# Linguistic Archives and Language Communities Questionnaire

Establishing (Re-)Use Criteria

Ilya Khait
Leibniz-Zentrum Allgemeine Sprachwissenschaft
Berlin Germany
ilya.o.khait@gmail.com

Leonore Lukschy
Endangered Languages Documentation Programme
Berlin-Brandenburg Academy of Sciences and Humanities
Berlin Germany
lukschy@bbaw.de

Mandana Seyfeddinipur
Endangered Languages Documentation Programme
Berlin-Brandenburg Academy of Sciences and Humanities
Berlin Germany
seyfeddinipur@bbaw.de

## ABSTRACT

Digital language archives hold vast amounts of materials in endangered or marginalised languages. However, due to limitations in technical infrastructure and the design of these archives, the materials are usually not easily accessible to speakers of the languages represented or their descendants. With the goal to establish best practices for researchers archiving linguistic data, this paper presents a questionnaire designed to assess how archival materials can be made more readily available to language communities.

## CCS CONCEPTS

• Information systems ~ Information systems applications ~ Digital libraries and archives • General and reference ~ Document types ~ Surveys and overviews

## KEYWORDS

Language archives, Endangered languages, Heritage materials

## 1 Introduction

For many years scholars have conducted fieldwork and have worked with speakers and communities describing and documenting their language and knowledge. Recordings created during these projects have been archived in large digital archives usually housed in the global North in academic environments. The advent of the digital held the promise of democratisation of access to knowledge for everyone. However, 20-30 years down the line communities cannot easily access their own recordings in these digital archives. Studying and overcoming this issue falls within the scope of QUEST (Quality - Established: Testing and application of curation criteria and quality standards for audiovisual annotated language data), a collaborative project that aims to establish curation criteria for digital language data for its subsequent use. In this paper, we present survey work on how communities access information digitally to understand how archives and researchers can ensure that the data they collect and preserve can be made accessible and discoverable to the very communities they come from.

We use the term ‘language archive’ to refer to digital archives of educational or memory institutions like universities or libraries, such as the members of the DELAMAN network, holding primary and secondary language documentation data, in the form of audio and video recordings, images, transcriptions, and other texts. The term ‘language community’ or simply ‘community’ is used to refer to speakers of endangered or marginalised languages and their descendants. When referring to researchers we use ‘outsider researcher’ for linguists or anthropologists who work with language communities they are not considered a part of.

The importance of making language materials such as video and audio recordings, as well as texts, available to the communities who provided the data has been recognised by a number of researchers (see [1]; [2]; [3]). Different attempts have been made to create archive interfaces that are designed for communities they serve [3]. AILLA, the Archive of the Indigneous Languages of Latin America, has a Spanish Interface making the materials discoverable in the most widely spoken language of Latin America. The Language Archive (TLA), holding the materials from the Volkswagen Foundation funded projects, developed portals for the public, and the Dane project developed a community portal [1].

Given the lack of electricity and digital connectivity, outsider researchers documenting languages have attempted leaving the recordings with the community on physical

storage devices such as tapes, CDs, hard drives, USB drives, or SD cards. The shortlivedness of the physical storage devices unfortunately meant that communities only have access to their own recordings for a short amount of time.

Long-term access can be provided through digital language archives, however the materials held there are not easily accessible and discoverable because of the way the interfaces are designed and the fact that the interface language is mostly in English. Several researchers have pointed out that large language archives are geared towards an academic audience rather than the broader public or the communities whose languages are represented in said archives (see [4], [5]).

Below we overview some known access biases and present the work on a questionnaire that surveys access to language materials and media usage in general by community members, then discuss the preliminary results obtained thus far and how archival materials can be made more readily accessible to communities in the future.

## 2 Restrictions

Informal interviews with linguists conducted between November 2020 and April 2021, who had carried out fieldwork in close collaboration with communities in Cameroon, Papua New Guinea, Peru and Colombia, and Vanuatu, highlighted the following obstacles to communities finding and accessing data in their languages.

### 2.1 Technical infrastructure

While internet coverage, particularly mobile, is spreading rapidly [6], communities speaking marginalised languages tend to live in areas with little access to the basic technical infrastructure needed to use digital media. The availability and affordability of electricity, digital devices and online access cannot be taken as a given. Even when the latter conditions are met at least to a certain extent, limits in web data volume, bandwidth or of the devices and software in use can also hinder access to media. There can be other obstacles as well, as e.g. in some areas, where mobile internet access is tied to certain platforms such as Facebook's Free Basics initiative [7], thus barring free surfing in the browser.

### 2.2 User skills and environment usability

The lack of computer literacy or the more basic written barrier, particularly in older generations, can hold language community members from discovering and accessing digital media. While generally more proficient in modern technology, younger people are likely to use mobile devices rather than computers. Practically, it means that in many cases data can be made available only through mobile-ready interfaces. It is also not uncommon that one's online experience is limited to a few popular applications.

Literacy is another barrier, discovering and accessing information takes place through reading and writing. To access information on the web literacy in a majority language is a requirement. Concerning digital language archives in particular, it is important to note the linguistic barrier with strong English bias tendency, the use of specific terminology, User Interface complexity and multilayerness that can sometimes be puzzling even for linguists. In addition, often additional steps like registering or requesting access for sensitive data makes data even more inaccessible. Although graded access has many advantages in terms of protecting speakers (see [8]), it also represents another barrier in navigating materials.

### 2.3 Discoverability

Discovering materials in archives is possible through text-based metadata. In addition both the catalogue and the metadata of large international digital language archives are mostly in English. Although there has been a push towards multilingual metadata in recent years, implementing this is still largely the responsibility of individual researchers. AILLA, the Archive of the Indigenous Languages of Latin America, stands out as an example of a large-scale language archive with bilingual metadata in English and Spanish.

Another obstacle to discovering materials is that the metadata categories and contents are grounded in linguistic categories which are relevant to mostly linguists. For descendants of speakers to find recordings of someone from their family, particular metadata such as a person's full name might be relevant. However, some informed consent protocols and recent data protection requirements lead to anonymisation of the speakers making their recordings not discoverable.

## 3 Designing the questionnaire

The primary aim is to develop an understanding what type of material communities would access through what digital medium. This understanding will allow in turn to build interfaces for e.g. social media channels serving the recordings relevant to the community as it is likely that recordings of an elder telling a story is of more value than some linguistic elicitation.

The secondary aim is to design a guide for researchers going into the field, so that they may address the question of how to make materials accessible to the community they work with at the start of the project rather than retroactively. This pertains to collecting community metadata which will be very different from metadata relevant to linguists.

In order to assess this, we designed a questionnaire divided into the following seven sections:

1) General Information

2) Materials of Interest

3,4) Existing Recordings

5) Recordings available online

6, 7) Connectivity, devices, and platforms

In order to obtain answers about communities who might currently not have access to the internet, we designed a slightly different version of the questionnaire for outsider researchers, such as linguists working with a community. The questionnaire targeted at language community members has thus far been translated into English, Russian, Spanish and French. It consists of 27 questions and takes ca. 10 minutes to complete. Links to the forms are provided in the appendix below.

It should be noted that there is a bias in who will answer the questionnaire as it is currently only being distributed online as a Google Form, and is only available in a limited number of major languages.

# 4 Preliminary results from the questionnaire

We will report preliminary data from 12 respondents (Three outsider researchers, 9 language community members (seven in English and two in Spanish). The respondents represent the following languages and regions: Yoruba, Igbo (Nigeria), Quechua (Bolivia), Bora (Peru), Punjabi (Pakistan), Shugnani (Tajikistan), Rejang (Indonesia), Khakas, Negidal (Russia, Altai and Far East respectively), Tsova-Tush (Georgia), Guernsey French (UK), and Irish (Ireland). Of the respondents who are part of a language community, five are female and four are male; five persons are in their twenties, three in their forties, and one respondent is in their late fifties. Five are native speakers, while three speak the language fluently and one speaks it a little. Below follows a brief summary with highlights of the results at hand, both from outsider researchers and community members.

## 4.1 Community interest in recordings

Popular genres of interest are

- dictionaries (twelve respondents)
- language learning materials (eleven respondents)
- knowledge about animals and plants, family and kinship and local history (ten respondents)
- stories, conversations and crafting knowledge (nine respondents)
- rituals (eight respondents)
- linguistic materials (seven respondents)
- knowledge about hunting, fishing and harvesting (six respondents)

As for media, text and video are convenient for most (ten each), leaving audio (nine) and images (eight) slightly behind.

## 4.2 Recordings shared with the community

All but one respondent report that recordings were made available to the language community. While website links are frequent (eight), digital carriers are used as often (a USB stick is mentioned four times and others such as hard drive, SD-card, DVD, CD appear once each). Materials on paper were distributed in four cases, and two respondents mention an analogue cassette. Nine respondents affirm that community members tried to access these materials and nine that they are aware of recordings of their language online (in most cases naming popular social media). Only in six cases people looked in digital archives and of those two encountered difficulties and one did not find any data.

## 4.3 Connectivity and communication

All respondents state that most people in the community have Internet access. Moreover, the internet is so good everywhere that they all can watch videos. It looks like most people have mobile phones, and seemingly in many cases these are actually smartphones. PCs are less common and tablets are rare.

Most popular platforms are

- WhatsApp (all twelve)
- Facebook (eleven) and Instagram (ten)
- Twitter and Skype (six)
- Tik Tok (four),
- Telegram (three)

Most language communities (ten) use phones to communicate, but almost as common are messengers and social media (nine) and text messages (eight). Somewhat

less widespread are emails (seven), video calls (six) and voice messages (four). The post is named only in two cases.

## 5 Discussion

The answers obtained thus far show active use of modern digital media among marginalised language communities all around the globe, with clear preference given to mobile devices and popular messengers and social networks.

As discussed in section 3, there is a bias in respondents due to the languages and means of distribution of the questionnaire as well as the small selection of responses collected so far. Moreover, community members filling in the questionnaire might be more likely to already have an interest in accessing materials in their language. In order to get a broader sample of answers, it would be ideal to distribute a printed version of the questionnaire, potentially via researchers conducting fieldwork.

Making materials accessible needs to be an ongoing process, particularly as communities' access to technical infrastructure is rapidly evolving. Rather than trying to teach speakers to access their data in complex archiving environments, materials should be made available to them via platforms they already know to navigate. A possible development in this direction would be, for example, designing language archive chat bot interfaces for popular messengers to search and deliver the data.

To all effects and purposes, researchers should discuss making materials available at the beginning of a documentation project to best serve community needs.

## 6 Appendix: Questionnaire links

Community members:

- English
  https://forms.gle/dqvyGNmVHA5uoBCMA
- French
  https://forms.gle/f5uxQuYKZpMhCDHW7
- Russian
  https://forms.gle/BNLm7BHe2CgYJecA9
- Spanish
  https://forms.gle/LKGWw2DHTzjYNcDq6

Outsider researchers (English):

- https://forms.gle/44HXXkGrcJJCq4xA8

## ACKNOWLEDGMENTS

The QUEST project is funded by the German Federal Ministry of Education and Research (BMBF). We are very thankful to the respondents for dedicating their time and information to assist us. We are grateful to our colleagues, Dr. Jocelyn Aznar, Prof. Dr. Manfred Krifka, and Dr. Frank Seifart for fruitful discussions and valuable suggestions. Dr. Aznar is also to be especially thanked for assistance with translation.

## REFERENCES

[1] Trilsbeek, Paul, and Dieter van Uytvanck. 2009. 'Regional Archives and Community Portals'. *IASA*32: 69–73.

[2] Wasson, Christina, Gary Holton, and Heather S. Roth. 2016. 'Bringing User-Centered Design to the Field of Language Archives', December, 641–81.

[3] Nordlinger, Rachel, Ian Green, and Peter Hurst. 2019. 'Working at the Interface: The Daly Languages Project'. Edited by Linda Barwick, Jennifer Green, and Petronella Vaarzon-Morel. *LD&C Special Publication No. 18: Archival Returns: Central Australia and Beyond*, Indigenous music of Australia, , 193–216..

[4] Holton, Gary. 2012. 'Language Archives: They're Not Just for Linguists Any More'. *Language Documentation & Conservation*Special Publication No. 3: Potentials of Language Documentation: Methods, Analyses, and Utilization: 105–10.

[5] Woodbury, Anthony C. 2014. 'Archives and Audiences: Toward Making Endangered Language Documentations People Can Read, Use, Understand, and Admire'. Edited by David Nathan and Peter K. Austin. Language Documentation and Description, no. 12: Special Issue on Language Documentation and Archiving: 19–36.

[6] Roser, Max, Hannah Ritchie, and Esteban Ortiz-Ospina. 2020. 'Our World in Data: Internet'. Our World in Data. 2020. https://ourworldindata.org/internet.

[7] Henning, Maximilian. 2019. 'How the Global South Can Protect Itself from Digital Exploitation'. *LATITUDE*, 2019. https://www.goethe.de/prj/lat/en/dis/21670998.html.

[8] Seyfeddinipur, Mandana, Felix Ameka, Lissant Bolton, Johnathan Blumtritt, Brian Carpenter, Hilaria Cruz, Sebastian Drude, et al. 2019. 'Public Access to Research Data in Language Documentation: Challenges and Possible Strategies'. *Language Documentation* 13: 545–63.