

# Collaborating with Language Community Members to Enrich Ethnographic Descriptions in a Language Archive

Mary Burke<sup>†</sup>

Department of Information Science

University of North Texas

Denton TX, USA

Mary.Burke@unt.edu

## ABSTRACT

Language archives connect users such as language communities, linguists, and other researchers, to language data. As the language archiving community develops, concerns have been raised about the ethics, ownership, accessibility, and context of archival materials. While there are no simple solutions to these questions, many language archives are seeking ways to involve language community members in these conversations as they continue. This presentation describes a pilot project undertaken at the Computational Resource for South Asian Languages (CoRSAL) which explores a collaborative archiving approach to enable language community members to tell their own stories by adding contextual information to archival materials.

## KEYWORDS

ethnographic metadata, collaborative archiving, contextual metadata, unknown-provenance language materials, language community partnerships

## 1 Introduction

Language archives are cultural heritage institutions serving as repositories of primary language data: material about or in a set of languages (audio and video recordings, transcriptions, translations, and linguistic annotations). Archival materials are meant to serve as a lasting record of the language, and the starting point for further linguistic analysis or creation of pedagogical materials [1]. Language archives connect users such as language communities, linguists, and other researchers to language data. As the language archiving community develops, concerns have been raised about the ethics, ownership, accessibility, and context of archival materials [1, 2, 3]. While there are no simple solutions to these questions, many language archives are seeking ways to involve language community members in these conversations as they continue. This presentation describes a pilot project undertaken at the Computational Resource for South Asian Languages (CoRSAL) which explores a collaborative archiving approach to enable language community members to tell their own stories by adding contextual information to archival materials.

## 2 Related Work

Language documentation is the subfield of linguistics dedicated to preserving linguistic diversity. The process of language documentation is closely related to other subfields of linguistics, such as language description, language revitalization, language archiving, and to other disciplines, like information science, archival studies, anthropology, and ethnobiology. This section briefly reviews recent work concerning the relationship between language communities and language archives.

As the field of language documentation re-reorients to prioritize the needs of language communities, language archives too are seeking ways to be maximally useful to language communities as well as academic audiences. We see this incorporated into the design of language documentation projects themselves (see [4] on Community Based Language Research) and in the way we think about language archives. For example, [5] encourages depositors to consider the potential audiences of their archival collection, and to describe the material in a way that is appropriate for those audiences, taking into consideration factors like their primary languages and domain knowledge. More recently, [6] recommend for language documenters to discuss language communities' unmet needs during the documentation process, and work with language archives to make collections accessible despite the target community's specific barriers to access (e.g., minimal internet access).

Recent work in this area has noted the integral role that community engagement and rich contextual descriptions play in facilitating access to archival materials [7, 8, 9, 10]. Through discussions of the ethics of appropriating materials and framing community stories in non-community perspectives, many have called for increased involvement of language communities in the archiving process [11, 12]. However, linguists and their research team are often the ones responsible for managing the data, creating metadata, and depositing material into an archive. Language archive metadata records typically include the following elements: Identifier, Title, Contributor/ Depositor/ Creator, Language, Date, Description, Format, Notes, Rights, and Related items [13], with a high degree of variability in free-text descriptions noted by [14]. So, the

metadata that accompanies items in a language archive is based on the information that the research team might use to identify an item. This includes information like the names of those in the video, the date it was recorded, a genre, or the name of a story or song, but may exclude crucial cultural context, like that this song is only sung at a particular festival or by certain individuals. See, for example, [15] for a recent project where a language community representative was hired to identify gaps and errors in metadata in PARADISEC legacy material from Papua New Guinea.

### 3 Project Description

In light of these developments in language archiving, we saw an opportunity to test out a workflow which allows language community members to add in cultural context to already existing metadata. In the summer of 2020, two students at the University of North Texas (UNT) were hired to add cultural information to two collections in CoRSAL, briefly described here.

#### 3.1 Burushaski Language Resource

Javid Iqbal, a Linguistics Masters student, is a Burushaski speaker. Before coming to UNT, he worked at the Burushaski Research Academy (BRA) as a research officer documenting cultural events and coordinating community meetings to raise awareness about the status of the language. He engaged with the Burushaski Language Resource, developed by Dr. Sadaf Munshi, which contains audio and video recordings of traditional, historical, and personal narratives, songs and poems, conversations, and recipes.

#### 3.2 Lamkang Language Resource

Sumshot Khular, from the Lamkang community, is currently a PhD student in Environmental Studies, and earned her Masters in Linguistics from UNT in 2018. She has been supporting the Lamkang language for decades in numerous capacities (e.g., translating the Universal Declaration of Human Rights into Lamkang; organizing workshops, documenting natural speech and community events). Khular contributed metadata to the Lamkang Language Resource, which was developed over the course of the Lamkang project (2008-present) by Dr. Shobhana Chelliah, Sumshot Khular, Rex Khullar, Daniel Tholung, among other Lamkang community members. The Lamkang Language Resource contains digitized printed material on Lamkang culture, and primary audio and video recordings of traditional narratives, procedural narratives, semi-guided narratives (pear stories), conversations, and songs.

Both collections include a large proportion of photographs of cultural events, community members, and significant places or items (e.g., plants, utensils, churches). Because Khular and Iqbal have both experience in language documentation work and expertise in their respective cultures, they were uniquely positioned to add contextual information to the material in these collections. Though the items in these collections were already accompanied by metadata, they added metadata in the Description and Subjects fields with cultural significance in ways beyond the ability of the

original collectors or current CoRSAL staff. After the pilot project, the metadata was reviewed and copy-edited by CoRSAL staff to ensure consistency. The following section summarizes the improvements that were made to the collections, including examples of the metadata records before and after the project.

## 4 Contributions Made

### 4.1 Cultural Context

The community consultants identified the salient aspects of items to highlight in the Description field, such as the significance of the colors or weaving pattern used in a shawl, or which occasions a garment might be worn on. Their additions are particularly invaluable for those items which were initially contributed with little or no metadata by community members. See for example, Figure 1, which compares the original metadata record and the record after the Lamkang community consultant updated it. The earlier version of the Description field, for example, states that this is a photo “illustrating Lamkang culture,” but it is not clear what aspect of culture is intended, or what context the photo was taken in. The new description expands on the traditional clothing items worn by the young dancer, the materials used to make them, and the event where the photo was taken.

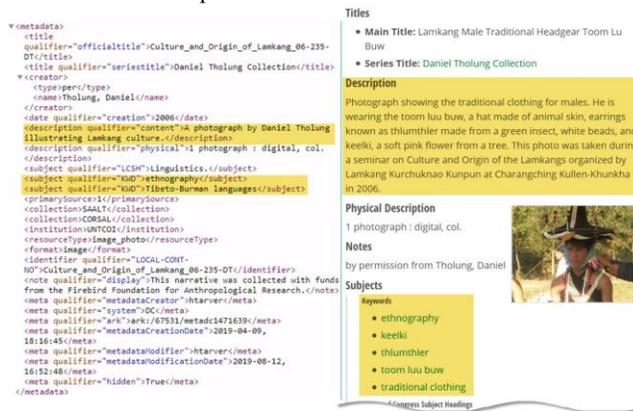


Figure 1: Example of a record before (left) and after (right) changes to the Keywords and Description fields

With this added information, the photograph is now connected to several others demonstrating the traditional clothing of the Lamkangs with the keyword ‘traditional clothing’. Further, the updated description notes that these photos were taken during an event for celebrating and educating about Lamkang culture in 2006.

### 4.2 Target Language Metadata

Language community consultants also added information in the target language; for example, for the photograph of Lamkang children in traditional dress (Figure 1 above), the Lamkang community consultant explained that the dancer is wearing a *toom luu buw* (hat) and *thlumthler* (earrings). For the same photograph, another metadata creator would likely have used generic subject

terms (e.g., ‘Ethnic costume’ from the Library of Congress Subject Headings). Though not inaccurate, a term like ‘Ethnic costume’ may not be as helpful as a Lamkang term to users whose primary language is Lamkang.

In some cases, CoRSAL staff were able to identify the scientific name for the plants and animals featured in the photographs with the help of community consultants, such that the final record contains the Lamkang name of the animal, the English name, and the scientific name. Compare again the metadata records before and after the pilot project for this photograph of the antlers of a hog deer.

**Figure 2: Example of a record before (left) and after (right) changes to the Keywords and Description fields**

While the initial metadata included only ‘antlers’ and the Lamkang term, adding the English term for the animal allowed us to identify the scientific name. The resulting record contains useful information for Lamkang speakers, as well as users interested in the wildlife of Northeast India.

In the Burushaski collection, the community consultant added the names of the recipes and dishes featured in photographs and audio recordings. See for example, Figure 3, which compares the metadata record for the same photograph in March 2020 and July 2021.

**Figure 3: Example of a record before (left) and after (right) changes to the Keywords and Description fields**

Note how the Burushaski name of the dish is included in multiple varieties of Burushaski, while the previous version of the metadata record did not have a complete Content Description.

## 5 Summary and Future Plans

As a result of this pilot project, the metadata for the Lamkang and Burushaski collections is more accurate, complete and culturally relevant. The contextual information added by the community consultants will improve the experience of Lamkang or Burushaski speakers using these collections, as well as those interested in the respective cultures. Given this positive experience, we intend to replicate this process with future incoming collections whenever possible.

## REFERENCES

- Henke, Ryan, & Berez-Kroeker, Andrea. (2016). A brief history of archiving in language documentation, with an annotated bibliography. *Language Documentation and Conservation*, 10, 411–457.
- O’Meara, Carolyn, & Guadarrama, O A. G. (2016). Accessibility to results and primary data of research on indigenous languages of Mexico. In *Language Documentation and Revitalization in Latin American Contexts* (pp. 59–80). De Gruyter Mouton. doi: 10.1515/9783110428902-003
- Wasson, C., Holton, G., & Roth, H. (2016). Bringing user centered design to the field of language archives. *Language Documentation and Conservation*, 10, 641–671.
- Czaykowska-Higgins, Ewa. (2009). Research Models, Community Engagement, and Linguistic Fieldwork: Reflections on Working within Canadian Indigenous Communities. *Language Documentation & Conservation* 3, 15-50.
- Woodbury, Anthony. (2014). Archives and audiences: Toward making endangered language documentations people can read, use, understand, and admire. *Language Documentation and Description*, 12, 19-36.
- Seyfeddinipur, Mandana, Felix Ameka, Lissant Bolton, Jonathan Blumtritt, Brian Carpenter, Hilaria Cruz, Sebastian Drude, Patience L. Epps, Vera Ferreira, Ana Vilacy Galucio, Brigit Hellwig, Oliver Hinte, Gary Holton, Dagmar Jung, Irmgarda Kasinskaite Buddeberg, Manfred Krifka, Susan Kung, Miyuki Monroig, Ayu’nwi Ngwabe Neba, Sebastian Nordhoff, Brigitte Pakendorf, Kilu von Prince, Felix Rau, Keren Rice, Michael Riessler, Vera Szoelloesi Brenig, Nick Thieberger, Paul Trilsbeek, Hein van der Voort, & Anthony Woodbury. 2019. Public access to research data in language documentation: Challenges and possible strategies. *Language Documentation & Conservation*, 13, 545-563.
- Greyling, Elizabeth, & Zulu, Siphso (2010). Content development in an indigenous digital library: A case study in community participation. *International Federation of Library Associations and Institutions*, 36, 1, 30–39. doi: 10.1177/0340035209359570
- Roeschley, Ana, Kim, Jeonghyun, & Zavalina, Oksana L. (2020). An exploration of contributor-created Description fields in participatory archives. In Sundqvist A., Berget G., Nolin J., Skjerdingstad K. (Eds), Sustainable Digital Communities. iConference 2020. Lecture Notes in Computer Science, 12051 (pp. 638-648). Springer, Cham. [https://doi.org/10.1007/978-3-030-43687-2\\_54](https://doi.org/10.1007/978-3-030-43687-2_54)
- Sullivan, Ryan. (2020). Archival description for language documentation collections. *Language Documentation & Conservation*, 14, 520-278.
- Burke, Mary, Zavalina, Oksana L., Chelliah, Shobhana, & Phillips, Mark (in press). User Needs in Language Archives: Findings from Interviews with Language Archive Managers, Depositors, and End-Users. *Language Documentation and Conservation*.
- Ormond-Parker, Lyndon, Corn, Aaron, Forde, Cressida, Obata, Kazuko, & O’Sullivan, Sandy. (2013). (Eds.) Information Technology and Indigenous Communities. Canberra: AIATSIS Research Publications.
- Shepard, Michael A. (2016). The Value-Added Language Archive: Increasing Cultural Compatibility for Native American Communities. *Language Documentation & Conservation*, 10, 458-479.
- Burke, Mary, & Zavalina, Oksana L. (2019). Exploration of information organization in language archives. *Proceedings of the American Society for Information Science and Technology*, 56, 1.
- Burke, Mary, & Zavalina, Oksana L. (2020). Descriptive richness of free-text metadata: a comparative analysis of three language archives. *Proceedings of the American Society for Information Science and Technology*, 57, 1.
- Harris, Amanda, Gagau, Steven, Kell, Jodie, Thieberger, Nick, & Ward, Nick. (2019). Making meaning of historical Papua New Guinea recordings:

Collaborations of speaker communities and the archive. *International Journal of Digital Curation* 14, 1, 136-149.