JOURNAL OF BUSINESS ANTHROPOLOGY

Designing for Diverse User Groups: Case Study of a Language Archive

Christina Wasson, Melanie Medina, Miyoung Chong,
Brittany LeMay, Emma Nalin, and Kenneth Saintonge

Abstract

This article explores the challenges of designing large-scale computing systems for multiple, diverse user groups. Such computing systems house large, complex datasets, and often provide analytic tools to interpret the data. They are increasingly central to activities in industry, science, and government agencies, and are often associated with "big data," data warehousing, and/or scientific "cyberinfrastructure". A key characteristic of these systems is the diversity and multiplicity of their intended user groups, which may range from various scientific disciplines, to assorted business functions, to government officials and citizen groups. These user groups occupy structurally different positions in local and global political economies, and bring different forms of expertise to the data housed in the computing system. We argue that design anthropologists can contribute to the usefulness of such systems by engaging in collaborative ethnographic research with the targeted user groups, and communicating findings to the designers and developers creating these systems.

Key Words

Language archives, design anthropology, multiple user groups, large-scale computing systems, indigenous groups

JBA 7(2): 235-267
Autumn 2018

ISSN 2245-4217

www.cbs.dk/jba

1. Overview: Designing Large-Scale Computing Systems for Multiple, Diverse User Groups[1] [2]

This article explores the challenges of designing large-scale computing systems for multiple, diverse user groups. Such computing systems house large, complex datasets, and often provide analytic tools to interpret the data. They are increasingly central to activities in industry, science, and government agencies, and are often associated with "big data," data warehousing, and/or scientific "cyberinfrastructure" (boyd and Crawford, 2012; Elish and boyd, 2017; Northrop et al., 2006; NSF, 2016; Wasson and Roth, 2015). The National Science Foundation recently stated that such computing systems are "a critical and increasingly important element of the science and engineering research enterprise" (NSF, 2016). We argue that design anthropologists can contribute to the usefulness of these systems by engaging in collaborative ethnographic research with the targeted user groups, and communicating findings to the designers and developers who create the systems. Up until now, it has been rare for the development teams of large-scale computing systems to engage closely with their intended users, and consequently, these systems have not yet achieved their potential usefulness (Goodale et al., 2012; Poore, 2011; Power et al., 2017; Ramakrishnan et al., 2014).

A key characteristic of large-scale computing systems is the diversity and multiplicity of their intended user groups, which may range from various scientific disciplines, to assorted business functions, to government officials and citizen groups. These user groups occupy structurally different positions in local and global political economies. They may also be geographically located in different parts of the world. Furthermore, they bring different forms of expertise to the data housed in the computing system.

As a result, members of the different user groups typically use the computing systems for dramatically different purposes. They bring different cultural practices of information access, sharing, and use to their engagement with the data and analysis tools. They also may face different technological capabilities and constraints. The design anthropologist's challenge is to ensure that interfaces, analytical tools, and database structures respond to the needs, practices, and constraints of diverse users. Such work takes user research and user-centered design to the next level of complexity.

---

[1] Authors are listed in the order of the sections they wrote; all contributed equally to the article.

[2] We gratefully acknowledge the generosity of the sixteen study participants who shared their time and insights with our Design Anthropology class. We also thank Shobhana Chelliah and the other members of the CoRSAL development team for supporting and guiding the class project, and for generously teaching us about language archives. Finally, we recognize the support of the National Science foundation through grants BCS-1543763 and BCS-1543828.

In this article, we focus on language archives as one particular type of large-scale computing system intended for multiple user groups. In doing so, we provide an example of design anthropology being applied in a non-business context. Our contribution is not so much the development of new theory, as it is the identification of an important new area of application for design anthropology, and aspects of research design required by the complexity of the domain. We offer readers a case study that describes how we started to develop methodologies to accommodate the complex ecosystem of user groups that surrounds language archives. Our approach can be generalized to other kinds of large-scale computing systems.

2. What is Design Anthropology?

There are several levels to our understanding of design anthropology. At the simplest level, we offer a definition provided for the *General Anthropology* journal in 2016:

> "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 (Wasson, 2016:1).

A second level to our understanding of design anthropology is the recognition that a key contribution anthropologists have made to the field of design, especially in the US, is their **concern with power**. Anthropologists tend to regard power as a central dimension in any social or cultural process. At the macrolevel, this could include the historical legacies of colonialism or the economic inequalities of capitalism (Edelman and Haugerud, 2005; Wolf, 1982). At the microlevel, it could include a consideration of power differences among researchers, study participants, and clients in the project context. The macrolevel and microlevel are intertwined.

The Participatory Design tradition that originated in Scandinavia was founded on a desire to empower workers who were facing managerial efforts to impose new technologies. It has thus displayed an explicit concern with power since its beginnings, advocating workplace democracy and supporting unions (Simonsen and Robertson, 2013). However, mainstream design in the US has not taken a similar stand. In practice, design in the US has implicitly supported the corporate goal of

selling more products while avoiding explicit engagement in political issues. In the US context, therefore, a major contribution that design anthropologists bring to the field of design is their explicit concern with power inequalities among stakeholder groups. Anthropologists tend to have a soft spot for the underdog, and work to create greater voice for marginalized groups. Like practitioners of Participatory Design, anthropologists collaborate with stakeholder groups through participatory research activities.

A third level to our understanding of design anthropology is the recognition that many members of this field highlight their **rigorous analysis methods**, and **use of social theory** to inform analysis, as things that set them apart from those who claim to conduct ethnography without much formal training. For instance, one design anthropologist interviewed by Wasson and Squires argued passionately:

> "What I rail against is… people with absolutely no background in the social sciences whatsoever claiming to be ethnographers… There are people out there who… claim that ethnography can be done without theory, which I think is… bankrupt as a concept… you can't just have the methods and then execute on them. You have to have a framework, a theoretical framework in which those methods are applied" (Wasson and Squires, 2012:27).

As a final comment, we do not agree with some observers who identify design anthropology as a subset of business anthropology. While design anthropology often takes place in business contexts, it also takes place in the nonprofit and public sectors. Design anthropology has the potential to bring significant benefits to sectors of society with fewer financial resources, in contexts where a profit motive is absent. Our case study provides an example of a project that was housed in a university, and whose stakeholders included members of multiple scientific disciplines as well as indigenous groups.

3. Language Archives

More than half of the world's 7,000 or so languages are at risk of disappearing before the end of this century (Evans, 2010). Starting in the 1990s, as this problem became widely recognized, members of the indigenous groups who speak these languages, as well as linguists, started to engage in a wave of language preservation and revitalization activities. As internet technologies became increasingly accessible in the early 2000s, online language archives began to spring up as a tool that could support language preservation and revitalization, as well as providing data on lesser-known languages that were valuable for linguistic analysis (Henke and Berez-Kroeker, 2016; Wasson et al., 2016a). Such online language archives are repositories of recordings, transcripts, and translations in a selected set of languages. They usually include linguistic

analyses of the languages, and may also contain various kinds of cultural data, such as field notes, photos, and recordings of music.

Most online language archives have been created either by linguists, or by members of indigenous groups. In this article, we focus on the former. Archives created by linguists have typically been designed with linguists in mind as the primary user group. However, they are considered cumbersome and often frustrating to use, and consequently very few linguists actually use them as a source of research data. Members of indigenous groups are interested in accessing materials on their languages that have been placed in language archives, but they also have a hard time navigating them.

These problems were illustrated in an exchange that took place at the 2016 *Workshop on User-Centered Design (UCD) of Language Archives* (funded by NSF grants BCS-1543763 and BCS-1543828). The workshop brought together members of the following stakeholder groups to map out UCD challenges facing language archives: language communities (i.e. indigenous groups), linguists, archivists, UCD practitioners, and funding agencies (Wasson et al., 2016a). The following exchange emerged during a discussion in which participants were reflecting on language archives developed by linguists. It started when one participant, Alexander, described the experience of language community members who try to access such archives. Participants in the exchange included: Alexander and Baldwin, language community members trained in linguistics; Seyfeddinipur and Holton, linguists who managed archives; Chelliah, linguist and Program Officer of the NSF Documenting Endangered Languages program; and Wasson, the facilitator.

Transcript Excerpt 1. The Big Aha (21 February 2016, Video 7)

Alexander: The academics are building these archives… and so you build it for people like yourself. So the door is an academic door, right? So other academics walk along and they say, "Oh! That's a—I know how to open this door, right? And it's for me! Everything in here is for me!" And it's like, um, for other, for other people who are not academics, right, they look at these, these archives and they're like, um, looking at tools from some kind of foreign thing, right, it's like a hammer with no nails, right, or like, um, a saw but there's nothing to cut, or something like that, right? The door isn't *made* for them. And if you go to like your search windows and your, some of the pages we've seen over the last day or two, where there's just crazy text, right.

All: ((laughter))

Alexander: I mean, I have a master's degree and stuff, but I'm like ((laughs))… this is kind of crazy, like, for somebody even of *my* background… One of the things that I really liked about the FirstVoices thing, is that the interface is obviously meant for any user…

Seyfeddinipur: I think it's a brilliant metaphor. I think it's a very beautiful metaphor. And the thing is, it's not even—Within the academic world, a linguist passes that door and says, oh, I don't know if I fit through that. And a historian says, oh that's a linguist's house, I'm not gonna go in there, it's gonna be all about syntax. So even within the academic world, don't think we are open, our door fits even them, right? They're like, what is ELAN, what is Toolbox? What is this, right?

Baldwin: I would take that one step farther. Not only are the institutions academic, the very archives that you think people want, are written by academics and can be read only by academics. So the average community member, even if they could get in the archive and find what they want, what do they do with it? So again, there needs to be a process that makes it usable to them.

Chelliah: Well here's the, here's the problem, is that, I see exactly what you're saying, but even as an academic I feel like I'm shut out a lot…

Wasson: And so one of the things that intrigues me about this discussion, is that, um, so as I think Wes said yesterday that, that there's always this sort of default assumption that linguists are the primary target audience for archives, and so it seems like we often discuss that user group differently from, say, like communities as user groups. In that we often don't even talk about linguists 'cause they're so *default,* that like most of the conversation, I think, has been more on language communities. And I, I would suggest that we try to level the, you know how we look at different user groups and, and talk about the needs of linguists more explicitly... we keep having this like default assumption…but then it seems they don't even work well for linguists!"

All: ((laughter))

Seyfeddinipur: Exactly!

Holton: That's why we don't talk about [linguists as a user group]… they're not using them, they're not the users.

In this exchange, it became clear that, in fact, **none** of the intended user groups – neither linguists nor language community members – was able to productively use language archives developed by linguists. The exchange provided strong evidence that conducting user research with language communities and linguists, and using the findings to inform the design of language archives, could greatly facilitate their use.

The Politics of Language Archives

As design anthropologists sensitive to the workings of power, our understanding of language archives is informed by an awareness of their colonial history, and recognition of the role played by colonial administrations in contributing to language loss in indigenous communities. In the U.S., for instance, the history of colonial policies and practices (including boarding schools) created significant hurdles for language use in the majority of Native American communities, greatly contributing to language decline (Austin and Sallabank, 2014; Jacob, 2013; Meek, 2010). Furthermore, museums, archives, linguists, and anthropologists historically collected material culture as well as linguistic and cultural knowledge from indigenous groups, stored these materials in museums that might be thousands of miles away, and displayed them in ways that were often culturally inappropriate (Lessard and Deal, 2015; Roy et al., 2011; Turner, 2015). Language archives may therefore be sites of struggle for power, control, access, and ownership (Wasson et al., 2016b). Language communities wish to exercise sovereignty over their cultural and linguistic heritage. Some forms of knowledge may not be considered appropriate to share.

Furthermore, the cultural logic of archiving is profoundly shaped by Western science, in particular the field of archiving (Foucault, 1982; Isaacman et al., 2005; Povinelli, 2011; Stoler, 2010; Zeitlyn, 2012), and in the case of language archives, the field of linguistics (Henke and Berez-Kroeker, 2016; Wasson et al., 2016b). The more a language archive adheres to predetermined definitions of what kinds of materials an archive should contain and how those materials should be organized, the less it is user-centered for language communities. We therefore recognize the need to remain open to indigenous ideas about what a language archive might look like, and what it might offer to users.

In our collaborative research with language communities, we are committed to avoiding the reproduction of colonial patterns of interaction as much as we can. We believe that a language archive has particular responsibilities toward the groups whose linguistic materials are stored in the archive; their needs as users should be given first priority. We strive to be aware of implicit biases that favor participation in the development process of some user groups, such as linguists and other academics, over other user groups, such as language communities.

4. Case Study: CoRSAL

As a case study to illustrate the challenges of designing language archives for multiple user groups, we describe exploratory user research we conducted for a planned language archive called the Computational Resource for South Asian Languages, or CoRSAL. This language archive is intended to become an online repository for materials on Tibeto-Burman

languages spoken mainly in Northeast India. It is the brainchild of Shobhana Chelliah, Professor of Linguistics at the University of North Texas (UNT). In summer 2016, Chelliah put together a team to create CoRSAL. Wasson was invited to contribute her expertise in design anthropology and UCD as a member of this team. Other team members included a computational linguist, a computer scientist with a focus on linguistic analysis, and an information scientist with expertise in database design.

Exploratory user research for CoRSAL was conducted by the sixteen students in Wasson's fall 2016 Design Anthropology course. Working in teams of two, the students conducted in-depth interviews with sixteen study participants. The project addressed this central research question: what were the needs of each user group with regards to this future language archive? This theme was divisible into various subthemes, such as users' relationship with current language archives, and what features they would desire in a future archive. Interviews lasted 1-1 ½ hours, and were recorded and documented with detailed field notes. Analysis of the interview data was a collaborative process that took place both in class discussions and via a qualitative data analysis program called Dedoose. The 90-page final client report documented both research findings and design implications (Al Smadi et al., 2016).

Four Intended User Groups for CoRSAL

The research was structured to compare the four main user groups planned for CoRSAL:

- Linguists who will use CoRSAL as a source of data
- Computational linguists
- Language communities
- Depositors, i.e. linguists and community members who will place their data in the archive

The following four sections document differences across the user groups, by providing brief ethnographic descriptions of their culturally shaped interactions with language archives. For the scientific user groups, we describe their epistemic cultures, meaning their disciplinary work practices and perspectives (Knorr-Cetina, 1999). In order to make the descriptions more specific, we highlight a single interviewee for each user group. The four sections all follow the same format: they start with a description of that interviewee, then summarize the culture of the user group, identify barriers to language use, and finally note design implications and recommendations. By providing a summary of those findings, Table 1 allows readers to scan at a glance the main differences across user groups.

Table 1. Summary of User Groups, Barriers, and Design Implications

	Linguists	**Computational Linguists**	**Language Communities**	**Depositors**
Group Description	• Typically main user group considered • Rarely use language archives as source of research data • Wide range of methodologies • Interviewee: Frank Seifart	• Relatively new field • Create speech/text processing systems or human machine translation and interaction systems • Usually work with more "common" languages that offer large data sets • Interviewee: Thelma Moore	• Speakers of languages whose materials are deposited in language archive • CoRSAL's focus is on Tibeto-Burman language communities • Exploratory research was conducted with 3 members of Lamkang language community • Interviewee: Sumshot Khular	• Deposit linguistic data into language archives • Multi-step process of preparing linguistic materials for deposit • Interviewee: Mark Post
Barriers	• Data difficult to find and access • Lack of standards for metadata and annotation	• Inconsistencies among file formats and annotation styles • Availability of relevant data	• Lack of standardized orthography • Lack of language learning materials • Inconsistent Internet and access to computers	• Time-consuming process of deposit preparation • Lack of professional recognition for deposit preparation • Difficulty of updating deposited materials • Fear findings will be pre-empted by others • Concerns with colonial legacy of archives
Design Implications and Recommen-dations	• Improve data accessibility through interface design and search • Develop thoughtful	• Avoid use of certain file types • Encourage depositors to provide annotation	• Encourage active role of community members in archive development • Support literacy efforts in	• Streamline language archive depositing process • Encourage universities to recognize deposits as

	standards for metadata • Encourage best practices for annotation	information • Allow for downloads of data subsets • Make data machine-readable	language community • Make hard copies of materials available to community • Develop mobile application for language archive	publications • Enable depositors to easily update their deposited materials • Specify usage rights for deposits • Empower language communities to gather and deposit data themselves

While there were important contrasts among the four user groups, there were also some needs that cut across user groups. Examples were ease of navigation and ease of search. These were two areas that interviewees found inadequate in most current language archives. Improving these aspects would benefit each user group by cutting down search time and aiding in finding relevant information. Additionally, improving the interface of a language archive could benefit the linguistic community as a whole. Interviewees expressed their belief that more users would choose to use an archive with an effective interface. This could include both depositors (language community members and linguists), and researchers, which could potentially facilitate more communication between these groups as well, if the archive were designed to encourage such interactions. With larger data sets available, this archive would also provide more resources to computational linguists. The interconnectedness of these issues among the four user groups emphasizes their importance in CoRSAL's design.

An additional need that user groups had in common was data protection. Language community members expressed concern regarding an archive's ability to protect sensitive data, such as sacred texts or private information. Depositors had similar concerns on behalf of the communities they worked with, and regarding their publication rights over in-progress data as well.

5. Linguists Who Will Use CoRSAL as a Source of Data

Traditionally, most language archives have envisioned linguists as their primary user group. Ironically, in practice linguists hardly ever use language archives as a source of data for their research. Only two out of the thirteen linguists who were study participants in the CoRSAL project

used language archives for research purposes. One of them was Frank Seifart.

Interviewee: Frank Seifart

Frank Seifart is an Assistant Professor of Linguistics at the University of Amsterdam. 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 and common ancestors. He is also interested in how languages influence each other due to language contact. He has specific questions regarding the temporal dynamics of speech in Amazonian languages and languages from all over the world.

Seifart's current work addresses language contact phenomena, particularly why languages borrow not only words, but also bound morphemes across words. He uses data from language archives for this project, and looks for particular formats of data. For example, he uses files from the language archive ELAR because they include time alignment between transcripts and recordings. Seifart has also used corpora from language archives for his research on language contact.

Work Practices of Linguists

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 tested their own intuitions about grammatical correctness, to data collected by field linguists who travel to different language communities and record naturally occurring discourse (Chelliah and De Reuse, 2011). In between these two extremes, linguists may work with speakers of languages who have migrated to the city where the linguists work. Linguists 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. This is partly due to the problems with language archives examined below, and partly a result of 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 (Henke and Berez-Kroeker, 2016).

Barriers to Language Archive Use

According to our interviewees, linguists are discouraged from using language archives due to barriers relating to the access and use of data. One of the most prevalent disadvantages is that the data in the archives are often difficult to find and access. Poor interface design, browse, and search functions are common barriers that contribute to such problems. As Seifart said, "There is a lot of user unfriendliness in the (user) interface."

Seifart described a common interface impediment as an "unfolding tree-structure system," which requires the user to click on every section to see the entire tree. Not only does this make it hard to find the data, but it is quite time consuming as well. He stated that when linguists have to spend an unreasonable amount of time digging through an archive 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 for another inconvenience to present itself: access restriction. He said 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." To circumvent this hassle, linguists can attempt to directly contact the depositor to gain access to the data. At this point, a linguist must resort to means other than language archives to gather research data. Therefore, Seifart emphasized, "The user interface is the crucial, crucial, crucial part to consider."

Linguists like Seifart need standardized and thoughtful metadata to quickly search for the types of data that would be useful to them. While navigating the DoBeS archive during the interview, Seifart said, "so what I would need to know for this question that we're addressing is, uh, how many data is there for a given language like Guarani, which is translated and transcribed and annotated?" He said that it is too complicated to search through an extensive list to discover the contents of transcription files. As he demonstrated the search function of DoBeS to interviewers, he concluded that he could not find out what he needed to know, and would have to request that information from its owner.

As with metadata, there is no standardization for linguistic annotation in the documents contained in a language archive. Linguists extensively mark up their transcripts with grammatical, morphological, and semantic analyses. These annotations are crucial for the transcripts' interpretation and use by other linguists. However, one linguist may use a type of annotation that is unfamiliar to another linguist, and the annotation system may not be well described. In that case, the linguist accessing someone else's data in a language archive would have to examine the depositor's publications in order to comprehend the

annotation style. Seifart believes that a purpose of archives is to provide non-depositors with access to the data; if one cannot locate the data because of poor metadata or understand the data because of incomprehensible annotations, one cannot truly access the data. This situation contributes to the linguists' hesitation to use language archives in their research.

Design Implications and Recommendations

The class report encouraged the CoRSAL development team to raise awareness among linguists that language archives could be a viable option for finding research data. 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 to avoid the frustrations of current language archives would also encourage linguists to consider CoRSAL as a source of data.

The design of language archive interfaces is important especially in terms of facilitating the visibility and accessibility of data. Ineffective navigational structures can be an impediment to the usability of an archive. If language archives provided more accessible overviews of their material, linguists could more easily determine whether the language archive contains useful data. Listing the size and format of the data would have a similar benefit.

A thoughtful, universal metadata system within the archive and optimally, across multiple archives, would make it easier for linguists to find data relevant to their research interests. This recommendation is generalizable to other kinds of large-scale computing systems. Standardized metadata systems will allow for more efficient data cataloging and searching within databases for user groups in any setting, from corporations to non-profits. The CoRSAL development team can take advantage of ongoing efforts to develop metadata standards for language archives (e.g. Drude et al., 2014). With regard to annotation styles, the class report recommended using findings from a 2009 Workshop on Cyberinfrastructure in Linguistics that identified the best practices for annotation in the following areas (Bender, 2009, 14-16): consistency/reliability, usability, resilience, accountability/responsibility, interoperability, and extensibility/adaptability.

6. Computational Linguists

CoRSAL is the first language archive that has identified computational linguists as an intended user group. Most computational linguists conduct their research with well-resourced languages such as English, for which they can find large data sets. However, there are some computational

linguists who conduct research with endangered languages, and there are even more who would be interested in doing so if they could compare data from languages that are in the same language family, yet radically different from the well-resourced languages they are used to analyzing (Agić et al., 2015; De Pauw et al., 2012; Garrette and Baldridge, 2013; Palmer et al., 2010). Researchers on our team interviewed four computational linguists. Among these four, two requested to be kept anonymous. To honor their wishes, researchers created pseudonyms for all four computational linguist interviewees, and their places of employment will not be disclosed.

Interviewee: Thelma Moore

Thelma Moore is a professor in the linguistics department at an American university. She has a Ph.D. in computational linguistics from the University of Texas at Austin. Her research interests include computational linguistics for low-resource languages and for language documentation. Low-resource languages are those that do not have large volumes of written texts. Moore described her fascination with these topics when she said:

> Language is this very human thing, so it is something that is close to all of us. We all feel like experts, we all are experts in language, in a way. When I learned about computational linguistics I thought well, this is even better. Because here is a way to do something really tangible and useful to make some kind of product that is connected to this area of study that my brain enjoys so much. So, I think we have the capacity to do some interesting things with computational linguistics, particularly since so much communication is text-based these days. Also… we have this real urgency behind documenting endangered languages and lower-resource languages, and I really believe that we can use computational methods to speed up and support that process.

Work Practices of Computational Linguists

The research team found surprising diversity in the research activities of computational linguists. 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 a linguist's area of study, background, and research questions, a variety of methods can be utilized to study, model, and analyze languages. The goals of computational linguists may range from building computational technologies to understanding a particular language.

The field of computational linguistics began with a focus on translation. Using computational methods saved researchers time and effort, produced reasonable results compared to manual methods, and

allowed researchers to analyze more data than ever before. Since then, the goals and methods of computational linguistics have diversified. Computational linguists are now using multiple tools and applications to write code for analyzing and processing their data. The most common tools are Python programming language, natural language processing (NLP) tools, and Natural Language Toolkit (NLTK). Additionally, some use text parsers to analyze sentences and generate output.

Barriers to Language Archive Use

Interviewees expressed that their most difficult challenges were inconsistencies among file formats and annotation styles. In recent years, it has become common for audio files to be uploaded to language archives as WAV or MPG files, which are useful to computational linguists. However, the format of text files containing transcriptions, translations, and annotations still varies across the board. Among many other things, the interviewees expressed difficulty using MS Word files, and extreme frustration toward PDF files. As Moore explained: "PDF files are not very useful, because a PDF is basically an image. So, then you can do optical character recognition, and try to get the text, but that often has mistakes."

As an additional frustration, annotation styles are not standardized, as discussed in the previous section. Interviewees noted that as a result, they often need to annotate the data themselves, which takes away from the time they can spend in high-level analysis. Moore described why computational linguists may be hindered from using language archives for their research when she said:

Often, it's not easy to see how other people's data can be applicable for the questions that you want to study. But then, very often, it's sort of this cost-benefit-analysis. You might see that someone has collected a lot of data on a particular phenomenon, and maybe it would be useful for you, but you would have to invest a lot of time and effort in order to get that data into some format that you can really work with it, or to learn the data. Or perhaps in cases of documented endangered languages, if you don't know how the language works, it would be difficult to use that data. Also, different linguists or different projects use different sets of labels for analyzing their data. So, one person might be interested in only how noun phrases work. So maybe they only label the noun phrases, but not the rest. Then maybe you want to analyze verbs, so this data has annotation that isn't so helpful for you.

In addition to the challenges of working with multiple annotation styles, several computational linguists also identified data availability as an issue. Many computational linguists do not conduct fieldwork, so they rely on other linguists to collect the data they analyze. At the same time, few field linguists who study endangered languages are making their data available for public use. The interviewees offered several explanations:

linguists are waiting until their annotation and analysis of data are complete (which may never happen); linguists do not have time to define metadata and get the data into the appropriate format; and linguists want to protect their publication rights. These issues will be discussed further in the section on depositors.

Design Implications and Recommendations

The first design implication for computational linguists is the need to avoid MS Word and PDF files. Rather, files should be formatted as XML, TXT, or CSV files. Moore emphasized the need for files to have an attached document explaining the format of the files: "One thing that data really needs to have is a clear explanation of what the format is. So, if I get a bunch of data that is in some format I've never seen before, but I also have a document that explains… I can work with that."

Although it is unlikely that linguists will adopt a universal style of annotation, it would be helpful for computational linguists to have information on the annotation. Moore addressed this when she said: "Something else that is important is that the metadata clearly states what kind of annotations there are, if any, who did the annotations, or even that the text was labeled by an unreliable annotator."

Due to the diversity of computational linguists' research activities, it would be useful to provide ways for computational linguists to download customized data sets, in customized formats. For instance, several 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 checkbox or dropdown menu controls were used to retrieve the desired data. Similarly, interviewees suggested that an interface that enabled users to run queries on a selected dataset would be useful. The computational linguists were interested in applying SQL queries, a function which is not facilitated by current language archives. They also pointed out the importance of having metadata available along with the data files.

Data sets on endangered languages are often small, yet many forms of machine learning require large data sets. To address this issue, CoRSAL should encourage depositors to place as much of their data as possible in the archive. Ways to achieve this goal are addressed in the following section.

In summary, Moore suggested the following design implications to make CoRSAL useful for computational linguists:

- Make data machine readable
- Harmonize label sets across corpora
- Build a model that supports multiple data formats

- Standardize metadata across languages to facilitate cross-linguistic research

To make finding data less cumbersome, CoRSAL should have 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). Lastly, it would be useful to create an online community for both the computational linguists who use CoRSAL and the developers who design the infrastructure. This would allow these users to quickly solve problems as they arise. Ultimately, the goal of these design implications is to decrease the time and effort computational linguists spend completing tasks that must be done prior to high-level analysis, so that they can focus on the vital aspects of their research, and contribute more in-depth findings to the field of linguistics. Furthermore, by replacing "computational linguistics" and "linguistics" with other fields that gather data from human populations, one can see how these design implications apply to large-scale computing systems used by a wide range of disciplines.

7. Language Communities

For our exploratory research, the Lamkang indigenous group represented the Tibeto-Burman communities whose languages will be included in CoRSAL. There are roughly 100-300 Tibeto-Burman languages in India (depending on definitions); about 10-15 languages will be included in the first iteration of CoRSAL (Post and Burling, 2017, 214).

The Lamkang are a scheduled tribe of approximately 39 villages clustered in the hill country of Chandel district, in Manipur, India. The Lamkang villages are primarily Christian, with over 20 Baptist and Roman Catholic churches. These institutions, and the prominent community members within them, have supported the language revitalization effort. Chelliah has worked with the Lamkang community for a number of years, so she was able to connect the student researchers with study participants. Researchers interviewed three Lamkang speakers via telephone.

Interviewee: Sumshot Khular

One of the community members interviewed was Sumshot Khular, a peace activist with graduate degrees in linguistics and human rights. She is dedicated to the preservation and promotion of Lamkang, and, among other honors, was awarded a Fellowship in Oral Literature in 2016 from the Firebird Foundation for Anthropological Research for her project *Documentation of the Lamkang Language*. Though her education and career brought her far from home—to Delhi, the United Kingdom, and Texas—Khular maintains her connection to Thamlakhuren village, using

her education and expertise to the benefit of her community. In Manipur, she has organized human rights training, theater workshops for young people affected by conflict, peace programs as executive member of the Naga Women Union, and has translated human rights documents into Lamkang.

The Lamkang Community

From the point of view of older members of the community, the Lamkang language is in decline. It is still the language spoken in the home and the first language learned by children. However, as younger generations participate in an increasingly English-speaking and Hindi-speaking world, there are fewer opportunities to practice speaking Lamkang in public. This is reinforced in the institutional and social spheres; education is conducted in English, and dominant languages Hindi and English are most common in entertainment media and on the Internet.

Khular's insights regarding the region's language revitalization movement revealed that the Lamkang language is a source of identity and cultural pride, connecting community members to each other and to their history: "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." Members of the language community are anxious to preserve their language in the face of intergenerational changes. CoRSAL will support this effort, collaborating with the community to create an archive that meets their needs.

Barriers to Language Archive Use

The first barrier to the community's use of CoRSAL is the lack of standardized orthography; printed materials are only useful if community members can read them. Linguists are currently developing an orthography for Lamkang. Some written works have been translated into Lamkang, including a New Testament and 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 Rights of Indigenous Peoples,* translated by Sumshot Khular. Concern remains that these materials are too grammatically advanced or conceptually dense to be useful in language learning. Khular said: "We have two Bibles but people find it difficult to read. And we have no base, like the alphabet…So without the alphabet we have two huge books that are too difficult for children to read or any adult even to read." Design anthropology can support the effort to develop written materials by aligning the expertise of linguists and translators

with the needs of language community users. An easy-to-learn orthography, standardized across materials and taught to community members, will be a critical component of the language revitalization effort.

There is a strong tradition in the Lamkang community of passing cultural knowledge through generations by practicing together. Khular remembers that in her childhood, it was her grandmother who taught her to weave. At the age of 2 or 3, Khular practiced weaving wild grass while she observed experienced weavers using yarn. In a similar way, children are taught to farm by accompanying their elders in the fields. This method of transmitting knowledge through practice could be used to support language learning within families. Lamkang is only spoken in the village setting, so family and the community are critical avenues for the dissemination of the Lamkang language. Khular describes: "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." At present, children first learn Lamkang in the home, but lose familiarity as they attend English-speaking boarding schools. With the introduction of pedagogical materials from the archive, children could actively maintain their knowledge of Lamkang both at home and at school.

This culturally-shaped approach to learning reveals a second barrier to using CoRSAL—materials from language archives are often formatted for use by linguists and other researchers, rather than for language learning. Though community members are interested in relearning their mother tongue, the resources to do so are inaccessible. One of the goals of this research is to demonstrate that language archives do not have to be designed around the needs of just one user group. Providing learning-oriented materials would not prevent CoRSAL from providing research-oriented materials—accommodating both groups is possible if we expand the cultural logic of archiving to include language community users.

The third barrier to language archive use is technology constraints. Lamkang villages have limited numbers of computers and unreliable internet connections. Khular describes a frustrating situation that is characteristic of the area's unreliable Internet:

> The whole day I was trying to write a mail and open and then it was like off and on, and it was not really possible. So unless I, maybe if I go to the city in the capital, that can be possible, but I am in the village right now as my sister was unwell and I have to be at home. So I cannot leave her and go, but…I cannot express though, it is really difficult.

This is a significant issue for language community access; because CoRSAL is primarily an online archive, the development team will need to

give thought to ways of facilitating access in the face of technology constraints.

Design Implications and Recommendations

By interacting with CoRSAL, community members can take an active role in the preservation of their language. This responds to the urgency that community members feel to preserve older generations' knowledge. Community members currently have a sense of being too dependent on outside initiatives, the results of which do not always make it back to the community. Access and deposit ability, combined with efforts to engage multiple generations of the community, have the potential to foster community support for the project. In the words of Sumshot 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."

Developing a standardized writing system for the Lamkang is an effort that extends beyond the purview of the CoRSAL development team. However, CoRSAL should be aware of the importance of creating literacy within indigenous groups whose language materials it collects; it should coordinate its activities with ongoing literacy efforts, and offer support.

Interviewees emphasized the importance of reaching the younger generation with targeted materials such as comics, storybooks, and animations. As Khular stated:

> 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.

While CoRSAL itself will not have the resources to develop such learning materials, the CoRSAL team could partner with various funding sources and local teachers, artists, and other contributors to encourage the development of such items. Furthermore, the CoRSAL portal for language communities should offer materials that could easily be used by teachers.

To address the lack of infrastructure and limitations of technology in the Lamkang villages, the CoRSAL team could partner with funders to set up a small building as a cultural center. This location could house a computer to provide access to the language archive, a printer to obtain hard copies of documents for circulation in the community, and a scanner that would enable community members to deposit photos and other documents.

The CoRSAL team could develop a mobile application, as smartphones are more common than personal computers in the

community. This addresses some of the community concerns about the villages' remoteness, and a mobile connection to the database could foster collaborative learning. For both digital interfaces, CoRSAL should provide tutorials for language preservation and database use.

8. Depositors

Many of the participants in our study had experience depositing data in language archives. Even the linguists who had been categorized as using language archives as a source of data turned out to use language archives primarily to deposit data. One of these interviewees was Mark Post.

Interviewee: Mark Post

At the time of the interview, Post was Senior Lecturer and Convenor of Linguistics at the University of New England, Australia; in January 2018, he moved to the University of Sydney. His research focus is evolution and typology of greater mainland Southeast Asian languages. He started his field study twelve years ago and has worked in the Eastern Himalaya region with languages of the Tani subgroup of the Tibeto-Burman language family. Post has collected data, written grammars, trained indigenous linguists, and worked on language maintenance and revitalization materials, including dictionaries and textbooks, working mostly in small communities with fewer than 40,000 speakers. Post has experience making deposits to ELAR and to PARADISEC. However, he does not use language archives as a source for his own research.

Work Practices of Depositors

The process of preparing linguistic materials for deposit in a language archive varies among individuals. However, the following tasks are typically involved.

Recording. Capturing spoken language is the primary focus of linguistic fieldwork. Recording tools have evolved with advances in technology; tape recorders have transitioned to digital recorders for audio and video files. Additionally, some researchers collect field notes and photographs of people and artifacts. Recordings are saved in a variety of formats.

Transcription. Transcripts textually represent the nuanced pronunciation of words. Formats vary depending on the capabilities of analytical software. Examples of such software include FLEx, PRAAT, and ELAN. Transcription quality improves with the depositor's language competency and, as such, transcripts can be revisited and improved.

Translation. The translation process includes both word-for-word translation of the transcribed data, and free translation, which conveys

the meaning behind the literal words. Like transcription, more valuable content is produced with greater language competency.

Annotation. The style of annotation is unique to both linguist and data. Analytic insights, descriptions of the data, identification and labeling of text segments, parts of words, morphemes, and semantics are some possible layers of annotation.

Application of Metadata. Metadata mark a file with specific information such as who created the file and when, which language is documented, in what location data were gathered, and other context. Software programs such as ARBIL and SayMore can be used to organize and apply metadata to files.

Barriers to Language Archive Use

Linguistic deposits take substantial time and effort. Mark Post estimated that full processing of four minutes of text – including transcription, translation, interlinearization, file preparation, and metadata management – takes two weeks of labor by someone who has a year's experience of analyzing the language. Part of the problem is that linguists usually need to use multiple software programs, and these programs rarely coordinate with each other. Post listed programs ELAN, FLEx, and Saymore as examples of programs he has used in the preparation of data for deposit. "Cumbersome" was a recurring adjective used by interviewees regarding this process. As Post commented:

> The procedure for archiving data is really very cumbersome and there is not enough of a focus, on sort of, there is too much of a focus, on, to get to the final stage and deposit, sort of thing. And nobody's even at the final stage. There is no such thing as the final stage. What you find in a situation of people like me, who has refined their analysis over ten or fifteen years is reluctance to say, Okay, I'm going to put all my effort into organizing my metadata and getting it exactly how the archive wants it right now so I can deposit it. Because you then have to undo that work later, go in there again.

Archival practices may demonstrate a generational divide. Post described three generations in the field today. The oldest generation often lacks technological expertise and struggles with software interfaces, to the point of not using them at all. Post's generation can pass with some effort, but still struggles with the more complex programs and sees technology leaving them behind. The youngest generation has learned programming almost as a second language. They adapt most quickly to different software interfaces, or reprogram them entirely. If program design continues as it does, Post worriedly commented:

> It might be a case of all of us [older linguists] dying or something

> like that and the next generation being able to manage everything. But we are talking about twenty to thirty years down the road, are we satisfied? You know, with seventy percent of linguists of the world not being able to use these tools?

A second barrier is that most linguists are not rewarded in university performance reviews for preparing deposits. The time they spend on this task is not acknowledged. During the interview, Post talked about being frustrated with the constraints on his time, and reiterated themes of time efficiency, challenging workloads, workflow obstacles, and cost/benefit analyses for unrecognized labor.

A third barrier is that the logic of archiving assumes that a deposit is complete, final, and will never need to be changed. At present, it is difficult to make changes to archived materials once they have been deposited. This logic of archiving may be appropriate for historical materials, but is not a good fit for linguistic deposits. Linguistic analysis is never complete. As Post noted: "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." Post explained that it takes ten to fifteen years for a linguist to develop a comprehensive analysis of a new language, and the analysis will continue to be revised for as long as the linguist works on the language. Thus, linguists never feel ready to make a deposit, since the expectation of the archive is that a deposit should be a finished product.

A fourth barrier is that linguists are often concerned about releasing data to the public before they have fully published their analyses, to avoid the possibility that other researchers might preempt their findings.

A final issue to consider is the potential barriers to deposits by language community members, rather than linguists. Post pointed out that the difficulty of learning to use data preparation and deposit programs is especially unfortunate given the legacies of colonialism. Community members could be excellent depositors and benefit most from access to the data. He made this powerful argument about ease of use: "Archiving should really be as easy as managing Gmail. It really should be, and if it is any harder than that then you have lost the battle. It should be as easy as Facebook, you know as everything that everybody in the world is using."

Design Implications and Recommendations

The barrier of time-consuming data preparation could be addressed through improvements to analysis software and streamlining the language archive deposit protocols. Ideally, the software programs

involved in data preparation and deposit could be integrated, similarly to the Adobe Suite or Microsoft Office Suite.

The second barrier, recognition, could be addressed by encouraging universities to value deposits as equivalent to publications. An NSF-funded project called "Developing Standards for Data Citation and Attribution for Reproducible Research in Linguistics" is already working on this issue; the CoRSAL development team could lend its support (Berez-Kroeker et al., 2017; Haspelmath and Michaelis, 2014; Thieberger et al., 2015).

CoRSAL can address the third barrier, traditional archival logic, by allowing depositors to easily edit their deposits and metadata. This involves a conceptual shift from regarding CoRSAL as an archive to envisioning it as a database that is regularly backed up to an archive. This conceptual shift has already taken place during the CoRSAL development team's 2016-2017 planning process.

The fourth barrier, depositors' concerns about protecting their right to publish, could be addressed by specifying usage rights for deposits. Perhaps CoRSAL could allow depositors to release certain parts of their data conditionally, or only after a certain amount of time.

Finally, the barriers for language community depositors could be addressed both by providing training and by simplifying the software. Post offers a workshop to indigenous communities in the Eastern Himalayan region called *Training and Resources for Indigenous Community Linguists* (sponsored by the Firebird Foundation for Anthropological Research). Its purpose is to empower indigenous groups to collect linguistic data and make deposits themselves. This could be a model for workshops offered to other groups as well. CoRSAL would be well positioned to support such efforts. Post's vision of a radically simplified suite of software programs, as easy to use as Facebook or Gmail, could inspire the CoRSAL development team, as well as the designers of data preparation software such as FLEx, to create more accessible software. Furthermore, other kinds of large-scale computing systems could benefit from these recommendations as well. Empowering a user group to use a product is key to that product's success, whether the users are indigenous peoples preserving their language, or small business owners looking for software to keep track of their finances.

9. Translating Research into Design

The research findings presented here have become the basis for a set of interface designs for CoRSAL. In spring 2017, a class at the Illinois Institute of Technology's Institute of Design translated our research insights into interface design prototypes. The class was taught by Santosh Basapur, who has collaborated with Christina Wasson on language

archive research and design since 2016. Basapur's design students thoughtfully linked our research findings to a broad set of design considerations for CoRSAL. They developed early-stage prototypes based on their analysis of potential workflows for the different user groups and key functionalities needed for each task. Two slides from their final presentation offer samples of these accomplishments. Figure 1 illustrates the search features designed for CoRSAL. Students designed a sophisticated set of search options that responded to linguists' frustrations with search functionalities on existing language archives. Separate slides also showed an option to enter a SQL query, for computationally sophisticated users.

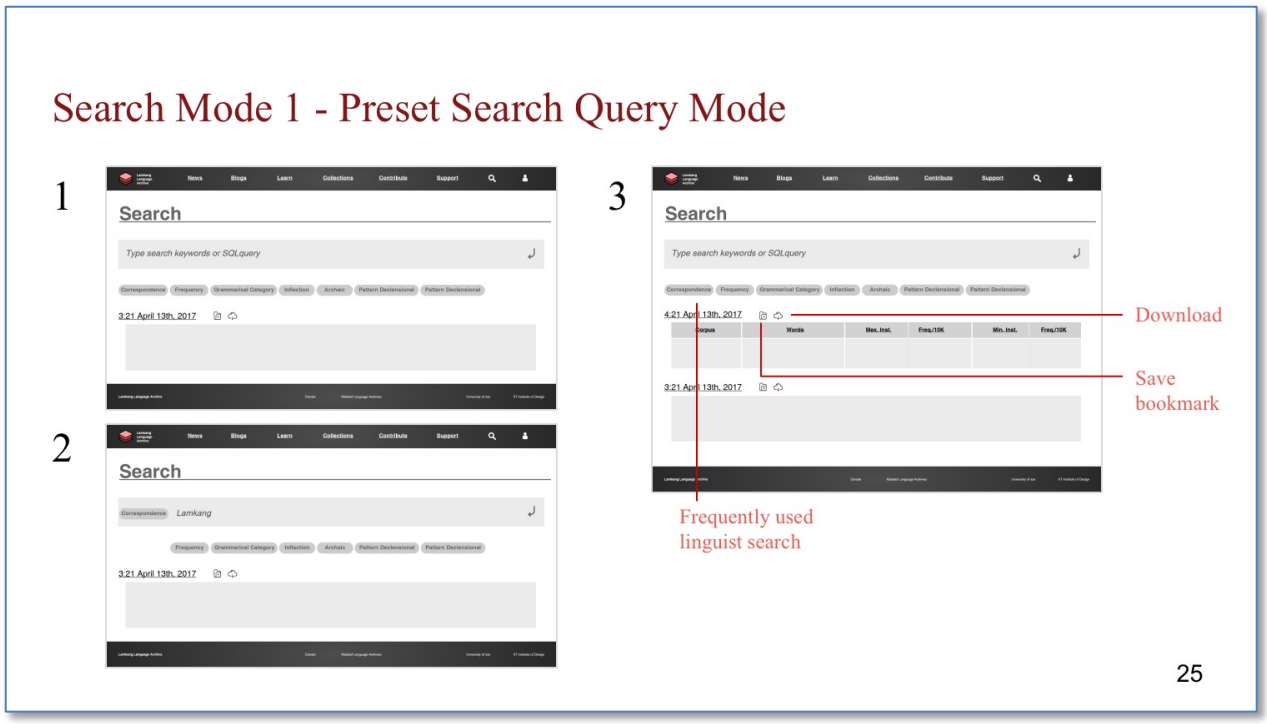

Figure 1. Search Mode 1

Figure 2 shows part of a deposit work flow that was designed for language community members. It was created for a mobile interface, since there are few desktop or laptop computers in the indigenous communities of northeast India. Mobile phones are much more common, and most use the Android operating system, as shown. This slide demonstrates how community members without linguistic training could be easily guided through the deposit process. Specifically, the slide shows users being invited to tag a recording with relevant metadata. Making it easy for community members to contribute their data could empower these communities to become active participants in the creation of

linguistic resources for their communities.

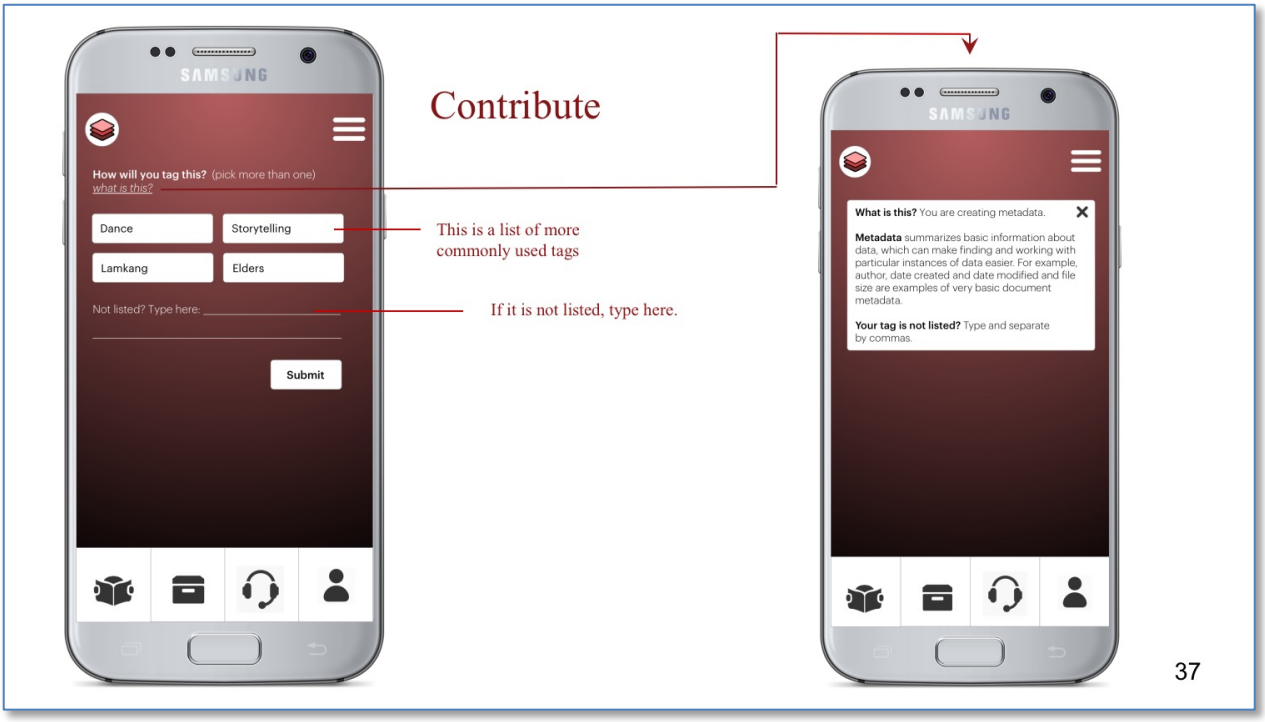

Figure 2. Deposit Process for Language Community Members

The work of Basapur and his students is ongoing. As the development of CoRSAL proceeds, Wasson's user research team and Basapur's design team will continue to weave their activities together to ensure the design of interfaces that are as useful as possible for each user group. They are also coordinating their work with those who are designing the database structure and other software and hardware aspects of CoRSAL.

10. Conclusions

The brief review provided here shows that for a large-scale computing system intended for multiple, diverse user groups, ethnographic research can reveal important differences in the needs and practices of these groups. There were many significant contrasts across the four user groups for whom CoRSAL was intended. For instance, while linguists seek transcripts that are densely annotated with linguistic analyses, indigenous groups engaged in language revitalization need easy-to-read versions of common sentences and stories. Likewise, while computational linguists need to be able to access linguistic data in a systematic, structured online format, indigenous groups facing technology constraints may prefer PDF files they can easily print out.

The approach developed for our research on language archives can be usefully extended to other contexts. In recent years, many types of organizations, including corporations, have turned to the collection and analysis of "big data," and have begun to develop computing systems to support these activities. In conversations with design anthropologists and IT professionals, Wasson has learned that extensive user research has not yet been conducted in business organizations, because the development of these novel forms of large-scale computing systems is so recent. Here, for instance, are comments by Natalie Hanson (see also article in this issue): "the technology is so complex and so new, everyone is still just trying to figure out how to make it work and extract business value out of it… Based on what I can see, it's still very early days for these types of systems" (Personal Communication 2017). It is unlikely that these computing systems will live up to their potential usefulness without user research that informs designers and developers about the practices and needs of their diverse user groups. The research design we created to accommodate the complex ecosystem of language archive users can easily be adapted for other kinds of large-scale computing systems.

We end with a call to include an investigation of power inequalities in all user research for large-scale computing systems. Regardless of whether a system is designed for scientific analysis or for industry purposes, there is likely to be an implicit or explicit power hierarchy across user groups. Development teams are likely to pay more attention to the needs of some user groups, and the needs of other user groups are more likely to be neglected. For example, Wasson and Roth (2015) conducted a user study for the design of a new, comprehensive data warehouse and set of analysis tools at the University of North Texas (UNT). The client was the team leading the development of this computing system. One of Wasson and Roth's key insights was that a "core-periphery" dynamic structured interactions among the four organizational units of the university (three campuses and an administrative center). In many situations, the UNT Denton campus occupied the most powerful position, i.e. the "core," while the other units were politically more peripheral. Thus most existing computing systems at UNT had been designed only for the needs of Denton campus users. This had created challenges for the other campuses, since they had different types of students and different organizational structures. Our analysis of the core-periphery dynamic was one of the most powerful findings for our client.

As design anthropologists, then, we suggest that mapping out power hierarchies across users and developers can be a useful exercise, revealing which user groups are likely to receive the most and the least attention in the development process. This recognition, in turn, can help the development team take steps to treat all user groups with more equal consideration.

References

Agić, Željko, Dirk Hovy, and Anders Søgaard. 2015. "If all you have is a bit of the Bible: Learning POS taggers for truly low-resource languages." *The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2015)*.

Al Smadi, Duha, 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, and Christina Wasson. 2016. Exploratory user research for CoRSAL. Report prepared for S. Chelliah, Director of the Computational Resource for South Asian Languages (CoRSAL). Department of Anthropology, University of North Texas. Available at https://designinglanguagearchives.files.wordpress.com/2017/03/exploratory-research-for-corsal.pdf.

Austin, Peter K., and Julia Sallabank, eds. 2014. *Endangered languages: Beliefs and ideologies in language documentation and revitalization*. Oxford: Oxford University Press. https://doi.org/10.5871/bacad/9780197265765.001.0001

Bender, Emily M. 2009. Cyberling 2009 workshop: Towards a cyberinfrastructure for linguistics. Workshop report. Seattle: University of Washington.

Berez-Kroeker, Andrea, Gary Holton, Susan Kung, and Peter Pulsifer. 2017. "Developing standards for data citation and attribution for reproducible research in linguistics: A project of the National Science Foundation" Accessed 10 May 2017. https://sites.google.com/a/hawaii.edu/data-citation/welcome.

boyd, danah, and Kate Crawford. 2012. "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon." *Information, Communication & Society* 15(5):662-679. https://doi.org/10.1080/1369118X.2012.678878

Chelliah, Shobhana L., and Willem J. De Reuse. 2011. *Handbook of descriptive linguistic fieldwork*. New York: Springer. https://doi.org/10.1007/978-90-481-9026-3

De Pauw, Guy, Gilles-Maurice de Schryver, and Janneke van de Loo. 2012. "Resource-light Bantu part-of-speech tagging." *Workshop on Language Technology for Normalisation of Less-Resourced Languages (SALTMIL 8-AFLAT 2012). European Language Resources Association*.

Drude, Sebastian, Paul Trilsbeek, Han Sloetjes, and Daan Broeder. 2014. "Best practices in the creation, archiving and dissemination of speech corpora at The Language Archive." In *Best practices for spoken corpora in linguistic research*, edited by S. Ruhi, M. Haugh, T. Schmidt and K. Worner, 183-207. Newcastle upon Tyne: Cambridge Scholars Publishing.

Edelman, Marc and Angelique Haugerud, eds. 2005. *Anthropology of development and globalization: From classical political economy to contemporary neoliberalism*. Malden, MA: Blackwell Publishing.

Elish, M. C., and Danah Boyd. 2017. "Situating methods in the magic of big data and artificial intelligence." *Communications Monographs.* Forthcoming.

Evans, Nicholas. 2010. *Dying words: Endangered languages and what they have to tell us*. Chichester: Wiley-Blackwell.

Foucault, Michel. 1982. *The archaeology of knowledge and the discourse on language*. New York: Pantheon.

Garrette, Dan, and Jason Baldridge. 2013. "Learning a part-of-speech tagger from two hours of annotation." *HLT-NAACL*.

Goodale, Paula, Paul Clough, Nigel Ford, Mark Hall, Mark Stevenson, Samuel Fernando, Nikolaos Aletras, Kate Fernie, Phil Archer, and Andrea De Polo. 2012. "User-centred design to support exploration and path creation in cultural heritage collections." *Proceedings of EuroHCIR2012*:75-78.

Haspelmath, Martin, and Susanne Maria Michaelis. 2014. "Annotated corpora of small languages as refereed publications: A vision." *Diversity Linguistics Comment*. http://dlc.hypotheses.org/691.

Henke, Ryan E., and Andrea L. Berez-Kroeker. 2016. "A brief history of archiving in language documentation, with an annotated bibliography." *Language Documentation and Conservation* 10:411-457.

Isaacman, Allen F., Premesh Lalu, and Thomas I. Nygren. 2005. "Digitization, history and the making of a postcolonial archive of Southern African liberation struggles: The Aluka Project." *Africa Today* 52 (2):55-77. https://doi.org/10.1353/at.2006.0009

Jacob, Michelle M. 2013. *Yakama rising: Indigenous cultural revitalization, activism, and healing*. Tucson: University of Arizona Press.

Knorr-Cetina, Karin. 1999. *Epistemic cultures: How the sciences make knowledge*. Cambridge: Harvard University Press.

Lessard, Kerry Hawk and Gregg Deal. 2015. "Real Indians, last Indians: Art, anthropology, and the museumization of indigenous lives." *Practicing Anthropology* 37(3):47-48. https://doi.org/10.17730/0888-4552-37.3.47

Meek, Barbara A. 2010. *We are our language: An ethnography of language revitalization in a northern Athabaskan community*. Tucson: University of Arizona Press.

National Science Foundation. 2016. "Dear colleague letter: Seeking community input on advanced cyberinfrastructure." https://www.nsf.gov/pubs/2016/nsf16090/nsf16090.jsp.

Northrop, Linda, Peter Feiler, Richard P. Gabriel, John Goodenough, Rick Linger, Tom Longstaff, Rick Kasman, Mark Klein, Douglas Schmidt, Kevin Sullivan, and Kurt Wallnau. 2006. *Ultra-large-scale systems: The software challenge of the future*. Report from the Software Engineering Institute, Carnegie Mellon. Pittsburgh.

Palmer, Alexis, T. Moon, J. Baldridge, K. Erk, E. Campbell, and T. Can. 2010. "Computational strategies for reducing annotation effort in language documentation." *Linguistic Issues in Language Technology* 3 (4):1-42.

Poore, Barbara S. 2011. "Users as essential contributors to spatial cyberinfrastructures." *PNAS* 108 (14):5510-5515. https://doi.org/10.1073/pnas.0907677108

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. https://doi.org/10.1215/10407391-1218274

Post, Mark W. and Robbins Burling. 2017 "The Tibeto-Burman languages of Northeast India." In *The Sino-Tibetan languages*, edited by G. Thurgood and R.J. LaPolla (second edition), 213-242. London: Routledge.

Power, Christopher, Andrew Lewis, Helen Petrie, Katie Green, Julian Richards, Mark Eramian, Brittany Chan, Ekta Walia, Isaac Sijaranamual, and Maarten de Rijke. 2017. "Improving archaeologists' online archive experiences through user-centred design." *Journal on Computing and Cultural Heritage* 10 (1):1-20. https://doi.org/10.1145/2983917

Ramakrishnan, Lavanya, Sarah Poon, Valerie Hendrix, Daniel Gunter, Gilberto Z. Pastorello, and Deborah Agarwal. 2014. "Experiences with user-centered design for the Tigres Workflow API." *e-Science 2014 Proceedings*.

Roy, Loriene, Anjali Bhasin, and Sarah K. Arriaga, eds. 2011. *Tribal libraries, archives, and museums: Preserving our language, memory, and lifeways*. Plymouth, UK: Scarecrow Press.

Simonsen, Lesper and Toni Robertson, eds. 2013. *Routledge international handbook of participatory design*. New York: Routledge.

Stoler, Ann Laura. 2010. *Along the archival grain: Epistemic anxieties and colonial common sense*. Princeton: Princeton University Press.

Thieberger, Nick, Anna Margetts, Stephen Morey, and Simon Musgrave. 2015. "Assessing annotated corpora as research output." *Australian Journal of Linguistics* 36 (1):1-21. https://doi.org/10.1080/07268602.2016.1109428

Turner, Hannah. 2015. "Decolonizing ethnographic documentation: A critical history of the early museum catalogs at the Smithsonian's National Museum of Natural History." *Cataloging & Classification Quarterly* 53(5/6). https://doi.org/10.1080/01639374.2015.1010112

Wasson, Christina. 2016. "Design anthropology." *General Anthropology* 23(2):1-11. https://doi.org/10.1111/gena.12013

Wasson, Christina, Gary Holton, and Heather Roth. 2016a. "Bringing user-centered design to the field of language archives." *Language Documentation and Conservation* 10:641-681. Available at https://scholarspace.manoa.hawaii.edu/bitstream/handle/10125/24721/wasson.pdf.

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

Wasson, Christina and Heather Roth. 2015. *Final report: Phase one user research for data warehousing/analytics/dashboards initiative. Prepared for D.A.D. Core Team, University of North Texas.* Available at https://static1.squarespace.com/static/587d7d9d29687f2d2feb8f4f/t/589f93422e69cf7248dd9559/1486852931229/Final+Report+for+Public+-+User+Research+for+DAD+Initiative.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 O. Butler, and Jacqueline Copeland-Carson, 251-284. Walnut Creek: Left Coast Press.

Wolf, Eric R. 1982. *Europe and the people without history*. Berkeley: University of California 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. https://doi.org/10.1146/annurev-anthro-092611-145721

Christina Wasson is Professor of Anthropology at the University of North Texas. She was trained as a linguistic anthropologist. After finishing her Ph.D., she worked for E-Lab, a design firm that used anthropological research to develop new product ideas. Here she developed an interest in the emergent field of design anthropology. Christina was a founding member of the Ethnographic Praxis in Industry Conference (EPIC) Steering Committee. At UNT, Christina teaches a course in design anthropology that prepares students for careers in this field. Clients for class projects have included the Nissan Research Center-Silicon Valley, Motorola, Microsoft, and the Dallas/Fort Worth International Airport. Christina's current focus on language archives brings together her interests in linguistic anthropology and design anthropology, and enables her to contribute to projects that address deeply felt concerns in stakeholder communities. For more information, see https://www.christinawasson.com.

Melanie Medina is an M.A. candidate in Applied Anthropology at the University of North Texas. She received her bachelor's degree in Anthropology from Florida Atlantic University. Currently, she is attending UNT in order to pursue a career in applied anthropology, specifically with an interest in implementing it in the video game industry and its surrounding cultures. Melanie recently worked with a partner to conduct a project with a retro video game store local to Denton, Texas in order to provide recommendations to the store's owners regarding its product layout and in-store gaming tournaments. Melanie is currently working as an instructional assistant for UNT's Anthropology department.

Miyoung Chong is a doctoral candidate in the College of Information, University of North Texas. Her research interests include media effect analysis, as well as new media theory including social media, film, and mass media. She is also interested in data applications of open data and computational linguistics. She has presented her research studies in international conferences including ASIST and iConference, and is associate editor for a special issue of the journal *Triple Helix*.

Brittany LeMay is an M.S. candidate in Applied Anthropology at the University of North Texas. She received her bachelor's degree in Anthropology from the University of Toledo. She is interested in environmental and urban anthropology, and plans to use her master's degree to take a solutions-oriented approach to environmental justice issues. She is currently consulting on a project for her local city government on how to increase community engagement in local environmental sustainability efforts. Brittany is also employed as an instructional assistant for the Anthropology Department at UNT while she continues working toward her master's degree.

Emma Nalin is an M.S./M.P.H. candidate in Applied Anthropology and Community Health at the University of North Texas. Her research interests include health equity, cultural conceptions of good health, and healthcare access in rural communities. Emma joined Dr. Wasson's language archiving project in January 2017 and conducted field research in Northeast India the following year. Other projects include her thesis research, "Strengthening Donor Commitment to a Free Clinic for the Uninsured" in upstate NY, and "Perceptions of the Water Energy Nexus in North Texas Households," directed by Dr. Jamie Johnson (UNT). Emma received a BA cum laude from the University of Notre Dame in 2016, majoring in anthropology and music. In addition to attending graduate school at UNT, she currently works freelance as a user research contractor for User-View, Inc.

Kenneth Saintonge is an M.S. candidate in Applied Anthropology at the University of North Texas, where he conducted research as a student in Dr. Christina Wasson's Design Anthropology course. He received his bachelor's degree from Eastern Connecticut State University in Art History and Sculpture. Presently, he volunteers with AmeriCorps VISTA in partnership with Texas Parks and Wildlife Department, developing methods that create greater access to the parks' resources and programs for underserved populations. Kenneth plans to use methods from design anthropology in future research, as he follows his interests at the intersection of educational and cultural institutions in the areas of art, history, material culture, digital culture, and identity.