endangered language communities to be actively involved in the preservation process. Additional tutorials should also be available including, but not be limited to: how to navigate the site, how to become a member - if that is the approach CoRSAL adopts in the portal development, how to communicate with a CoRSAL representative, how to deposit (broken down by type of material), and how to export materials from the CoRSAL archive. Such tutorials would facilitate positive user experiences and streamline directional use of the CoRSAL archive based on user group.

Use English for CoRSAL Interface Language

We recognize the concern by community members for literacy and believe this also gives rise to the question of what the language(s) of the CorSAL interface should be. We learned from our interviewees that it is common among the educated within the community to know several languages, typically between three to five languages, and that even those who do not yet know Lamkang in its written form can read and write in other languages, typically the state language, Hindi and/or English. Given this information, we believe that using English for the CoRSAL interface is the best solution. From our research, we believe English is the most widely spoken. Further research may reveal the need to make the paper copies for the community available in other common South Asian languages, as the interviewees mentioned that most

community members also speak Manipuri or Hindi. This type of information should also be established for the other language communities.

Protect Sensitive Materials

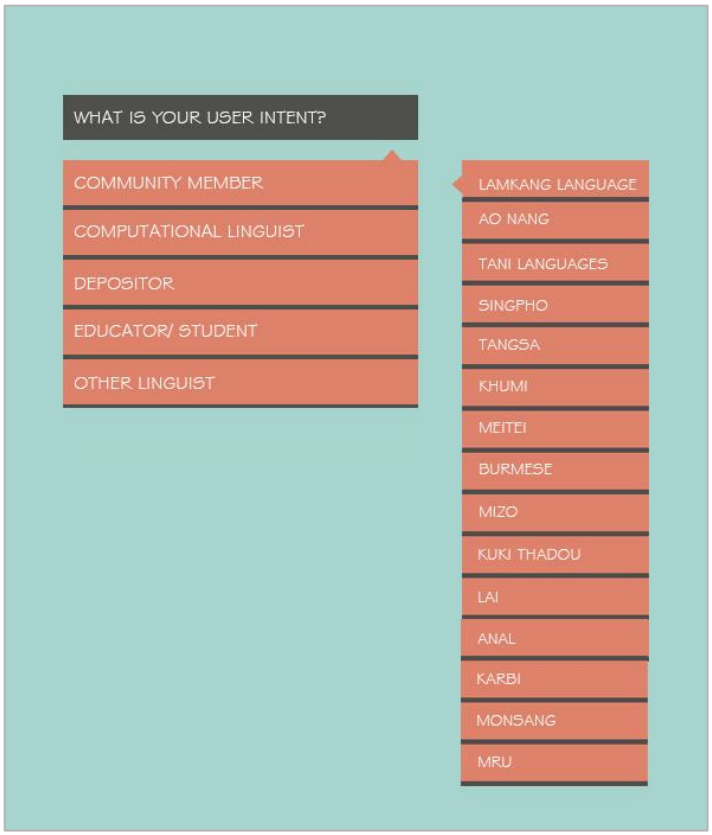
A common concern for indigenous communities is the ability to protect private, sacred, or otherwise sensitive materials in language archives from being accessed by outsiders (cites). We recommend that this concern be addressed as a major interface design consideration. Furthermore, the treatment of sensitive cultural materials is not isolated to the Lamkang community, but should be assessed on a community by community basis for all fifteen future languages projected to be included in the CoRSAL archive.

It would be possible to create a password restriction or an interface setting that separates materials based upon categorization of the users. Some artifacts, dances, arts, specific oral narratives, or other aspects of the cultural traditions might be restricted to specific language community members, with no access available to the general public. Recognizing that the purpose of CoRSAL for the community has a key goal of language and cultural preservation, it is our recommendation that these valuable materials not be excluded from the archive.

A concern is that a password requirement may restrict too much of the information available on the archive. We recommend further research in the implications both negative and positive that password protection could bring to the archive interface, retaining the integrity of the context of the archival material yet providing protection and preservation of this endangered language for future generations.

Interfaces Customized to User Groups

An option for the CoRSAL landing page would be directional pivots that request information on the user's intent in CoRSAL. The user's responses would lead them to interfaces customized to different user groups. Selection options include community member (with a further drop down question indicating which culture/language group), computational linguist, other types of linguists, educator/student, and or depositor. This is shown in the example to the right for visual clarity only. The example is not intended to be the final use or development of the site organization.



Information on Lamkang

The following list of resources may provide useful insights to members of Santosh Basapur's design class at the Institute of Design/IIT, as they investigate previous work

generated on the Lamkang community, language, and culture.

- Dr. Chelliah, Rev. Daniel Tholung, and Sumshot Kuhlar were heavily involved in documenting the Lamkang culture on this site: http://lamkanglangaugeresource.weebly.com/ .
- A good source to learn of the history of the culture of the Lamkang community: www.lamkaang.com
- The Lamkang Spelling Workshop run by Dr. Shobhana Chelliah focusing on the grammar and orthography of the Lamkang language: https://www.facebook.com/LamkangSpellingWorkshop/?fref=ts

Participatory Research and Design

Finally, we strongly recommend for the CoRSAL team to make it possible for Lamkang and other language community members to contribute to the archive while also being active participants in the design and development of CoRSAL interface features. Based upon feedback from community members, previous dependence on others to preserve and document the Lamkang language has been perceived as unsuccessful for the community thus far. It is our recommendation that steps be taken to collaborate and encourage active community member involvement as language preservationists throughout the development of CoRSAL. As shown in the pivotal work by Shilton and Srinivasan (2007), archival work should be approached as a participatory endeavor, thus preserving important contextual meaning in addition to the language materials themselves. Our design implications can be distilled to the recommendation for the creation of an archive that will successfully meet all of the needs of the community based upon the scope of the resources available and the project mission and vision inherent at the conception of CoRSAL. In the eyes of the community, CoRSAL is more than just a language archive. It is an educational tool in language, culture, and tradition, a museum, and a traditional archive. It is our vision after interacting with the Lamkang community members that CoRSAL become a means by which the work of language preservation and analysis will continue to thrive while also bringing back to life the culture, heritage, and language of those who fear for its very future.

3. Linguists

Duha Al Smadi, Miyoung Chong, and Anh Vu

This chapter describes the research findings based on our interviews with members of three of the intended CoRSAL user groups: computational linguists, other linguists, and depositors/archive managers. These linguists were located across diverse institutions, inside and outside of the United States. After conducting the interviews, we realized that there were many similarities across these three user groups. We therefore placed an initial description of all three user groups together in this chapter, while Chapters 4-7 take a deeper dive into particular challenges faced by linguists.

In the next section, the findings from the interview sessions are presented, and lastly, design implications based on the findings will be provided for the CoRSAL project.

Research Findings

Computational Linguists

CoRSAL's decision to target computational linguists as a user group is one of the innovative aspects of this planned language archive. The majority of computational linguists work on widely spoken languages, such as English, because of the availability of large data sets for these languages. There are some computational linguists who study endangered languages. However, no language archives for endangered languages have been specifically designed to present data in a way that is useful for computational linguists. It is therefore important for the CoRSAL team to identify the needs of this user group.

Interviewees

In this group, four professors working in the computational linguistics field were interviewed. Two of the four requested anonymity. In order to accommodate this wish, we created pseudonyms for all four, and do not reveal their current university affiliations or provide pictures of them.

Franklin Boss

Our first interviewee was Franklin Boss, who is a professor in the department of computer science and engineering at an American university. Boss holds a PhD in computer science and cognitive science from the University of Colorado, Boulder and he is the director of a lab on language technologies. Boss's research is primarily in the areas of natural language processing, machine learning, and cognitive science, with an emphasis on educational technology, and health and clinical Informatics.

Thelma Moore

Thelma Moore is a professor in the linguistics department of an American university. She has a Ph.D. in computational linguistics from the University of Texas at Austin. Her research interests are in computational discourse and semantics; computational linguistics for low-resource languages and for language documentation; active learning; automated short-answer scoring; discourse structure and coherence; modes of discourse; and distributional, lexical, and formal semantics.

Tara Grant

Our third interviewee was Tara Grant, a professor of linguistics at another U.S. university. She is specialized in grammar engineering and application of endangered languages. Her main project identifies syntax and morphology within constraints and essentially builds grammars.

Jessica Hill

Finally, Jessica Hill is associate professor in the linguistics department at an American university. She has a Ph.D. in computer science from Massachusetts Institute of Technology. Hill's research focuses on speech and text analysis.

Findings

We found a surprising amount of diversity in the research activities of computational linguists, even though this is a relatively new field. Some computational linguists aim to create working speech or text processing systems, while others aim to build human machine translation and interaction systems. Depending on the area of study, the linguists' background, and their research questions, a variety of methods can be utilized. These methods that could span from research to applied work are used to study, model, and analyze languages. Generally, the goals of computational linguists may range from building computational technologies to understanding a particular language. Every area in computational linguistics is important to a specific kind of people (i.e. computer scientists, engineers, etc.) who may have diverse goals compared to one another.

Computational linguistics is a somewhat new field which originated in the U.S. by integrating efforts to use computers to automatically translate text from source to target language. Using computational methods saved researchers time and effort, and produced reasonable results compared to manual methods. Also, CL methods made it possible to analyze more data than would be possible without computers. The CL field is distinguished by rapidly evolving in terms of research areas, tools, and methods used.

Although the methods might vary, some techniques are commonly used by most computational linguists. Machine learning (also considered supervised learning method), unsupervised learning, statistical models, and neural networks are the most common. However, some of our interviewees indicated that some of these methods if not all are a subset of the artificial intelligence area. These techniques are used to build models in order to recognize patterns and engage in labeling (i.e. predicting) in order to constitute themes.

Multiple tools and applications are used by computational linguists. Usually, they use applications to write code for analyzing and processing their data. The most common applications used are Python, natural language processing (NLP) tools, and Natural Language Toolkit (NLTK). Additionally, some of them use text parsers to parse sentences and generate output. For example, Grant uses an application called "incr tsdb()" to parse text in order to build language grammars.

Types of Data Used

Data format is a common concern among computational linguists. Potentially useful data may be useless if it doesn't have a usable format and annotation. The dominant type of data used by computational linguists is text files. Our interviewees agreed that PDF and Word files were terrible, but beyond that, text format and annotations were based on project type. Most of the data used in computational linguists' projects are not collected by them. For example, Hill indicated that she used a variety of corpora from Language Data Consortium (LDC). This may take an amount of strain off computational linguists because they don't have to develop and

parse text files. Sometimes, audio files may be used by computational linguists, especially WAV files and video such as MP4.

Size and Availability of Data Sets

While some of our interviewees expressed concern about how much machine learning could be applied to the small data sets available for endangered languages, it was not a universal concern. Grant said that small data sets were not a problem for her approach to machine learning.

Several computational linguists identified data availability as an issue. Not too many linguists are making their data available for public or research use. Our interviewees offered several explanations, including linguists waiting until their annotation and analysis of data are complete (which may never happen), not having time to define metadata and get the data into the appropriate format, and wanting to protect their publication rights. These concerns are explored further in Chapter 5.

The Time Sink of Preparing Data

A major challenge facing computational linguists is the time and effort required to prepare data for analysis. Many of the tools and techniques they use overlap with those of other types of linguists, and are described in Chapter 4.

Computational Linguists Can Help with Annotation

According to our interviewees, one of the potential contributions computational linguists could make to CoRSAL would be to speed up the annotation process by partially automating it. Boss noted that currently, linguists spend a lot of time on annotation that prevents them from having that time available to engage in more high-level analysis. This issue is discussed further in Chapter 4.

Desired Characteristics of CoRSAL

CoRSAL has the potential to address multiple needs of computational linguists. According to Moore, who is involved in the planning process, CoRSAL will be different from other archives in several ways. First, CoRSAL data will be machine readable, and label sets will be harmonized across corpora. Additionally, the CoRSAL team plans to build a model that support multiple data formats. Finally, Moore noted that the CoRSAL depositors will work together to come up with common tag sets, and this will make it possible to perform cross-linguistic research more easily. These features are all new for language archives, which have not addressed the needs of computational linguists in the past.

Hill mentioned a few other issues relating to ease of use. She noted that previews will be critical to the success of CoRSAL as a way of helping researchers quickly see if a corpus meets their needs. She also indicated that it would be very helpful if CoRSAL provided online accessibility to data files such that linguists could edit these files without needing to download them.

Finally, Hill argued that an active developer community is more important in choosing software than almost anything else. She explained that when she runs into issues with software she often has to code her way out. Being able to get quick advice from an online forum is extremely useful. This suggests that a computational linguists forum might be a good idea for CoRSAL.

Other Linguists

Traditionally, most language archives envisioned linguists as their primary user group. The CoRSAL team also considers linguists to be an important user group, although only one of several.

Linguists use archives as a source of the data which form the basis for their analyses and theoretical models. For instance, linguists may use archival data to conduct cross-linguistic comparisons that lead to new discoveries about language typologies. We use the term "other linguists" in this report in order to distinguish this group from computational linguists.

We also distinguish between linguists who use language archives *as a source of data* from linguists who use language archives *to deposit their data*. The latter are addressed in our third user group, "depositors and archive managers."

Interviewees

Five linguists shared their experiences in using language archives and, more generally, in linguists' cultural practices of data access and use. They also provided valuable advice on developing CoRSAL. The five linguists were Mark Post, Stephen Morey, Haj Ross, Robert Henderson, and Frank Seifart.

Mark Post

Post works as a Lecturer in the Linguistics Department at the University of New England, Australia. His research focus is evolution and typology of greater mainland Southeast Asian languages, and he specializes in the Tani subgroup in East Himalaya. He started his field study in Southeast Asia twelve years ago and has worked in the Eastern Himalaya region that has a branch of the large Tibeto-Burman language family. Post has collected data, written grammars, and worked on language maintenance and revitalization materials, including dictionaries and textbooks, working mostly in small communities with 30,000 to 40,000 speakers. Post has some experiences making deposits to ELAR and to PARADISEC. However, he does not use language archives as a source for his own work. He said that although PARADISEC is very helpful and flexible, the lack of automation of the process is cumbersome due to its low functionality.

Stephen Morey

Stephen Morey is a Senior Lecturer at the Department of Language and Linguistics at La Trobe University in Melbourne, Australia. His concentrations are in Asian studies and linguistics, language documentation, linguistic typology and syntax. Morey has deposited data in archives such as DoBeS, ELAR, and PARADISEC. While sharing his experiences in depositing to these archives, he said that dealing with metadata is the most significant challenge.

Morey has studied numerous languages from Northeast India, including those of the Singpho and Turung peoples. He has also examined Tangsa, which is a part of the Northern Naga Subfamily having 80 subtribes. Morey has recorded songs and explanations about the songs from the musicians or singers because he has been interested in traditional songs, particularly the tonal and grammatical aspects of the Tangsa language. He has used ELAN software for his studies because it can play the recording data with the WAV file, which makes it possible to see the recording spot during the extent of the play.

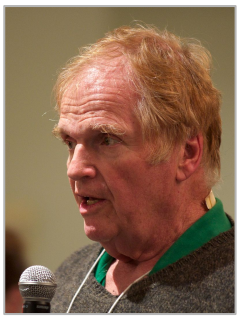
Morey has thirteen deposited collections in the PARADISEC archive. He deposited them seasonally and mostly with WAV and MPEG files. However, he said that because of the way he archived materials in PARADISEC, the metadata are not as easy to identify as one might wish. He noted the significant time commitment required to add metadata to deposits, and argued that without a research assistant, an academic in a teaching and research position would never find

enough time to create the metadata. On the one hand, Morey noted that the value of a language archive is limited without the metadata that describes what its collections are about. On the other hand, Morey said that PARADISEC is much easier for linguists to deposit in than DoBeS, which provides more extensive metadata.

Morey noted that ELAR had recently adopted the IMDI metadata standard developed by DoBeS. With IMDI, it is necessary to create a separate "bundle" of metadata for each recording or set of recordings. This involves not just entering the basic metadata, but attaching to it additional "nodes" containing information about speakers and the content of the recordings. This is done through the ARBIL program. Everything then has to be uploaded through the LAMUS program and then the recordings attached. It is immensely time consuming.

During the transition from the old to the new depositing system, ELAR allowed Morey to submit a well-structured XLS spreadsheet rather than going through the full IMDI process. Morey recommended that language archives that use IMDI should always enable depositors to submit such spreadsheets, to keep the time involved in preparing deposits to more manageable levels.

Haj Ross



Haj Ross is a Distinguished Research Professor in the Linguistics Department of the University of North Texas. His primary research areas are "semantax," which is an interfield that sees syntax and semantics as inseparably interpenetrating, and poetics, the study of verbal art with the aid of specific linguistic analyses of texts. He analyzes hand written data using his own expertise and rarely uses software for analysis. Ross said that he does not use language archives in his research.

Ross has been working about 50 years in syntax, but also studying the grammar of emphasis in German, Portuguese, and especially in English. He said he also has been interested in fast speech in English.

Ross has also examined poems as "placed language." For example, he said that most poems have a vertical left margin, and it turns out the poets place each word in a poem. He has found what he calls "corridors" around which the poet will place words, which are poetically important. They may follow a straight line or a curved line, or even an elliptical line. He said that poetry is not merely for the mind and for the ear, but also for the eye.

Ross has a history of creating "squibs," short notes about phonological, morphological or semantactic phenomena which defy analysis in current theoretical frameworks. These squibs have been partially digitized and uploaded to a website. This is his closest connection to archiving.

Robert Henderson

Robert Henderson is an Assistant Professor at the University of Arizona and specialized in Linguistics and Latin American studies. Henderson has used language archives as a source of data for his research, primarily the Archive of the Indigenous Languages of Latin America (AILLA).

Henderson said that he primarily works on language documentation, while also asking theoretical questions about linguistics. He has just received an NSF grant to work on basic phonetic and documentation of the endangered Mayan language, Uspanteko. He said this includes documenting narratives, conversation, and setting-controlled conversations.

The purpose of this research is to build a phonetic and phonological corpus.

Henderson stated that Uspanteko is interesting because it has a tone system and a pitch accent system that no other language in Guatemala has. So he developed questions such as "Where did this language come from" and "What are some of its properties?" Henderson hopes that this project will help us to better understand these systems and how they work in human languages and help to build this corpus. Henderson also said that he currently has a project on idiophones in Mayan languages.
In his research, he often employs elicitation, which he sits with a microphone and asks people a number of questions. For the Uspanteko project, he asked people to translate words and read lists to understand and record the language. He has also applied story boards when he need to narrate a story or ask people contextual questions.

Henderson utilizes WAV files to save stories and audio. To store transcriptions as text files, PRAAT text grids were used for annotating phonological data. He primarily deposits in AILLA for his studies, including the Uspanteko project, and mentioned that most of his data in AILLA has not been annotated or translated or digitized.

Frank Seifart

Frank Seifart is Assistant Professor of Linguistics at the University of Amsterdam. He has much experience in depositing data into language archives, and periodically uses language archives for his research. He also frequently uses experimental methods, including video stimuli, to elicit the words that different speakers of different languages use.

Seifart's research focus is South American Amerindian, in particular Amazonian languages. He tries to understand and describe the grammatical structures and their relations in terms of linguistic families, common ancestors. He is also interested in how they influence each other from language contact, and how that influences their grammatical structure. He has specific questions regarding the temporal dynamics of speech in Amazonian languages and languages from all over the world. For example, he is looking for reasons why the use of nouns causes people to slow down in their speech, while use of verbs causes people to speed up.

Seifart is currently working on a project about language contact phenomena by asking questions about how and why languages borrow, not only words, but bound morphemes across words. He looks for particular formats of corpus data. For example, he said that he uses DoBeS files because they have the ability of taking time into consideration, and many of them have associated video files.

Seifart has used corpora from language archives for his research on language contact. For other typological research, he said that he uses larger databases.

Findings

All the participants addressed several important issues in using archives. They described limitations, and many of them expressed frustrations in using archives. The most serious obstacles for them were primarily related to data and accessibility problems including interface of the archives. Beyond this, however, were the cultural practices of data sourcing in the field of linguistics, which often do not include consideration of language archives.

How Linguists Obtain Data

The discipline of linguistics encompasses a broad range of methodologies and ways of obtaining data. Data sources range from the "introspection" of syntacticians such as Chomsky, who use themselves as the native speaker whose intuitions about grammatical correctness they test, to data collected by "field linguists" who travel to different language communities and record naturally occurring discourse such as stories. In between these two extremes, linguists may work with speakers of languages who have migrated to the city where the linguists work. They may use structured elicitation methods, such as asking native speakers to translate specific sentences from English into their own language.

The majority of linguists do not look for data in language archives. Partly this is because of the problems with language archives that Chapter 6 examines in more detail. But partly it is due to the history and culture of linguists. Until recently, language archives were not online, so it was often cumbersome to travel to the physical location of a language archive and then make copies of recordings and paper documents. Instead, linguists relied on more accessible data sources. These historically grounded norms still persist.

While we selected most linguist study participants because of their engagement with language archives, we selected Haj Ross to represent the viewpoint of linguists who do not turn to language archives as a source of data. Ross said that he collects much of his data by listening closely to the speech around him.

> "He said that he is very good at listening. He listens to KERA, NPR, and conversations by people. Whenever he hears something strange, he writes down right away. If he listens to something that catches his attention, he wrote it down. He always has two pens and a piece of paper with him." (OL Ross)

Most of Ross's research is conducted on languages he speaks well, so that he can draw on his own intuitions. "He starts in English, but he is very interested in examining other languages for the same meaning; he is very good at German and pretty good at Portuguese. He also learned French and Russian."

At times, Ross has also elicited data from native speakers of languages he does not speak. He and his wife, who is a linguist, were able to work on a language from Mozambique by finding a native speaker when they lived in Brazil. They were able to study Xichangana, the third most popular language in Mozambique, because a native speaker was one of his wife's students.

Limitations and Frustrations Regarding Data in Language Archives

The linguists we interviewed who do have experience with language archives identified a number of concerns related to their ease of use. Chapter 6 examines these issues in detail, but a summary is provided here.

Morey commented that one of the major problems in archives is that the data and descriptive information are difficult and time-consuming to edit, although many data, including the names of groups and the spellings of the languages, change as time goes by. Ross said that he hardly trusts data in archives because he cannot assure the credibility of the transcribed data. He strongly insisted that transcribing has to be done by trained staff because it is a very difficult task.

However, Seifart provided different opinions when asked why linguists don't use archives as a source of data. He said that it is because it's easier to contact the researcher personally. For example, when researchers set up a project that examines ten languages, they just can ask ten people who have compiled relevant data. The researchers know that they want to work with

the ten people, and it is more appealing than going to some online archive to find the data they need.

Limitations and Frustrations Regarding Accessibility or Interface Issues in Language Archives

Most participants strongly emphasized that accessibility and interface of archives are significant factors in using archives. Post insisted that archiving should really be as easy as using Gmail or Facebook. Morey also emphasized that the most significant aspect in using archives is accessibility, and Seifart described that the weakness with archives in general is that they don't have a user-friendly interface. For example, Seifart said that he had to navigate the entire archive to get an idea of what kind of languages are stored because he could not get an overview about the collection of archives. The complicated interface prevented him from effectively using a language archive despite his need.

Furthermore, Seifart said that for every single archive file, he wants to see a variety of information, including author name, year, title of session, date of publication, and publisher or institution, to easily create a citation for the individual records in the archive. While navigating the DoBeS archive, Seifart criticized that it is too complicated to search through an extensive list to discover what is in the transcription files. Seifart said that he could not find out what he was looking for and realized that he would have to request that information from the owner.

Depositors and Archive Managers

This user group includes linguists who deposit their materials in language archives, and the people who manage archives. Most often language archives are managed by linguists, although they could be managed by someone trained as an archivist. Depositors have a different set of needs than linguists who use language archives as a source of data for theory development.

Interviewees

For this user group, we interviewed four members of Chelliah's Lamkang research team. The team is preparing materials on the Lamkang language for deposit in CoRSAL. Among the four participants, Chelliah has the most experience with language archives. She has looked at language archives, but never used them for her own research, or for teaching. The main reason she initially engaged with language archives was for her work as Program Director at the NSF Documenting Endangered Languages program, where she evaluated corpora and archives funded by her program. Reiman has limited experience using language archives, which comes mostly from his work with SIL. Robinson and Utt have no prior experience with language archive.

Shobhana Chelliah

Shobhana Chelliah is a Professor of Linguistics at University of North Texas, and from 2012-2015 she was Program Director for the Documenting Endangered Languages program at the National Science Foundation. Her primary research area is documentary linguistics.

Chelliah plays multiple roles in relation to CoRSAL. First, she is the director of the CoRSAL development effort. She therefore represents the role of the archive manager. Second, she leads the research team that is preparing Lamkang language deposits for CoRSAL. In that sense, she is a depositor. Third, she expects to use data deposited in CoRSAL for research purposes, which puts her in the "other linguists" category.

Chelliah's interest in lesser known languages goes back to her dissertation research, which focused on Manipuri. As mentioned above, her interest in online language archives as a way of sharing data developed through her work as Program Director for the NSF Documenting Endangered Languages program. There, she realized that the materials collected by grant recipients were not useful to the broader linguistics community or to the language communities being studied unless those materials were made publicly available by being put online.

Will Reiman

Will Reiman is Adjunct Faculty and Research Catalyst in Applied Linguistics at the Graduate Institute of Applied Linguistics. He was employed for many years at SIL International, working on language documentation in rural areas of Africa and Indonesia. His primary role on the Lamkang team is ordering a large amount of data into a database, including audio, video, and written data. He works mainly with FLEX and SayMore.

Melissa Robinson

Melissa Robinson is currently a Research Assistant in the Linguistics Department at the University of North Texas. Her work focuses on phonology, analyzing the sounds of words and phrases in the Lamkang language for deposit into CoRSAL. The software she works with is PRAAT.

Tyler Utt

Tyler Utt has a master's in Linguistics from UNT. He has extensive experience working with the Lamkang language. He is interested in understanding how the language works. Most of what he has done has been related to FLEX.

Findings

The Lamkang language documentation project began in a non-traditional way. It started with the community reaching out to Chelliah, rather than vice versa. The chief of a Lamkang village, Beshot Khular, had started to collect language materials such as songs and folktales, because the traditions were dying out. A Lamkang person sent those materials to Chelliah. Chelliah applied for funding to document Lamkang, and after the funding came in, a long-term project developed.

The workflow for preparing the Lamkang data seems to be well-established. Each member of the team is assigned a distinct task. The team starts by collecting materials from the native speakers and backing up the files. There are various types of materials that are collected – stories, songs, conversations, monologues, video recordings, etc. The team then go into transcribing and analyzing what they collected. Depending on the kinds of data that need to be processed, there are different software programs to meet specific tasks.

Many software programs are involved in the research process. For the team, the data is first imported to SayMore. After the metadata is added, everything will be gradually moved to FLEX and other programs. FLEX is used to analyze text files for lexical patterns. ELAN deals mainly with transcription and annotation of video and audio recordings, and PRAAT is used to analyze sound files. Certain data from FLEX can also be exported to SayMore and ELAN. Eventually, everything that belongs to a corpus will be uploaded to a language archive at one take. Chelliah mentioned that "two bottlenecks" of the process are transcription and analysis. This is

an iterative process; the two phases happen together. Each time they come back and examine the file, they find something new.

With regard to language archives, study participants made recurring comments about the need to accommodate a lack of standardization in annotation styles. For instance, Chelliah stated that her desire is for CoRSAL is to let community members contribute, so they could enter data in whatever way they want. Each linguist also has a different way of coding. Therefore, there needs to be a common set of metadata so that everyone can access and understand the uploaded materials.

From the perspective of the archive manager, the sustainability of an archive is of primary concern. While this topic did not come up explicitly in the interviews, we know that it is guiding Chelliah's actions, as seen in her ongoing search for funding opportunities. Chapter 8 discusses the issue of financial sustainability in more detail.

Another issue that is likely to be of concern to an archive manager is the ability to track users. This would help the archive manager ensure that the archive is designed with the actual users in mind, and could identify parts of the site where users encountered problems.

Design Implications

This section identifies design implications that emerge from our research findings. These implications demonstrate how ethnographic thinking provides an interpretive lens that offers new ways to see how linguists' cultural worlds are organized, and offers a framework for thinking about CoRSAL might accommodate their cultural practices of data access and use (Hasbrouck, 2015).

Computational Linguists

Types of Data Used

In terms of file formats, all interviewees agreed that PDF and Word documents were terrible. These formats should be avoided at all costs.

The diversity of computational linguists' research activities means that it would be useful to provide ways for computational linguists to download customized data sets, in customized formats. For instance, multiple computational linguists indicated that it would be useful if CoRSAL created an interface that allowed them to choose a subset of data to download by selecting information type prior to extraction. This could be accomplished if a checkbox or dropdown menu controls were used to enable the desired data to be retrieved. Similarly, Boss suggested that an interface that enabled users to run queries to select a subset of the available dataset would be useful. Generally, the computational linguists were very interested in an interface that would enable them to apply SQL queries, which is not facilitated by current language archives. Grant also pointed out the importance of having metadata available to be downloaded along with original data files.

Size and Availability of Data Sets

The CoRSAL team should keep in mind that some, although not all, forms of machine learning require large data sets. For computational linguists, therefore, bigger should be regarded as better. Improving data availability is addressed in Chapter 5.

Preparing Data and Helping with Annotation

The time sink of preparing data and the opportunity to partially automate annotation are addressed in Chapter 4.

Other Desired Characteristics of CoRSAL

Moore's vision for CoRSAL seems highly desirable for computational linguists. Specifically:

- Make data machine readable
- Harmonize label sets across corpora
- Build a model that supports multiple data formats
- Develop common tag sets across languages to facilitate cross-linguistic research

Other ease-of-use issues that would contribute to CoRSAL's usability for computational linguists include a preview option to allow users to quickly see if a corpus meets their needs, and the ability to edit files online without downloading them (by users who are not the depositor).

Finally, we recommend creating an online community for computational linguists who use CoRSAL, as well as the developers who design the infrastructure. This will make it possible for them to help each other solve problems quickly, which will make CoRSAL more attractive to them.

Other Linguists

The limitations and frustrations our linguists described concerning data and accessibility issues are addressed in Chapter 6. The one issue we address here is the culture of linguists. We encourage the CoRSAL team to contribute to efforts to raise awareness among linguists that language archives are a viable option for finding research data. Perhaps sessions could be organized at the Linguistic Society of America meeting and other relevant venues, showcasing exciting uses of data from language archives. At the same time, designing CoRSAL in a way that avoids the frustrations of current language archives would also help to encourage linguists to consider CoRSAL as a source of data.

Depositors and Archive Managers

We suggest that the CoRSAL team create a separate portal for depositors, due to the big differences between uploading data and accessing data for analysis purposes. The portal would give depositors their own channel to access and modify their data and metadata. We also recommend that depositors be able to easily update and modify their materials, as described in Chapters 5 and 9. A clear set of guidelines for depositing would also be valuable to help the users understand how to prepare their data, as well as how to use the archive. It also keeps all the data on the same boat. ELAR and AILLA are great examples for creating a depositor channel. They have clear guidelines for depositors that include such information as the types of data to be accepted, how they should be prepared, how and where to send them to. Depositors then have their own portal to manage their metadata.

In addition, Robinson suggested that it would be helpful if deposits were accompanied by information about the depositor's annotation system, since there is so much variation in annotation styles.

With regard to the needs of archive managers, design solutions could include a system that helps track users' activity. The system could show how many people have visited the page, or downloaded materials. To obtain more information, we can have users create an account, at least if they want to interact with the archive. The account may also enable users to personalize their interface, which potentially will make for better usability and attract more users. These

methods may also help deepen our understanding of users, like where they are from, what they do, what kind of data they are interested in, and so forth.

COMMON CHALLENGES FACED BY USERS

4. Preparing Data for Deposit

Jenny Hooker and Corderon Jackson

Computational linguists, data depositors, archive managers and other linguists shared a common concern with preparing their materials, be it for analysis or for deposit to archives. Each of the linguist user groups we researched focused on the preparation of data, since this is the quintessential foundation of any field of research.

Understanding how linguists prepare data before depositing or analyzing it is important because it allows us to understand the difficulties of the process, which will in turn allow us to help to improve it and the user experience of CoRSAL. For instance, if CoRSAL could assist in making the preparation of data less time-consuming and less difficult, it would encourage more linguists to deposit more of their data.

Research Findings

Overview: Common Steps in Preparing Data for Deposit

The process of preparing materials was extensively discussed by our interviewees because of how complicated it can be for each individual to find a method that suits their means and their needs. Each person who does research has their own personal methods and style, along with their own programs they are familiar with, as well as their own agenda as to what their data is to be used for.

At the same time, it is possible to provide a general overview of the common aspects of the data preparation process. The following tasks are typically involved:

Recording

Capturing spoken language is the primary focus of linguistic field work, but cultural context plays an important role in future analysis. Recording tools have evolved with advances in technology, with the most common being tape recorders for both audio and video, and more recently digital recorders for audio and video files. Additionally, handwritten notes, film, photographs and digital photographs of people and artifacts are collected by some. The ways in which they are stored and organized depends on the format of the recordings.

Transcription

In order to transcribe, linguists often have to merge separate audio files and video files to have the best quality data to work with. After this step, transcription of audio recordings can include textually representing the nuanced pronunciation of words, as well as writing them in a standardized written form. Some linguists use word documents for transcription, while others prefer software that offers transcription abilities, such as Transcriber, PRAAT, and ELAN. Finding ways to time-align the transcription timestamps with the recordings is often important, but there is no single tool for this and linguists must manually align the timestamps. Transcription quality improves with language competency and, as such, transcriptions can be revisited and improved.

Translation

The translation process includes both word-for-word translation of the transcribed data, and free translation, which entails conveying the meaning behind the spoken words. Like transcription, this step produces more valuable content with greater language competency.

Annotation

Annotatation is an iterative process that is done simultaneously with analysis, which by some estimates can take 10-15 years. The style of annotation is unique to the data, and refers to the act of writing notes about the data. Analytic insights, descriptions of the data, identification and labeling of text segments, parts of words, morphemes, and semantics are all examples of annotations that may be done depending on the goals of the research.

Application of Metadata

The use of metadata is of importance to linguists because it allows those browsing for online resources the ability to more easily understand the potential information which a particular file may offer them. Metadata is associated with a specific file(s), providing information such as who created the file and when, which language is documented, in what location was it gathered, and much more potential information. Software including ARBIL and SayMore are used to organize and associate metadata. Unfortunately, the trend seems to be that depositors find metadata to be time consuming to add manually, while the lack of metadata renders countless files undiscoverable and unused.

Variation in Data Output Formats Across Linguists

Individual linguists each prepare their data focused on different aspects of language, depending on what they're personally researching, and as a result, a wide variety of file formats are used in various processes. Depositors not only upload the raw data files, commonly audio or video files, but also upload their personally done transcriptions and annotations in various formats. File formats are important for both uploading and downloading data, and the multitude of software needed for creating an analyzed corpus requires a lot of file conversion. Different tools provide different file outputs, and different places for deposit require different formats.

> "Basically, I've got these dinosaur versions of all this stuff that would be very useful to people, and it would be wonderful if somebody would say 'well look I've developed this conversion tool, now all you have to do is press this button and it's going to magically put all your stuff there." (OL Post)

> "Morey originally had a Word document containing the full linguistic text with line by line transcription. They [ARBIL] do not allow Word documents but accept PDFs." (OL Morey)

> "Jessica explains that when working with large sets of language data some of her biggest frustrations comes from the different formats. Audio is most commonly WAV files or MPEG, but annotations are largely variable. Text files should never be PDF but otherwise there are a number of tools. She also explains that most of the tools listed below are musical score style interface. This means there will be audio with transcriptions below that are tied to times of the audio files." (CL Hill)

> "The different tools all have different outputs. Each requires a different reader. She explains that she has been on a project in which she switched between 3 different transcription systems because they each had advantages. Started in Word because it is easiest to write in Arabic, moved to PRAAT because it is easiest for syncing time, finished in ELAN because it works best with video." (CL Hill)

Variation in Annotation Styles Across Linguists

Because annotating is subjective, there are a countless number of ways to go about doing it, which makes any attempt at standardizing it difficult for archivists and stifling and limiting for

potential depositors. Among archive depositors and managers, there are repeated references to problems with inconsistency in annotations, and an expressed need for established, effective guidelines on how to identify and label text. Those with experience annotating data described how theirs and other annotations are very specific to the projects during which they were made, making standardization difficult. Linguists generally understand that making use of others' annotated data will likely not suit the purpose of their research and often apply their own, personally appropriate annotations to already annotated data, which, of course, saves each linguist no time.

Several examples from our interview fieldnotes illustrate this variation in annotation styles. For instance:

> "Another of the main points that Dr. Boss focused on that he thinks CoRSAL would do well to work with is understanding the fact that all linguists and all linguistic data is unique, meaning that all linguists will annotate and look for certain data that may not pertain to others' research or standard of annotating/formatting data. Dr. Boss believes that the lack of activity in use of and depositing to language archives is caused by both linguists inability or unwillingness to share data they believe to be of use to other linguists, as well as the difference in linguist's studies rendering others' data irrelevant." (CL Boss)

From Moore, another computational linguist:

> "For the problems of annotation and consistency, she already assume a degree of noise in data and the model used is robust to some amount of noise. Then, she may work on tools for cleaning up the data before moving forward." (CL Moore)

From computational linguist Hill's interview notes:

> "When asked about why most linguists don't use archives she explained it could be related to the problem psychology has that is explained above. She thinks many linguists want data done differently than how others have provided. Collecting data is part of the research process and the types of annotations or assumptions made when collecting the data may influence its value." (CL Hill)

Variation in Metadata Requirements Across Language Archives

Among the linguists who had experience with language archives and archive managers, a common theme that emerged was the usefulness and importance of meaningful metadata being associated with data files. When browsing an archive, metadata can allow researchers to learn more about a certain file in a shorter amount of time than it takes to download and look through it. But, due to time constraints or tedium, depositors tend to only include the minimum of required information to make deposits, such as file name and date created, rather than attempting to include as much information about the file as possible. More willing depositors have also grown frustrated with certain archives and their strictly enforced deposit procedures that require certain information in a certain way. As it stands now, depositors find adding descriptive metadata to each of their files to be time-consuming. As a result, however, their files may be disregarded by linguists who don't have the time to look through them to determine their usefulness. There is also a considerable amount of data on the Internet which is practically inaccessible due to a lack of substantive metadata, and there is no one with the incentive to do this work, which essentially leaves potentially groundbreaking findings out in the ether.

Hill provided us with what she felt to be a list of the most important pieces of metadata that a file can be associated with:

- Language
- Language family
- Genre (story, conversation, etc.)
- Media type (text only, audio, video)
- What annotations are available
- Quantity of data

Representative comments about metadata from our interviews include the following:

> "Regarding a workshop about archives held in January, Morey stated, 'we need to spend far more time talking about metadata preparation.' It was described as 'cumbersome,' 'tedious,' and 'arduous' by multiple respondents, and often the unseen aspects of the labor feel wasteful." (OL Morey)
>
> "He compares ELAR to holding the webpage hostage, where he is being demanded to spend the time to completely redo his metadata, annotations, etc. and with a frustrated laugh calls it "insanity" that he needs to spend so much time to do what amounts to reading off of a spreadsheet. 'Their intransigence in not facilitating that on their side is just insane… and who is being served by this, exactly? I really don't know.'" (OL Post)
>
> "He says he has been more active with PARADISEC because they are more flexible. They will upload items with minimal metadata, so he can enter title and rudimentary metadata. The lack of automation of the process is 'just so cumbersome it's unreal.'" (OL Post)
>
> "When she was asked about databases that she has been in touch with, she described Uspanteko. This database was supported by developers and her colleagues, but has different degree of annotations and needs frequent updating. Also, she admitted that the metadata is necessary to specify what kind of annotations is used in the file and information about annotator." (CL Moore)
>
> "Archive managers are really interested in metadata. Their main thing is whether it can be searchable. Users might also be interested in searchability." (DM Chelliah)

Preparing Deposits Takes an Inordinate Amount of Time

In describing the process of preparing data for deposit, the wish for efficient time investment was another common theme across linguist user groups. Across the various methodologies and processes that each of the linguists practiced, the amount of time that gets sunk into the preparation of data was unsatisfactory. Often, among depositors, the amount of time required to appropriately deposit the data deters many from depositing information all together, meaning a potentially endless amount of linguistic data goes undocumented and unanalyzed. Mark Post, a member of our 'Other Linguists' user group and a professor who works with typology of Tibeto-Burman languages, specifically Tani, estimates that processing four minutes of transcribed text will take two weeks of labor by someone who has a year's experience of analyzing the language. Clearly, the amount of time put into preparing a set of data for analysis is disproportionate to the amount of useful information which can be extrapolated through analysis.

Many of the issues that linguists run into while attempting to make use of language archives are in regards to the amount of time and labor required to prepare less-than-satisfactory amounts of data, as well as the lack of incentives to complete this time-consuming work. This includes inconsistencies and miscommunication between linguists through incomplete or insufficient metadata, the subjective nature of annotation, and the frustrating process of converting files

from one format to another. The poor usability of most archives' user interface forces linguists and depositors to either spend time to learn how to navigate and use the archive or deters them from making use of it altogether. There is little incentive for anyone, other than research assistants, to put the amount of hours needed to write transcriptions or prepare data for deposit as employers tend to not recognize their experience. These transcriptions must be checked for inaccuracies as well, otherwise, other linguists will have to invest time into correcting them or jeopardize the accuracy of their analysis.

> "Linguists don't archive their data immediately as they are 'supposed to' because 'the amount of labor that goes into a single deposit is such that you're not going to want to do it more than once,' said Post. The analysis and transcription are the bulk of the job." (OL Post)

> "Depositors simply cannot keep up with the amount of work involved in depositing." (OL Morey)

Other Disincentives

Our interviewees also identified a variety of other disincentives to deposit their data in language archives. Many of these had to do with career management. For instance, most linguists are not rewarded in performance reviews for preparing deposits. The time they spend on this task is not acknowledged. They often have so many other tasks on their plate that it is hard to make time for data preparation. They are often concerned about releasing data to the public before they have fully published on it, to protect important findings from being extracted and published on by other researchers without any benefit to the depositor. Finally, it is often not even clear to them that anyone will use their deposits.

> "He is frustrated with the constraints on his time, and over the course of the interview, he reiterated themes of time efficiency, unreasonable workloads, workflow obstacles, and cost/benefit analyses for unpaid labor. His position as an academic, juggling teaching, fieldwork, family, and life in general situates him and others with many obligations already." (OL Post)

> "Post recalls his first experience with using an archive which discouraged him considerably. Though attached to a grant, he was discouraged from depositing right away by ELAR because was understood that annotated and analyzed metadata holds more value than raw. Processing took him more time than he had anticipated, but by the time it was ready ELAR had changed the format requirements, requiring that he try to learn Arbil (which was extremely difficult) and then re-do the work. When he submitted again, he was told that there was a relatively simple issue with the format again, and was asked to re-do the metadata for a third time. He reiterates that this is something that must be done over weekends without credit, but he also carries around guilt because of his grant obligations. 'So my Minyong data is now sitting on a hard drive somewhere in ELAR and there's an empty webpage in ELAR that is supposed to have my data there, and we're sort of in this Mexican standoff.'" (OL Post)

> "He expresses that most archival situations feel like throwing data into a 'bottomless dark pit.' One does not know if anyone will access the data, or if they themselves will have access to the data again." (DM Reiman)

> "As a senior researcher Morey has so many tasks that he is responsible for, he does not have time for the meta data entry, but he has to because otherwise it will not get done. Most people in his position will not do it if they do not have to. This is resulting in a large portion of the recording not being archived." (OL Morey)

Preparing Deposits Reduces Time for Analysis

A disappointing reality that many of the linguists brought to our attention was that the amount of time that goes into preparing data for analysis reduced time for performing the analysis that actually interests them. Linguists pour countless hours into transcribing audio and video files, preparing translations, annotating files, and attempting to familiarize themselves with software that crashes. For instance, here is a summary of the obstacles that keep computational linguists from meaningful analysis:

> "There is a lot of intangible labor that goes into the preparation for depositing. Needing to contact software programs directly about features that are not working correctly and trying to learn new software unsuccessfully are but some examples of frustrations one linguist faced. Software (namely, FLEX) crashes frequently, different tools provide different outputs which will require separate readers, and then there is the added time investment of trial and error to find out that PRAAT is good for syncing time. Different formats is one of the recognized frustrations across linguist groups. Poor user interface and unfamiliarity with software left one linguist 'clicking around and hoping you land on what you need.'" (CL Henderson)

Design Implications

Having discussed our research findings, we've come to a better understanding as to the process of, difficulties and frustrations associated with preparing data either for deposit in an archive or for analysis. The issues that were most uniformly discussed by the linguists were inconsistencies in how different people choose to annotate their data and metadata, the unpleasant task of converting data from one format to another, and most predominantly, dissatisfaction with the amount of time and effort that must be put into the preliminary part of research that is preparing the data to be analyzed.

With those main issues clearly identified and found to be widespread the linguistic community, we are able to provide informed suggestions as to how CoRSAL can be designed and developed in a way that can address those issues, promote usability, retain its relevancy as it continues to accumulate data, as well as assisting the language communities in their growth, and, above all, preventing endangered languages from being forgotten.

Considerations for Standardizing Data Format (Output)

Allow Data Deposits to Be Uploaded in as Many Formats as Possible

- CoRSAL shouldn't require that depositors format their data in a particular way, as this can deter busy and unwilling potential depositors from sharing their data due to the difficulties or an inability to convert files.
- The only thing which should be required of data to be deposited is substantial metadata; as long as there is identifiable metadata, linguists will be able to decide to spend time looking into the data on their own. When metadata isn't present and files are unidentifiable or depositors don't know what to expect, the data won't be used.
- Making CoRSAL both easy to deposit to and easy to find data within will ensure that it continues to have data deposited to it and that linguists continue to utilize its corpora. The more easily and often that data is deposited to the archive, the more its community can grow and the more viable for financial backing it appears.

As computational linguist Hill noted:

> "Trying to work within a specific file format will not work for everyone… An active development community is more important than flawless software… A strong UI will attract users, which will in turn lead to more data" (CL Hill)

Enable Users to Upload and Edit Data Using a Browser-Based, Editable Interface

If CoRSAL allows people to work on an interface integrated with the archive and computer model(s), people will be able to upload incomplete data as it's gathered in real time. An activity log would allow users to keep track of edits, would be optimal if backup versions were saved as edits were made periodically, creating different versions of data as time goes on. Would certainly be optimal if compatible with cell phones, particularly for the Lamkang and other communities of the sort.

Offer Guidelines to Future Depositors Before They Prepare Data

Provide general guidelines as to how depositors can best prepare their data to be deposited into CoRSAL, especially beginners and those less technologically savvy.

Allow Users to Select Specific Annotations, Filter Data, and Export in the Data Format That They Wish

Allow users to specify what sort of data they hope to extrapolate from documents, filtering out what's unnecessary to them out and allowing them to download the data they'd like in the format most suitable to them would help linguists save a considerable amount of time wading through information that's irrelevant to them.

Ensure that the format(s) in which the browser-based interface creates documents, as well as the format(s) that the computer models store documents in is as flexible as possible and able to be converted into a variety of data formats would be very important.

As Boss noted:

> "What I work with is from the comp. linguists perspective, it would be nice if people could be given an interface asking what information they'd like to extract, if the person using it can use a checkbox system to say this is what I care about, this is what I want, and this is the format I'd like it in that'd be helpful, much better than forcing people to take data as it comes." (CL Boss)

Considerations for Standardizing Metadata

Require That Certain Fields of Metadata Are Filled for Deposits to Be Added to Archive

Requiring that certain fields of metadata are filled before deposits can be added to the archive will help the frustrations that linguists face when searching for viable data, come across a file with no identifiable traits, and either spend time determining whether it can pertain to their research or not or immediately disregard it. Neither situation is preferable, so ensuring that all data on the archive has metadata associated with it will be an improvement for linguists.

This ties into the potential for computer-mediated automatic processing of data; through feature vectors, individual files of data could be analyzed as they are uploaded to the archive and attempt to extrapolate any metadata that it can. If the computational linguists' models are able to do so, this could prevent depositors who feel as if adding metadata is time consuming from being deterred from sharing their data by doing the busy work for them.

Suggested Metadata Requirements

The most important metadata that can be associated with a deposit, according to computational linguist Hill:

- Language
- Language family
- Genre (story, conversation, ect.)
- Media type (text only, audio, video)
- What annotations are available
- Quantity of data

Specify usage rights; having depositors document how they wish to have their data shared when first uploading it allows other linguists to know from the get-go whether or not they should put time into analyzing other's data. It also allows depositors who have concerns about not being recognized or reimbursed for their research, or privacy concerns on behalf of their community, more autonomy in how their data is released to the public.

> "It'd be great to consider if CoRSAL could allow people to upload their data linked to them and only release certain parts of their data conditionally, saying only certain types of research can be done after a certain amount of time, people need to be able to access most of the information about the data for it to be useful but without running into the original issues." (CL Boss)

> "She also believes that a useful language archive will include usage rights upfront. She explains that it is not uncommon to find data that seems interesting that she isn't allowed to use." (CL Hill)

Allow for Depositors to Create Their Own Fields of Metadata
Provide guidelines and suggestions for customized fields.

Considerations for Facilitating Annotation

Develop CoRSAL's Supervised Computer Model(s) to Actively Learn and Improve at Annotating, Labeling and Sorting Linguistic Material as It Is Deposited into the Archive
The models that power CoRSAL will need to be able to identify, categorize and analyze "very fine grained things", morphologically or phonologically based, grammatical structures, words on their own and in relation the others, etc.

> "CoRSAL really needs to be able to label chunks of data on their own and also different chunks in relation to one another and even across data documents, determine that things spoken/written about in different ways are still identified as a single object/concept" (CL Boss)

Allow Browser-Based Edits and Annotations to Update Data in Real Time

Allow for the Annotations Linked with Data to Be Toggle-able
Allowing users to choose between seeing pre-made annotations or not can aid in their research in either way. Allowing for multiple sets of annotations to be visible at once would also be helpful.

Have Model(s) Focus on Error Analysis to Develop Them to Be as Efficient in Processing Data Deposits as Possible

Provide Broad, Basic Guidelines for How to Create Annotations
If CoRSAL could create guidelines that are broad enough to allow individuals to annotate in their own way, but also make those annotations useful to others, it would facilitate research and networking among linguists.

Also Allow for the Link Between an Annotation Set and Its Creator to Be Retained and Apparent
Along with making broad guidelines on how to annotate, making the creator of an annotation set apparent promotes collaboration among linguists and makes it much easier to get information about the annotation from the source.

Facilitate Edits to Deposits

Allow Depositors to Easily Edit Their Deposits and Metadata
This will allow depositors to make changes to their data and the associated metadata as they go. For some linguists, it would be an advantage if they could work online rather than downloading large corporate before editing them

> "He would like it if SayMore could communicate directly to an archive, and update files the way that Dropbox and other services do 'in the background' once he's changed some settings, so that when he does an annotation it will keep the metadata infrastructure intact, or if he changes the metadata structure, it updates on the archive, then 'now we're talking. Because this is really easy for me to do. Everything that I am doing here is work. Everything is contributing to a final project. None of it is time wasting, none of it is busy work, everything stays there and it can be changed in any way that I want – updated in any way that I want.'" (OL Post)

> "Hill explains that the data collections are so large that she has to do all of her work remotely. She suggested it would be great if you could edit the files online without downloading. She said that at the very least she wants to be able to preview it online. She also explains it could help to solve the issues of people posting to the database in the first place." (CL Hill)

> "An issue regarding the language documentation field for professional linguists, he says, is that 'we're never done with our analysis. Never done.' He has noticed an increasing assumption in the field that it is possible to document a language before it is described 'and to be done with that documentation before you analyze the entire language. This is a fiction with a capital F! That's fiction with all caps, as a matter of fact.' He lightly pounds his hands on his desk, making an audible thump when making this point. He goes on to explain that granting agencies, teachers of language documentation, and students too, often insufficiently appreciate the fact that analysis and documentation need to happen simultaneously." (OL Post)

Opportunity for Computational Linguists to Automate Annotation and Metadata

The linguists with the most to offer as far as potential design implications for CoRSAL were the computational linguists, because of their familiarity with and vision for the future of supervised machine learning and natural language processing. In supervised machine learning, these computational linguists work to develop algorithms that allow models to be given sets of data, create annotations, labels and classifications on that data, and then have a human linguist review and correct the results, which in turn improves the ability of the model to analyze data sets. The same methods could be applied to the application of metadata, allowing depositors to upload their data and fill in the blanks of what the model could not identify and correcting its mistakes, potentially saving time and making depositing feel less labor-intensive.

However, computational linguists cannot process linguistic data until it has been transcribed by linguists, and transcription involves at least some degree of analysis. So the annotation process cannot be fully automated; linguists will continue to play a role.

Nonetheless, if CoRSAL were to develop and implement a model which was able to annotate and label transcribed data as it was uploaded to the archive, as well as assign metadata to it, and that model were to continue to improve itself as more and more data was added to the archive, it would save linguists countless hours of monotonous work which could then be put into gathering more data to deposit or performing analysis on what's available.

> "Boss mentioned multiple times that it's frustrating for human linguists to use their precious time performing tedious, repetitive tasks to annotate and classify data and have less time available to use on actual, high-level analysis on said data. Boss is a huge proponent of moving away from supervised machine learning and moving towards semi-supervised and active machine learning by making use of domain adaptation and other methods which allow the models and programs linguists create to adapt to new, unfamiliar sets of data, making the models better and better at annotating and sorting data." (CL Boss)

Among the methods the model to process data could implement are:

Supervised Machine Learning
Boss described supervised machine learning as creating a model based on algorithms, teaching that model to analyze a representation of the sorts of data you will be working with, having it perform analysis on other instances of the same type of data and making the necessary corrections to improve the model incrementally. At first it requires human effort, but it will quickly reduce the massive amounts of data that once needed to be looked at down to smaller, much more manageable amounts.

Active Learning
Boss described active learning as providing the aforementioned model with a news article it's unfamiliar with, labeling it as such, and notifying the supervising human that it requires assistance. The model sorts articles by theme and places those it cannot into a "maybe" pile, which the human will later sort through, allowing the model to continue to improve itself and minimizes human input.

Experimental Methodology
Boss explained that experimental methodology builds on supervised, active machine learning by providing the model with data that it hasn't been familiarized with yet in order to test its current effectiveness. This certainly requires human input, but the time spent improving the model will save linguists an unimaginable amount of time annotating and processing comparatively.

Domain Adaptation
Domain Adaptation is having a model that has been built to work with data from one domain of information and applying it to another to test its effectiveness, i.e. a model built around sports articles being tested on a political article or a cooking recipe.

Error Analysis
Error Analysis is what it sounds like and straightforward: providing the model being developed with a set of data, seeing what the model got incorrect, identifying why it was incorrect, then determining if there is a feature vector which could be extracted and applied to the model to help improve it. Again, requires human input but has the potential to save human time in the long run.

Feature Extraction

Boss explained feature extraction as working with a data set, essentially a bag of words or any document, and using the fundamental algorithms the model is built on, such as semantic role labeling, predicate-argument structure, etc., to find out how individual features are defined within the context of one another, and assigning meaning to that feature, which in turn, continues to improve the model. For example, determining that the feature 'until' is a preposition, then determining that 'until' is most often used in reference to temporal situations

Feature Vectors

The introduction of 'feature vectors' teaches the model to determine whether or not a certain word or phrase is present in a document, then based on that feature vector, learning to classify documents based on the totality of associated feature vectors and associated labels, identifying patterns and common themes within a document and between documents. This can be a huge help to automatically assigning metadata to files and requires little human interaction once the model has begun to develop.

Other Considerations

Clearly Link Deposits to Their Authors with Citation and Contact Information

- Encourage users to cite deposits like publications
- Make easy for users to reach out to depositors with questions about their data

> "Seifart says that… he wants for every single archive file to have a button for suggested citation, author name, year, title of session, date of publication, publisher or institution." (OL Seifart)

Create an Online Forum for Users to Connect, Discuss Linguistics, Network

Creating an online forum and branding CoRSAL as an language archive with social network capabilities could help CoRSAL expand its niche to different sorts of user groups, as well as provide a home for linguists worldwide to conduct their research, deposit their data and connect with fellow linguists as well.

The use of an online forum system allows individual users to be connected to a unique profile which can be helpful in keeping track of their compiled corpora and their universal updating of it, which saves time that would have to be spent going to each file individually.

It would also be one of the best formats to allow users to communicate with one another in regards to each other's' data, annotations and projects, as well as unique, individual profiles allowing users to gain recognition for what may elsewhere be considered to be incomplete or insignificant contributions to the linguistic community.

Depositors who have gathered culturally relevant information or artifacts could also make use of their profiles and storage space on CoRSAL to store other findings not immediately related to linguistics, as linguists have expressed the need for a location of this sort.

> "However, she identified that there can be issues with annotations that don't match, but noted that through shared tasks, the annotations can be checked by field linguists and updated in synchrony with the computational linguists' use of the data sets." (CL Grant)

> "He would like to archive other kinds of objects. He has thousands of still photographs. He makes a point to say he is not trained in anthropology, but thinks that he should have a place to put his cultural documentation. He believes he has, by far, the largest annotated corpus of photographs pertaining to Tani material culture that has ever been collected, and

knows the function of the objects represented, where they were made, and by whom. He finds it “bizarre” that he cannot archive it. He mentions wanting to include his field notes as well. All of these things being archived, he argues, would be valuable to any community linguist who speaks the language, or would be valuable 100-200 years in the future in case these languages may no longer be spoken. He suggests that extinct Native American languages would be better off in their reconstruction efforts if this type of archive were available.” (OL Post)

5. Linguists Hesitate to Deposit Data in Language Archives

Kenneth Saintonge

Research Findings

This chapter will look at why so few linguists deposit their data in language archives. Out of the thirteen linguists who were interviewed for this research, four had deposited data. They are Chelliah, Morey, Henderson, and Post. All four are documentary linguists who work with endangered languages. Chelliah is motivated to deposit data in order to connect the collected linguistic materials back to the communities they came from. Morey sees depositing as the most important thing that he can do, making the best translations that can make the best descriptions of how languages work. He is helping safeguard materials by archiving analog fieldnotes from languages that have already gone extinct in addition to his own fieldwork. Post thinks it is a good idea to archive if fieldwork is involved and because of grant obligations. Henderson was depositing material for an NSF grant when the interview took place.

Personal Connection to Data

A context for the findings presented in this chapter is the personal connection that many linguists seem to have with the data they have collected. Linguists spend countless hours working on a small piece of a much larger puzzle. This work is often solitary and intimate, creating a bond between the linguist and the data they are studying. A large amount of time and energy are expended for small yet monumental steps forward, with little reward given for the fruits of their labor. As a result, a bond develops between linguists and their data. Due to this intimate nature of their work ,the linguist feels almost as if the data is their child and has to take responsibility for it. This dedication, focus, and total immersion can produce research which linguists see as an extension of themselves, because of their connections to the language, the people, the sounds and the context.Chelliah stated in the interview that linguists often do not really think about using somebody else's data to write something, because data is considered "your data" or "my data"; it feels awkward to take someone else's for one's use (DM Chelliah).

Language community members with training in linguistics may also prepare linguistic materials for deposit. This produces the most intimate of relations to the data, as it is the life blood of their culture and community. Post encountered a situation where language community members resisted depositing their language materials in an archive because it felt to them like data theft: "the whole process of archiving is not only too difficult, it is scary to people, at least in the area we work in. Because it looks to them like theft of knowledge and data" (OL Post). However, as Chapter 2 describes, the Lamkang participants in our study are eager to contribute to a language archive.

Waiting Until Collections Are Complete

One reason linguists hesitate to deposit their materials in language archives is that they want to wait until they are sure a collection and its analysis are complete. Much like an artist who does not want their masterpiece seen until its completion, the linguist is wanting to get their research into a finalized state before being viewable by anyone else. This creates the "archive depositing paradox". Post states:

> "He 'had a couple of problems with archives, in that the procedure for archiving data is really very cumbersome' and 'there's too much of a focus on the get to the final stage and then deposit, and nobody's ever at the final stage. There's no such thing as a final stage.' Yet data are always coming in, needing to be annotated, corrections to be made; a linguist's work is never really done" (OL Post).

Difficulty of Updating Deposits

Linguists' wish to delay deposits until their analyses are complete is partly due to the difficulty of making changes in deposits once they have been uploaded. Almost all language archives have the operational hurdle of once the deposit has been made there is no easy way to change or update information. To change anything, the process has to start from the beginning with the data being uploaded and a request made to take down that old information. This becomes a big problem as language information is constantly changing and needing to be updated. Morey gives this example in the interview:

> "As words are recorded their spelling and pronunciation may change as more information is collected. An informant may not have the correct way to say it and thus this misrepresentation can misrepresent the language as whole. As more research is done and these issues come to light they need fixing. To do this he would have to change the listing of his word where ever it shows up in each actor, individually. In addition, there are general descriptions of every language that Morey has dealt with, this will include regions that the language has been encountered, and socioeconomic information. In some cases, as his research has progressed, this information becomes out of date. Again issues with labeling certain aspects of the metadata. 'The tribes that are being studied have names that are contested, they have names for themselves. There are debates on how to spell these things, this is an issue when the whole idea of archiving is about permanence, structured organization that does not change too much. Once something is entered it should not be changed.'" (OL Morey)

This major limitation of language archives was frustrating and disincentivizing to linguists, stopping them from depositing in the first place. Chelliah has encountered this; "The language community want to see materials from five years ago when they collected it, but linguists are still holding on to it because there is stuff that they need to work out before putting those up" (DM Chelliah). Post also weighs in passionately. "A big, big, big mistake — a colossal mistake" is how he critiques the idea that one-time fieldwork can be documented and archived. Data should be "reworked, refined, looked at from a different angle, and reprocessed. This process can take five to fifteen years" (OL Post).

Protecting the Fruits of Their Labor

Linguists usually obtain recognition for their accomplishments and breakthroughs by means of their publications. They may worry that exposing their data to be viewed by others in the linguistic field prior to publication creates the potential for research to be copied or stolen, wasting hours, maybe years, without any recognition for their effort. Chelliah mentions that "the owner might get upset if somebody else used their data for further research, because at this point in time, there isn't a culture where both sides get credited" (DM Chelliah). Boss, who is not a depositor, is also aware of this fear: "There's a lot of people worried about their data being 'scooped' and written/analyzed about by others and having their papers published before them, it's understandable" (CL Boss). Research being accessed before completion can also benefit another linguist's work, making the former's irrelevant.

It Adds Up to Time

Time is one of the main discouraging factors when it comes to depositing in language archives. For many of the reasons our participants hesitated to make deposits, time was an underlying issue or connected in some way. Many linguists do not even think about depositing because they barely have enough time to complete their own work; why would they want to create more work for themselves? "Why spend time organizing extra your data to be used by others when you can spend that time on your own work?" (OL Post). A single minute of data can equal an hour plus of preparation. This does not benefit the linguist because of all the other reasons mentioned in Chapter 4. Post reiterated this time drain:

> "He explains that for people like him who spend 10-15 years refining analysis, there's 'reluctance' for people to put their efforts into organizing metadata into a format for someone else so that it can be deposited 'because you'll have to undo that work later and go in and do it again.' He states that he and many others are in the situation where 'data is piling up and piling up' with only a small percentage ever being archived" (OL Post).

Linguists are very much for contributions to posterity and future existence of languages, but they have limited time and resources. Furthermore, they may wonder if the effort involved in depositing is worthwhile. "So I would be amazed, frankly, if very many people of at least my generation, have an ability to do an incredible amount more in terms of documentation in addition to whatever else that they do, than I can. Because really we don't get any recognition at all for doing this kind of thing" (OL Post). They would like to have it deposited for themselves at least, but it is also doubtful they will ever return to access it again. "We need to be able to design an archiving process that does operate more like… a backup service and less like a library" (OL Post). Post believes that it comes down to determining the greatest value to both the research and the linguist. "The transition is extremely expensive. I mean, there's a huge amount of data representing years and years of effort in about 4 or 5 of these toolbox projects that I've got, and bring them all into Flex. We're talking about many many days of continuous labor" (Ol Post).

Preparing Deposits for a Particular Language Archive

As Chapter 4 described, the preparation of data is a complex process no matter what field of linguistics you are participating in. Most archives have a particular way that a collection must be set up in order to deposit it. This means that the depositors must arrange the data into the prescribed format set by the archives. Many data do not even make it all the way through the preparatory phase, let alone to the archive, because of the time commitment involved.

One of the major issues is learning the different programs. Post noted that the older generations are less apt to learn this.

> "Another problem with constant changing programs, will older linguists want to keep learning new programs to deposit or give up." (OL Post)

> "People from earlier generations, or who are from places in the world that haven't been as exposed to technology, face similar challenges in terms of their ability to look at a piece of software and immediately understand how it works." (OL Post)

Adding Metadata

Existing language archives have different metadata requirements for deposits. DoBeS is one of the more rigorous ones. DoBeS requires data to be processed through Arbil or ELAN software before it can be transferred to another program which will then uploaded it to the archive. This

process allows the maximum amount of metadata to be uploaded. Metadata entry takes a considerable amount of effort, which acts as a deterrent for potential depositors.

On the other hand, archives that have minimal metadata requirements are less useful to potential users, leading to a different set of disincentives for potential depositors. An archive like PARADISEC is much more simple; it only requires a spread sheet to be sent and they will upload it for you. Post says that PARADISEC is "very helpful and very flexible and so on, but leaves considerable work up to the depositor." Given its limited resources, it also has limited functionality (OL Post). He describes PARADISEC as low-functionality and a "data graveyard" as far as he can tell. He understands that they have limited resources (OL Post). This simplicity is reflected in the lack of metadata present on the archive, which impedes searchability. Metadata entry is difficult for those that are not familiar with the programs and how the process works. "Sometimes metadata functions are hard to use for native speakers, linguists, or people who are just starting out" (DM Chelliah).

Design Implications

To try to remedy the negative experiences mentioned by our participants and to increase deposits in language archives, the following major design points should be addressed.

Make it Easy for Depositors to Update Deposits

Depositors should have more control of the data they deposit. This includes the ability to update or take down data at their will. The updating process should be quick and simple. This would be a profound change to current archiving practices. Chapter 9 discusses our recommendations for reconceptualizing language archives in more detail.

Ensure That Depositors Receive Credit for Their Work

If deposits counted as publications, they might be considered a better use of time by linguists employed as faculty at universities. We recommend supporting efforts to create this kind of recognition for deposits. Facilitating communication between depositors and potential users of the deposits might also help to create more recognition for the depositors.

Protect Publication Rights for Depositors

Develop a plan to protect depositors' rights to be the first to publish analyses of their data. For instance, being able to control who can see their data in the archive would give piece of mind to those who are still working on the data or are not ready to have it fully released yet.

Streamline Annotation, Metadata Tagging, and Uploading

A easier work process is needed for annotation, metadata tagging, and uploading deposits. The interfaces should be malleable to fit the need of multiple types of depositors, from seasoned linguists with a team of research assistants to the language community that may not have that much experience with technology. An easier work process will combat the time drain that was mentioned as one the greatest deterrents to depositing. Focus will need to be put on how to make multiple changes at once, and quicker data manipulation. A greater variety of import and export formats for both annotation software and archives would facilitate many user activities.

Finally, coordination among the annotation software, metadata software, and language archive is necessary for seamless data preparation and upload for deposit. With better program

compatibility, time and frustration would be saved. Though not perfect, a great example is the cohesion of the software families of Office Suite and especially Adobe. Such program compatibility would not only make the depositing process more efficient, but might be helpful to the depositors by creating innovative ways in which annotation could occur. This would benefit language archives in the long run, making them easier for depositors to use and more useful to researchers and language communities.

Add Linguistics Students as a User Group

The CoRSAL team could consider adding linguistics students to its list of targeted user groups, both as depositors and as researchers. Their role as researchers is addressed in Chapter 6. With respect to students' role as depositors, we note that linguistics programs are starting to offer courses on language documentation, such as the one that Shobhana Chelliah is teaching this semester (fall 2016). In Chelliah's course, students learn how to prepare linguistic materials for deposit in language archives through hands-on data preparation activities. Similar courses are offered at other universities, and are becoming more common.

Build Community with Depositors

Finally, we recommend for CoRSAL to build community with depositors through a forum and perhaps also an advisory board. A depositors' advisory board would encourage depositors' engagement with CoRSAL and ensure that their needs were recognized in CoRSAL's ongoing development. The board might even be able to take on some responsibilities in terms of managing the archive.

An online forum could bring together depositors, the CoRSAL team, and users of the archive. Users' access to archival materials could be coupled with their ability to communicate with depositors and the CoRSAL team. A bonus to this structure would be the creation of a living archive that would maintain its usefulness and relevancy to users and depositors as it evolved over time.

6. Linguists Don't Use Language Archives to Obtain Research Data

Brittany LeMay and Melanie Medina

As a key user group, it is essential that the design of CoRSAL addresses the common challenges linguists face when using language archives as a source of data for their research. This study included interviews of thirteen linguists. The majority of them have not used language archives as a tool for their personal research; only two of the linguists have collected data from archives. In order for CoRSAL to be as useful to this potential user group as possible, it is essential not only to analyze and understand why lack of use is a trend among linguists, but what the design implications might be for encouraging future use of CoRSAL data. Those who were interviewed gave a variety of reasons why they did not utilize language archives as a source for gathering data for their research. After expressing their perceived limitations of language archives, several of the interviewees gave recommendations of what they believed would make a language archive a more useful research tool for them and other linguists. An array of solutions was given, ranging from interface design, to search functions, to systems of metadata. In order to make a more useful language archive, it is important to take into consideration what each user group needs.

Research Findings

The Two Who Have Used Language Archives to Obtain Data

Only two out of the thirteen linguists we interviewed had used language archives as a source of research data. These two were Robert Henderson and Frank Seifart. Henderson uses AILLA (Archive of the Indigenous Languages of Latin America) to access audio files and clause-by-clause transcriptions. He downloads mp3 files from AILLA (since the WAV files are too large to be stored on the site) and then converts them to WAV format in order to work with them in PRAAT. Henderson's area of study is Latin American indigenous languages, and AILLA functions as a language archive that specializes in this area. Seifart utilizes corpora from TLA (The Language Archive), ELAR (Endangered Languages Archive), DoBeS (Documentation of Endangered Languages), and AILLA. With the data from these archives, he does cross linguistic comparisons and studies language contact. As a depositor, Seifart has a working knowledge of archives that helps him get the data he needs.

Difficulty of Finding Relevant Data

Our interviewees indicated that linguists are discouraged from using language archives in the place of other sources due to the disadvantages associated with them. One of the most prevalent of these disadvantages is that the data in the archives are difficult to find and access. Several interviewees noted how poor interface design made it difficult to navigate within archives. Not only does this make it hard to find the data, but it is quite time consuming as well. When linguists have to spend an unfeasible amount of time digging through an archive in order to find data relevant to their research questions, they end up deciding to return to their usual sources instead. It can be especially disheartening when the linguist has to go through the extensive time and effort trying to locate relevant data, only to learn that they are being denied access to it for a variety of possible reasons.

There are two features of language archives that can contribute to these inconveniences: the archive's interface and its search function. These challenges are explored in more detail in Chapter 7. In his interview, linguistics and Latin American studies specialist Robert Henderson (one of the two out of thirteen interviewees who have used language archives for their research) described a language archive that he has used for his work that has incurred these issues. In his experience collecting data for his research, the AILLA interface did not enable him to view the available materials for each language prior to selecting and loading the page for one of the languages. In the words of Henderson, this interface involves "a lot of clicking around and hoping you find what you need." This can be a time-consuming process that results in the researcher going back and forth between webpages to find if the archive has what they are looking for.

Linguistics assistant professor Frank Seifart (the second interviewee who has used archives in the past) described another interface feature that he feels can be an impediment to linguistic research. He refers to it as an "unfolding tree-structure system." This system requires the user to click on every section in order to be able to view the entire tree, which serves as an additional example of a language archive interface being difficult to navigate and therefore being time-consuming for the user to operate. With interfaces such as these, even if the linguist does manage to ultimately find the resource they were hunting for, another inconvenience often presents itself: access restriction. Seifart echoed Henderson's feeling regarding complicated interfaces in stating that language archive users frequently have to "click through some complicated structure, then you'll get to some final node where you expect the session, and then you don't have access or there's nothing in there." In order to circumvent this hassle, linguists can attempt to directly contact the depositor to gain access to the data. At this point, a situation has been created in which a linguist must resort to means other than language archives to gather research data.

Another problematic feature associated with interfaces is the search function. Linguists have expressed in their interviews that the search functions of language archives tend to be inadequate tools for navigating data. Even if an archive did have data relative to their research question, it would still be of little use to the linguist if they were unable to effectively search for and locate that data. According to the interviewees of this study, it is challenging to have a search function that is simultaneously easy to use and all encompassing. Applied linguistics faculty member Will Reiman conveyed that archive search functions should be easy to use (which would benefit members of the language speaking community) yet still able to do in-depth searches for research purposes. Jessica Hill, a computational linguist who uses data from the Language Data Consortium (LDC) in her research, found the LDC's search function to not be effective and chooses to instead conduct her own search by sorting the data by year and utilizing her browser's find tool. During her interview, linguistics professor Tara Grant disclosed that she would prefer a SQL (Structured Query Language) interface over a GUI (Graphical User Interface), but she was not aware of any language archives that currently have that. An insufficient search tool also amplifies the aforementioned issue of time consumption. If a search tool does not adequately meet the needs of its user, the user will have to spend notably longer amount of time finding what they came for.

Another reason why linguists don't use language archives for research purposes that should be considered (aside from the characteristics of the archives) is the content stored on them. Several interviewees stated that they do not use language archives as data sources for their research because the data in the archives is not relevant to their studies. In fact, when asked about his past experience with language archives, Tyler Utt (a linguistics graduate student assisting with data preparation for CoRSAL) simply replied that he is unaware of any language archives that would be pertinent to his current work. Computational linguist Thelma Moore supported this

statement in her belief that it is not easy for linguists to find data from other researchers that is relevant to their own research questions. Hill gave further credence to this by comparing linguists to the field of psychology in the sense that they do not usually share data due to the fact that they had a particular research question in mind when collecting that data.

In addition to the relevancy of data, the quantity of data is of concern as well. Hill commented that most endangered language archives do not have enough data to be of use in research, and those that do are still subject to formatting issues.

Attitudes Towards "Ownership" of Data

Computer science and engineering professor Franklin Boss proposed what he believes to be the reason behind the paucity of deposits in language archives: linguists' lack of willingness to share data, and the perception that differing language foci render their data irrelevant to others. Additionally, Shobhana Chelliah, linguistics professor and principal of CoRSAL, pointed out that the current culture of linguistics does not equally credit depositors and authors of publications, so the owner of a set of data may be displeased if another linguist used their data to advance their research. During her interview, she exemplified this when articulating "I guess we don't really think about using somebody else's data to write something about. Because it's often thought about as 'your data' or 'my data' or 'your data' and like how dare I go and write on Anh's data." This may discourage linguists from using someone else's data even if they come across linguistic data relevant to their own research interests.

File Format and File Size

The interviewees in this study have spoken of the file sizes of data being an additional hesitation in their use of language archives as a research source. Henderson pointed out that AILLA refrains from storing WAV (waveform audio file format) files on its website due to the large file size of this format. For his research, Henderson uses the software PRAAT to analyze sound files. However, this program does not allow for the annotation of MP3 files, which means that he needs to download the MP3 file available on AILLA and then convert the file to WAV, and then finally import it into Pratt. If he wishes to get the WAV file directly, he will have to contact AILLA. Either way, the archive is presenting roadblocks that are slowing him down from obtaining the data he wants for his research. Linguist Mark Post voiced a similar concern with file size. He does not use language archives as data sources for his research because of large file sizes. He goes on to explain how this issue can inhibit field linguists as well. Poor Internet connections in the field make it nearly impossible to download these massive files or at least download them within a reasonable amount of time. Once again, we are seeing the trend of language archives inconveniencing linguists and being time-consuming.

Metadata, File Naming, and Annotation Standards

Numerous linguists said that metadata is an integral part of their work, so the lack of protocol for metadata in language archives can be a deterrent for using them as a data source for research. Linguist Stephen Morey provided a description of what he believed metadata should consist of in a language archive. He characterized metadata as being the "name of the item, then the sort of who, where, what, how, when type of thing." He went on to explain that the "when" and "where" are easy because they tend to be captured by the recording device. The "who" is something that can be recorded in the field notes by listing who is present. He feels it is also beneficial to ask the speakers other relevant information about themselves, such as their age, where they're from, and what languages they know. Morey believes that the "what" can be somewhat broader; it can be anything from the title of the story that was told by the speaker, to a full annotated transcription of the event. He also stated that it is better to collect more metadata at the time of the event, but the more you collect, the longer it takes to sort through

all of the metadata. It may be more cumbersome to collect and sort through large amounts of metadata, but Morey holds the conviction that it is better to have these additional data in case they can be beneficial in the future, as opposed to neglecting to record the data for the sake of time. Furthermore, uniform systems of composing and entering metadata can encourage linguists to deposit their work into archives. This increases the quality and quantity of data found in language archives, which positively contributes to their viability as research sources for linguists.

Another important area for standardization is file naming. Morey informed us that he has used 4-5 different file naming systems over the course of his career. He believes that this warrants more discussion within the linguistic community. Morey advised that "we need to spend far more time talking about metadata preparation," and that this process starts with file naming.

Linguistics graduate student Melissa Robinson explained in her interview that in her experience, there are different guidelines for data annotation and that linguists have to explain them in their publications. This lack of universal annotation and metadata protocol makes archive work all the more complicated. A linguist accessing someone else's data in a language archive would have to dig through the depositor's publications in order to comprehend the annotation style and gain the information they are looking for from the data. Additionally, this lack of protocol can deter linguists from depositing their work into archives, thus contributing to the lack of quantity of data in some archives, which in turn also contributes to the difficulty of linguists finding data in archives that is relevant to their research. This supports Morey's notion that the manner of annotation and metadata entry employed by language archives can dictate how successful they are among linguists. He believes that the purpose of archives is access to the data; if one cannot locate the data because of lack of useful metadata, or fully understand the data because of incomprehensible annotations, one is not able to truly access the data, therefore contributing to the hesitation of linguists to use language archives in their research.

Usability

The usability of a language archive is largely dependent on making the data and features linguists need for their research easily accessible. Post declared that "archiving should really be as easy as managing Gmail. It really should be. If it's any harder than that, you've already lost the battle." As described above, a contributing factor to an archive's ease of use is its approach to metadata. Having standards that meet the needs of all user groups, including both the language communities and linguists, can greatly aid users in finding what they are looking for.

Design Implications

Our research findings lead to the following design ideas that would encourage more linguists to use language archives for their research.

Enable Users to Easily Find Relevant Data

As Chapter 7 examines further, the visibility and accessibility of data are important to a successful language archive. These issues can be addressed through the design of language archive interfaces. Ineffective navigational structures such as the "unfolding tree-structure system" that Seifart mentioned can be an impediment to the usability of an archive. It would be beneficial for language archives to provide more effective overviews of their data. This would allow linguists to more efficiently determine whether or not language archives have data relevant to their area of research. Making the size and format of the data clear would have a similar benefit as well.

Facilitate Citation of Deposits

We recommend making the citation process much easier for linguists who are using data from a language archive as a research source. Many academic publication databases, such as university library websites, have a button to click that provides citation information for a selected resource (often providing options for different citation styles as well. Seifart recommended that language archives employ a citation feature similar to this one.

Communication Between Researchers and Depositors

Seifart and Henderson (the two linguists in this study who have previously used language archives for their research) have both contacted a depositor directly for data, rather than downloading it from the archive. In fact, Seifart stated that he felt it is simpler to just contact the depositor. If language archives presented an easy way to directly contact a depositor for data that a linguist wished to have access to for their research, many of the problems plaguing archive use would be addressed. Large, cumbersome files (and different file types) have made the process of using language archives for research difficult for linguists. Being able to contact the depositor would make it easier for researchers to obtain access to the exact file they want in the file type they need. This would ease the burden of archives having to host large files on their websites as well. Encouraging contact between depositors and researchers in this manner could also increase interaction and data sharing among language archive users, thus positively contributing to the quality and quantity of data in these archives. If archival interfaces provided an easily accessible manner in which to contact depositors with questions and requests, linguists might be more likely to look to language archives as a means of finding others with similar research interests.

Metadata and File Naming

When addressing the issue of file size, it is also important to take metadata into consideration. Time is a precious resource for linguists — metadata should be able to efficiently indicate to them whether a set of data is relevant to their research prior to going through the time and effort of downloading and sorting through it. Having a universal file naming and metadata system within an archive (and optimally, across multiple archives), would benefit all aspects of the archiving process. It would be easier for depositors to determine what metadata to use when making deposits, would save researchers many valuable hours that would otherwise be wasted digging through unrelated data, and would make it easier for linguists to find data relevant to their research interest. Facilitating this process will make language archives a more effective tool for those working in the field of linguistics.

Annotation Standards

We recommend utilizing the findings of the 2009 workshop on cyberinfrastructure in linguistics that identified best practices for annotation. The following best practices were identified at the workshop (Bender 2009, 14-16):

Consistency/Reliability

- A uniformity of the range of annotations used in the archive. All of the annotation types used in the archive are a part of this set of annotations.

Usability

- Having effective annotation tools available for depositors to work with when analyzing and depositing their work.

Resilience

- Holds up to dispute among annotators regarding different interpretations of annotation rules.

Accountability/Responsibility

- Having a transparent connection between annotations and the associated data. Encourages and allows for those who use the work to credit annotators of the data.

Interoperability

- The annotations are still useful when taken out of the archive's context and being used for analysis.

Extensibility/Adaptability

- The annotation style is applicable to data from sets other than the archive it is typically used in.

Add Linguistics Students as a User Group

Finally, our interview findings suggest that the CoRSAL team should add linguistics students to its list of targeted user groups. Robinson, who is involved in the CoRSAL project, is a graduate student herself and emphasized the importance of language archives being usable to students. In her past educational experience, she had little experience looking at actual linguistic data and said she would have greatly valued a chance to "take something that we learned at book level and use it on a real data level to see why it is important to learn about this stuff" and to learn "what does it really look like when dealing with a language?" Chelliah gave an idea of how an archive could provide this experience to students: "For something usable for the American classroom, it could be something that aids the professor on how to teach their students, because the language archive should know their materials really well, and they can tell you how to use it to train the students." Educating students in the use of language archives would provide the next generation of linguists with a useful new tool in their repertoire. It would benefit the linguistic community as a whole developing a new set of linguists familiar with the use and benefits of language archives. This familiarity would help them provide better data to language archives and encourage more data sharing among the community.

7. Navigation Needs in CoRSAL – Interface and Search

Molly Blair and Sebastian Barnes

Introduction

When developing any product it is critical to investigate the how the interface will impact the usability of the product. When developing CoRSAL we feel it's critical to consider how the information design will impact all users. The goal of CoRSAL is to serve both language communities and the depositors and linguists that want to study their languages. Our analysis suggests that each user group has different needs, and if we hope to create the best archive possible we should work to meet the needs of every group. This chapter will begin with our findings about the interface and search tools on the site, and then we will follow up with the design implications we have developed based on these findings.

Research Findings

Linguist Interface Needs

All three groups of linguists emphasized the need for an effective interface. There were a number of different ways linguists felt the interface could be improved. Both Grant and Hill suggested that a strong user interface would attract more users over time. They explained that a good user interface would encourage more language community members to participate in CoRSAL, and the large data set and user interface would attract linguists to study those languages. Therefore we think it is crucial to devote time to creating an easily used interface.

We begin by discussing our findings in regard to linguists. The linguist groups continuously echoed their frustrations with looking through language archives to find what they needed. Morey, Grant, and Reiman all indicated frustration while looking for data on archives. Their complaints all focused on the lack of information about the files available for download. Metadata is at the heart of many of these issues, and all of the solutions will require strong metadata to work appropriately. The major issues we found fall into three areas: the complicated interface, challenges in communication, and inefficient downloads.

Complicated Interface

Many of our interview participants suggested that the interface of large archives required knowledge of advanced tools to be useful. Many linguists are not tech experts, and we should make sure that anyone would be able to easily access the entire site. Creating an effective interface would ensure that people could use the resources regardless of their technical ability. Ross suggested that he wanted to be able to explore resources without knowing what he wants. He says that the archives he has used before either were challenging to use or mostly were based around search. Without effective browse tools it is much harder to find new resources to study.

Communication Challenges

Another reason the majority of the participants had not used archives for research was because they felt they were more complicated to use than communicating with field researchers. The

prevailing opinion was that a field researcher can quickly explain the intricacies of their notes. Grant specifically explained that when "working with something in an archive if I don't have a way of communicating with the linguist who produced it, it is much more mysterious." When using an archive, any confusion about the content or annotations must be resolved by the person who downloaded it.

Inefficient Downloads

When Hill was assisting a graduate researcher with a project exploring language archives, she noted that before her student could evaluate the content, she had to download it. Often times this download process required her to go to individual pages to download each file. Even when she navigated to each file, there was no information about the files, so it was impossible to judge if the file was worth downloading in the first place. As a result, to test if any of the archives had useful content, this student was forced to jump through a number of hoops before even analyzing the data. To improve our archive, we need to make it easy to download multiple files at once while providing information about the files being downloaded.

Language Community Interface Needs

The needs of the language communities are significantly different from those of the linguists. Therefore it is critical to provide different resources for each user group. Many linguists including Post, Robinson, Reiman, Hill, and Grant recognized this issue. They all suggested solutions ranging from different portals for each user group, to different browsing tools based on preferences.

Their suggestions seem valuable for future language communities, and would be very useful for non-linguists who are interested in these languages. Unfortunately the language community we worked with, the Lamkang, has very limited internet access. As a result, a fantastically designed site will have very limited effect on the community itself. Therefore, if we intend to service this community, we must also develop a means to create offline materials.

Linguist Search Needs

All three groups of linguists identified that search is a necessity, in some form. Of the thirteen linguists interviewed, eight mentioned the need for active search. Three others identified necessary metadata relating to what needs to be searched, but did not explicitly state how they wanted the metadata to be searched.

Ease of Use

Essentially, linguists identified that the most important aspect of a search function was its ease of use. The searchability of a language archive can be a deciding factor in whether or not the linguist will use the archive or move on to other resources. Hill described a situation where one of her students was gathering data from language archives. The student used the most searchable archive because it was easier to find the information needed. Hill noted that other archives might have more suitable information, but without being searchable, the information stored in an archive is stagnant.

Henderson told us that currently, language archives are not easily searchable. Users are subjected to interfaces where they are "clicking around and hoping to find what [they] need," rather than having the information easy to find and search. This method is not only time consuming, but inefficient. Essentially, the current interface of language archives is akin to trying to find a specific product on an e-commerce site without a search option.

Search Function Usage

What makes search easy to use varies among linguists. Chelliah, Hill, and Reiman recommended a search interface that allows users to see the extent of searchable content. Hill recommended a live search that allows users to see results as they're typing.

However, Grant and Ross identified queries as the easiest function. These queries include SQL queries, as suggested by Grant. A SQL query is a piece of code that finds content within a data set using four parameters (Goldstein 2005).

- **Select:** Identifies which categories a user wants to view (ex. depositor, format, length)
- **From:** Identifies the group of content a user wants to view (ex. Lamkang)
- **Where:** Narrows the field by only including content requested (ex. Format = audio)
- **Order By:** Organizes the content based on user preference (ex. Order by length)

While SQL queries ensure that the user is getting exactly what they are looking for, the issue is that users who are not familiar with that input are limited in their searches. One challenge with search is to balance the various needs and search styles of different types of linguists, but this can be overcome by offering an interface with multiple search options (as we will discuss in a later section).

Search Challenges

The biggest challenge in developing a search function is that in order to ensure deposited data is searchable, the process of depositing data becomes inconvenient. Metadata needs to be thoroughly applied during depositing. This makes uploading harder for depositors and contributors because sufficiently adding this metadata takes time and effort. However, this doesn't mean that ease of depositing has to be sacrificed for better search. Rather, CoRSAL needs to implement a solution that automatically tags certain metadata fields.

Lamkang Search Needs

Unlike the community of linguists, the Lamkang community did not mention search features, likely because the Lamkang community does not have regular Internet access and would prefer data to be housed in a brick-and-mortar archive. If the archive is not digital, then the search functionality is widely different. Before we can assess solutions for search, we need to better understand whether members of the Lamkang community will use the archive to find specific sources, or if it will be more for browsing.

Design Implications

Interface Design

Making CoRSAL a usable tool will require powerful browsing options that service both the language community and the linguist groups. To create an effective interface we need browsing tools driven by metadata. Ensuring we have workable metadata will be crucial in ensuring that we have an interface that can be useful to anyone. While linguists generally showed more interest in search there were still calls for a system to look through the data on the site. After we have determined the breadth of metadata we intend to use, we could conduct a card sort to better understand how people would naturally navigate the site. A card sort consists of giving potential users notecards with the pages planned for the interface. The participants then sort them into groups that make sense to them. By basing our interface on the feedback of our users we can ensure it is easy for them to use (Usability.gov n.d.).

In addition to using metadata to inform our architecture, we need to make sure metadata is easily accessible. The most common complaint about language archives was the amount of work required to determine if the content was even usable. By including information about the quality of the audio and the annotations we can assist linguist in finding useful data. A quicker, more informative interface would allow linguists to find exactly the data they want without wasting time looking for it. To facilitate an easier way to browse all of this metadata we suggest that linguists have access to a checklist browse system (Figure 1). Such checklist systems are common on shopping websites. This would allow linguist to refine what they are looking for without being limited by the information architecture we develop.

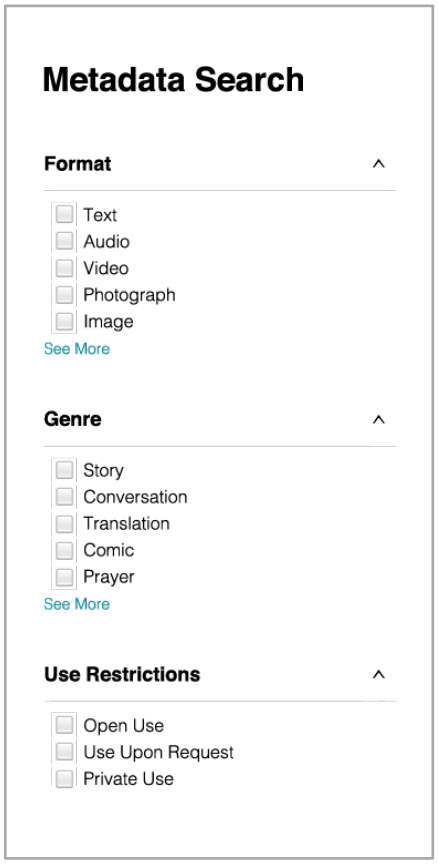


Figure 1: Checklist based browse

While making data easier to find and analyze on the fly is very important, there are other reasons linguists are uninterested in using language archives. Because it is common for linguists to have different opinions on the best practices for annotation, communication between researchers is critical. Three interview participants suggested that they prefer to work with field researchers, because it is easy to communicate with them. To facilitate this process we suggest creating a messaging system within CoRSAL so that researchers can easily communicate between one another. To keep people accountable for their uploads we also suggest a rating system that tells users how quickly depositors respond. This will make it much easier to feel confident in research before downloading it.

While these features are great for linguists, they don't address our greatest interface challenge. The Lamkang language community needs some way to access the data we've collected. Community members had a number of suggestions about the best tools to facilitate language

and cultural learning in their communities. The first requirement would be to create a backend that can process CoRSAL's data in a number of ways. If our data is rigidly tied to the files it was uploaded with it would be very challenging to create workable files for community use. Using an XML database or something similar to control data uploads would make it much easier to output useful files.

Some of the suggestions of the community include libraries with learning materials and comic books. To enable creators to develop such tools, we need to make it easy for users to explore concepts that would relate to language learning materials. If a user could easily search for specific types of verbs it would be much easier to find good examples for a lesson. To enable the Lamkang and other communities to use our resources we need to consider how we can make CoRSAL's data usable off-line.

Search Design

Since search is essential to using CoRSAL, the search function should be easy to access as well as easy to use. Since finding data on the archive is an integral part of using the archive, the search bar needs to be available on every page of CoRSAL. Generally, users look in the header for this search bar, so to make it easily accessible, it should be in a place users expect to look (Morrison 2015).

Advanced Search Function Design

Since linguists want various search options, CoRSAL needs an advanced search page. The page needs to include listed metadata search terms, a live search option, and an option for inputting a query rather than a basic search feature. Figure 2 is a wireframe to identify the elements needed to create this page. The intent of the wireframe is to explain how these various search formats can work on one page, rather than being a precise image of the final page.

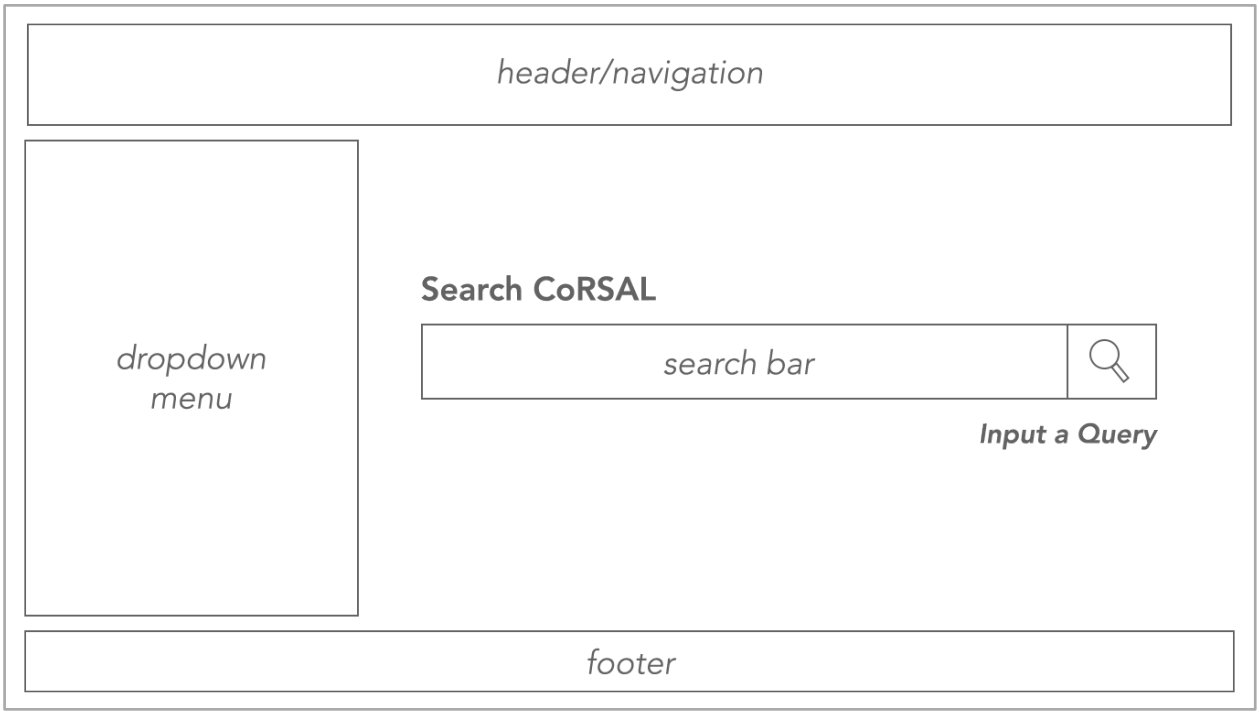


Figure 2: Search UI Wireframe

The purpose of the side navigation is to show all the available metadata. Drop down menus would provide clear categories for each type of metadata, while checkboxes would allow users to search by only selected parameters. Figure 1 from earlier in this chapter shows an example of what this drop down menu should look like.

Essentially, the search tool needs to be easy to use, but shouldn't tie users to one search method. Some prefer a more guided search while others want to be direct. This search function combines different users' searching preferences without cluttering a page of content.

Streamline Metadata Tagging for Easy Uploads

One potential solution to solve the ease of use/searchability issue is to utilize supervised machine learning to automate certain types of metadata tagging. Supervised machine learning entails training a program to take input variables and create output variables (Brownlee 2016). In this case, the input variable is the file, and the output variable is a certain piece of metadata such as format, length, file size, and depositor. This essentially means more time is spent on descriptive metadata that increases the searchability of the deposited data. The challenge then becomes developing a program that can screen input files and output accurate metadata based on a set group of variables.

Improve Quality of Metadata

In order to help users find data that is useful to them, it will also be important to improve the quality of the metadata. A thorough list of metadata fields should be developed that takes into account the needs of each user group, in particular the typical range of types of searches carried out by members of each user group.

CONCLUDING REFLECTIONS ABOUT CORSAL

8. The Financial Sustainability of CoRSAL

Aaron Davis

The topic of this chapter is a bit different from the others in that it was not an explicit focus of our research questions. Nonetheless, concerns about the sustainability of CoRSAL clearly underlay comments from some study participants, especially language community members. Financial sustainability was also a significant focus of discussion at the February 2016 NSF-funded Workshop on User-Centered Design of Language Archives that was co-organized by Wasson (Wasson et al. 2016).

Research Findings

One of the central goals of archives that house materials on endangered languages is long-term preservation of those materials. Ensuring the safety and preventing the loss of this valuable information is paramount. Preservation is a key value for both documentary linguists who collect such data, and for the communities whose languages are at stake.

Preservation depends on stable, long-term funding of language archives, and unfortunately, this can be a challenge. Archivists have noted that "a language archive's need for guaranteed, long-term sustainability is a poor fit with the dominant funding model of short-term grants for specific projects" (Wasson et al. 2016, 29). This issue was examined at the workshop on User-Centered Design of Language Archives.

> Workshop participants identified several limitations of the U.S. approach to funding the development of language archives. One issue was that funding takes the form of "soft money", meaning short-term grants, usually for three years. Gary Holton pointed out that "within the U.S., at least, this is the way we fund science… this *is* the model." Yet the concept of preservation is central to the notion of a language archive, and for preservation, the temporal horizon is not three years but hundreds or thousands of years. Mandana Seyfeddinipur argued that "the problem [with soft money] is that's not sustainable, this is not something that will save the archive. This is something that gives you money for a certain amount of time" (Wasson et al. 2016, 33).

At the same time, the U.S. lacks a central archive where NSF-funded scientists could deposit their materials. The Smithsonian collects some kinds of artifacts, but not language data. So the responsibility for long-term preservation is put in the hands of each language archive.

We might say, then, that the "elephant in the room" for CoRSAL is the long-term sustainability of the project. Since the goal of this language archive is the continued preservation and documentation of endangered languages, it is a project that will need to be supported in perpetuity. Sustainability is largely a financial question, but it also relates to the project's organizational structure and governance.

Funding

Funds are needed to support the following functions:

- Server space and software to ensure the availability of the collected data
- Technology support staff for trouble-shooting and continued updating of the infrastructure
- Creation of physical documents, as per earlier design recommendations, which are printed and available for use within language communities due to their infrastructure restraints

- Archiving staff to process and supplement data as it is deposited, respond to user questions, and possibly help in the creation of educational materials
- Grant management staff to apply for funding, manage grants, and engage in other fundraising activities

We advocate that funding should not sought from users, as creating a pay wall would discourage linguists from using the resource, and is unfair to the language communities that have specifically asked for our help in preserving their language.

It is likely that the funding of the CoRSAL infrastructure will be an ongoing activity, and not one with an end goal. While the maintenance of the infrastructure is likely to be considerably less than its start-up costs, there will be a continuous need for upgrades, staffing, and the addition of new projects.

Organizational Structure and Governance

CoRSAL's long-term sustainability will also require a well-functioning organizational structure. As more people become involved in the project, there will be a need to define roles and organizational relationships among team members. One danger is putting too much responsibility into the hands of one person, where if that person for some reason becomes suddenly unavailable, the work process of the whole team flounders.

Another issue to consider is the value of guidance and advice from a variety of stakeholders, as well as ensuring open communication with those groups. For instance, a concern from the Lamkang community that has been raised was that of wondering where the work is going. With the archiving process unclear to the community, there has been a lack of transparency about how the CoRSAL team may be working toward the revitalization goals of the community.

Design Implications

Funding

It appears that funding for language archives usually combines support from a host institution and the repeated receipt of grants from various sources. Host institutions range from single universities, to consortia among several universities, to private foundations. There are also single-language archives that are hosted by the indigenous community whose language is being preserved and revitalized.

The obvious host institution for CoRSAL is the University of North Texas, since the four central members of the CoRSAL team are all UNT faculty. We suggest that the CoRSAL team might eventually approach UNT for expanded financial support, as the project develops further. CoRSAL could be positioned as a source of positive publicity for UNT, enhancing the university's reputation for cutting edge research and technology.

Another long-term possibility to consider would be developing a consortium among several universities. PARADISEC provides an example of such a consortium.

> PARADISEC is a consortium of three universities: the Universities of Sydney, the University of Melbourne, and the Australian National University. Operational functions are distributed across the participating campuses.

> PARADISEC is directed by a Steering Committee of representatives from these three universities, with Dr Nick Thieberger as the PARADISEC Director, and Prof Linda Barwick as the Sydney Director.
>
> At the University of Sydney PARADISEC is hosted by the Sydney Conservatorium of Music, our University of Melbourne base is in the School of Languages and Linguistics, and at the Australian National University PARADISEC is hosted by the ANU College of Asia and the Pacific (PARADISEC 2016).

A consortium might be useful because of its relative stability and the wide array of project goals for CoRSAL. With the resources available through multiple universities, depending on the commitment required of a host institution, it might be possible to alleviate some of the operating cost of the archive (through hosting servers, having departmental resources or staff available to process data).

We suggest that grant support be sought from a mix of private and public sources. It might be most efficient to have a staff member dedicated to managing incoming grants, as well as soliciting private donations and applying for new grants. Below is a starter list of possible sources for grants to support CoRSAL:

- National Endowment for the Humanities
- National Science Foundation
- UNESCO (under their partnerships for building capacities to protect, promote and transmit heritage)
- Foundations such as the Gates Foundation
- Private donors

Especially for foundations and private donors, it would be a good idea to dedicate a web page to the finances of CoRSAL, and how that money is being spent. Showing the operating costs of specific projects would allow potential donors to have a better idea of what their money would be used for, and the needs of CoRSAL.

Organizational Structure and Governance

As the CoRSAL team expands, we recommend the development of a clearly defined team structure, and articulating roles and responsibilities for each team member. We suggest avoiding excessive dependence on a single person.

We also suggest that an advisory board with representatives from all major user groups and other stakeholders would be valuable. The advisory board could provide insight and advice about directions CoRSAL should take, problems encountered by user groups, and so forth. It could function as a communication tool to ensure that CoRSAL's actions were transparent to all user groups, especially the language communities. How representatives for such a body would be chosen from language communities might best be left to the communities, but it seems appropriate to ensure that they have representation.

A final question to consider is "Who owns CoRSAL?" This is particularly relevant to language communities, who may worry about losing ownership of and property rights to their data. The issue has come up for many indigenous groups whose linguistic and cultural materials have been deposited in archives and museums. The CoRSAL team might wish to consult experts on how to ensure that CoRSAL protects the rights of language communities.

9. What is a Language Archive? A Reconceptualization

Robin Cole-Jett

The traditional view of a language archive is of a controlled and static repository with a primary function of preserving a language. Words and their pronunciations, uses, and attendant relevant information (such as recording, photographs, and stories) become artifacts that the archive handles as separate objects. However, while this approach works well for historical materials, whose authors and creators lived in a past time, it is not a good fit for language data. The problem with this approach is that a language is not an artifact. If it is still spoken, it is a living entity - not unlike an organism - that is constantly evolving. To treat it as a finished product is counter-productive to both linguists and language communities. Linguists continuously re-assess their transcripts, translations, and annotations as their analysis evolves, and community members would like to be able to engage in an ongoing process of collaboratively annotating language data with their insights (Wasson et al. 2016).

This chapter is informed by the author's background as a historical archivist.

Research Findings

Engaging the Language Communities

Many studies have documented the implicitly colonial practices of archives that historically disenfranchised indigenous communities (First Archivists Circle 2006, Povinelli 2011, Zeitlyn 2012). Traditional archivists operate in a "top-down" approach. Mathur (2000, 92) describes that historically, archives served as repositories of small contributions that aided in the understanding of "large-scale social and cultural systems." This viewpoint is predicated on the European idea of the nation-state: all knowledge frames group identity within the prevailing power structure. Leach and Wilson (2014) explain that because of this Western view, the knowledge of indigenous groups is deemed insignificant because their contributions do not fit into the established parameters. Since indigenous knowledge may not align with the knowledge systems of those in power, archives may treat indigenous contributions as a series of unconnected artifacts rather than as elements of a coherent cultural narrative that sits apart from the Western epistemologies.

Traditional archivists assign significance to the items not based on what the object means to the community from which it was collected, but how the potential user of the artifact might gain knowledge from viewing it. This method relies on the archivist as the final arbiter "to consciously or unconsciously assert chosen narratives as truth while ignoring or reframing others" (Shilton and Srinivasan 2007, 88). The colonialism apparent in this method is obvious – the archivist becomes the person who assumes the power over the collection's purpose, meaning that they assert authority over the kind of memory that is then distributed and interpreted by the artifacts.

One goal of the CoRSAL project is to counter the "systematic disenfranchisement" of indigenous groups (Shilton and Srinivasan 2007, 89) that characterizes traditional archives. Indeed, the research team's engagement with the Lamkang community was initiated by the community itself.

> "Usually, the linguist or anthropologist goes and collect things, and they know exactly where things are – they have it transcribed, and have it in their notebooks. The Lamkang project was built from a different perspective, which was untraditional. It began with the community reaching out to the team and sending in materials." (DM Chelliah)

Yet the focus of the CoRSAL team on creating an online language archive might inadvertently have disenfranchising effects if it does not take the local context of the indigenous groups into account and commit to meeting their needs on their ground. Our research with the Lamkang community found sharp differences between their needs and Western linguists' assumptions of what a language archive should be. Perhaps the most basic difference was the *form* that language materials should take. To Western linguists, an online archive seems like the obvious form. Yet, as Chapter 2 described, an online language archive may not be of much use to the Lamkang because of their limited Internet access. As D. Tholung points out, access to technology is highly restricted.

> "The main concerns of the project in terms of this sector are: the lack of internet access and resources or equipment… There is one internet shop in the entire Chandel district, they must pay to use it." (LC D. Tholung)

Our interviewees told us that physical locations in which they could learn and study their language and culture would be more useful to their community. Their vision aligned with the category of Type 4 language archive described in Wasson et al. (2016), where an indigenous group creates a community center that typically houses not only a language archive but also cultural heritage materials, computers, meeting rooms, and other resources for the community.

For instance, D. Tholung expressed his desire to create an actual building to store and retrieve cultural artifacts:

> "The resources we would need to create a brick and mortar building to house these resources would need volunteers to map out and go from village to village to collect resources to build." (LC D. Tholung)

In either an online environment or in a physical building, the Lamkang community might want to include "educational tools" to study their language and culture. Chelliah noted:

> "If depositors could create small educational tools that go along with their archive that it would be much better sold. It's up to us to create… usable tools from archive, but they haven't thought about that yet." (DM Chelliah)

Our exploratory research with members of the Lamkang community revealed important insights into the perspectives and wishes of this language community. At the same time, it was limited to three study participants. It is possible that Internet access will improve for the Lamkang community in upcoming years, and it is also possible that Lamkang community members other that our interviewees may be more oriented toward using the Internet. We note that the Facebook group "Lamkang Spelling Workshop," created by Shobhana Chelliah, has attracted 370 likes. While it is not currently very active, a 2015 post received 10 comments, most of which appeared to come from community members. Chelliah regularly communicates with a few community members via WhatsApp.

We encourage further user research with additional Lamkang and beyond that, with the other communities whose languages will be deposited in CoRSAL over the period of the PIRE grant. Other communities may have the same needs as the Lamkang, or they may have different

needs. For some, mobile apps might be the most useful way to access linguistic materials. Others might have needs that we have not yet conceptualized.

In terms of the development process for CoRSAL, we should keep in mind that the language communities' lack of access to technology and reliable Internet services may limit their ability to fully collaborate with the Denton-based CoRSAL team in a participatory research and design process. To the extent that the Denton-based CoRSAL team is committed to such a process, meeting this challenge may require creative ideas and effort.

Engaging Linguists

Archives as Counter to the Workflow of Linguists

The linguists interviewed for the project did not share enthusiasm for the traditional view of an archive, which does not seem to accommodate their workflow. As Chapter 4 described, our interviewees explained how their *work is constantly in progress* through continuous annotations and re-evaluations. The logic of archiving requires that the data that are created and deposited are authoritative and final; however, this is not how linguists work.

Post, a linguist, emphasized that analysis is never complete.

> "We're never done with our analysis. Never done… and to be done with that documentation before you analyze the entire language – this is a fiction with a capital F! That's fiction with all caps, as a matter of fact." (OL Post)

Robinson explained that annotations change when new guidelines are developed. This re-iterates the point that annotating "is never done."

> "I have had to re-annotate because we set up different guidelines. The last one I did had sixty words, and each one of those words was done three times, so it sixty times three. And that's just one recording." (DM Robinson)

Since linguists do not see their deposits as finished or complete, the amount of work that requires a "finished product" to be deposited can be daunting, as Morey indicated.

> "Depositors simply cannot keep up with the amount of work involved in depositing." (OL Morey)

Consequently, the linguists interviewed expressed a desire for their deposits to be endlessly editable. The need for refining the data they deposited (i.e., maintaining control of the workflow) was very important to them.

Chelliah desired a system in which a depositor could modify and refine her deposit.

> "She would prefer a system that allows us to go back to the file that they have uploaded and edit, or do a different version. There is always a typo or reanalysis or rehearing. So it would be nice to pull things down, make another version and upload again." (DM Chelliah)

Post wants to extend an easy process to all linguists who want to refine the objects/ artifacts deposited.

> "We need to design an archiving system that has a built-in expectation for progressive refinement on various levels… He would like to see depositing in stages: Initial deposit that

> is "very easy to update", an intermediate, and maybe a final deposit where you state it is finished." (OL Post)

The Concept of an "Archive" Might Be the Wrong Model for CoRSAL

In using the term "archive" to describe CoRSAL, the implication becomes that the project will be developed and run according to standard archiving practices. In this model, the role of the archivist is to determine what will be deposited into the archive, and conversely, what information can be removed from the archive.

To ensure that the processes involved in accession (depositing) and deaccession (removing) are applied in an ethical manner, the archivist maintains authority over the structure of the archive, follows procedures congruent to her profession, and advocates for the deposit itself. Upon placing the deposit, the linguist loses control of the work and cannot edit or refine the contribution.

The term "archive" might, therefore, put off scholars who do not want a final resting place for their contributions. The logic of archiving implies that their scholarship becomes "dead" until a researcher brings it to life when their data is used in some form or another. In other words, the depositor (in our case, the linguist) loses control over the contribution once it is accepted by the archivist.

Metaphors of What Linguists Would Like CoRSAL to Be

The linguists we interviewed described their own ideas of a more interactive approach to a language archive using metaphors, analogies, and examples. They favored approaches that allow for the sharing, refining, and control of data.

Post wanted the archive to be intuitive and user-friendly, and compared the notion of archives to commercial communication tools.

> "So one of the things that I tried to hammer home at this meeting and I'm not sure that it really was heard, was that archiving should really be as easy as managing Gmail. It really should be. If it's any harder than that, you've already lost the battle… It should be as easy as Facebook, as everything that everybody in the world is using right now." (OL Post)

Robinson used books in her research, but saw an online archive as more of a database that could provide more detailed information than a book.

> "We have books… which document a language…. it is important to have other people's information, what they have studied. We go to this book… if we had a database, with access to a lot of languages and a lot of details, it would probably be easier to find the information. It allows me to further my research in another language." (DM Robinson)

Seifart recommends taking the "Switchboard" corpus of English as an example of a much-used corpus that also worked well with his studies and that challenged the notion of a traditional archive.

> "It would be very useful for designers of archives to look at the Switchboard corpus of English… Switchboard is a way to set up a language archive successfully and entice researchers to use the data… Its strength is that it is so accessible and modular that it has been used repeatedly for linguistic studies." (OL Seifart)

Documenting, or Curating, the Workflow

According to the linguists interviewed, language archives do not always document how the materials in their collections have been captured, preserved, or annotated. Archivists call the documentation of the work deposited "curation" – the deposit is described as thoroughly and meaningfully as possible.

Curation should include the following information (at minimum):

- Name, professional status, biography of the depositor
- Formats of the deposit (file formats)
- Scope of the deposit (number of entries, languages documented)
- Methodology of annotations (a statement of methods, including reference guide)
- Finding aids (tags to identify sections of the collection)
- Easy retrieval: downloads and uploads
- Tracing logs that generate a historical file of use

The linguists interviewed indicated that curation was lacking in language archives, though they did not specifically use this term.

For instance, Morey noted:

> "It is necessary to have a document explaining the format of data." (OL Morey)

Ross believed extensive tagging of a deposit was essential to access.

> "Knowing how to tag each squib so that it is accessible for all different kinds of queries. You never know when you [are] making an archive. You don't know all the questions you are going to want to ask about it." (OL Ross)

Seifart, who contacted linguists who deposit archival materials directly in order to ascertain their methodology, desired extensive documentation so that all the information he required would be available immediately.

> "[Seifart] talks to the person whose data is on the archive and gets the data from them… it's been easier to contact researchers who do deposits personally…
> He wants overviews. He wants that if he clicks on a language from the left (clicks on Guarani) to say that there are 300K words of text corpus which are transcribed into Latin script and translated into English and glossed by morpheme and with pathological speech or something along those lines. That's what he needs to know." (OL Seifart)

Like Seifart, Grant also desired in-depth documentation of a deposit.

> "Grant identified that aside from annotation styles, computational linguists want formats and explanations of what is included (for example, audio should include how much is transcribed, how long, what is included, how much looks like IGT, etc). There should also be further information and demographics on the speakers. It should also include percentages of language use." (CL Grant)

Seifart knew that the more information a deposit (object, artifact) revealed, the more useful it was to the researcher.

> "Seifart says that there are clear indications for individual records for the archive is that he wants for every single archive file to have a button for suggested citation, author name, year, title of session, date of publication, publisher or institution." (OL Seifart)

Design Implications

By acknowledging, documenting, and sharing the workflow of linguists as well as for community members, CoRSAL can avoid the colonial nature inherent in creating archives. CoRSAL should also create a system of curation/documentation of deposits that are extensive, vital, and do not require linguists to make assumptions or guesses on the materials. Thus, in considering the implications from the research and potential design of CoRSAL, a few key recommendations should be considered.

Position CoRSAL as a New Type of Resource That Is Not a Traditional Language Archive

Clearly state that CoRSAL does not follow the logic of archiving; that it instead permits endless, easy annotation of deposits. The model of a relational database is more appropriate. This will mitigate linguists' reluctance to use CoRSAL. CoRSAL may also be re-conceptualized as an "exchange," "tool," or "initiative."

Be Aware of Inherent Colonialism

CoRSAL should consistently and constantly "check" itself to ward against the implicit colonial biases of archiving and indeed linguistics. In designing CoRSAL, the needs and desires of the language communities should be identified and furthered. To do so, user research should be conducted with all communities whose languages will be deposited in CoRSAL during the PIRE grant period, and designers should be reflexive in the design process. This reflexivity should be noted in the design specifications as well as in the site documentation.

Accommodate the Workflow of Depositors

CoRSAL should function as a relational database in which linguists can easily and repeatedly deposit, curate, annotate, and edit all of their work. A log that indicates revisions should be automatically created and unchangeable. The log should also be reflexive, meaning that the depositor must explain what changes were made to the initial deposit.

Curation of CoRSAL is Key to Its Success

Curation is a total and deep documentation of each deposit, and is vital to making the exchange useful and meaningful.

- Depositors should be in complete control of the curation using a standard input format, such as a fill-in form with open-ended questions and the ability to upload data that supports methodology and annotation.
- A "tag" generator should identify key words that then can be used in the search function. This tagging function should also be available "free form" to the depositing linguist as well.
- Information that must be contained in the curation documentation (at minimum) are: name and affiliation of depository; language; types of annotations; annotation methodology; file formats, including versions; demographics of language users (a map can be helpful); research interests of depositors; glossing. This list is in no way comprehensive, and should be treated as an initial brainstorm.

CoRSAL has the ability to re-define the concept of a language archive into something that is more usable and meaningful for all user groups. Accommodating the workflow of linguists,

consciously avoiding colonialism, and deeply curating the deposits will enable a dynamic and useful exchange of knowledge that goes far beyond current models.

10. Summary of Design Implications

This chapter pulls together the design implications from Chapters 1-9 into a single list, summarizing the most important points. Many of these points were addressed in more than one chapter.

1. Replace the "Language Archive" Concept with a New Model

We suggest that CoRSAL should pioneer a new model for language archives. The concept of an "archive" and its associated practices are a poor fit with the work practices of linguist depositors. While the logic of archiving requires the deposit of a completed, unchanging artifact, linguists engage in a never-ending process of updating and revising their transcriptions and annotations.

CoRSAL (and ideally all language archives) should permit endless, easy annotation of deposits. A model of data storage that is dynamic and interactive, such as a relational database, would be more appropriate.

2. Create Portals for Each Major User Group

Given the contrasting needs of the major user groups, we recommend creating three main portals. The portals would target:

- Language community members
- Researchers
- Depositors

Within these main portals, there could be further levels of customization, for instance for:

- Different language communities
- Different types of researchers

3. Design Implications for Language Communities

We invite the CoRSAL team to make a sincere effort not to unwittingly reproduce colonialist relationships that are, frankly, the historical norm between archives and indigenous communities. The following recommendations can greatly assist in this effort.

Participatory Research and Design

- We strongly recommend for the CoRSAL team to make it possible for language community members to contribute to the archive while also being active participants in the design and development of CoRSAL interface features
- Include linguistic and cultural materials of interest to language communities in CoRSAL
- Empower community members to develop materials
- Engage both older and younger generations

Accommodate Local Technology Constraints

- Provide hard copies

- Partner on seeking funding for local space and tools

CoRSAL Infrastructure

- Provide tutorials on language preservation and use of CoRSAL
- Use English for CoRSAL interface language
- Protect sensitive materials

4. Design Implications for Researchers

Enable Researchers to Easily Find Useful Data

The most important design aspect for researchers is making it easy for them to find data that are useful to their research. This is a major weakness of many existing language archives. The ease of finding useful deposits depends on CoRSAL's interface design, search function, and preview capabilities. Chapter 7 provides detailed recommendations.

Computational Linguists

Types of Data Used

- Avoid PDF and Word documents at all costs.
- The diversity of computational linguists' research activities means that it would be useful to provide ways for computational linguists to download customized data sets, in customized formats.

Size of Data Sets

The CoRSAL team should keep in mind that some, although not all, forms of machine learning require large data sets.

Other Characteristics Desired by Computational Linguists

Moore's vision for CoRSAL seems highly desirable for computational linguists. Specifically:

- Make data machine readable
- Harmonize label sets across corpora
- Build a model that supports multiple data formats
- Develop common tag sets across languages to facilitate cross-linguistic research

Other Linguists

We encourage the CoRSAL team to contribute to efforts to raise awareness among linguists that language archives are a viable option for finding research data.

5. Design Implications for Depositors

Accommodate Workflow of Depositors

- As stated above, CoRSAL should function as a relational database in which linguists can easily and repeatedly deposit, curate, annotate, and edit all of their work. A log that indicates revisions should be maintained.
- Allow deposits to be uploaded in as many formats as possible.
- A clear set of guidelines for depositing would be valuable, especially for novices. For instance, depositors should describe their annotation style as part of their deposits.

Protect Symbolic Capital of Depositors

- Ensure that depositors receive credit for their work. Include citation information for each deposit, and encourage the field of linguistics to treat deposits as equal to publications.
- Protect publication rights for depositors.

Partially Automate Annotation and Metadata

- There is an opportunity for computational linguists to at least partially automate annotation and metadata assignment of deposits. This would save depositors time and ultimately increase the volume of deposits.

6. Design Implications for CoRSAL Manager

Ensure Thorough Curation of Deposits

Curation is a total and deep documentation of each deposit, and is vital to making CoRSAL useful. Metadata is a key tool for curation.

Track User Activity

The design of CoRSAL could include a system that tracks users' activity. This would help the manager know what parts of CoRSAL were the most useful and where users were encountering problems.

7. Add Linguistics Students as a User Group

The CoRSAL team could consider adding linguistics students to its list of targeted user groups, both as depositors and as researchers. With respect to students' role as depositors, we note that linguistics programs are starting to offer courses on language documentation.

Educating students in the use of language archives as a source of research data would benefit the linguistics community developing a new set of linguists familiar with the use and benefits of language archives.

8. Build Online Community Among Stakeholder Groups

We recommend creating an online community to facilitate communication among depositors, researchers, language communities, and technical support staff. Advantages include:

- Researchers would be able to contact depositors if they have questions about the data
- Participants would be able to help each other solve problems quickly, which would make CoRSAL more useful
- It would help ensure that CoRSAL's evolving design was targeted to the needs of its users

9. Long-Term Sustainability of CoRSAL

The "elephant in the room" for CoRSAL is the long-term sustainability of the project. Unfortunately, there is a conflict between the very long time horizon of language preservation,

and the very short time horizon of grants. Chapter 9 provides suggestions for how to start addressing this issue.

Appendix: Interview Guides

The following pages contain interview guides for these four user groups:

- Lamkang community members
- Computational linguists
- Other linguists
- Depositors and archive managers

Interview Guide: Lamkang Community Members

- Briefly explain project
- Go over informed consent notice, answer any questions

- Please tell us a little bit about yourself. Where do you live? What is your town like? What do you do? Do you have a family? [What is the most interesting thing that has happened to you in your life/in the last year…? Small talk to build rapport…]

- What does the Lamkang language mean to you? Why is it important to you?
- Do you use Lamkang in your daily life? If yes, what contexts do you use it in?
- What other languages do you speak?
- Who do you talk to in Lamkang?

- What sources do you currently use to learn about Lamkang language/culture or engage in Lamkang language/cultural activities?
    - Books
    - Websites
    - Local organizations and events
    - Education
    - Friends/family
    - Etc.
- Is there a library where you can access materials about Lamkang?
- Who are the main users of these resources, aside from yourself?
- Overall, what resources and learning tools do you think are currently the most effective for members of the Lamkang community?

- What are the most important issues about the Lamkang language facing your community today?
- How could Dr. Chelliah's research team be useful to your community with regard to these issues?
- Are there resources for language learning or language maintenance that you wish you had, that Dr. Chelliah's team might be able to provide? [e.g. grammars, stories, spelling, etc.; books, video/audio recordings, websites, mobile apps…]
- [Explore spelling issue, whatever else emerges]
- Who do you see as the main users of these future resources?

- What are the technology constraints for Lamkang speakers? Do people have internet access? Do they have computers? Do they have smartphones?

Interview Guide: Computational Linguists

- Briefly explain project
- Go over informed consent form/notice, answer any questions
- If face-to-face, obtain signature on one informed consent form; and give the other copy to the research participant to keep

- What is your background and training in computational linguistics?
- What do you think is fascinating about computational linguistics?
- What is its importance?
- How would you define computational linguistics?
- What are some of the overall goals of computational linguistics?
- What kind of research questions does this field focus on?
- Describe methods used in computational linguistics

- What are some projects you have worked on?
- What were major "ahas" from these projects?

- What do the linguistic data you use have to look like? What format? What annotations?
- Do you use data other people have collected, or have you also used data that you collected yourself? Please describe the different kinds of data you have used in your work.
- What characteristics do the data you use have to have? Are there specific kinds of annotation you need? If so, what are they?

- What databases or language archives have you used?
- Do you remember the first time you used a language archive or database? What stands out in your memory?
- What languages have you examined?
- Have you done cross-linguistic comparisons? If so, please describe.

- What kinds of software do you use for your analysis?

- We are interested in learning about the strengths and weaknesses of the databases and analysis software you currently use.
- Let's start with the databases. What are strengths and weaknesses? What do you wish you had that you don't? What kinds of problems do you encounter, and how do you work around those problems?
- Then the software. What are strengths and weaknesses? What do you wish it would do that it does not? What kinds of problems do you encounter, and how do you work around those problems?

- We would like to do a walk-through of your methodological process, where you show us the different databases and software programs you use.
- So, to start with, what database would you like to show us? Maybe the one you use the most, or that you are using currently?
- [have them do a walk-through of the database; ask for clarifications as needed; strengths, weaknesses]

- Now let's navigate the most common analysis software programs you use
- Which ones would you like to show us? [maybe ask for description of one particular research project and the software programs that were involved]

- [have them do a walk-through of the programs; ask for clarifications as needed; strengths, weaknesses]
- [if relevant] What were the key research findings from this project?

- We understand that not many (non-computational) linguists are using language archives as a source of data. Why do you think this is?

- How will CoRSAL be different from the databases/language archives you've used before?
- Do you think a language archive interface could meet your needs easily, while still catering to language community members?
- What do you see as the potential benefits of CorSAL?
- What advice would you give to the team that develops CoRSAL, to make sure it's as useful as possible?

Interview Guide: Other Linguists

- Briefly explain project
- Go over informed consent form/notice, answer any questions
- If face-to-face, obtain signature on one informed consent form; and give the other copy to the research participant to keep

- What is your background and training in linguistics?
- What do you think is fascinating about linguistics?
- What is its importance?
- How would you define linguistics?
- What kind of research questions do you focus on?
- What kinds of methods do you use in your research?

- What do the linguistic data you use have to look like? What format? What annotations?
- What languages have you examined?
- Have you done cross-linguistic comparisons? If so, please describe.

- Have you used language archives before in your research? If so, which language archives? What did you use them for?
- What were their strengths and weaknesses? What did you wish you had that you didn't? What kinds of problems did you encounter, and how did you work around those problems?
- Do you remember the first time you used a language archive? What stands out in your memory?

- Do you use analysis software? If so, please describe. What are strengths and weaknesses? What do you wish it would do that it does not? What kinds of problems do you encounter, and how do you work around those problems?

Whatever is relevant of questions below

- We would like to do a walk-through of your methodological process, where you show us the different language archives and software programs you use.
- So, to start with, what language archive would you like to show us? Maybe the one you use the most, or that you are using currently?
- [have them do a walk-through; ask for clarifications as needed; strengths, weaknesses]

- Now let's navigate the most common analysis software programs you use
- Which ones would you like to show us? [maybe ask for description of one particular research project and the software programs that were involved]
- [have them do a walk-through of the programs; ask for clarifications as needed; strengths, weaknesses]
- [if relevant] What were the key research findings from this project?

- We understand that not many linguists are using language archives as a source of data. Why do you think this is?
- What features would entice you or other linguists to use a language archive? How could it be designed to meet your needs?
- Would you be interested in using a language archive like CoRSAL if it was designed to meet the needs of computational linguists, other linguists, and language community members?
- What advice would you give to the team that develops CoRSAL, to make sure it's as useful as possible?

Interview Guide: Depositors and Archive Managers

- Briefly explain project
- Go over informed consent form/notice, answer any questions
- If face-to-face, obtain signature on one informed consent form; and give the other copy to the research participant to keep

- Please tell us a little about yourself and the project you are working on. What is your background (in linguistics)? What is the goal of the project? What is your role on Shobhana's research team? How long have you been doing this? What are your plans for the future?

- We are interested in learning about the experience of linguist depositors and archive managers. We understand that you are preparing Lamkang data to be deposited, so we want to ask you about your experience

- Overview of research team's work: what is involved in depositing the data? What are all the different steps in preparing the deposits?
- [SayMore, FLEX, PRAAT, ELAN, file naming, etc.]

- Your own work activities: what are you working on? What types of files? Audio, video, photos, transcripts, other?
- Can you walk us through the parts of the work process that you work on, and show us what you do with different software programs?
- [do walk-through]
    - What are strengths and weaknesses of current systems?
    - What problems do you encounter, and how do you work around those problems?

- Can we spend some time observing you as you work on a current task?
- If it's OK, we'd like to ask you to narrate what you are doing
- [observe, ask questions whenever you are not sure what they are doing]
- At end, ask, what was easy? What was hard?

- We understand that not many linguists are using language archives as a source of data. Why do you think this is?
- What other language archives have you worked with? What did you think were the best/worst? Why?
- Do you remember the first time you used a language archive? What stands out in your memory?

- What advice would you give to the team that develops CoRSAL, to make sure it's as useful as possible to depositors and archive managers?

- Do you remember the first time you used a language archive? What stands out in your memory?

References

Bender, Emily M. 2009. "Cyberling 2009 Workshop: Towards a Cyberinfrastructure for Linguistics: Workshop Report."

Brownlee, Jason. 2016. "Supervised and Unsupervised Machine Learning Algorithms." *Machine Learning Mastery*. http://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/.

First Archivists Circle. 2006. "Protocols for Native American Archival Materials." http://www.firstarchivistscircle.org/files/index.html.

Goldstein, Jackie. 2005. "Writing SQL Queries: Let's Start with the Basics." *Microsoft TechNet*. https://technet.microsoft.com/en-us/library/bb264565(v=sql.90).aspx.

Hasbrouck, Jay. 2015. "Beyond the Toolbox: What Ethnographic Thinking Can Offer in a Shifting Marketplace." *EPIC Forum*. 10 March. https://www.epicpeople.org/beyond-the-toolbox-what-ethnographic-thinking-can-offer/

Leach, James, and Lee Wilson. 2014. "Anthropology, Cross-Cultural Encounter, and the Politics of Design." In *Subversion, Conversion, Development: Cross-Cultural Knowledge and the Politics of Design*, edited by James Leach and Lee Wilson, 1-26. Cambridge: MIT Press.

Lewis, M.P., G. Simons, and C.D. Fennig. 2015. *Ethnologue: Languages of the World,* 18th edition. Dallas: SIL International.

Mathor, Saloni. 2000. "History and Anthropology in South Asia: Rethinking the Archive." *Annual Review of Anthropology* 29:89-106.

Ministry of Tribal Affairs, Government of India. 2013. "Statistical Profile of Schedule Tribes in India 2013." http://tribal.nic.in/WriteReadData/userfiles/file/Statistics/StatisticalProfileofSTs2013.pdf.

Morrison, Craig. 2015. "How to Design a Usable Search Function that Keeps Users Coming Back." *Usability Hour*. http://usabilityhour.com/how-to-create-a-usable-search-box-that-makes-your-users-happy/.

PARADISEC. 2016. "About Us." *PARADISEC*. http://www.paradisec.org.au/about-us/.

Povinelli, Elizabeth A. 2011. "The Woman on the Other Side of the Wall: Archiving the Otherwise in Postcolonial Digital Archives." *Differences* 22 (1):146-171.

Sankhil, Anjana. 2012. "A Brief Account of the Lamkang Naga Tribe: An Insider's View." https://www.scribd.com/doc/109805873/A-Brief-Account-of-the-Lamkang-Naga-Tribe.

Shilton, Katie, and Ramesh Srinivasan. 2007. "Counterpoint: Participatory Appraisal and Arrangement for Multicultural Archival Collections." *Archivaria* 63:87-101.

Usability.gov. n.d. "Card Sorting." *Usability.gov*. https://www.usability.gov/how-to-and-tools/methods/card-sorting.html.

Wasson, Christina. 2000. "Ethnography in the Field of Design." *Human Organization* 59 (4):377-388.

Wasson, Christina. 2016. "Design Anthropology." *General Anthropology* 23 (2):1-11.

Wasson, Christina, Heather Roth, and Gary Holton. 2016. "Findings from the Workshop on User-Centered Design of Language Archives: White Paper." https://designinglanguagearchives.files.wordpress.com/2016/04/wasson-et-al-2016-white-paper.pdf.

Wasson, Christina, and Susan Squires. 2012. "Localizing the Global in Technology Design." In *Applying Anthropology in the Global Village*, edited by Christina Wasson, Mary Odell Butler and Jacqueline Copeland-Carson, 251-284. Walnut Creek: Left Coast Press.

Zeitlyn, David. 2012. "Anthropology in and of the Archives: Possible Futures and Contingent Pasts. Archives as Anthropological Surrogates." *Annual Review of Anthropology* 41:461-480.