



Organised jointly by
ELAR and PARADISEC

Language Documentation and Archiving

During the
International Decade of
Indigenous Languages

BOOK OF ABSTRACTS
2026



2022-2032 | INTERNATIONAL DECADE OF
Indigenous Languages

CONTENTS

Talks 2

Posters..... 88

PAPER & TALKS PRESENTATIONS

The ELM Protocol: Remote Language Documentation and Community-Led Archiving in Conflict-Zone Realities

Sajda Abdelaal (SudaLingua Project)

In the context of the 2023 Sudanese conflict, traditional in-person field linguistics has become functionally impossible, threatening the survival of highly endangered Nilo-Saharan languages. This paper introduces the Emergency Language Management (ELM) Protocol, a decentralized framework developed by the SudaLingua Project to facilitate language documentation during active telecommunication blackouts.

The ELM Protocol utilizes a tiered technological hierarchy: Starlink satellite connectivity provides the backbone for “Camp Champs”—local speakers within refugee settlements—to record high-fidelity audio using the Lingua Libre mobile ecosystem. These data packets are then managed by urban “Remote Workers” who perform metadata enrichment and quality control using the Meta-Wiki infrastructure. This system allows for the continuous documentation of phonetic and lexical data without requiring the physical presence of external researchers.

This study analyzes the implementation of the protocol, focusing on the preservation of Sudanese Arabic dialects and minority languages. We present an assessment of data integrity collected via satellite vs. traditional networks and discuss the ethical implications of the “Camp Champ” stipend model (\$900) as a form of professional integration. By integrating intellectual property protections with open-access archiving, the SudaLingua Project demonstrates a resilient, community-owned model for linguistic stewardship in crisis. The results suggest that the ELM Protocol offers a scalable blueprint for documenting endangered heritage in “Zero-Connectivity” zones globally.

Multilingualism as a determining factor for the linguistic revitalization of the Yawalapíti of Alto Xingu (Brazil)

João Carlos Albuquerque Souza de Almeida (University of Brasília)

This study aims to describe the process of linguistic revitalization of the Yawalapíti in recent decades. This group is situated in the Upper Xingu, state of Mato Grosso, Brazil, in the transition between the Cerrado and the Amazon. Their language belongs to the Arawak family, though it is distinct from the other three Arawak languages in the region (Mehinaku and Wauja), one of which is extinct (Kustenu). In addition to Arawak-speaking groups, Tupí-speaking (Kamayurá and Awetí) and Karib-speaking (Kuikuro, Kalapalo, Matipu, and Nahukua) groups inhabit the Upper Xingu. This multilingual and multi-ethnic cultural system involves different peoples participating in shared rituals and social networks. Historically, the dissolution of the Yawalapíti caused by past epidemics and the subsequent regrouping, combined with a policy of linguistic exogamy, created a multilingual configuration in the main village where the Yawalapíti language eventually lost ground to Kamayurá and Kuikuro.

Since the 2000s, the Yawalapíti have planned to revitalize their language by teaching it to children and strengthening intergenerational transmission—which had been previously interrupted—utilizing methods similar to the Master-Apprentice methodology (Hinton et al., 2002). This initial effort produced several semi-speakers who now seek to replicate the process with their children. This movement gained urgency after the COVID-19 pandemic, which resulted in the death of the former village chief and one of the primary speakers. What began as family-led intergenerational transmission projects, with full speakers teaching their children, evolved into a community project, expanding to other families and the village school. This teaching process required the production of didactic materials which, in turn, necessitated the documentation of the last three remaining full speakers. Today, a language committee oversees short, medium, and long-term goals, supported by an ongoing documentation process and database creation (Adamou, 2024).

From this ongoing process, I highlight essential questions regarding language and identity within this dynamic, multi-ethnic context (Dorian, 1989). Ethnic belonging, language proficiency, ritual dynamics, and linguistic performance are central points of debate. Furthermore, the development of an orthographic convention is vital for affirming Yawalapíti identity, especially in contrast to Portuguese and other indigenous languages of the Upper Xingu. I propose that the multilingualism of the Yawalapíti, and their familiarity with the diverse languages of neighboring peoples, facilitates the learning process and serves as a powerful motivation to reclaim their linguistic identity.

Remote Fieldwork in Multilingual Environments: The Limits of Observing Repertoire Management in Zargari (Iran)

Ioana Aminian Jazi

(Austrian Centre for Digital Humanities, Austrian Academy of Sciences)

Repertoire-based models of multilingualism conceptualize language contact not as interaction between discrete systems but as the dynamic selection and inhibition of feature sets within a shared communicative repertoire (Matras 2009/2020; 2018; 2021; 2024). Convergence, compartmentalisation, matter and pattern replication emerge from context-bound selection and inhibition processes within a multilingual repertoire rather than from transfer between bounded linguistic systems. Yet many documentation frameworks remain structured around identifiable linguistic systems and speaker- reported language choices.

This paper builds on insights from recent work on Zargari contact features (Aminian Jazi & Matras 2026), which examined structural outcomes of simultaneous contact with Azeri Turkish and Persian in Zargari, an endangered Romani variety spoken in north-western Iran (Windfuhr 1970, 2002; Baghbidi 2003). Speakers distribute features across domains, interlocutors, and affective registers in ways that resist system boundaries. Intergenerational stratification further complicates this ecology: elders acquired Persian later in life; middle generations maintain triadic domain differentiation; younger speakers show Persian dominance with partial Zargari and Azeri competence. Individual biographies reveal repertoire configurations in which Zargari indexes intimacy, secrecy, and emotional intensity, Persian institutional authority and mobility, and Azeri extended kin networks and cross-village interaction. These are not stable “language choices,” but shifting activations of feature clusters.

Following nationwide protests that began in 2022 with the “Woman, Life, Freedom” movement and subsequent periods of intensified political repression and communication blackouts, sustained in- situ participant observation became impossible. Documentation shifted to remote collaboration under conditions of internet shutdowns, sanctions, power cuts, and uneven digital literacy. The loss of co- presence exposes a methodological paradox: repertoire theory depends on observing interactional ecology, yet remote documentation fragments that ecology. This paradox becomes particularly acute in interviewer positioning. The linguistic and social profile of the interviewer conditions repertoire activation. A Zargari insider may elicit default, unmarked repertoire pools in which background knowledge and local references remain implicit; an external assistant may trigger accommodation and hyper-articulation. Gender, age, and social alignment further modulate access to domains and speech styles. The choice between emic and etic positioning is therefore not merely ethical or logistical but epistemic: each configuration reshapes the very repertoire one seeks to document.

The design of research instruments compounds this issue. System-oriented questions (“Which language do you use at home?”) risk reifying separations that repertoire theory problematizes. Biographical narratives must instead trace domain-sensitive feature distribution without presupposing stable linguistic systems, yet remote elicitation may prompt speakers to rationalize fluid practices into categorical identities.

The Zargari case foregrounds a broader challenge for language documentation under conditions of remote fieldwork: if what counts as a distinct “language” emerges from processes of selection and inhibition within a shared multilingual repertoire, how can documentation capture repertoire dynamics when co-presence is lost and interviewer positioning, political constraint, and methodological framing actively reshape the repertoire under observation?

I'm working hard, but it's hardly working: Insights from User Experience methods on digital infrastructure in language documentation

Sunkulp Ananthanarayan (University of Washington)

There is a tradition of (rightful) concern regarding the future-readiness of language data within the documentation and archiving community (cf. [1,2,3]). However, the workflows that produce archival deposits are shaped by a fraught and volatile ecosystem of under designed digital tools to prepare data. Documentary linguists routinely assemble workflows from specialized linguistic software (e.g. FLEEx), generalized productivity tools (e.g. Excel), and improvised pipelines while striving toward best practices expected from archives, funders, and institutions for well-structured digital deposits and theoretically significant publications. Despite the importance of these documentary workflows for strong archival records, little empirical study has been undertaken to understand how documentarians interact with their tools.

This paper reports results from a two-phase user experience (UX) study investigating determinants of utility and usability within the technological ecology of language documentation and description, limiting scope to processes before data reaches the archive and discussing how that affects archival workflows. I combine a user survey with qualitative context from user interviews with documentary linguists at different career stages. Within UX studies, utility is defined as a tool's capability of executing requested tasks, and usability refers to the ease and smoothness with which it is used. These come together to give an app's usefulness [4].

Survey results suggest that the structure of linguistic data has a strong influence on software choice. Text corpus workflows incorporate more specialized tools than other language data-related tasks (such as lexicography and audio data, as opposed to metadata-related) because annotation requires editing across various points in otherwise hierarchical data, and users appear to have a strong preference that a text preparation tool be helpful by linking more data, which can go awry with cutting across levels of hierarchical data.

We also see more users abandoning specialized tools, compared to generalized tools, due to cycles of de facto deprecation resulting from insufficient resources for maintenance. These patterns have direct downstream consequences for the entire ecology of language documentation, archiving, and description by shaping the very structures we (stakeholders and outsiders alike) deposit, access, and analyze. To explain these patterns, this paper introduces a typology of documentation tools defined along two dimensions: (i) was the tool designed for this task?, and (ii) what is the epistemic distance between the developer and end user (i.e., is the developer also the user, a linguist, or a bona fide software developer?)

Interview data further reveals two user personae characterizing the ways users situate themselves across their software. Technologist-users construct their own pipelines, either by building tools themselves or constructing elaborate workarounds to get past existing software's limitations. Commissioner-users request external developers to create a more robust or professional tool. Technologists often achieve high task-specific utility through iterative customization, augmenting usability as habits develop in their new systems. Commissioners end up with highly usable applications, developing utility through iteration and discussion with developers. Unfortunately, better outsourcing is difficult to arrange—tools created by linguists external to the project often beget major usability pains, now due to habit mismatch when a user is working with a tool created by a domain expert. The reason for this is that the external linguist is bound not to have the same idiosyncrasies as the user nor the training of a software professional (nor certainly a designer).

Taken together, the findings suggest that usability and utility in language documentation—and therefore the sustainability of its downstream shape—are shaped by three interacting factors: (1) the structural properties of

the data being managed, (2) the extent to which a tool is built bespoke for a task, and (3) the distance between developers and end users. I conclude that sustainable archival infrastructure requires not only better tool design but also institutional recognition of the time, effort, and money spent on labor involved in infrastructure design for the creation and maintenance of documentation workflows which constitute how we interact with the data we seek to fortify. Rather than argue against reinventing wheels as [5] does in service of building effective archive interfaces, instead my study supports becoming more explicit about the expanded purview of documentary linguistics as technology has grown more capable, acknowledging that upstream tooling for documentarians¹ is of a different nature from stakeholder-facing structures. Creating the space for inclusion of design practices when we build digital infrastructure can strengthen portability and allow for better institutional memory, in user-commissioned and -built tools.

1 Including stakeholder/insider documentarians—the focus is on which hat they wear for a process and who is determining their resource allocation

Linguistic studies of indigenous sign languages in the Brazilian Amazon: An overview

Paulo Jeferson Araújo (Universidade Federal de Roraima)

Due to the International Decade of Indigenous Languages (2022-2032) and the creation of a Working Group focused on Indigenous Sign Languages in Brazil (GT-LIS), Indigenous and non-Indigenous researchers have carried out descriptions and documentations of sign languages used by deaf Indigenous people in various Brazilian states. In this work, I present an overview of the research and some discussions that have been raised in events and instant messaging groups, such as the role of Indigenous researchers in the description of these languages, the language status of the gestural systems used by deaf Indigenous people, and the political-academic effects of the Working Group on Indigenous Sign Languages in making them more visible in the research agenda focused on Amazonian languages. As a starting point, I use the case of the Macushi Indigenous Sign Language (LISMac), currently being documented and described in the state of Roraima, as an example, and then draw a parallel with other researches in other countries with similar agendas. From issues of language policy, such as the teaching of sign languages in indigenous communities in relation to the national sign language, the Brazilian Sign Language (Libras), to more specific theoretical issues in the field of research, such as the gesture-sign continuum, these are highlighted as essential for establishing research on these languages as an integral part of the linguistic and cultural diversity of the Amazon.

Underrepresented Languages in AI: Language Documentation and Language Preparation for Digital Inclusion

Gratien Gualbert Atindogbe, Rachel A. Ojong Diba (University of Buea)

Language inclusion is essential for fairness, accessibility, and global technological participation. Yet African languages remain severely underrepresented in human language technology, often absent from machine-readable audio datasets and the large language models (LLMs) used to train speech recognition, synthesis, and language analysis systems. As a result, millions of speakers of African languages are effectively excluded from emerging AI-driven technologies and digital services. If this gap persists, African languages risk further marginalisation in the rapidly expanding AI ecosystem. This position paper argues that language documentation is not only a response to the loss of languages, cultures, and social memory, but also a crucial step toward addressing the persistent absence of African languages in AI systems. Documentary linguistic data, including records of linguistic practices, everyday speech interactions, and speakers' metalinguistic knowledge (Himmelmann, 2006) provide essential resources for building robust language technologies capable of maintaining performance, consistency, and reliability across diverse prompt conditions (Zhang et al., 2024). However, documentation alone is insufficient. Many African languages still lack well-annotated audio datasets, diverse speech samples, parallel corpora, and large monolingual text collections necessary for training modern AI systems. We therefore introduce the concept of Language Preparation (LP), defined as the systematic evaluation of the linguistic and technological readiness of a language for integration into LLMs. LP involves assessing literacy levels, the availability of annotated speech datasets, representation in commercial technologies, participation in academic or industrial research, and the broader digital resources available for the language. Drawing on documentary research and the Theory of Digital Transformation (Mergel et al., 2019), we argue that LP represents a necessary stage in the development of inclusive language technologies. By identifying gaps in standardisation, digital resources, institutional support, and technological representation, LP helps guide the development of AI-ready linguistic datasets. Finally, because AI systems must generate human-like speech and text, documentation efforts must capture speakers' full linguistic repertoires rather than focusing solely on a single "target" language such as a first or dominant language. As AI continues to reshape knowledge production and communication, improving language documentation practices to produce comprehensive and robust training data for technologies such as automatic speech recognition (ASR) and machine translation (MT) should become a central priority.

Community Access to Language Documentation: Digital Dissemination of Oirat-Kalmyk Audiovisual Materials

Chingis Azydov (Commission Vanishing Languages and Cultural Heritage, Austrian Academy of Sciences)

Language documentation projects increasingly face the challenge of making archived materials accessible and meaningful to the communities whose linguistic and cultural heritage they represent. This paper presents a case study of digital dissemination strategies developed within the work of the Commission Vanishing Languages and Cultural Heritage for audiovisual materials documenting the Oirat-Kalmyk language and folklore.

Field expeditions conducted in regions of the Oirat-Kalmyk diaspora (including Kalmykia, Western Mongolia, Kyrgyzstan, and Xinjiang) have produced a large collection of interviews, songs, and oral histories recorded in audio and video formats. The most complete form of these materials is published on the commission's website, where recordings are presented as video clips with subtitles containing a phonetic transcription of the Oirat language and an English translation, accompanied by a description of the video clip and a PDF file of the subtitles. However, experience has shown that the majority of interested viewers encounter these materials through the commission's YouTube channel.

To further expand outreach, I launched an Instagram account in 2020 to connect with younger members of the Kalmyk community and increase awareness of the project. Audience reactions revealed an important challenge: many viewers found the Latin-based phonetic transcription and English translation used in the videos difficult to follow. To improve accessibility, I decided to create alternative versions of selected clips with Kalmyk Cyrillic subtitles and Russian translations. This led to an increase in the number of views, shares, and accounts reached.

The paper discusses the practical and conceptual implications of adapting documentation outputs to different digital platforms and audiences. It also outlines ongoing plans to expand the production of Kalmyk- and Russian-subtitled videos for dissemination through Instagram, Facebook, and VKontakte, reflecting on how social media can support community engagement with language documentation archives.

Rapid Tooling with AI-Assisted Coding: A Workflow for Streamlining ELAN-Based Transcription in Language Documentation

Wolfgang Barth (Australian National University)

The transcription and annotation of field recordings is one of the most labour-intensive bottlenecks in language documentation. This paper presents a practical, replicable workflow for addressing this challenge. Using large language models as an interactive coding collaborator, documentarians can rapidly develop lightweight, purpose-built tools. No advanced programming expertise is required.

The core contribution of this paper is the workflow pattern itself: an iterative cycle of problem identification, AI-assisted code generation, and deployment. This approach fundamentally changes what is feasible for individual researchers or small teams. Tool development for repetitive tasks is compressed from weeks to hours, freeing researchers to focus on the steps that require human judgment.

We ground the discussion in concrete examples from an ELAN-based transcription pipeline. These include tools for automatic identification of annotation segments (Fig. 1) and bulk management and editing of ELAN files across a corpus (Fig. 2). We also address a key ethical challenge. Many language documentation corpora are held under access restrictions that prevent sharing with third-party AI services. Our solution is synthetic data cloning, generating surrogate data that replicates the structural and statistical properties of restricted materials without retaining actual content. This allows AI tools to be used for more tasks without compromising data sovereignty obligations.

We reflect on the broader implications for documentation practice, including what kinds of problems suit this approach, how to communicate effectively with an AI coding collaborator, and how tools can be shared and adapted across projects. We close by arguing that rapid AI-assisted tooling deserves recognition as an emerg-ing documentation skill in its own right.

Digitally Responsible Documentation: Archiving and Empowering Kwegu and Nyangatom Communities in Ethiopia

Tesfaye Negash Bayou (Kotebe University of Education)

As a documentary linguist who has been conducting fieldwork on the endangered Kwegu and Nyangatom languages of Southern Ethiopia since 2024, I have assembled a substantial corpus of primary data. To date, I have collected, transcribed, translated, and ELAN-annotated 5 Kwegu narratives and 5 Nyangatom narratives, and compiled a lexical database of over 2000 words for each language using FLEEx software. This work forms the Ethiopian contribution to a five-year international collaborative project (2024–2029) led by Japanese institutions (Yamaguchi University, Tokyo University of Foreign Studies, Osaka University, Kobe University) and supported by Kotebe University of Education (Inui et al. 2024). This paper outlines a methodological framework for the next phase: mobilizing these documentary materials to directly benefit Kwegu and Nyangatom speakers through ethical, community-controlled, and context-appropriate tools. The framework responds to the critical need identified by Seifart et al. (2018) to address the “transcription bottleneck” and ensure that documentary materials serve community priorities alongside scholarly goals.

The role of ethnograrchaeology in language documentation: a case study of pottery manufacture

Bernat Bardagil & Peter Coutro (Ghent University)

As orally transmitted means of knowledge and communication, material culture manufacture traditions are very closely connected to language and are often sought within the scope of language documentation, as well as of reclamation and revitalization initiatives. This paper discusses the contributions of ethnoarchaeological approaches to the documentation of pottery in the context of documentary linguistics. Being dependent on suitable combinations of environmental and cultural factors (Arnold 1985, 2000), it is no surprise that pottery is relevant both for the perception of cultural identity in one's community and in others, and for establishing a connection between cultural maintenance and the natural environment (Garcia-Soto et al. 2024). It has been pointed out by Silva (2021) that pottery and ceramics are "important in the construction and maintenance of identities, communicating multiple symbolic meanings, and their capacity to embody cosmological understandings and otherness beyond human beings." Pottery is also a relevant area for observing and documenting teaching-learning processes (Roux 2016). From a linguistic perspective, high-resolution studies of pottery have proven to be crucial in historical linguistics, both for tracking patterns of lexical and cultural diffusion and for establishing sound correspondences with different branches of a language family, as in Bostoen (2007). This paper illustrates the polyhedric value of thorough ethnolinguistic documentations of pottery manufacture with a focus on two areal case studies, namely the Guaporé-Mamoré region in Amazonia, and the Kwilu-Kasai and Congo-Ubangi regions in central Africa. While creating quality documentation records of the process of pottery manufacture alongside its social, spiritual, environmental and familiar contexts is in itself valuable, we show that these processes are especially suitable for the organization of linguistic and cultural reclamation activities, in the sense of community practice-based revitalization programmes (Amaral 2020:31), for establishing a conversation with cultural heritage originated in the archaeological record, and for stimulating local sustainable economic practices, as well as addressing challenges in the institutional preservation of documentary materials. An exhaustive study of the presence of pottery manufacture in language documentation repositories (e.g., ELAR, CLA, DoBeS), via the metadata included in the file bundles, reveals a much smaller presence of ethnoarchaeological elements than one might expect, and especially so in the case of pottery manufacture, with as low a presence as 3%-8%, depending on the areal focus. We believe that the smaller than ideal presence of pottery in language documentation collections (and initiatives) is a result of, on the one hand, a general feeling of unpreparedness on the part of linguists when faced with the task of doing justice to the documentation of a complex chaîne opératoire, and, on the other hand, a relative misinformation about the crucial role that pottery manufacture plays in linguistics (e.g., lexicon, procedural discourse, language-environment interaction) and other domains, among other causes. With this paper and the accompanying guided documentation questionnaire, we hope to remedy that and contribute to a larger presence of pottery manufacture in language documentation.

Language documentation training efforts for members of Micronesian communities in Hawai‘i

Kelsey Bialo, Violet King, Jillian Breithaupt, Yumi Standlee, Hunter Procter, Jaeyong Yang, Gary Holton, Leah Pappas, Bradley McDonnell (University of Hawai‘i at Mānoa)

Over the past five decades, people from ‘large ocean states’ of the western Pacific (commonly known as Micronesia) have migrated to Hawai‘i in relatively large numbers, bringing with them a rich linguistic and cultural heritage. However, rather than fostering a sense of pride, this heritage often results in linguistic discrimination, leading many children to not learn their heritage language. At the same time, despite decades of research dating back to the 1960s which produced an impressive number of grammars and dictionaries, the vast majority of Micronesian languages lack modern documentary corpora, and few of these publications have effectively supported development of pedagogical and curricular materials which support language maintenance efforts, particularly within the diaspora context. Finally, very few members of these communities have emerged as language researchers or activists, thus emphasizing a need for targeted linguistic training.

In an effort to increase both the quality and quantity of documentation of Micronesian languages and provide training and capacity-building for members of various Micronesian communities in Hawai‘i, faculty and graduate students at the University of Hawai‘i have developed several interrelated initiatives. The Language Documentation Training Center (LDTTC) has offered community-facing workshops which focus on building capacity in language documentation skills among members of the public (Ajo et al. 2010; Sou et al. 2023). Workshops are hands-on and result in documentation that is immediately useful, such as audio recording for podcasts, stop-motion videos for narratives, and online audio-visual books. More recently, LDTTC has developed partnerships with a local non-profit organization serving Chuukese and Marshallese communities with the goal of creating documentation that connects elders to younger members of these communities.

The Micronesian Undergraduate Research Apprenticeship in Linguistics (MURAL) operates with the same principles but is targeted towards undergraduate students at the University of Hawai‘i who are from Micronesia or of Micronesian heritage. MURAL workshops are held weekly, led by faculty and graduate student mentors from the Department of Linguistics, and cover a wide range of topics related to linguistics. Apprentices learn the methods and tools of language documentation through collaborative, hands-on activities, ultimately developing a linguistics research project and creating a curated documentary collection of audiovisual recordings of speech events that are transcribed, translated, and glossed. Currently in its second year, the MURAL program has had two cohorts of apprentices representing three speech communities across Micronesia, each producing original documentation and applied research projects.

In this presentation, we detail the initiatives described above and demonstrate how they are locally situated and able to integrate language documentation efforts with pressing needs that are unique to the diaspora in Hawai‘i, while recognizing the value of diasporic language communities for language documentation efforts. In the end, we demonstrate how these efforts provide yet another model for collaboration in language documentation that critically involves training and capacity building (e.g. Fitzgerald 2018). We also show how these efforts contribute to our understanding of models of language work among diaspora communities (Kaufman 2018, Bax et al. 2024).

Distributing Legacy Reading Materials in the Virtual Age: Digitizing Traditional Yapese Stories Using ReadAlong Studio

Jillian Breithaupt (University of Hawai'i at Mānoa)

Yapese (Glottocode: yape1248) is an Austronesian language spoken primarily in Yap State, Federated States of Micronesia (FSM) (Blust 2013). The speaker population is estimated at about 5,000 within the state, with additional speakers in diaspora communities located primarily elsewhere in Micronesia and in the United States (Yap Statistics Office 2010). Despite the presence of several indigenous languages in the country, the official language and lingua franca of the FSM is English. In Yap, it is the primary language used in both public and private schools, while Yapese is only taught marginally during elementary and lower secondary education. This prevalence of English has caused concern among educators and family members regarding the acquisition of Yapese. In order to address this, the Yap Department of Education (DOE) partnered with Pacific Resources for Education and Learning (PREL) from 2011-2012 in an initiative to improve the curriculum for both English and vernacular (Yapese) language arts in order to promote bilingualism (Yap DOE 2015). However, the outcomes of this initiative have only been implemented in public schools, and do not benefit children who are enrolled in private school or living outside of the country. Therefore, there remains a need for educational materials available to Yapese speakers outside of the public education system.

To date, the only written Yapese language materials available to the public are a collection of stories and pedagogical materials developed by PREL in the 1980s and a series of children's books created by the Yap DOE in the early 2000s. However, they are not easily accessible. The online domain that previously hosted the collection of PREL stories has recently become unavailable, so they now only exist as hard copies at Yap DOE and digital scans housed on Google Drive. The Yap DOE children's books are also kept in these locations, with additional copies available at a handful of libraries. Therefore, the current project was developed in the summer of 2025 as an ongoing collaboration between the author and members of the Yapese community in an effort to address the limited accessibility of existing reading materials, particularly for children living outside of Yap State. Furthermore, we hope to improve the utility of these resources by making them multimedia with the addition of an audio element.

This presentation will detail the development of the current project in collaboration with community stakeholders, and its ongoing implementation. Traditional Yapese stories and those of cultural relevance from the Yap DOE and PREL collections were chosen by collaborators. The chosen stories were then read by an adult Yapese speaker and recorded using a Zoom H4n audio recorder with a Shure SM35 headset microphone. The recordings were then imported into ReadAlong Studio, a program developed by National Research Council Canada and Carleton University to create indigenous audiobooks. Two read-along versions of each story were created: one that used the orthography from the original source, and one where the text has been rewritten by the speaker recording the story. This was done to account for differences between the orthography used in education and what is used in daily life. The process of distributing these read-along books to the community, and challenges that have arisen throughout the project will also be discussed.

The colonial archive and the moribund languages Bom-Krim and Sherbro

Jennifer Brunner (University of Graz)

Finding archived material recorded in the 1950s and 1960s that is barely known, inaccessible, not digitized, and related to languages that are probably already extinct is a linguistic treasure. Walter Pichl, an Austrian linguist who worked very actively in West Africa – especially in Senegal and Sierra Leone—for around 20-30 years in the last century, left behind more than 200 recordings, thousands of pages of unpublished manuscripts, a collection of more than 250 ethnographic objects of unknown provenance, boxes of slides, photographs, films, official documents, and personal correspondence, including letters from Léopold Sédar Senghor, the first president of Senegal. For our purpose focus on Pichl's recordings and written material concerning the South-Atlantic languages of Sierra Leone, namely Sherbro and Bom-Krim. The latter are two highly endangered varieties, perhaps already completely extinct. This material raises several central questions: How can such heritage be used in a decolonial, empowering, community-driven way? How can it be made accessible to descendants and local researchers within a framework that respects ethical and legal restrictions? And how might research on these materials contribute to the restitution and restoration of cultural knowledge and artifacts removed within a colonial framework?

Digitization is the first step toward accessibility and addressing these questions. It enables the creation of an international and interdisciplinary research collaboration that can recontextualize the material and build an environment to retell the story of these language communities. From the perspective of language documentation and description, further concrete tasks emerge: Bom/Krim have not been extensively documented or described. The resources of Pichl that have been published and made accessible have been extremely valuable to the few researchers who have worked on these languages in the following decades. As Tucker Childs puts it in his 2025 posthumously published Sherbro grammar: "His (Pichl's) data merits review" (Childs 2025). This is especially true since Childs—who worked on each of these three languages and conducted ELDP projects on Bom-Krim in 2007 and on Sherbro in 2014 – found that only "a few score speakers use Kim while even fewer know Bom" (Childs 2010).

This linguistic reality differs markedly from Pichl's earlier observations. In the 1970s he reported that "Krim was only spoken by a few hundred people" (Pichl 1972: II), still distinguishing it from Bom. Although Pichl already perceived it the languages as highly endangered in the 1960s, the forty separating his work from that of Childs represent a significant period in the trajectory of language attrition.

Comparing the archival material of Pichl with that of Childs provides insights not only into the processes of language shift and loss. Pichl's recordings may also offer a glimpse into the past cultural practices that had already disappeared by the turn of the millennium and are absent from Childs' recordings. At the same time, Childs' audiovisual documentation may reveal practices that Pichl could already observe but could record only in audio form. The coexistence of completed ELDP projects on moribund – or possibly extinct – languages alongside much older, largely forgotten and inaccessible archival materials on the same languages constitutes a rare case. Our project therefore asks how colonial archives can be integrated into the broader aims of contemporary language documentation and preservation. Pichl's materials, together with Childs' ELDP projects on Bom-Krim and Sherbro, provide a case we aim to bring into discussion as a possible model for working with historical linguistic archives.

Navigating Skepticism and Solidarity: Solo Fieldwork Realities in the Malavedan Community

Bhavya C P (University of Delhi)

Documenting a severely endangered language requires navigating through a complicated network of differing expectations from the government, the researcher, and the speakers themselves. This paper reflects on the methodological and ethical realities of a self-funded, solo female researcher conducting fieldwork to document Malavedan, an endangered Dravidian language within the highly restricted, state-protected forest regions of Kerala, India (Eberhard et al., 2025). The Malavedan community faces profound linguistic attrition and broken intergenerational transmission (Author, 2025). Yet, they exist at a difficult intersection of external and internal expectations: the government demands conversational fluency in Malavedan to issue crucial tribal certificates, questioning their integrity and enforcing a rigid ideology of linguistic authenticity as a prerequisite for indigenous identity (Bucholtz & Hall, 2004) while the community itself earnestly desires language preservation and frequently meets to discuss the creation of an original script.

Simultaneously, deep-seated skepticism toward outsiders, fueled by fears of linguistic extraction, exploitation of tribal benefits, and a history of negative external interactions, created an immediate barrier to traditional documentation (Smith, 2021; Czaykowska-Higgins, 2009). The initial access to the settlements, facilitated by government “promoters” and after a long and difficult bureaucratic verification process, showed that standard elicitation tools were not enough. The pre-prepared questionnaires didn’t work very well because the language was too forgotten to answer strict questions.

As a result, the methodology had to change from formal interviews to deep participant observation, spontaneous speech recordings, and relationship-based elicitation (Chelliah & de Reuse, 2011). This was done with low-resource workflows like borrowed departmental equipment and smartphone recordings. Navigating this highly guarded environment as a solo female researcher proved to be a distinct advantage, as noted by Abbi (2001). Trust was built slowly, not through language prompts, but through shared humanity: listening to the problems of community members, especially a large group of single mothers who were pushed to the edges of society by out-marriage, sharing daily gossip, and playing with their kids. The research was ultimately enabled by the Malavedan community’s profound hospitality, which significantly contrasted with their initial wariness. Despite the community’s own anxieties and limited financial resources, this care was evident even in small details, such as vegetarian families cooking and packing non-vegetarian meals specifically to ensure the researcher’s well-being. This hard-won trust yielded fragmented but vital linguistic data. Because Malavedan lacks an orthography (Kumar et al., 2023), the dominant regional Malayalam script was used as a bridge for quick field notes. These notes were later compared to raw audio to transcribe unique, non-Malayalam phonetic inventories into IPA.

This paper argues that in extremely vulnerable contexts, the conflict among state mandates, community preservation objectives, and the practicalities of solo fieldwork calls for the abandonment of rigid methodologies in favor of deeply empathetic, relationship-oriented documentation. It highlights how navigating differing expectations requires a researcher to integrate into the community, honoring their experiences as much as their spoken word.

Documenting Karara: an extremely endangered language in a region with indigenous people living in voluntary isolation

Carmen Brasolin (Instituto de Filosofia e Ciências Humanas da UNICAMP)

The Karara language has only recently entered academic and public knowledge following a language documentation project that began in 2023 and is still ongoing. The language's last remaining speaker is the sole survivor of an ambush led by the Indigenous Bureau Service in Brazil. In that event a family was captured from an isolated village on the Carará River in 1962. Studies now show that karara is parte of the Pidjanan subgroup of the Arwak family, historically documented in the region but spoken by minority groups that many believe no longer exist.

In recent years, there are reports of indigenous communities living in isolation nearby the Hexkaryana-speaking villages of the region where the Karara family also lives. In other words, indigenous communities in isolation, like the Karara family once was, still inhabit the area. Their territory in the Brazilian northern Amazon region is still undergoing a legal process to be recognized as both an indigenous land inhabited by people in voluntary isolation and an indigenous land for the Hexkaryana-speaking community. How does a community engaged with the protection of people living in voluntary isolation engage in the documentation of an extremely endangered language spoken by a last survivor of a former isolated community?

In this paper, I will discuss how the documentation of the Karara language and culture has been instrumental in encouraging the surrounding community to protect those who remain isolated in the forest, who may be Karara language speakers or speakers of other unknown languages. As documenting and qualifying these isolated communities requires a non-contact approach according to national and international laws, I will address how the community views endangered language documentation as part of an effort to raise awareness of their language and culture and secure recognition for them.

A Practical Guide to Immersive Language Documentation

Whitney Cary (SIL Global), Miles Cary (SIL Global)

This paper describes the creation of an immersive language documentation corpus of the Tibeto-Burman language Pahari (phj). We reflect on the documentation process used and provide recommendations to practitioners, communities, and archives for including immersive media in documentation corpora. Immersive media formats increase the holism of documentation corpora by situating speech content within a wider representation of the physical and sociocultural environment. While these technological innovations bring new concerns regarding ethics, efficacy, and accessibility, they also have the potential to provide many advantages to researchers and communities.

Beginning in 2023, we created a documentation corpus in two varieties of the Pahari language using four types of immersive media. The corpus includes over 15 hours of 360° video with ambisonic audio, 245 360° photos, and 16 3D models. In accordance with the desires of the Pahari community, the corpus is easily accessible to community members and the general public through publication of 360° videos with annotations on a dedicated YouTube channel. Additionally, 360° photos and 3D models have been published in an online interactive virtual tour with embedded language and cultural information.

We present the recording methods, data processing workflow, development of presentation formats, and challenges experienced with immersive media in this project. We conclude with suggested guidelines for the capture, processing, archiving, and publication of immersive media in language documentation projects.

‘Yuyayninchikta Yachachinallapuni Manaraq Chinkapushaqtin’ Women Documenting Their Language and Life Narratives

Noemy Condori Arias (University of California Santa Barbara)

South Bolivian Quechua Oral Traditions plays a crucial role in understanding Quechua culture, knowledge, and the language itself. In the past, it was passed down through daily activities like farming, herding sheep, bedtime storytelling, or when people from different communities gathered to exchange food. However, due to migration from rural communities to big cities and other countries, many traditional Quechua contexts have changed significantly.

In response to this, I want to share our project: *Yuyayninchikta yachachinallapuni manaraq chinkapushaqtin* ‘Sharing our knowledge before it’s too late’. This project began in the city of Cochabamba, within a three-generation family, Grandmother, mother and granddaughter reflecting about our own oral traditions within the communities of Vacas. As Quechua community members, mothers, teachers and researchers, our main goal is to document folk stories, legends about the creation of animals and nature, horror stories, traditions, women life stories as a way of resistance and the history of the town. During the recording sessions, children and youth are encouraged to listen to their parents narrating these tales.

In Quechua communities, women have historically played a central role in intergenerational language transmission and the preservation of oral traditions. Mothers and grandmothers not only raise children but also shape their linguistic and cultural identities, teaching community values and preparing them to live as Quechua people. Within this context, women-led language documentation emerges as a powerful and intentional act of cultural continuity. By consciously recording oral traditions, life histories, and local knowledge, women transform everyday practices of transmission into explicit forms of language activism and revitalization.

As a Quechua daughter, I grew up hearing my mother encourage other mothers to continue transmitting our language. At the time, I did not recognize these conversations as acts of language activism aimed at maintaining our *Qhishwa simi*. Today, I understand them as challenging efforts that resist language shift and cultural erasure. Such movements are urgently needed to support community-driven revitalization and to confront the ongoing loss of Indigenous languages.

In this presentation, we share our locally grounded methods for documenting oral traditions and women’s life stories entirely in Quechua. My mother, Indalicia, leads the engagement with elders, and we shift roles depending on the social and generational context. While documenting within our own community offers important advantages: such as trust, shared knowledge, and linguistic intimacy; it also presents unique ethical, emotional, and methodological challenges. Finally, as part of our documentation project, we are also developing materials and making them accessible to Quechua communities in Bolivia (printed audiobooks), as well as to speakers, heritage speakers, and learners worldwide (online). Some of our recently published materials include a story-book with audio recordings, translated into Spanish and English, and a *Lottery* game that was used to engage children from the Vacas-Cochabamba communities.

Training Future Information Professionals with Digital Language Archives

Serge I. Coronado (University of North Texas, Texas A&M University)

The LAMlangArc Training Project (<https://ci.unt.edu/lamlangarctraining/>) develops experiential training for emerging information professionals working with digital language archives. Funded by the Institute of Museum and Library Services (RE-254860-OLS-23) and led by the University of North Texas College of Information in collaboration with linguistics and archival partners, the project addresses the growing need for professionals equipped to steward archival materials that support community memory, language documentation, and cultural revitalization.

This paper presents outcomes from the Summer 2025 pilot offering of the graduate course Community Language Archiving and Curation for Information Professionals. The course integrates theoretical foundations of community archives with hands-on engagement with language archival materials, including digital collections, metadata design, and preservation strategies. Students worked with real-world archival resources and case studies to explore how archival materials can be mobilized for research, education, and creative reinterpretation.

The pilot course served as a testing ground for an interdisciplinary training model that bridges library and archival studies, linguistics, and digital humanities. Through project-based assignments, participants developed practical skills in curating community language materials and designing access pathways that support both scholarly and community uses. Outcomes from the pilot - including student projects, feedback, and curriculum refinements - demonstrate how archival training initiatives can expand the educational and cultural impact of language archives.

By highlighting lessons learned from the Summer 2025 pilot, this paper contributes to discussions on how archival materials can be activated within educational contexts and community-centered initiatives. The results inform future iterations of the course and broader efforts to build sustainable capacity for community language archiving and cultural heritage stewardship.

Extending oral annotation: Integrating automatic speech recognition and multi-speaker workflows in language documentation

Christopher Cox (Carleton University (Canada))

In oral annotation workflows, interpretative information is added to audiovisual documentation in the form of verbal annotations, with language consultants providing audio-recorded repetitions and/or spoken translations of timestamped sections of source recordings (Reiman 2010, Boerger 2011). While oral annotation techniques have been argued to have considerable advantages over written-first workflows in many common documentary contexts, lowering barriers to participation in annotation activities sometimes posed by writing and sidestepping aspects of the transcription bottleneck (cf. Khagdi 2017), as Berez-Kroeker, Gabber, and Slayton (2023: 202) note, these methods have not yet found broad adoption among documentation practitioners, in part for having not “demonstrated that [oral annotation] eventually results in widely repurposable textual transcriptions.” Moreover, current software applications aimed at supporting oral annotation, such as SayMore (Hatton 2013), currently provide only a single, non-overlapping layer of time-aligned annotations as the basis for oral annotation, offering limited support for oral annotation to integrate naturally into the multi-tier annotation structures that are common in documentation projects that include conversational discourse and interactional language use.

In this paper, we introduce a suite of open-source software tools that seek to address the above issues, extending the use of current oral annotation software to documentation involving multiple speakers (cf. Cruz 2016) and facilitating the integration of automatic speech recognition (ASR) into oral annotation workflows. Using these tools, it is possible to:

- Convert sets of oral annotations recorded in SayMore into stand-alone audio files with matching, pre-segmented ELAN transcripts to which existing ASR tools can be readily applied; then re-integrate automatically recognized text back into the original SayMore project;
- Transfer transcribed text between single-tier SayMore oral annotation projects and multi-speaker transcripts of a source recording in a round-trip fashion, allowing the transcribed text of oral annotations managed in SayMore and the corresponding time-aligned annotations in multi-speaker transcripts of the source recordings to remain in sync;
- Export sets of transcribed oral annotations in formats amenable to archiving, creating audio files with matching time-aligned transcripts that preserve textual transcriptions and provenance information for all oral annotations. While current versions of SayMore can export sets of oral annotations as multi-channel WAV files, neither the transcribed text of these annotations nor any information about their relationship to the source recording is preserved, leaving the exported audio without any written annotation and limiting its use in further stages of documentation.

We demonstrate the use of these tools through the example of a multi-year documentation project focused on conversational discourse, showing how both these extended, multi-speaker oral annotation techniques and accompanying use of ASR have supported the creation and archiving of richly annotated master recordings and oral annotations with interlinking, time-aligned transcripts.

The Afterlife of a Documentation Project: Challenges, Progress, and Lessons Learned

Vijay D'Souza (*NEILAC*)

This presentation traces the aftermath of an ELDP-funded IGS documentation project among the Hrusso Aka community of Arunachal Pradesh, India. While the literature on language documentation methodology, revitalisation, and collaborative documentation with endangered and indigenous communities is well-developed (Austin, 2021; Himmelmann et al., 2006; Hinton et al., 2018; Leonard & Haynes, 2010; Smith, 1999; Woodbury, 2011 among others), scholarship exploring what happens after a documentation project concludes remains scarce. This presentation is a contribution to that underexplored space, reflecting critically on what a documentation project can set in motion beyond its formal conclusion.

The project was carried out between 2015 and 2020 by a team comprising the presenter and Hrusso Aka native speakers working in close collaboration. In discussions leading up to project implementation, the team decided to adopt an integrated approach incorporating the following five goals: (1) Documentation, (2) Language description, (3) Revitalisation, (4) Responding to community needs through respectful linguistic intervention, and (5) Capacity building (equipping community members with documentation and revitalisation skills).

This integrated and collaborative approach led to a language revitalisation movement within the community. Team members have continued documentation, literary production, recovery of lost narratives and mythologies, and have produced podcasts, textbooks, children's literature, and news broadcasts. A nascent but significant movement has emerged in which educated community members are now choosing to transmit Hrusso Aka to their children at home — a reversal of the near-total shift to Hindi among educated young couples in the recent past. A particularly important institutional development was the invitation extended to team members to form the literary wing of the Aka Shotuko Kunu, the apex council of the Hrusso Aka community, leading to the formation of the Aka Language Academy in 2021, now entrusted with all language-related matters.

The presentation will be delivered in three parts: first, the five goals and how they shaped the longer-term aftereffects of the project; second, post-project strategies, initiatives, and challenges; and third, a reflection on how these developments represent a hopeful phase of revitalisation and how, despite persistent challenges such as continued shift to Hindi, dialectal uncertainties, and funding gaps, they reflect a meaningful attitudinal shift in favour of Hrusso Aka.

The presentation will conclude by drawing methodological, ethical, and institutional lessons for similar documentation projects, before reflecting on what the Hrusso Aka experience contributes to a broader theoretical question: whether documentation and revitalisation should be reconceived not as separate endeavours but as two dimensions of a single, community-oriented practice.

The home of language: Documenting multimodal language use in interaction

Josh Dahmen (Australian National University)

Language documentation aims to create and preserve a “lasting, multipurpose record of a language” (Himmelmann, 2016). Advances in portable documentary technologies now make it possible to capture language use within face-to-face interaction – the home environment of language (Sidnell, 2016) – in multi-party settings with high-resolution images and sound.

The benefits of including recordings of authentic conversations in language documentation projects have long been noted (see, e.g., Childs et al., 2014; Evans & Dench, 2006). However, analyses of recording types in several language archives suggest that conversational data are still relatively underrepresented. Mihas and Santos Pérez (2017) note that the likely reason is that speakers of endangered languages tend to use more widely spoken languages in everyday interactions, but they also argue that it could indicate “the depositor’s reluctance to tackle the messiness of the naturally occurring conversation” (p. 2).

Such conversational messiness (e.g., mispronunciations, corrections, cut-offs, repair, non-lexical vocalisations, dialectal and idiosyncratic variation, codeswitching and codemixing) indeed poses challenges for language documenters whose goal may be to capture the underlying context-free structure of a particular language variety (see also Dahmen, 2022). While a context-free structural focus is essential for understanding the linguistic system and is often more easily transformed into language resources for the broader community, it rarely provides a true representation of the language as it is actually used in face-to-face interaction.

This paper advocates for the systematic inclusion of unelicited interactional and conversational data in language documentation and discusses how such recordings can be integrated into documentary and descriptive projects. Drawing on fieldwork experience in Australia and New Guinea, it addresses (i) practical techniques, including state-of-the-art camera and microphone setups, strategies for minimising the observer’s paradox (Labov, 1972), and tips for transcribing multi-party talk; (ii) ethical considerations, such as informed consent, post-recording review with speakers, and the management of sensitive content (e.g., talk about third-parties); and (iii) the range of community outcomes that are supported by this type of documentation.

I argue that recordings of interactional data that are void of any particular research agenda allow the widest range of future uses – from reliable quantitative analyses of token frequencies to qualitative studies of gesture and multimodal interaction, as well as applications for community resources focusing on conversational practices.

IPA-based text-to-speech for documenting the endangered Hong Kong Tanka

Dai Yanmei (Department of Chinese Language Studies, The Education University of Hong Kong)

The Tanka language of Hong Kong, traditionally spoken by maritime communities known as the Tanka or boat people, is an endangered Sinitic variety undergoing rapid language shift toward Cantonese (Wang et al., 2025). Despite its historical and cultural significance in the maritime societies of South China, systematic digital documentation of Tanka remains limited. A major challenge for computational approaches is that the language lacks a standardized writing system, making conventional grapheme-based speech technologies difficult to apply.

This study explores the use of IPA-based text-to-speech (TTS) technology as a tool for documenting the phonological, prosodic, and lexical features of Hong Kong Tanka. Fieldwork recordings are used to construct a speech corpus including narrative discourse, lexical elicitation, and culturally embedded speech genres. The corpus is transcribed using detailed International Phonetic Alphabet (IPA) representations and aligned with multiple annotation layers capturing segmental structure, tonal patterns, prosodic phrasing, and culturally specific vocabulary related to maritime life and traditional fishing practices. These multilayer annotations provide a linguistically grounded representation of the sound system and cultural lexicon of the language. To enable speech synthesis for a language without a standardized writing system, the study adopts a phoneme-driven modeling strategy in which IPA sequences serve as the primary input to the TTS system. This approach builds on recent dialect speech synthesis frameworks that utilize unified IPA representations and dialect-aware modeling architectures (Chen et al., 2025). Such models employ techniques such as mixture-of-experts architectures and parameter-efficient adaptation to support speech synthesis for extremely low-resource varieties using limited training data. For languages without writing systems such as Tanka, IPA-based TTS provides a practical approach for preserving phonological structure and expanding the methodological toolkit of digital language documentation.

Goyatì Eyits'q Gonàowò Ilè Zq Ne; The Interplay of Language Documentation and Digital Archiving in Tłıchų Yatì

Danielle Dacanay (Tłıchų Government), Renee Saucier (Tłıchų Government),
Rebecca Bourgeois (Western University)

Tłıchų yatì (ISO:dgr, formerly Dogrib) is a Dene language spoken by ~2000 people in the Northwest Territories, Canada (Statistics Canada 2021). Academic documentation of the language has been sparse in comparison with its neighbours (i.e. Cook 2004; Rice 1989), and what little has been done has come primarily in the form of sporadic fieldwork conducted by unrelated groups of external organisations (such as universities (i.e. Saxon 1986), museums (i.e. Howren & Howren 1983), and missionary collectives (i.e. Davidson 1959)). As a result, most detailed documentation of the language lies unpublished, or internally published, in various archives (i.e. Howren 1967; Lafferty n. d.) and private collections (i.e. Ackroyd 1982). These collections are overwhelmingly located outside of Tłıchų territory, and as such, remain inaccessible (and often entirely unknown) to the great majority of the Tłıchų population.

Community-led documentation of Tłıchų yatì suffers a similar problem. Since the latter quarter of the twentieth century, various Tłıchų governmental organisations have become involved in the production of language resources, coinciding with a heightened interest in Indigenous culture and self-government. Much of this 'auto-documentation' consists of audio recordings of elders, basic educational materials, and transcriptions of legends and oral history, produced on an idiosyncratic, project-to-project basis. Although these materials have generally remained locally stored, their piecemeal origins have made them difficult to access for most interested parties, scattered across the collections of local schools, governmental offices, and communal institutions. To date, no exhaustive catalogue of existing community-made language materials has yet been produced, nor does any standard pathway exist for accessing these materials.

To address this, since 2022, the autonomous Tłıchų Government has provided permanent funding for both a Language Revitalisation Division, responsible for first-party language documentation and resource creation (i.e. Anonymised 2026, forthcoming), and for a stewarded archival collection, the Tłıchų Nàowò Deghats'etsaa (or Tłıchų Cultural Commons Digital Archives). Although founded separately, both entities share significant common cause with respect to language materials; namely, both seek the identification and compilation of existing materials, the digitisation and repatriation of materials held outside of Tłıchų territory, and the storage of these materials in centralised, accessible repositories.

Our paper therefore discusses the emergent institutional collaboration between the Tłıchų Language Revitalisation Division and the Tłıchų Cultural Commons Digital Archives, providing an outline of the full process for identifying potentially relevant language materials in the field, obtaining (or producing) archival copies, making those materials accessible to the public, and incorporating their contents into the development of novel language resources such as dictionaries and corpora. In doing this, we also provide an overview of substantial overlap (and occasional contrast) in long-term goals between the language division and archive, and outline future steps to expand and formalise the collaborative process, which has, to this point, largely arisen naturally.

Transmedia Yokot'an (Chontal Mayan) project

Amanda Alejandra Delgado Galvan (Dirección de Lingüística-INAH), Paola Gutiérrez Aranda (Laboratorio de Lingüística - ENAH), Bernardino Montero (Community Researcher) and Miguel Dominguez (Laboratorio de Lingüística-DL-INAH)

This presentation examines the application of transmedia storytelling, a multi-platform and multi-format technique, to expand the reach of the Yokot'an language documentation project, initiated in 2016. Language documentation relies on sustained relationships and collaborative efforts, which continually reshape and redesign the project (Twigg (2024)). Yokot'an is a Mayan language spoken in the state of Tabasco, Mexico, by around 30,000 speakers. It has been described far less often than other related Mayan languages and is considered endangered; fluency is mainly among adults, and children do not use it. Consequently, traditional practices are at risk of not being transmitted.

Our transmedia project aimed to: (1) honor and preserve the memories of Yokot'an traditional ritual specialists within the practice of Utäsijob ni matän 'gift' traditional pilgrimage festivities; (2) create a virtual space for our community-driven, collected data material in digital archives that are more accessible and easier to navigate—also providing a solution for better access in low-internet environments; and (3) develop a system for documenting the genuineness of the documentary linguistics process, in which multiple methodologies and practices are used, such as training sessions in visual arts, annotation, language description, and methodologies for ethnographic data collectors. These serve as moments when community members step into the roles of co-creators and co-researchers (Phillips (2012); Thieberger and colleagues (2016); Delgado and Pico (2021)).

Thus, a transmedia approach was the most effective way to achieve these objectives, as these experiences and stories can be delivered across multiple platforms and formats using current digital technologies. We demonstrate that an Indigenous-led documentation project fosters interest in memory and language revitalization, where participants actively shape the documentation process by learning new skills, contributing their perspectives, and developing materials alongside the research team.

From undisplaced to diasporic to refugee communities. Adapting LD design principles to the Minawao UNHCR refugee camp (Cameroon)

Pierpaolo Di Carlo (University of Naples)

Recent discussions within Language Documentation—see, e.g., the 2025 LD&C conference—have foregrounded diaspora communities and urban multilingual settings. While this shift has productively expanded the field’s attention beyond its original focus on “ancestral homelands,” the diaspora contexts commonly discussed to date involve digitally connected, self-organizing urban communities embedded in relatively stable institutional environments (e.g. Kaufman & Perlin 2018). Refugee camps and IDP settings differ structurally. Camps, in particular, are spaces shaped by humanitarian administration, aid dependency, mobility restrictions, and overlapping formal and informal authority systems (Turner 2015). In such contexts, residents navigate chronic uncertainty not only due to scarce material resources, but also to “project fatigue” generated by short-lived external, ephemeral initiatives. Documentation cannot rely on assumptions of researcher presence, institutional continuity, and community stability. Is the field equipped to operate under such conditions? Drawing on an ongoing documentation project on two endangered languages spoken in the UNHCR refugee camp of Minawao (Far-North Region, Cameroon), the paper shows how project design features often stemming from ethical commitments—such as sustained community collaboration and direct access to project outputs—in fact emerged as methodological necessities in the camp environment.

First, reduced foreign presence and the camp’s tense social dynamics require redistribution of expertise across a locally based team. Documentation is organized as coordinated sub-teams focus on recording, metadata, annotation, and infrastructure maintenance. Transparent task allocation and continued training aim to reduce perceptions of unequal benefit and to mitigate tensions that may arise when externally funded projects intersect with fragile social equilibria. Community collaboration thus moves toward a team-science model (Bennett et al. 2010) where the production of the documentary record is structurally dependent on the coordinated contribution of distributed expertise—no single actor, local or external to the camp, can independently reproduce the workflow or its outcomes.

Second, these training demands and social conditions in an environment where internet connection is costly (when available) require rethinking reliance on internet-based services. The project integrates Internet-in-a-Box (IIAB), an offline, locally hosted Wi-Fi system that distributes fully customizable content at no cost for users. Operating without internet connectivity, this infrastructure supports an expanding repository of tutorials and project resources while enabling a living language archive updated in near-real time and accessible on smartphones and tablets at key locations within the camp—the latter being crucial also in mitigating the risk of community-internal tensions.

The project has also extended collaboration beyond academia, engaging NGOs and IIAB activists to stabilize infrastructure and sustain training amid international aid cuts radically affecting life in the camp. This suggests that recalibrating LD design principles for refugee communities entails not only methodological adjustments but also broader institutional alliances capable of supporting research in humanitarian contexts.

Far from being an isolated case, Minawao reflects broader global tendencies. About 120 million people worldwide are currently forcibly displaced, with Sub-Saharan Africa hosting a disproportionate share of both refugees and internally displaced persons (UNHCR 2024). Conflict, climate stress, and protracted instability increasingly produce long-term displacement rather than temporary humanitarian episodes (IDMC 2024). As displacement-affected settings become structurally common environments in many African regions, language documenters are likely to encounter such conditions with growing frequency. Experiences such as that of

Minawao may therefore offer practical and methodological preparation for a near future in which documentation routinely operates under humanitarian constraints.

Documenting endangered linguistic diversity in the Brazilian Amazon: The view from Ka(g)wahíva

Wesley dos Santos (University of Wisconsin-Madison)

This talk concerns the language endangerment and language documentation of Ka(g)wahíva, a critically endangered language spoken by several Tupí-Guaraní Indigenous communities from the Brazilian Amazon. Currently, the language is no longer being acquired by children, and if this trend continues, the present time represents the last opportunity to fully document the language.

In this talk, I report on a collaborative project with three Ka(g)wahíva communities aimed at achieving this goal by creating a multipurpose, long-lasting collection of audiovisual recordings for their language.

In particular, the presentation will provide a background on the Ka(g)wahíva people and language ecology, including its context of endangerment; present and discuss the effort and challenges to document several etnovarieties of this language through a collaborative project between the author and the three communities involved, namely Juma, Jupaú, and Parintintin; and present the outcomes.

As part of the latter, I present the tangible results of the project to both the native and scientific communities. This includes accessible language records for the communities involved, in the form of wordlists, field notes on several grammatical phenomena, and verbal arts; annotations; language learning materials; training for a community member to document their languages; and opportunities to connect with and practice the language. Additionally, I demonstrate how this work extends our existing knowledge of the available information about Amazonian linguistic diversity with a systematic documentation of several varieties of a single language.

From paper to digital: transition from dictionaries to a multilingual database for promoting Kanak languages

Anne-Laure Dotte, Jannai Tokotoko, Gaël Digard (Université de la Nouvelle-Calédonie)

Among the challenges facing the promotion of Kanak languages (New Caledonia), the accessibility of up-to-date, freely available, and expandable lexicographical resources is a recurring demand from both speaker communities and non-speakers (ALK, 2021).

In this context, our project aims to develop a multilingual database based on lexicographical data from some twenty existing dictionaries (bilingual Kanak-French, or even multilingual several Kanak languages-French), whose formats (typescript or compuscript), types (thematical or alphabetical), layouts (double-column or full-page), and the granularity of the information.

This paper aims to present a methodological and technological innovation that makes it possible to extract OCRised content from digitized pages of these dictionaries in order to feed the multilingual relational lexical database. The challenge is to assign a label to each identified lexicographical entry that characterizes the type of information (headword; translation; example; translation of the example; metalinguistic information, etc.).

In this purpose, we developed an application designed to train a model using a sample of labelled pages, in order to automate the extraction and identification of information types for the entire dictionary. More specifically, a generic JSON-driven pipeline transforms labelling work (manual annotation) into a complete and reproducible chain of hierarchical lexicographic data (datasets → training → inference → structured entries export). This JSON pipeline is then duplicated in the application's declarative parameters for other dictionaries, enabling automation and considerable time savings, while minimizing recognition errors and interference.

In addition, this paper also addresses the challenges of the protocol next steps, namely the cross-correlation of lexicographic data between different dictionaries/languages, in order to achieve the expected end result: a freely accessible online tool enabling multilingual lexical queries.

Finally, this tool allows to incorporate spelling reforms resulting from the gradual standardisation of Kanak languages. It also has the advantage of giving new life to lexicographic works that are often out of print, difficult to consult, or contain obsolete lexical information. Their digital transition allows to update them, including by making it possible for speakers themselves to participate online to lexical data updating. In the end, by converting printed-only dictionaries to digital database, this project participates to continuing the work of linguists who have been committed to promoting these languages for more than 50 years.

The Source and the Resource: Insights from the Livonian Speaker Engagement Programme

Valts Ernštreits (University of Latvia Livonian Institute)

Successful language documentation relies on two interdependent elements: the source of documentation and the resources required for its processing. For critically endangered languages, speakers are the most crucial component of both — yet in communities with an extremely limited number of proficient speakers, securing either element presents exceptional challenges. This is particularly acute in the case of Livonian, an Indigenous language of Latvia and one of the most endangered languages in Europe, with approximately twenty fluent speakers.

This presentation offers insights from the Livonian Speaker Engagement Programme, a virtual initiative launched by the University of Latvia Livonian Institute in 2022, designed precisely to address this dual challenge. The programme engages proficient Livonian speakers in a wide range of documentation-related tasks carried out within the institute's scientific projects, including transcription and translation of existing documentation, speech data collection, lexical development, digitization of archival sources, teaching, and the creation of educational and cultural content. A central motivation for the programme was to address the critical shortage of language proficiency required when working with pre-existing Livonian documentation — materials that frequently survive in handwritten form or poor-quality recordings and demand a level of linguistic competence that typically exceeds that available among researchers alone.

After more than three years of operation, the programme has demonstrated its effectiveness as an approach to documentation and processing of materials for a critically endangered language on multiple levels. Beyond providing the research institution with additional capacity for processing both pre-existing and newly collected materials, the programme has generated significant broader outcomes. Participating speakers have demonstrably improved and reactivated their language skills through sustained, goal-oriented engagement with Livonian texts and speech — an outcome that has in turn expanded the pool of individuals capable of contributing meaningfully to future documentation work.

A particularly significant documentation outcome of the programme has been the creation of a contemporary Livonian speech corpus, recorded by eight community members, which is currently being used to develop speech synthesis tools and is intended to support future speech recognition systems — technologies that will enable more efficient computational processing of existing historical spoken Livonian recordings.

This presentation argues that in the context of critically endangered languages with extremely small speaker communities, close collaboration with speakers is not merely beneficial but essential to a successful documentation process. Furthermore, such involvement can be mutually beneficial: providing research institutions with irreplaceable linguistic expertise, while simultaneously supporting speakers' language maintenance and contributing to broader revitalization efforts. The Livonian case demonstrates that when approached collaboratively, documentation can serve not only as a record of a language, but as an active instrument for its sustainability.

Navigating Legacy: Insights and Challenges of Mobilizing Language Documentation Archives in Siberia and Beyond

Andrey Filchenko (Nazarbayev University)

This paper discusses the multifaceted process of integrating legacy language materials into contemporary language documentation and description (LDD) projects. Drawing primarily from the Tomsk Field Archive of Siberian Indigenous Languages (TFASIL) - established in the 1950s to document endangered Siberian languages among others - Southern Selkup, Eastern Khanty, Chulym Turkic, and Enets - featured on ELDP projects - this talk explores the technical, methodological, and ethical complexities of “mobilizing” historical data for modern academic and community use.

The transition from analog to digital formats reveals significant technical hurdles: legacy audio recordings often suffer from physical degradation and inconsistent metadata, while handwritten manuscripts pose challenges for automated recognition (OCR) due to idiosyncratic abbreviations and varying orthographic conventions. Beyond these technical barriers, the discussion highlights a “recontextualization” process, wherein modern researchers ascribe new utility to data originally collected under different linguistic paradigms. A critical methodological concern in this process is the “clean-up”, where archival entries represent normalized rewrites that obscure primary features of spoken discourse, such as prosody, self-repairs, contact phenomena, and allomorphic variations. This normalization essentially constructs a “written norm” for languages that were originally documented in a purely spoken state.

Ethical and community dimensions further complicate the mobilization of these archives, where risks are attached to making sensitive content accessible, such as private community gossip or culturally taboo subjects, necessitating tiered access protocols. Furthermore, while “giving back to the community” through language data repatriation is a recognized best practice, community interests often diverge from academic expectations. In Siberia, grass-roots initiatives, such as student-led Selkup workshops and online learning platforms, demonstrate that communities may prioritize practical, contemporary utility over the mere possession of raw archival data.

The insights and methodological gains of TFASIL LDD projects are subsequently applied to a project on a seemingly less endangered language - the Multimedia Corpus of modern Spoken Kazakh Language (MCSKL). Although the MCSKL utilizes state-of-the-art digital tools, it remains a “constructed reality” shaped by specific theoretical commitments and stakeholder needs. By comparing a mid-20th-century legacy archive with a 21st-century multimedia corpus, the talk argues that no language resource is a neutral or transparent representation of some “pure” linguistic state.

Ultimately, the mobilization of language data should imply a reflexive approach, a critical examination of the evolution of decisions, participants, and historical contexts that shape a language archive. This critical awareness is essential for drawing authoritative insights and ensuring that language documentation remains ethical and relevant to both the scientific community and the source heritage speakers.

Indigenous access to archives: supporting Indigenous communities to reclaim their community's knowledge safely

Alice Gaby (Monash University), Ebony Joachim (Living Languages)

Traditionally, archives of Indigenous language and culture—such as the Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS)—were established to preserve (almost exclusively white) outsider's records of the exoticised Indigenous other. Increasingly, these archives are reorienting to serve the communities represented in their collections. Both archives and platforms for engaging with archival materials—such as Nyingarn—are developing mechanisms and programs to repatriate and rematriate Indigenous materials and knowledge, though they may face many hurdles and roadblocks along the way (Croft et al. 2020; Thieberger et al. 2023). These include programs such as Paper and Talk that train community members in linguistic analysis and other research skills necessary to identify, interpret and transform archival resources into materials that can be used for language learning and teaching in their communities (Thieberger 1995; Marmion et al. in press).

This paper considers what we can learn from two Australian models for returning Indigenous language materials to their communities, Paper and Talk and Nyingarn, as well as pitfalls that can be avoided in returning the language knowledge held in archival materials to communities. This paper draws on our combined experiences as participant, organiser, trainer and linguistic support in the Paper and Talk program, as well as engagement with other archives and collections (e.g. through Nyingarn). We discuss examples of the incredibly rich information community researchers have been able to reconnect with through archival work, and the meaning this holds for their communities. However, we also consider some negative experiences encountered in archival work, including racist and traumatic content uncovered, as well as inter- and intra-community conflict that can be stirred up by outsider researcher's records and interpretations. We argue that the solution is not to restrict access to these materials, but to provide culturally safe contexts in which people can engage with difficult and at times distressing materials. We report on a range of changes made to the Paper and Talk program in response to the problems above, and make recommendations that could be translated to a wide range of other archival contexts in order to provide Indigenous communities with safer and freer access to their heritage records.

Documenting the Northern Alta Language in Motion: From Walking Conversations to Mobile Documentation Corpora

Alexandro Garcia Laguia (University of Valencia & Lund University)

In the last decade, the incorporation of action cameras into linguistic data collection has resulted in the emergence of innovative on-the-move documentary methods (Larsson et al. 2021, Obert to appear) and collections of mobile, walking-based recordings. Such innovations expand the content of language documentation corpora, whose text types traditionally privilege stationary interactional settings. As potential expansions of documentary corpora, mobile collections of walking recordings align with the aim of language documentation to create a broad record of communicative practices (Himmelmann 2006), providing the chance to document landscape-embedded walking interaction, a real-life domain traditionally under-represented in documentary corpora.

While this represents a clear improvement, the growing use of mobile collections also raises new questions regarding their design, structure, and role within documentary corpora. What kind of communicative events or text varieties are captured in mobile interaction (Mosel 2014)? Can a collection of walking recordings form a language documentation corpus or, rather, a subcorpus complementary to a stationary one (Cox 2011, Mosel 2018)? Are such collections entirely opportunistic, or is there room for design, sampling, and structure? To what extent are such collections comparable cross-linguistically?

This talk reflects on these questions through a recently compiled mobile collection as a case study. The Northern Alta mobile collection includes about two dozen walking recordings collected since 2023. The presentation explores methodological variables in this collection and how these shape its content and structure, including speaker configuration, recording technology, prompting strategies, and researcher presence.

Handling 20th-century Indigenous language legacy materials in the 21st century

Inge Genée (University of Lethbridge)

In April 2022 I discovered that my colleague professor Donald G. Frantz had left his entire archive in his office on campus before he passed away. The materials cover his entire career from the late 1950s to the late 2010s. The most valuable are Blackfoot fieldwork recordings and transcriptions from the 1960s through 1980s, notes and recordings for Blackfoot language courses, and notes for a Blackfoot dictionary (Frantz & Russell 1989, 1995, 2017) and grammar (Frantz 1991, 2009, 2017). There is also some material for Cheyenne and Plains Cree. The Frantz family and the university left it to me to decide what to do. I have since been working with local Blackfoot research assistants, fluent speakers and Elders to inventory, catalogue, digitize and describe the materials in preparation for archiving and community distribution.

A colleague recently called it “a crisis in documentary linguistics,” referring to similar collections of legacy materials at risk of being lost if not handled appropriately (Johnson 2004). A central challenge in deciding what to do with such materials is that what was considered appropriate in the mid-20th century is not what is considered good practice in 2024 (Smythe Kung et al., n.d.; Brandão et al. 2023). In this talk I will discuss some of the practical and ethical dilemmas we encountered while working with the materials and how we addressed them.

Preservation. Physically, the audio materials consist of reel-to-reel tapes, cassette tapes, CD ROMs, and VHS tapes; the written materials are mainly handwritten and typed notes, transcriptions with morphological analysis, and prints of computer files of which the originals no longer exist. The student research assistants digitized all audio-recordings and scanned all paper records.

Metadata. The recordings have no metadata at all. If there is a matching transcription this sometimes contains an acronym which we could trace back via the acknowledgement sections in the published dictionary and grammar. The transcriptions do not include any ethnographic notes or other information other than sometimes a date and location.

Rights holders. The main rights holders are the speakers on the recordings or their families if they have since passed, and the Blackfoot language communities. (The Frantz family relinquished all rights.) The most pressing task is to identify these speakers, but given the lack of metadata it is unlikely we will ever be able to identify them all. No consent forms have been found – they probably never existed. Listening to all the recordings has only occasionally yielded an additional name or a recognized voice.

Archiving for preservation and access. The physical materials are now housed at the University of Lethbridge, where they can be accessed by appointment (<https://library.ulethbridge.ca/aschome/uofffonds>). We are investigating depositing digital copies in one of the DELAMAN archives that offer levelled access and are actively seeking organizations in the Indigenous communities to donate copies to. In consultation with Blackfoot community members and on the advice of Elders, it has been decided that the digitized materials cannot be made freely available on the internet, but can be returned to the community on memory sticks, even when individual speakers have not been securely identified. At an official launch in March 2026, we distributed 200 memory sticks to community attendees, and we freely give out additional copies to community members who ask.

Funding and time. Archiving is not sexy. I received a small internal grant in 2022 but have since then mostly funded the work by redirecting funds from my other grants and devoting significantly less time to my other research work.

AI-Assisted Grammar Documentation for Senhaja Berber: Conversational Questionnaires Meet Hybrid Grammar Inference

Evgeniya Gutova (University of Tromsø), Sebastien Christian (University of French Polynesia)

This paper presents an interdisciplinary methodology combining fieldwork-based corpus collection with computerized grammatical inference for the documentation of Senhaja Berber, a Berber variety spoken in the Rif region of northern Morocco. Our collaboration brings together expertise in Senhaja linguistics (Author 1) and the Digital Inferential Grammars for Endangered Languages (DIG4EL) software platform (Author 2). Together, we address one of the central challenges raised in current debates on AI in language documentation: how to harness the power of computational tools responsibly, without bypassing communities or replacing the irreplaceable role of linguists and speakers.

Our data collection follows the Conversational Questionnaire (CQ) method (François 2019), designed to elicit naturalistic, contextualized speech while targeting specific grammatical properties. In this project, CQs are conducted in both directions between Senhaja Berber and Moroccan Arabic, enabling systematic comparison of tense-aspect-mood (TAM) systems across these contact varieties. The resulting parallel corpus is glossed using FLEx (FieldWorks Language Explorer) and supplemented by historical data from Renisio (1932). This corpus then feeds directly into the DIG4EL pipeline.

DIG4EL (Christian 2025) is a hybrid grammar induction method — it combines rule-based and probabilistic approaches. It is a free, open-source software that assists linguists and language teachers in producing grammatical descriptions from minimal data. It integrates: (i) large typological databases (WALS, Grambank) as probabilistic priors; (ii) automatically augmented CQ corpora with graph-based grammatical representations; (iii) controlled Large Language Model (LLM) agents strictly constrained to reason on existing data; and (iv) Loopy Belief Propagation over a Bayesian network of grammatical parameters. The system has demonstrated a mean inference accuracy of 76% and a median of 85% across 116 typologically diverse languages in the example domain of canonical word order, consistently outperforming a WALS-baseline by a wide margin.

For Senhaja Berber, DIG4EL processes the CQ-based parallel corpus to infer grammatical parameters, generate structured grammar drafts, and produce teacher-ready lesson materials — outputs that can be directly reviewed and corrected by the linguist. Preliminary experiments confirm that DIG4EL can make meaningful grammatical inferences from small, controlled corpora of the kind produced by CQ fieldwork. This is particularly significant for endangered languages where large corpora are unavailable and traditional computational methods are inapplicable.

We situate our approach within the broader conference theme of AI challenges and opportunities in language documentation. We engage critically with concerns raised by researchers such as Bender and Hanna (2025) regarding AI hype, arguing that responsible AI use in documentation requires: (a) transparency and explainability of all inferential processes; (b) strict data sovereignty — DIG4EL data is not used for LLM training and is managed by trusted community caretakers; (c) compliance with FAIR and CARE principles and the EU AI Act; and (d) a human-in-the-loop design in which AI outputs are always provisional and require expert validation. Rather than replacing linguists, DIG4EL is designed to reduce the time burden on the limited number of researchers working on the world's endangered languages.

The Senhaja case is particularly instructive because of the contact dynamics between this Berber variety and Moroccan Arabic, and because existing documentation (Gutova 2021; Renisio 1932) provides a basis for evaluating DIG4EL's inferences against established descriptions. By combining rigorous fieldwork methodology with principled AI assistance, our project demonstrates a scalable, ethical model for grammar documentation during the Decade of Indigenous Languages.

Documentation and Revitalization of Languages of Northern Pakistan: Challenges and Approaches

Amir Haider (Nazarbayev University Kazakhstan)

Northern Pakistan is home to remarkable linguistic diversity, with more than 31 indigenous languages spoken across the mountainous regions of the Karakoram, Himalaya, and Hindu Kush ranges. These languages belong primarily to the Dardic and Tibeto-Burman families, along with at least one language isolate, and are spoken in geographically isolated pockets. The region also possesses a rich cultural heritage, including ancient rock carvings, oral traditions, and deeply rooted indigenous norms and values. However, over time—particularly following the spread of Islam—attitudes toward local languages and cultural practices have changed in some communities. In certain areas, religious orthodoxy has created challenges for language documentation and cultural preservation efforts.

As the ex-leadership team member of Forum for Language Initiatives (FLI), this paper elaborates our documentation efforts for documenting 22 oral languages into writing. As an ex-training coordinator with FLI (current PhD scholar in Nazarbayev University) the author has been directly involved in the documentation of 22 languages to date. The paper outlines the specific language documentation activities undertaken, including fieldwork procedures, community engagement strategies, and data collection methods. It also discusses major challenges encountered during documentation, particularly those related to community attitudes and religious sensitivities.

In addition, the paper highlights the ongoing language revitalization initiatives, including awareness-raising efforts and Mother Tongue–Based Multilingual Education (MTB-MLE) pilot projects. Selected case studies are presented to illustrate effective approaches to community participation and attitude change. Finally, the paper offers recommendations for sustaining future language documentation and revitalization efforts in Northern Pakistan.

Lessons from the Osage Language and Lifeways Project

Dylan Herrick (University of Oklahoma)

Osage belongs to the Dhegiha branch of Siouan languages once spoken widely in the greater Mississippi valley (La Flesche 1932; Quintero 2004, 2010). Although there are no longer any L1 speakers (Quintero 2010), Osage Nation has a language department that oversees a pre-K – 5th grade school ‘Táaposka ákóhtapi’ (Our School) and offers language classes in four local high schools. It also engages local communities through community classes, online classes, and other electronically delivered options for Osage citizens, regardless of their place of residence. This presentation discusses a collaboration between two members of Osage Nation and two linguists that resulted in *Osage Language and Lifeways*, a twenty-chapter descriptive and teaching grammar that follows language revitalization goals by seeking to meet the needs of both community members and linguists (see e.g., Hinton & Hale 2001).

Despite Osage Nation providing a variety of learning opportunities, the development of linguistically informed pedagogical materials is almost entirely absent. Our project’s primary goal was to provide support to Osage language teachers and students by directly addressing this lack of materials. We discuss our goals and some of the challenges we faced in terms of writing for both community and linguistic audiences as well as navigating differing expectations from Osage Nation leadership, the Osage Nation Language Department, and community members.

From the outset, we worked to balance several distinct goals. Our primary goal was to create an introductory grammar that would help teachers, learners, and linguists learn the basics of the Osage language. Given the breadth of our intended audiences, we strove to make our work as accessible and accurate as possible. We accomplished this by using a minimum of technical linguistic terms and favoring terminology used by community members and teachers (Yamada 2007) and by using plain-language descriptions and copious examples (Cox 2015; Herrick & Hirata-Edds 2015; Hirata-Edds & Herrick 2017). Our second goal was to ensure that the book focused on how to form phrases and productive, useful language from as early on as possible. A third goal was to teach the Osage alphabet and sound system in a way that allowed us to link to and build on existing work such as Quintero’s *Osage Dictionary* (2010) and her highly technical *Osage Grammar* (2004). As a final goal, we included short dialogues and personal stories from our Osage co-authors. This allowed us to provide something of value and interest to community members who are less likely to invest time in explanations of grammar.

In addressing these goals, we faced a number of challenges that we present here. In particular, we discuss our decision to adopt a phonemic orthographic system, as this decision proved to be the most contentious one we made. The Osage sound system contains unaspirated, glottalized, pre-aspirated, and post-aspirated stops, vowel length, nasal vowels, and phonemic accent placement (Quintero 2004, 2010). Early versions of the Osage alphabet cannot distinguish all these sounds, but a later version, adopted into a unicode font (Everson et. al. 2014), uses diacritics to allow all phonemes to be distinguished. The originator of the original Osage alphabet objected to the use of diacritics, and this led to a number of challenges, only some of which were surmountable.

In addition to issues of orthography, we touch on lessons we learned from other challenges. This includes how teaching the basics for a Siouan language is not basic, particularly if the starting point is English, how to best write for multiple audiences and purposes, how to best include cultural information, how to collaborate during a pandemic, and how to carry out our project given the absence of true L1 speakers.

Integrating Automatic Speech Recognition into Community-Based Documentation Workflows: Practical Gains and Responsible Constraints

Nathan Hill Rolando Coto Solano, Afsana Ferdous and Juan Pablo Mendoza (TCAS, Trinity College Dublin)

Language documentation projects are increasingly exploring the use of automatic speech recognition (ASR) to support transcription and annotation, including recent work on end-to-end ASR pipelines for endangered and extremely low-resourced languages (e.g., Meelen, O'Neill & Coto-Solano 2024; Coto-Solano et al. 2022). Building on this emerging line of work, this paper examines how locally deployable ASR systems can improve and restructure existing community-based documentation workflows. Specifically, we investigate how ASR integration affects transcription effort, annotation cycles, labor distribution, and consistency within ongoing documentation projects for Rengmitca (not currently assigned an ISO 639-3 code) and Yonghe Rma (ISO 639-3: ylo).

We describe the design and implementation of ASR-assisted workflows within these ongoing documentation projects, two low-resource languages with limited prior digital infrastructure. Rather than positioning ASR as a replacement for linguistic expertise, we frame it as a semi-automated support tool embedded within established documentation practices. We outline the workflow architecture developed for local deployment, including model training, human-in-the-loop correction cycles, and integration with orthography development and translation processes.

Our ongoing workflow indicates that ASR can reduce initial transcription effort under certain data and hardware conditions and may accelerate annotation cycles and allow more focused allocation of expert time. However, these gains are accompanied by new technical dependencies, correction demands, and infrastructural requirements that must be explicitly managed. We discuss constraints related to limited computational resources, data scarcity, orthographic instability, and community priorities, emphasizing that technological adoption does not automatically translate into sustainable practice.

By embedding ASR within established documentation workflows, this study examines both the practical benefits and the responsible limits of AI integration in low-resource language contexts. It contributes a practice-based account of workflow redesign that supports documentation improvement while maintaining community agency and long-term sustainability.

One Decade, Many Routes to Participation: A Global Engagement Framework for Indigenous Languages in International Contexts

Jafari Saeed (University of Kurdistan)

Urgent international attention to the loss, endangerment, and marginalization of Indigenous languages has increased, yet one major problem remains unresolved: visibility does not automatically lead to transmission. Too often, public-facing language initiatives remain short-term, symbolic, or institutionally isolated, generating awareness without creating sustained participation, repeated use, or intergenerational continuity. Documentation may remain in archives, outreach may not move beyond one-way dissemination, and educational efforts may fail to connect with the cultural, historical, artistic, and knowledge worlds in which languages actually live. This paper argues that the central challenge is not simply how to elevate Indigenous languages internationally, but how to design participation so that recognition becomes use, and use becomes resilience.

In response, the paper presents a global engagement framework, structured around a small set of interconnected flagship projects and expandable collaborative initiatives to be launched in 2027. The proposed projects: Open Indigenous Languages Schools, From Word to the World Exhibition, Discover Indigenous Worlds! and The Portal to the Linguverse are conceived not as standalone programmes, but as linked public pathways. Together, they connect community-based learning, travelling exhibition practice, interdisciplinary discovery across science, art, culture, and history, and a shared digital participation infrastructure. Their purpose is to create multiple routes into engagement for different publics while building stronger bridges among Indigenous communities, local knowledge holders, linguists, ethnographers, anthropologists, artists, technologists, decision-makers, science communicators, and international university-based partners working on linguistic diversity, documentation, and revitalization.

This presentation offers the design and three-year outcome of the Indigenous Language Transmission and Resilience Ad-hoc Group in the UNESCO's International Decade of Indigenous Languages at its the mid-term, and asks what an international engagement architecture for Indigenous languages should look like if its goal is not only awareness, but durable language use. It therefore focuses on the practical and conceptual obstacles such a model must address: fragmented publics, uneven institutional capacity, weak coordination across sectors, and the persistent gap between outreach and everyday language practice. By framing participation as a continuum - from encounter, to involvement, to repeated use, to transmission - the paper proposes a scalable model for coordinated global action that has benefited the UNESCO IDIL network, both by the increase in the number of countries involved and the increased levels of engagement. Its broader implication is that public engagement should be understood not as a peripheral supplement to revitalization, but as one of its most strategic instruments.

Legacy materials-based multipurpose corpus of South Ryukyuan languages

Aleksandra Jarosz (Adam Mickiewicz University)

In this paper, I will introduce the idea of a legacy materials-based multipurpose corpus of South Ryukyuan languages (Japonic family), highlighting its importance, content, and challenges predicted along the way to its realization.

The online corpus of the four South Ryukyuan languages – Miyako, East Yaeyama, West Yaeyama and Yonaguni (cf. classifications of the Japonic family in Jarosz 2024) – is devised to consist primarily of a range of legacy materials, and namely: collections of fieldwork texts, example utterances from research reports, descriptive or dialectological sources, and pieces of oral literature such as songs, stories, proverbs and riddles. It will be available in two metalanguage interfaces: Japanese, the targetship being the South Ryukyuan communities and second language learners, and English for academic and popularization/representation/worldwide visibility purposes.

The corpus responds to an array of needs procured by the fact that all South Ryukyuan languages are under-documented and endangered. The estimated absolute number of speakers ranges from 200 (West Yaeyama) to no more than 20,000 (Miyako; cf. Aso, Shimoji and Heinrich 2014, Shimoji 2017, Jarosz 2023, Jarosz 2024 *inter alia*), and the typical score on the the UNESCO language endangerment assessment scale pertaining to documentation and literacy (UNESCO Ad Hoc Expert Group on Endangered Languages 2003: 12, 16–17) does not exceed 2. The existing South Ryukyuan documents span over a roughly 130-year-long time range and include research reports with narrow circulation (e.g. Uchima 2002, Uchima 2004), unpublished fieldwork notes (such as Nikolay Nevskiy's notebooks stored in Tenri University Library archives), rare or otherwise unreachable publications (e.g. Hokama 1965), works whose titles or subject matter do not imply any content specifically related to South Ryukyuan (e.g. Nohara 1998), incidental South Ryukyuan texts collected amidst other languages or other types of content (e.g. Okinawaken Taibei Seikyūken Jigyō Kyōkai 2010, 2011), or self-published attempts at creating texts by the community members (e.g. Kawamitsu 2019–2023). The awareness of their existence among the potential target audience is low due to their limited distribution, obscurity and overall inaccessibility.

The corpus will therefore respond to a perceived need of providing a centralized directory of the existing South Ryukyuan language documentation. Two major challenges foreseen at the moment are: obtaining permissions to reproduce the existing records in the corpus so as to ensure its legal and ethical acceptability; accurate linguistic analysis and translations of some of the more archaic/obsolete structures, due to the language endangerment having progressed for many decades and there being fewer and fewer proficient speakers available. The latter point, however, is precisely why the task of producing the corpus is urgent.

Developing a framework for easy-to-build text display websites for language documentation

Rebecca Jarvis, Andrew Liu (University of California, Berkeley)

Developing and deploying online tools to foster community access to language-documentation materials is highly appealing in the digital age of language documentation. However, creating web-sites and other online resources from scratch requires fairly substantial technical knowledge, which most language documentation practitioners are not trained in. Developing tools from scratch with such constraints can require large group projects with substantial funding requirements that is not feasible in many language-documentation settings. In this project, we strike a middle ground by working towards a modern, semi-flexible website tool that is aimed to enable future “plug-and-play” adaptation by other language documentation practitioners. Specifically, our tool focuses on presenting glossed and translated texts in an interactive manner.

Our own project (involving a language documentation group and linguistics/computer science undergraduate students) began in collaboration with the Atchan language community in south-eastern Côte d'Ivoire, for which we developed Version 2 of the Atchan Song and Story Corpus (ASSC, tinyurl.com/atchan2; a link to the Github repo will be added to the site homepage) in 2025-2026. This site adopts many display conventions from our earlier Version 1 of the ASSC (tinyurl.com/atchan), as well as from the Moro Story Corpus ([link here](#)). In Version 2 of the ASSC, we rewrote the underlying codebase to use Zola, a static site generator (SSG), resulting in a static, easy-to-load site, and lowering the technical knowledge needed to author and publish content. We have been adding features of interest to the Tchaman community, including mobile optimization, embedded videos for each text, easy toggling between several text display methods, and a fully bilingual website interface. Importantly at a practical level, the website is easy and cheap to host; our site is currently hosted for free on GitHub Pages (with the Education Plan), with videos hosted on Youtube to minimize the size of the site.

From a design perspective, the template is designed to separate the content-writing experience as much as possible from the technical backend of the website. This is accomplished in three main ways: (1) automating all the technical tasks required to go from content to deployment with CI/CD workflows, (2) taking advantage of static site generators to convert user-friendly markdown files into HTML, and (3) automating conversion of data files directly exported from tools intended for linguists into webpages. Functionally, this means that the language documentation practitioner engages only minimally with the technical part of the site once the website shell is built. New texts can be added by directly uploading individual texts to the site's GitHub repository, and adding metadata about the uploaded text to a spreadsheet also inside the GitHub repository. Once these steps are complete and an accompanying pull request is submitted, the website automatically recompiles to include the new material. This all can be done either locally or via the GitHub web interface.

In addition to our Atchan-specific work on the ASSC website, we also aim to create documentation and resources to enable adaption of the ASSC website framework in other language-documentation settings. One major goal in this regard involves developing support for texts in different file formats. The site currently supports uploads in a human-readable txt format, as well as in Cross-Linguistic Data Format (CLDF)-compliant files; we have also developed basic support for FlexText format, though the robustness of this support is not clear at this point. More generally, we have developed extensive documentation of how the website works and its various features. We are currently in the process of developing a “quick-start guide” that consists of (1) a set of user-friendly instructions and (2) a content-free shell site that others can clone in GitHub to serve as the basis for future projects. We welcome discussion of ways to lower barriers to entry in creating digital resources like this one for language documentation.

Outreach and the Relevancy of the Archival Loop: PARADISEC's Approach to Collaboration and Engagement

Jodie Kell and Steven Gagau (University of Sydney, PARADISEC), Julia Colleen Miller (The Australian National University, PARADISEC)

PARADISEC (the Pacific and Regional Archive for Digital Sources in Endangered Cultures) is a digital archive preserving materials from a wide range of small cultures and languages worldwide. This presentation will highlight three approaches used by PARADISEC team members to enhance outreach, improving the discoverability, accessibility, and community engagement of the archive's digital collections.

The first speaker is an Indigenous community member who will demonstrate how his role within the Melanesian community has enabled the creation of two collections: the WM2 archive from the Wantok Musik Foundation and the SG2 cassette collection from ENB Province villages in Papua New Guinea. He will highlight how collaboration supports Indigenous-led archival initiatives through access to funding and resources.

The presenters will then outline PARADISEC's tertiary education programs, which teach practical skills in digitisation, metadata creation, and long-term preservation. By embedding archival expertise into higher education, these programs empower students to protect cultural and linguistic heritage while gaining real-world experience in digital preservation.

Finally, the presentation will discuss how PARADISEC outreach programs enable community engagement and participation by describing two case studies: the Toksave: Culture Talks podcast that enhances accessibility by prioritising Indigenous voices and community empowerment and the PARADISEC-developed Raspberry Pi Wi-Fi devices that overcome issues due to lack of regional internet infrastructure and provide access to the archive for communities.

Together, these case studies demonstrate how PARADISEC harnesses digital archives to educate, foster collaboration, and promote community engagement, ensuring that cultural and linguistic heritage is preserved, accessible, and actively shared.

Documenting language transmission in a critically endangered language: the case of Livonian

Gunta Kļava (University of Latvia Livonian Institute)

Language transmission in endangered language communities rarely follows the pathways described by classical models of intergenerational transfer. Where mainstream sociolinguistics has long centred the family as the natural and primary domain of language continuity, critically endangered languages frequently exist in conditions where this domain has been profoundly disrupted or has ceased to function altogether. In such contexts, transmission does not simply slow down – it transforms, finding new carriers, new spaces, and new social logics that conventional frameworks are ill-equipped to capture. As Hornsby (2022) observes, how transmission plays out in practice varies considerably across communities, and the attitudes and ideologies speakers associate with their language ultimately emerge as determining factors in its survival. Transmission is thus not merely a demographic or structural phenomenon, but a deeply social and identity-bound process, shaped by the specific historical and community circumstances of each language situation. Livonian offers a particularly illuminating case. One of the world's most critically endangered languages, it has survived not through the preservation of traditional transmission environments, but in spite of their loss. Over the past fifty years, transmission has shifted predominantly towards a horizontal pattern – from speaker to speaker, outside of family boundaries – shaped by historical disruption and community fragmentation. Nevertheless, certain families have maintained vertical transmission across generations, and the broader community has demonstrated considerable resilience in finding new pathways for the language. This dual pattern reflects both the fragility of Livonian and the adaptive capacity of its speakers, raising important questions about how identity and solidarity continue to sustain a language when its conventional social infrastructure has largely collapsed. To document and analyse these processes, a methodological framework – the Mapping of Language Transmission – were developed integrating archival sources, historical speaker records, biographical interviews, and community ethnography, with particular attention to the conditions under which the language has been maintained. The framework traces transmission trajectories across at least one hundred years, identifying recurring and emergent patterns through which Livonian continues to be used, learned, and passed on today. Preliminary findings identify several key transmission modes operating beyond traditional sociolinguistic domains: speaker-to-speaker acquisition outside family structures, participation in shared community events, and engagement through digital platforms – environments that have acquired a vital sociolinguistic function precisely because conventional domains such as family communication, public interaction, and formal instruction are no longer viable for such a small and geographically dispersed community. The presentation argues that documenting these practice-grounded processes is essential both for language documentation and for developing meaningful instruments of preservation and revitalisation responsive to the needs of contemporary speakers. Idealised transmission models, designed with larger and more stable communities in mind, are of limited applicability here. By centering the lived experience of a modern community navigating an extraordinary linguistic situation, this work demonstrates that critically endangered languages can sustain continuity through resilience, solidarity, and adaptation – contributing to broader efforts to maintain global linguistic and cultural diversity.

GlossKit: An Integrated Protocol for Documentation, Data Processing, Archival, and Publication

Shuan Osman Karim (University of Würzburg)

GlossKit is a web-based tool in Documentary Linguistics designed to streamline the transformation of linguistic data from recording to archival deposit and publication. Its design is guided by three constraints: (C1) A documentarian's first responsibility is to create a multi-purpose record that benefits both scholarly and speaker communities. (C2) The average documentarian does not have advanced training in morphology, syntax, phonetics, phonology, programming, or Machine learning. (C3) Documentary work is inherently collaborative, involving native-speaking consultants and multiple scholars.

Language documentation relies on tools for collecting, organizing, analyzing, and sharing linguistic data. Together these tools constitute a project's "workflow," "pipeline," or "data life cycle" (Mattern, 2022). A widely adopted processing pipeline is the ELAN-FLEX-ELAN protocol championed by ELDP. Users produce time-aligned transcriptions in ELAN (Wittenburg et al., 2006), import them into FieldWorks Language Explorer (FLEX) (SIL International, 2022) for morphological parsing and lexicon building, and export the results back to ELAN for archival in EAF format. Recordings and annotations are then enriched with persistent identifiers and metadata to satisfy the FAIR principles (Findable, Accessible, Interoperable, Reusable; Wilkinson et al. 2016), which are central to the creation of a multi-purpose record.

Despite its widespread use, the ELAN-FLEX-ELAN pipeline presents several structural limitations: (I1) Segmentation and transcription in ELAN require on average 40 man-hours per hour of recording (Foley et al., 2018, 205–206). (I2) FLEX is designed primarily for single-user workflows. (I3) FLEX prioritizes monolingual, monologic, concatenative data, and its parser is optimized for string-based agglutinative morphology (Skilton et al., 2025, e141). (I4) FLEX is resource-intensive, limited to Windows/Linux, and performs poorly on low-RAM systems, excluding many potential users (Skilton et al., 2025, e140). (I5) Export fragility: dictionaries export only to rigid XHTML/XML formats, and interlinear text export often disrupts spacing or metadata, making corpora difficult to repurpose (Skilton et al., 2025, e148–e149). (I6) Although archival is possible after re-export to ELAN, the resulting materials are not publication-ready.

Constraint (I1) has been partially addressed by advances in computational speech processing. Transformer-based ASR systems such as wav2vec2 Baevski et al. 2020, Whisper Radford et al. 2023, and ESPnet Watanabe et al. 2018 leverage self-supervised pretraining on large multilingual corpora and can be fine-tuned on small endangered-language datasets. However, their practical adoption is constrained by (C2): most documentarians are not programmers. GlossKit integrates ASR model training directly into the documentary workflow. For a previously undescribed language, the system automatically segments audio and generates a rough IPA transcription. Users refine segmentation, map IPA output to an orthography of their choice, edit text, assign consultants to utterances, and annotate multilingual or codeswitched segments. Confirmed mappings generalize to subsequent uploads, while a language-specific ASR model trains continuously in the background. Transcription accuracy improves over time without requiring knowledge of the computational backend. Resulting models are accessible within GlossKit and publicly available on the HuggingFace platform.

GlossKit is natively collaborative (addressing I2). It supports string-based and agglutinative morphology comparable to FLEX, with ML-supported non-concatenative models in preparation (I3). As a web application hosted on AWS, it removes device and geography constraints (I4).² Integration of segmentation, transcription, and morphological parsing modules eliminates export fragility and ensures consistent formatting and metadata across outputs (I5). Crucially, exports are generated in publication-ready formats, addressing (I6). GlossKit

² Note that Adams et al. (2021) suggested web services as a possible improvement for the ELPIS platform.

thus reconceptualizes the documentary workflow as a single integrated, web-based pipeline from recording to publication.

KaLO: A community-focused legacy archiving project

Yumi Standlee, Violet King, Armando Molina Gómez, Blaine Billings, Bradley McDonnell, Leah Pappas (University of Hawai'i at Mānoa)

The Kaipuleohone-LDTC Legacy Ongoing (KaLO) Archiving Project is a new archiving initiative at the University of Hawai'i at Mānoa (UHM), which brings together students and faculty to archive legacy materials in Kaipuleohone, the University of Hawai'i Digital Ethnographic Archive. Launched in Fall 2025, this student-faculty collaborative workshop series trains student linguists in archival processes with donated language materials. In this paper, we will discuss the motivation behind the project, current activities, future directions, and the funding and organizational structures that make it possible.

KaLO is a collaboration between Kaipuleohone and the Language Documentation Training Center (LDTC). Kaipuleohone is an open-access archive (Albarillo & Thieberger 2009, Berez 2013) housed within the larger UHM repository ScholarSpace. As a training archive, Kaipuleohone's goals are to digitize and describe documentation materials, as well as to make preparing and depositing audiovisual documentary materials "an integral part of the graduate curriculum" (Berez 2013:263), both of which are integrated in KaLO's archiving workshops. LDTC (Sou et al. 2023) is a student-run organization founded in 2004 at UHM. LDTC's goal is to organize training workshops for emergent scholars and language workers to produce documentation and revitalization materials for low-resource languages. LDTC's contribution to KaLO is threefold: a cultivated community-oriented method-ology, long-term relationships with community and student linguists, and funding sources for in-person events.

KaLO's current activities aim to fulfill various aspects of Kaipuleohone and LDTC's missions. In the fall of 2025, the inaugural KaLO archiving workshop introduced Rejang language legacy materials (1972-2007) collected by Dr. Richard McGinn (McDonnell 2025). This workshop trained UHM students in various stages of preparing physical and born-digital materials for archiving, including formatting file-type transfers, inputting metadata, scanning fieldnotes, and compiling PDF files of physical books. In the spring semester, we plan to continue training students with Rejang legacy materials through archiving workshops, regular small-group sessions to complete archive preparation stages, and instructional lectures on preparing new fieldwork documentation materials for archiving. Through these workshops, legacy materials are made accessible to academics and communities while providing students with hands-on experience with archiving to supplement courses in language documentation.

KaLO workshops will provide a space for cohorts of new linguists to practice properly depositing and navigating archival materials. By bringing the Archive into the regular practices of the Linguistics Department, students will engage more deeply in the methodologies and ethics of producing an accessible and multi-purpose language archive (Himmelman 2006, Nathan 2011). Additionally, we are initiating outreach at community events to local speakers of under-resourced languages as the first step of KaLO's future goal to build capacity for community members to archive their own legacy materials and navigate their language's archival record. We hope that this direction will, by bringing together students and community members, train linguists who understand how community members may use and interact with their archived materials and who organize and manage their data responsibly.

To apply a Creative Commons license or not: A decision tree for language documentation

Susan Smythe Kung (University of Texas at Austin)

Many widely used content-sharing platforms and digital repositories give content creators the option to apply a Creative Commons License (CCL) to digital media files when they upload them to the platform. When used, these built-in mechanisms facilitate open licensing and thus support the R (for Reusable) in the FAIR Guiding Principles for scientific data management and stewardship. However, the ease with which an open license can be applied to a media file can cause inadvertent and unintentional violations of any one of the CARE Principles for Indigenous data governance. Due to a general lack of available training opportunities (in almost all disciplines) on copyright, opening licensing, and traditional Indigenous protocols, open licenses, such as the Creative Commons suite of licenses, are often applied inappropriately to media content by well-intentioned individuals of all backgrounds.

In response to this generalized lack of training in the face of wide-spread easy options for applying CCLs, I have developed a decision tree designed to walk anyone involved in language work--including documentation, revitalization, reclamation, teaching, and advocacy--through various questions they should consider before applying a CCL to a digital work. The decision tree poses specific targeted questions that determine the path to one of two final outcomes: to apply a CCL or not. As teams or individuals work their way through the decision tree, they are pointed to relevant information they may consult to help them answer the question at hand and facilitate a deeper understanding of the relevance of both the question and their answer to it; see Figure 1 for an excerpt from the decision tree. Depending on how users answer the questions, the decision tree might or might not lead them to the Creative Commons License Chooser at the end. Thus, the decision tree itself does not provide a recommendation about which CCL to use, but instead, it points people to the easy to use, online license chooser created and maintained by Creative Commons.

During this presentation, I will explain why and how this decision tree is important for language documentation and archiving in the decade of Indigenous languages; and I will walk the audience through the different questions to be considered, the linked learning resources relevant to each question, and the possible outcomes of each decision.

AILLA: The first 25 years

Patience Epps, Susan Smythe Kung, Anthony C. Woodbury (University of Texas at Austin)

In March 2026, the Archive of the Indigenous Languages of Latin America (AILLA) celebrated the 25th anniversary of its initial launch. In this presentation, we reflect on AILLA's 25-year legacy of archiving linguistic and cultural resources of Indigenous languages of Latin America, the Caribbean, and parts of the United States. We provide a short history of AILLA's founding and a summary of its current holdings; and we show how key institutional collaborations have contributed to AILLA's longevity and success. We share specific examples of how AILLA's multilingual graphical user interfaces in English, Spanish, and Portuguese have facilitated community participation and engagement in the archiving process and how dedicated title and description fields reserved for Indigenous languages have enriched the metadata in many of the collections. We highlight several early AILLA collections that engendered subsequent collaborative projects and new AILLA collections, many created and conducted by Indigenous researchers about their own languages. Finally, we end with our vision for AILLA's next quarter century.

GeoELAN: A companion tool for geo-referencing action camera footage in ELAN

Jens Larsson (Lund University), Laurits Stapput Knudsen (Lund University), Alexandro Garcia Lagua (University of Valencia, Lund University), Karolin Obert (University of Texas at Austin)

Over the last two decades, the field of language documentation and description has seen a consolidation of methodological standards and theoretical expectations. Alongside this development, there has been a growing demand for documentary records that are longlasting, multipurpose, and adaptable to the diverse needs of communities, academic researchers, and policymakers (e.g., Himmelmann 2006; Austin 2007; Woodbury 2011). While this goal is widely recognized as essential, meeting such expectations in practice remains a significant challenge, particularly when it comes to integrating heterogeneous types of data into coherent, accessible, reusable, and meaningful collections (Author 2 forthcoming). While the advent of action and 360° cameras has facilitated the recording of multiple data types useful for documentation agendas (i.e., video, audio, geospatial), their seamless integration for linguistic analysis and communityrelevant purposes has so far been a laborintensive endeavor.

In this work, we introduce a new companion tool with a graphical interface that builds on GeoELAN (the authors' earlier commandline version; see Author 1 et al. 2020) with added support for 360 cameras. GeoELAN's main purpose is to geo-reference linguistic annotations made in ELAN. To do so, the tool is designed to unpack, align, and integrate geospatial data within ELAN timelines. To achieve this, the tool automatically identifies and consolidates all clips belonging to a recording session from any single reference clip, extracts all associated telemetry streams (including GPS and sensor data), and provides integrated previews that allow users to inspect spatial trajectories. GeoELAN further generates ELAN files with fully linked media sources and simultaneously exports GPS logs (GEOJSON, KML, GPX formats) to facilitate interoperability with mapping software. The shift from a commandline to a graphical interface constitutes a substantial advance: it removes the barrier of familiarity with command-line interfaces and makes GeoELAN readily accessible to a wide range of users.

Beyond enriching the documentary record, this functionality enables finegrained spatial analysis of any observable linguistic or nonlinguistic category, such as lexical variation, gesture patterns, environmental interactions, or communitydefined cultural domains. As such, GeoELAN not only sets up an efficient environment for integrated data analysis, but also enhances the adaptability to community priorities, particularly by facilitating the production of cartographic material as a direct output of documentation activities. We demonstrate the tool's core functionalities, discuss its relevance for multimodal and multiscalar documentation practices, and reflect on its potential to expand both analytical possibilities and communityoriented applications in contemporary language documentation.

Documenting temporal expressions on the move

Laurits Stapput Knudsen (Lund University)

Much language documentation and description have traditionally relied on data collected in stationary and controlled settings, leaving natural and dynamic contexts like during motion underexplored. This methodological gap not only limits the scope of linguistic documentation but generally constrain our understanding of how linguistic resources are used in everyday life (see Obert & Burenhult, 2024). Recent technological and methodological developments – such as lightweight action cameras, portable audio equipment, and geospatial methods for synchronizing linguistic data with movement trajectories – allow researchers to document language as it is used on the move offering new opportunities to enrich both language documentation and description (Cialone, 2020; Larsson et al., 2021; Possemato et al., 2021).

This paper demonstrates how these innovations can be applied to the description of Wik-Mungkan (Paman, Australia), focusing on the description of temporal reference in motion. By synchronizing linguistic data with movement trajectories, speakers' temporal expressions are mapped onto their physical paths, revealing potential correspondences between spatio-visual factors and tense marking. The results show a clear alignment between tense use and physical proximity to a motion goal, with future tense predominating at greater distances and present tense near the destination. This spatialized structuring of temporal meaning highlights the value of embodied, context-rich data for linguistic description. The relation between spatial and temporal language has been widely discussed, and temporal reference is widely understood as grounded in abstract spatial schemas and metaphors (Gaby, 2012; Majid et al., 2013; Matlock et al., 2005). In a moving context the mapping is less abstract: When moving toward a goal, for instance, the future is literally ahead, and the spatial experience provides an immediate, continuously updated analogue to temporal relations. Language use captured in motion thus offers a naturalistic testing ground for links between temporal semantics and spatial cognition.

Beyond the theoretical insights presented in this paper, these methods have practical applications for language documentation. They allow for the creation of dynamic maps, the documentation of ethnobiological and toponymic knowledge, and the preservation of linguistic practices tied to movement and landscape. This study thus exemplifies how using state-of-the-art geospatial methods both contribute to theoretical discussions on tense-semantics, embodiment, and the ecological context, while also providing direct value to the community by enriching the documentation of the language.

Centering Community Expertise in Language Resource Development: A Human-Centered Design Approach

Cedar Lay (University of Hawai'i at Mānoa)

In the recent decades since Himmelmann's 1998 definition of documentary and descriptive linguistics, language documentation and revitalization have increasingly emphasized the importance of addressing the needs of the communities whose languages are being studied. As linguists recognize an ethical responsibility to shape documentation projects around community priorities, community empowerment has become a central component of contemporary language work (Dobrin 2008; Henderson 2013). This shift reflects a broader understanding that linguistic data represents the people we work with and must therefore be managed in ways that respect their dignity, autonomy, and goals (Berez-Kroeker et al. 2022).

This project explores human-centered design (HCD) as a methodology for developing language resources that more effectively respond to community-identified needs. Widely used in fields such as healthcare, education, and public health, HCD provides a structured process for designing programs and resources through continuous collaboration with community members (Bartlett 2021; Hagan 2018; Hart et al. 2017). Because communities are geographically and culturally distinct, there is no universal solution for language resource development (MacDonald et al. 2020; Wasson et al. 2016). By centering community participation throughout the design process, HCD offers a framework for ensuring that language resources are grounded in the lived experiences, priorities, and cultural contexts of the communities they are intended to support.

This talk presents a case study applying HCD to the development of language resources for the Palauan community in Palau and abroad. Following the phases of the HCD process (Weeby 2018), the project begins with the empathize and define phases through qualitative interviews with Palauan community members to identify needs, challenges, and priorities related to language learning and use. These interviews are followed by focus group discussions in which participants collaboratively reflect on emerging themes and ideate potential solutions. Insights generated through these stages are then used to guide the prototyping and testing of new language resources.

As a concrete example of this approach, findings from interviews and focus groups are being applied to the redesign of an existing Palauan–English online dictionary and language learning portal. Community feedback has informed both the functional features and conceptual organization of the resource, allowing the redesign to better reflect how Palauan speakers and learners seek, interpret, and use language information in everyday contexts. Through iterative testing and refinement with community participants, the project aims to produce a resource that more closely aligns with community expectations and practices.

By positioning community members as collaborators and experts throughout the design process, HCD has the potential to function as a decolonizing methodology for language resource development, promoting Indigenous ways of thought to counteract colonial power structures often inherent to Western academia (Czaykowska-Higgins 2018; Leonard 2017; Wilson 2008). While this work focuses on the Palauan community as a case study, it ultimately demonstrates the broader potential of human-centered design as a flexible and community-responsive framework applicable across diverse languages, communities, and resource types.

Hubs-and-nodes: Decentralized fieldwork for language documentation

Jakob Lesage (Humboldt University, Berlin), Olga Olina (Humboldt University, Berlin)

This paper introduces the hubs-and-nodes model, a distributed approach to language documentation. While pandemic-era adaptations have expanded remote fieldwork (e.g. Williams et al. 2021; Grzech & Tisalema Shaca 2022; Rice 2021), documentation efforts still tend to rely on the presence, mobility, and funding of a single researcher. This model makes documentation vulnerable to disruptions in funding, mobility, or political stability, and limits the achievable workload.

Drawing on documentation projects in Taraba State, Nigeria, we show how documentation activities can be distributed through existing community networks and strengthened through strategically placed hubs. This approach decentralizes expertise, equipment, and responsibility. Community collaborators, like academic researchers, act as nodes in a distributed network, while hubs give all actors infrastructure, coordination, and networking support: equipment storage and repair, data transfer and backup, training, coordination, and networking across different language communities. Village-based hubs support documentation in rural settings, while urban hubs enable data processing, storage, and cross-community collaboration.

We report on projects that have produced two ELAR-archived collections and demonstrate how the model enabled a third, community-initiated project through inter-community collaboration. These cases show that the model works under intermittent access conditions, enables year-round documentation work, and shifts ownership of decisions toward local participants.

From Archives to Online Courses: A Case Study of Andwa Revitalization

Sydney Ludlow (University of Hawai'i at Mānoa)

This paper discusses the process of creating a revitalization language project from archived records for Andwa language. Andwa is an Amazonian language located near the Ecuadorian and Peruvian border. The language is categorized as dormant (Lewis, 2008); however, in Ecuador, there is growing interest to practice Andwa culture and learn the language. This presentation details the difficulties and successes of creating a program from archived records and navigating the various barriers linked with this process.

The Andwa community of Killu Allpa, located along the Bobonaza River in Ecuador, has begun revitalization efforts. The community, though ethnically Andwa, are speakers of Pastaza Kichwa. There is a deep sense of longing to speak the Andwa language. This desire began a community-linguist collaboration to find materials and documentation projects carried out in previous years. Using the Arazhuanza et al (2020) archival record in the CLA collection we have begun adapting the records into language courses. The process has come with many difficulties including language barriers, lack of context, and inaccessibility to technology. Solutions to these problems include translation efforts from English to Spanish as well as Spanish to Kichwa, finding free online platforms that allow for downloadable lessons, creating lessons from scratch in Kichwa language medium, and finding laptops for the community. These efforts have led to several functioning Andwa language lessons which can be downloaded by community members and implemented for usage.

This research is important as Killu Allpa, along with other Andwa communities, are forming a formal Andwa nation organization. As part of their development they are hoping to fund Andwa language teachers and work to collaborate with Andwa people on the Peruvian side of the border. This work is crucial in these early stage efforts to support the language and culture as it continues to grow and find new life.

Language Documentation Toolkit - A Command-line tool for language documentation and linguistic analysis

Justin Mackie (Arizona State University)

Currently, 42% of the world's total languages are endangered languages. Documentary linguists and language communities create language resources, such as annotated and glossed texts, dictionaries, grammars, glossaries, etc., to slow the process of language endangerment and support linguistic analysis. Language documentation projects often face annotation and data analysis bottlenecks that slow the development of language resources and linguistic analysis (Rangel, 2019). Natural Language Processing (NLP) researchers have developed applications to address bottlenecks that hinder language documentation projects (Wittenberg et al., 2006; Berez, 2007; Moe, 2008; Rogers, 2010; Lane et al., 2021; Gessler, 2022).

Despite the development of applications to aid the language documentation process, documentary linguists and language communities rarely use them. Gessler et al. (2025) found that reasons for the lack of use of NLP applications include difficulty setting up the system, difficulty integrating the system into workflows, and a lack of technical knowledge. In addition, internet access is a barrier to the adoption of NLP applications, as language communities in rural areas may lack reliable access (Rangel, 2019). As a result, the main issues with language documentation projects remain, which can slow the development of resources to slow language endangerment.

This presentation focuses on a system demonstration of the Language Documentation Toolkit (LDTK), a command-line tool designed to improve language documentation and linguistic analysis workflows and help reduce the annotation bottleneck. To set up LDTK, a user specifies the location of the directory containing a corpus of parallel texts in plaintext or Word document format. After the user provides the location of the corpus of parallel texts, the corpus is processed, and the user can then interact with it in various ways. Users can retrieve specific information, including all instances of specific word sequences, word concordances, closely related words, and subword sequences, via a simple search. LDTK automatically creates various linguistic resources for users, such as wordlists, glossaries, initial morpheme lists, and annotated corpora, including part-of-speech-tagged, glossed, and morpheme-segmented versions of their initial parallel text corpus, via unsupervised machine learning models. Users can customize which models they want to use for which tasks, and LDTK provides documentation on how the models transform the initial corpus. Users can export all resources that LDTK creates, including the answers to their search queries, as plaintext, JSON, or CSV files.

The main benefits of LDTK are its simple setup and interface, its ability to help start language documentation projects by providing initial versions of annotated corpora and requiring only a collection of parallel texts, and its ability to retrieve specific information users want for creating language resources or conducting linguistic analysis using a corpus. In addition, LDTK can be used in areas with unreliable internet access, as it does not require connecting to any external APIs or webpages. This presentation will showcase how LDTK assisted linguists in creating annotated corpora and in morphosyntactic analysis of Takelma, a low-resource indigenous language spoken by the Umpqua tribe in the Northwest United States.

Converting legacy media to text to increase findability of cultural records

Aso Mahmudi (University of Melbourne), Nick Thieberger (University of Melbourne), Raphaël Merx (University of Melbourne), John Ferlito (INODES)

Many recordings made in the past have little metadata and so are difficult to find by normal catalog searches. We have been exploring the use of language models for automated speech recognition (ASR) with very small training sets, in one case (Bislama) 29 hours of transcribed audio, and, in the other (Nafsan) 10 hours. We will show that the error rate is remarkably low (5% for Bislama, 10% for Nafsan) and that, having developed the model, the ASR can run on a laptop computer with no internet connection. The tool, written by Aso Mahmudi, called Easper, produces Elan files, segmented and transcribed. A second tool, written by Raphaël Merx, takes a Bislama transcription in Elan format and translates it from Bislama to English. In each case some correction is still needed, but even the raw transcript provides a point of entry to the recording. The plan we are working on now with PARADISEC is to test language models against items in other languages in the collection in order to create a first-pass transcription. We will then have an ingestion pipeline for untranscribed media that automatically creates a textual transcript which will provide a searchable index of the media. We are building an online manual transcription system that will allow users to correct automatically generated transcripts.

More Than Mere Documentation: Uncovering Indigenous Political Voice in Early Ethnolinguistic Text Collections

Eleanor Nevins, Ziqi Yu (Middlebury College), Helen Carle and Tom Woodward

Our central argument is that responses to colonial linguistics by speakers of indigenous languages are recoverable within the transcripts of ethnolinguistic text collections. A set of premises built into the practice of text collection make this possible. The first is that creating them requires contact between a language researcher and an indigenous speaker, with attention to empirical details of the latter's extended speech. Second, in contributing their oral "texts," indigenous speakers in fact address themselves to researchers and to anticipated audiences. Third, because language is intrinsically reflexive and metacommunicative, some metacommunicative cues in the speaker's address are recoverable in the transcript. As a result, alongside the ethnological frame established by the researcher, another frame is communicated, albeit underrecognized, by the speaker. For example, the temporal terms in which speakers represent themselves and their subject matter in transcripts are often at odds with the past-oriented temporal frame of the ethnolinguistic text collection as established by disciplinary standards of the time. Transcripts reveal that speakers often address themselves to possible futures, or voice political critique of some aspect of felt colonial history. Examples from transcripts of North American indigenous languages, such as the Maidu texts of Roland Dixon (1912) and William Shipley (1962), as well as the Chiricahua and Mescalero texts of Harry Hoijer (1938), readily illustrate this, as would any other colonial text collection given adequate attention. An implication is that the legacy of Americanist text collections extends well beyond the horizon of language documentation to include histories of indigenous political voice (Nevins 2017, 2013).

To bring our argument to a wider public, including indigenous language heritage communities, we are developing an open-source Word Press web application. Our focus for now is the Dixon (1912) and Shipley (1962) Maidu language text collections. We have included Shipley's audio-recordings from UC Berkeley's California Language Archive, which we reference with . Our website provides several points of entry to the texts, recordings speakers, places and speech situations represented in the collections. The open source user interface pairs line by line scrolling Maidu source transcript and English translation. Where audio is available, it is included in the paired scrolling format and its item identifier in the California Language Archive displayed. We are developing search functions, a companion map with places and landscape photos, as well as short historical and language content mini-essays designed to enrich reading and searching the texts. We are hoping to develop moderated community participation functions. Our intent is to share the specific project and the digital toolkit built around it, with the idea that others might find the whole or some of its parts applicable to other text collection projects. We present this application, designed to recognize political voice in early ethnolinguistic text collections, as one way to bring legacy documentation to bear upon contemporary indigenous empowerment initiatives, including language renewal and reclamation programs (Garrett 2023).

Archivist and corpus linguist perspectives on the archival loop: Matukar Panau as a case study

Julia Colleen Miller (PARADISEC at The Australian National University), Danielle Barth (Australian National University)

The archival loop, or dialogic archiving, is a cyclical, collaborative process that connects speech communities, original researchers, new collectors, archivists, and archives through ongoing relationships rather than static, “one and done” deposits. Within this loop, communities and researchers co-create documentation, build capacity, support revitalisation, and produce shared outputs such as dictionaries, literacy materials, films, and publications. Original collectors mentor new researchers, community members, and students. They provide access to legacy materials, which are revisited and enriched through new recordings, refined transcriptions, and additional perspectives, while new collectors continue documentation aligned with evolving community priorities. Materials therefore remain open to correction, reinterpretation, and reuse over time, increasing their longevity and impact. For example, songs may be re-recorded by younger generations or annotations added through continued fieldwork. The archive functions as an active partner by enabling engagement, supporting community safety by improving access conditions, facilitating metadata enrichment, incorporating new and related content, and expediting repatriation. While this dialogic approach allows collections to remain dynamic and responsive, it also presents challenges, including version control, DOI persistence, the absence of a single “final” analysis, and the complexity and data growth that accompany multiple perspectives and evolving best practices.

Matukar Panau is an Oceanic language spoken in Madang, Papua New Guinea, and has been documented for over 15 years as a born digital project. Data has been archived in multiple language repositories, allowing comparison of archival processes. Long-term collaboration between an outsider-linguist and members of the speech community has enabled deep engagement and sustained co-creation. As documentation and analysis have continued, people from Matukar and Surumaran have changed names and clan affiliations, transcription conventions have evolved, analytical interpretations have been refined, and multimodal annotation has been added to many items. Technological change has also made the use of multiple recording devices easier and increasingly recommended. These developments have strengthened the documentation rather than undermined it. A single deposit at either the beginning or the end of the project would have (had) limited scope and analytical depth. Instead, the Matukar Panau collection has grown into one of the largest Oceanic language collections in terms of speaker representation and levels of annotation.

Continual enrichment of the Matukar Panau collection has only been possible through long standing dialogue with archivists around file naming, file types, metadata, and versioning, requiring compromise from all parties. Our presentation demonstrates that the benefits of the archival loop outweigh these challenges and outlines practical strategies for mitigating them, drawing on our shared experience in language documentation, corpus building, and archiving.

Creatively combining equipment for language documentation while swimming and boating

Josiah Murphy (University of Hawai'i at Mānoa)

Methods of language documentation have shifted over the decades from pencil-and-paper notations to include digital audio-video recordings and internet-farmed surveys. A bigger toolbox has enabled linguists to record real-time data from a variety of locations; however, some locations still limit which tools can be employed. These limitations can include extreme heat and cold, high winds, heavy moisture, and salt or other mineral accretion. This is a problem for my research in particular, which documents canoeing cultures in the Pacific Ocean. Interviews on land were easy to conduct, but it took several attempts to find a combination of equipment that could adequately record audio and video language data of speakers who were paddling a canoe or swimming in the ocean. My presentation will explain the challenges of documentation in an ocean environment, review the limitations of various equipment, and demonstrate the equipment combination that eventually succeeded in producing live language documentation suitable for use in research. My solution for recording on the water involves utilizing two different types of digital recorders, marketed recorder mounts, and materials available at hardware stores. Samples will be provided of recordings made on canoes near a Polynesian island as proof-of-concept. My self-written user manual will be available for those looking to replicate my equipment methods, and is clear enough for community members to apply without professional oversight. I will also explain how to transfer the data files to a smartphone or to a computer for integration with software such as ELAN. This presentation will be of interest to those documentarians working in oceanic, atoll, rainforest, and other humid environments who may still be seeking workable solutions. It will also be helpful for equipment designers and technologists to better understand the types of products linguists and anthropologists are looking for.

Training Language Documentation using Smartphones: Pitfalls and Possibilities

Kelsey Neely, Mandana Seyfeddinipur (Endangered Languages Documentation Programme)

In the last decade, smartphones have become increasingly affordable and durable, contributing to their near-ubiquity around the world. Smartphones also have emerged as powerful tools for recording video and audio. For good reason, language documenters have taken a keen interest in leveraging these tools for documentation. Much of the work on the use of smartphones for documentation has focused on technical specifications (e.g., McDonnell et al. 2024, Sanker et al.) or presents work on how smartphone apps can be leveraged for documentation (e.g., Neely 2023 on WhatsApp). These works often point out that the familiarity of smartphones is a potential advantage of these tools for engaging new documenters, particularly language community members who may not have access to other types of recording devices. However, there is limited guidance on the technical limitations and pitfalls in using smartphones and how to navigate them. This paper addresses the potential challenges and how to manage them when utilizing smartphones for documentation.

Two trainings—a one week in-person training with a diverse group of Ni-Vanuatu and a six week online training with Maya digital activists—allowed us to identify specific pitfalls and challenges and develop strategies addressing them.

Three areas require special attention: making the most of the available tech; video framing and capturing language in use; and sharing and preserving materials. First, given the wide variety of devices participants used, we identified two areas of concern: storage space on the device and the availability of external microphones. Some trainees had newer phones, access to MicroSD memory expansions, and/or external microphones (often Bluetooth lapel mics), but other trainees had phones with damaged internal microphones and/or very limited storage for recordings. It was important to anticipate these challenges and address them directly in training with concrete, practical advice on making the most of limitations and by providing information on affordable upgrade options.

Secondly, when using smartphones, we observed a tendency to film social media style, i.e. with a vertical framing, as this is the habitual viewing and recording practice for platforms like TikTok or Instagram, with a single subject's face framed quite tightly. This preconception can lead to a monologic, auditory-centered bias that neglects to record multi-participant interactions including the full gestural space, so we found it necessary to address social media recording habits and their potential drawbacks.

Thirdly, sharing and preserving the recordings is another area that needs to be addressed. Participants were familiar with YouTube and easily uploaded recordings with their phones for sharing. Here it is imperative to discuss the ethical issues and intellectual property rights considerations inherent in using a commercial platform like YouTube and in making recordings available to the public in general. At the Vanuatu training, ELAR downloaded and archived the recordings, allowing the participants to consider the issues around long-term preservation and sharing on social media.

Drawing on what we learned from these two very different training contexts, this paper offers practical guidance for language documenters working with smartphones, focusing on topics like assessing device capabilities, adapting recording practices and navigating the complex decisions around preservation and access. Ultimately, we argue that with appropriate training and awareness, smartphones can serve not just as convenient recording tools but also as catalysts for more inclusive, community-driven documentation practices.

A Light at the End of the Tunnel: Indigenous-Led Documentation and Intergenerational Synergy in the Mijikenda Community

Nancy Ngowa (Pwani University)

The rapid erosion of linguistic and cultural heritage has sparked an alarming realization among indigenous populations, shifting the paradigm of language documentation toward community-led initiatives. In the coastal region of Kenya, the Mijikenda community has pioneered proactive strategies to safeguard their intangible heritage, moving beyond traditional academic observation to grassroots activism.

The primary goal of this study is to analyze the shift toward indigenous-led documentation within the Mijikenda community. Specifically, the paper seeks to:

1. Examine the institutional and social frameworks used by the Mijikenda community, such as university MoUs and cultural festivals, to systematize language and heritage preservation.
2. Evaluate the role of intergenerational dialogue and capacity-building workshops in transferring documentation skills from elders to the youth.
3. Assess the impact of modern technology and social media in reviving cultural lifestyles and creating awareness of linguistic heritage among the younger generation.
4. Determine the sustainability of community-driven models in ensuring that cultural materials are preserved and passed down through successive generations.

Indeed, the light at the end of the tunnel is the deliberate integration of the youth in the documentation process. Through intergenerational dialogues, capacity-building workshops, and active participation in the recording process, the Mijikenda are bridging the generational gap. By empowering the youth to lead documentation efforts, the community ensures that these preservation practices are not merely temporary interventions but sustainable traditions passed down through lineage. This paper will show that when documentation is indigenous-led, it fosters a sense of ownership that is vital for the long-term survival of endangered languages and cultures.

Linguistic Documentation and Archiving of Tjwao: A Community based Approach

Shumirai Nyota, Believe Mubonderi (Zimbabwe Open University)

This study analyses the documentation and archiving of Tjwao (also known as Tshwao), a critically endangered Khoisan language spoken in southwestern Zimbabwe, mainly within the Tsholotsho District. Current sociolinguistic data indicates that Tjwao is on the verge of extinction, with less than 20 active speakers remaining, most of whom are elderly individuals between the ages of 60 and 90. Despite its recognition in the 2013 Constitution of Zimbabwe (often erroneously listed as “Khoisan”), the language lacks substantial institutional support, and intergenerational transmission has effectively ceased. This decline of Tjwao is rooted in a long history of systemic marginalization and forced displacement. The establishment of the Wankie Game Reserve (now Hwange National Park) in 1928 forcibly removed the Tjwao from their ancestral hunting and gathering grounds, pushing them into contact with the agrarian Bantu, specifically the Ndebele and Kalanga. The Tjwao were not agrarian and in addition they had no farming implements. This forced them to do menial jobs for the Ndebele and Kalanga such as herding cattle and as farm laborers. This situation forced the Tjwao to shift to their masters’ dominant languages to increase chances of getting access to the menial jobs available. To survive economically and escape social stigma, many Tjwao shifted to these dominant languages, which were perceived as necessary for employment and social advancement. Early colonial language policies, such as the Doke Recommendations (1931), did not help the Tjwao language as they prioritized English, Shona, and Ndebele, relegating former minority languages and Tjwao, excluding it from education and administration. This Dokean recommendation entrenched a sense of “linguistic inferiority” that persists in the post-independence era to date. To counter this trajectory, the documentation project employs a community-driven approach that leverages digital tools to facilitate intergenerational skills transfer. These include Annotation Platforms that enable Tjwao youth to be trained in using browser-based annotation tools such as Label Studio for natural language processing (NLP) tasks. These platforms allow for tagging and transcribing audio-visual recordings of elders, turning raw data into structured linguistic resources. Mobile and Social Media tools, like Actionbound are also used. This initiative uses mobile-based tools to gamify language learning and community outreach. Collaborative Whiteboards are also used where Platforms like Miro are used for virtual brainstorming and community workshops, allowing dispersed community members to participate in deciding which cultural materials should be archived or kept private. By centering Indigenous leadership, the study ensures that the resulting archive is not just a “museum” of a dying language, but a living, community-owned resource. This approach prioritizes trust-building and culturally grounded dialogue, balancing professional archival standards with Tjwao knowledge systems to safeguard the Tjwao unique linguistic heritage for future generations.

Preserving Rohingya Language and Oral Traditions in Displacement: The Role of the Rohingya Cultural Memory Centre

Olufunke Ogunderu (Project Officer (MHPSS), International Organization for Migration),
Muhammad Kamruzzamann (Copywriter (Content Development), Rohingya Cultural Memory Centre)

The intangible cultural heritage (ICH) of the 1.1 million forcibly displaced Rohingya currently residing in 33 camps in Cox's Bazar and on the island of Bhasan Char, Bangladesh, faces multiple UNESCO-identified threats, including weakened transmission, decontextualization, and economic pressures. Within the world's largest refugee settlement, the Rohingya language and its oral traditions remain particularly vulnerable, especially as approximately 52% of the displaced population are children, among whom literacy in the mother tongue remains limited due to restricted educational opportunities. This paper examines how the International Organization for Migration-supported Rohingya Cultural Memory Centre (RCMC), starting in 2018 as a community-led initiative, integrates language documentation and digital archiving within a protracted displacement context. Using a mixed-methods approach combining qualitative interviews with participants, volunteers, and facilitators across multiple camps, alongside quantitative analysis of participation records and programme monitoring data, the study explores how documentation practices are embedded within community cultural programming. The RCMC develops and curates digital archives of oral histories, traditional songs, storytelling sessions, awareness dramas, and language materials, while linking archival preservation to non-formal learning platforms, publications, and mobile outreach initiatives. These efforts demonstrate how documentation and archiving can move beyond static record-keeping to become participatory, locally accessible, and intergenerational practices. The paper argues that community-driven archival models in low-resource, high-displacement settings expand conventional approaches to language documentation by foregrounding collaboration, accessibility, and cultural sustainability, offering insights for scholars and practitioners working in similarly constrained environments.

Dual Archives as Moral Infrastructures: Access, Ownership, and Circulation in the Digital Documentation of Nsiteu (Cameroon)

Rachel Ayuk Ojong Diba (University of Buea)

Language documentation is often presented as an ethical response to endangerment: record, archive, preserve. Yet preservation infrastructures frequently assume users with stable electricity, broadband internet, and high digital literacy. This paper examines what happens when those assumptions fail. Drawing on fieldwork conducted in 2022–2024 in rural western Cameroon (Kouba, Tongo, and Bazou), I describe the moment of attempted return after archiving Nsiteu materials in an international digital repository. Community members had anticipated access to the recordings they helped produce. However, the chief of one village responded: “Je ne parviens pas cliquer sur le lien.” Another participant could not identify her interview among numerous files; another exhausted her prepaid data before navigating the site. The materials were technically open access, yet practically inaccessible. I argue that this is not a minor technological inconvenience but a structural mismatch between preservation infrastructures and the lived digital realities of rural communities. Digital repositories are designed for global researchers, not for low-literacy, low-connectivity users. In this context, preservation without circulation risks reproducing extractive knowledge flows: linguistic materials travel outward but do not return in usable form. To address this asymmetry, I propose a dual-archive model: (1) a scholarly repository ensuring long-term preservation and global research circulation, and (2) an offline, community-based archive using Raspberry Pi kiosks and mobile applications designed for audio- first, low-literacy access. Rather than presenting this as a technical fix, I frame dual archiving as moral infrastructure, a deliberate attempt to balance digital inclusion, data sovereignty, and ethical circulation in a context of profound infrastructural inequality. The paper concludes by arguing for a shift in language documentation paradigms: from preservation to circulation. If infrastructure determines whose knowledge travels, then documentation must be designed not only to store language, but to return it.

From Language Identifiers to Language Differentiation in Machine Supported Contexts: The Romanch Languages of Switzerland

Ignacio Pérez Prat (Lia Rumantscha), Hugh Paterson III (Institute for Advanced Study, Princeton)

Language development (Ferguson 1968, Fishman 1968, Eberhard 2017) is the intentional activities taken to make a specific speech variety function for the needs of its user community. Such activities induce the language, in situ, to become more instrumental rather than merely an object of science, barter, or identity (cf. Bourdieu 1977). We present the struggle for ISO 639-3 recognition of the Romanch languages (speech varieties) of Switzerland. Our presentation relates this narrative to the efforts of language documentation, the development of technological resources using statistical inferencing, and mutual intelligibility within the Romanch languages.

The Romanch languages (speech varieties) of Switzerland, since the days of the ISO 639-1 have been represented by a single ISO code, even though they have had—for over 200 years—5 different orthographies (Grünert 2024, 158, 161), 4 different Bible translations, and unique spheres of social identity. Additionally, there is a sixth constructed language used in official publication contexts. This has created a problem for the cataloging of language resources in libraries, and the development of digital tools (spell checkers, speech-to-text, automated translation, etc.).

Factors guiding community-external support for language documentation and conservation (revitalization, development, etc.) largely depended on an assessment of vitality (including across domains of language usage, cf. Grenoble & Whaley 2005), and/or a determination of uniqueness of “language” (e.g., Ethnologue/ISO 639-3) or “speech variety” (e.g., Glottolog). These efforts sought to provide a foundational set of documentation for endangered languages. Popular models of language vitality (e.g., EGIDS; Lewis & Simons 2010) completely ignore the societal impacts and interactive nature of language in digitally mediated communication—or as Grenoble and Whaley (2005) does: treat digital communication as a single domain. However, others (e.g., Kornai 2013) postulated that the lack of digital support is a significant burden and community limiting factor in language vitality more broadly. This has caused a reassessment of the notion of relevance of Digital Vitality (e.g., Gibson 2015, Ceberio, et al 2018, Simons, et al, 2022) with a growing consensus that the lack of language development in digital contexts is an impediment to language vitality and the lack of digital autonomy is an endangerment to a community’s ability to determine the future of its own language (cf. Varoufakis 2023).

The determining factor for language “uniqueness” has been impacted by ISO 639-3, a foundational component in the IETF’s BCP-47(RFC5646). ISO 639-3 is used in a variety of contexts from language resource cataloging (and libraries more broadly) to text rendering contexts in web-browsers and text editors. The foundational ISO 639-3 code set was donated by SIL to the ISO TC37 working group. These codes assisted in the identification of unique language development contexts (usually Bible Translation projects) established on the basis of mutual intelligibility via lexical wordlist comparison (a lower than 80% similarity threshold needed to be established to indicate different languages). This deciding factor has often been referenced under the ambiguous term of “Mutual Intelligibility”. Unacknowledged in “Mutual Intelligibility” discussions (for the purpose of language identification) has been the role of syntax, morphology, and information structure (within dialogic discourse) in creating mutual understanding between interlocutors. Now in the era of generative text from statistical models (often branded Generative AI) these issues in naturalness or accuracy become more salient as a discriminating factor in “intelligibility” with digitally mediated communication interactions. Therefore, with digital language resource development—an essential part of language development—a new, and yet to be acknowledged, emphasis on intelligibility exists.

Clear cataloging of language documentation materials also depends on ISO639-3 codes. For example, how should archivists differentiate training data sets? (MARC and Dublin Core both use BCP-47.) How should

translated texts be indicated (specifically using the ' -t' flag via IETF RFC6497)? Following Vamvas, et al. (2025) we provide a framework for investigating mutual intelligibility in digital mediated contexts when there are 6 different independent orthographical representations. We further lean into Gillis-Webber's (2023) framework of viewpoints to suggest that other language codes from the time of the ISO 639-1 code set actually represent ISO 639-3's macro-language concept. The Romanch experience serves as a model for other language community organizations seeking to facilitate language development in digital contexts so that their communities can flourish within their own communicative norms.

Citing Oral History: A Framework for Low-Resourced Language Audio-visual Documentation

Subhashish Panigrahi (OpenSpeaks)

The majority of the world's language communities use oral histories as the primary mode of knowledge transmission (Mahuika, 2021). Yet they remain systematically excluded from being cited in Wikipedia (Reagle & Koerner, 2020a) and conventional publications as sources. Citation policies widely require sources to be written, formally published, and institutionally authorised, reinforcing epistemic hierarchies and colonial knowledge systems (Reagle & Koerner, 2020b). Thus, carefully documented, community-reviewed oral histories in low-resourced languages are discarded as unverifiable or self-published material.

This paper presents the OpenSpeaks Oral History Framework (Panigrahi, 2026) to operationalise the citation of oral history in low-resourced languages. The framework has evolved from learning from two phases of the OpenSpeaks Archives initiative (2024–2026), a Wikimedia-supported project documenting oral histories in 20 low-resourced languages and dialects from India, Nepal and Sri Lanka.

The framework recommends three operationalisable parameters:

Parameter 1 — FAIR–CARE alignment: All recordings must adhere to FAIR (Findable, Accessible, Interoperable, Reusable) and CARE (Collective Benefit, Authority to Control, Responsibility, Ethics) principles.

Parameter 2 — Community assessment as peer review: Depending on available resources, documenters can use a structured review process to document reviewers' concerns, disagreements, and consensus. This assessment complements provenance metadata and enables external audits. This is analogous to editorial peer review, in line with Wikipedia's reliability criteria. This process draws on decolonial and feminist data practices.

Parameter 3 — Multilingual transcription with time-coded references: Documenters can aim to create transcripts, or, depending on resources, subtitles in at least one neighbouring official language and English. Subtitles have two key additional benefits over transcripts. They make multimedia accessible for people with disabilities. Secondly, time codes in subtitles can function like page numbers in textual citations, helping cite multimedia to precise locations.

The framework is an evolution of OpenSpeaks (Panigrahi et al., 2019), a 2017 toolkit for language documentation, built on lessons from the documentation of over 20 languages from India, Nepal, and Sri Lanka, most of which are Indigenous. Some examples of practices include language documentation by native speakers, interview questions and premises co-created with community members, participation of multiple speakers before and during the interview, interviews of community-nominated elders, and sharing recordings and remuneration with the community. While grounded in decolonial and feminist data practices of multiple independent community language archivists and in epistemic justice, this framework also needs to be tested across more diverse communities and field documentation scenarios.

Kwapa Community Teaching Project: Documentation and Linguistic Description Methodology

Etna T. Pascacio Montijo (Universidad Autónoma de Baja California)

The main goal of this presentation is to share a linguistic documentation and descriptive methodology developed for a community teaching project of the *Kwapa* language in Pozas de Arvizu, Sonora, Mexico.

The *Kwapa* people use this name in their own language, but the language is commonly known as Cucapa or Cocopah. It is part of the Yuman branch of the Cochimí-Yuman linguistic family. *Kwapa* people are located in the Mexican states of Baja California and Sonora and in Arizona in the United States. *Kwapa* is an endangered language due to its small number of fluent speakers and the fact that younger generations do not speak it. According to some *Kwapa* people, there are around 20 fluent speakers in Mexico, and fewer than 50 in Arizona.

In March 2024, a group of five *Kwapa* speakers from Pozas de Arvizu, Sonora, had the opportunity to lead a community workshop to teach their language to *Kwapa* children and young adults. Two external researchers are supporting the project with a collaborative and co-creative approach: Jorge Urbano specializes in language teaching, and I, a linguist with experience in descriptive linguistics with the Matlatzinca language, as well as some previous projects with the *Kwapa*. The speakers had no interest in beginning a documentation project because of previous experiences with government institutions and other researchers who did not return the products of those projects to them. Given this background, we prioritized the workshop.

First, we identified their concerns and needs. Then, we established some theoretical bases for language acquisition and teaching. After that, they provided cultural subjects to discuss, and we defined linguistic topics to teach based on those cultural topics. For each session, we all together registered words related to the cultural subject, as well as some phrases, sentences and dialogues. Sometimes, we included a story for children to provide continuous oral and written input. After that, we designed a lesson plan and didactic materials for the activities.

The workshop sessions were on Saturdays, with a short break, and we shared a meal at the end to strengthen the social bonds. Since March 2024, there have been 54 sessions, plus some additional sessions for planning and data collection. Two women of the speakers group are the teachers and the others support the activities.

The speakers are very interested in the learning process of the students, who are their own relatives, so even when the speakers don't know about linguistics or grammar, they always refer to the word order, the meanings, the pronunciation, etc., and they review very carefully all the data. We believe that our methodology allows us to create progressive documentation and linguistic descriptions according to the cultural and linguistic needs of the teaching and learning process.

This is a collective project. As part of the team, I would like to share the methodology and the linguistic strategies to register and verify the data with the speakers, to provide guidance on the morphological and syntactic characteristics required for student's knowledge level, and to analyze the language grammar.

Advantages of Community-Oriented Documentation for academic outcomes

Polina Pleshak (Hungarian Research Centre for Linguistics), Irina Burukina (University of Florida, Gainesville)

In recent years, the “decolonization” of linguistic fieldwork has attracted increasing attention (Hudley et al. 2024), highlighting the tension between prioritizing community interests and upholding scholarly rigor and productivity. This paper presents a documentation project on Patzún Kaqchikel (Chimaltenango, Guatemala) and shows that a community-oriented approach can yield a sociolinguistically more diverse and linguistically richer corpus than a traditional documentation model.

Kaqchikel is a Mayan language classified as threatened (Eberhard et al. 2022) or vulnerable (Moseley 2010), spoken by approximately 410,000 bilingual (Kaqchikel-Spanish) speakers. Our project, launched in 2019, began as a collaboration with the Patzún Women’s Cooperative Aj Su’m, with the initial goal of producing a trilingual (Kaqchikel–Spanish–English) book of recipes and oral histories. Topic selection and recruitment of speakers were made jointly with members of the Cooperative to ensure local relevance. We recorded approximately 10 hours of spoken Kaqchikel from 17 speakers representing diverse occupational, educational, and age backgrounds. Unlike many documentation projects, we did not aim at exhaustive coverage and deliberately avoided genres that are used less frequently or that are less representative of contemporary speech (e.g., traditional songs or folktales). Instead, we prioritized four key genres suggested by community members: everyday meal recipes, narratives about modern life, local customs that are still practiced, and important local institutions. During the fieldwork, we lived with local families, participating in daily activities and community events. This immersion fostered trust and enabled interactive recordings. Because the intended audience was community members themselves, the speakers produced vivid, detailed narratives that differed markedly from elicitation-driven data. The COVID-19 pandemic halted further in-person fieldwork and made the preparation of a book infeasible. We shifted our primary goal toward creating an annotated, open-access online corpus. The result collection comprises 7.25 hours of transcribed audio, translated into Spanish and time-aligned in ELAN. It adds to the body of existing Kaqchikel corpora (Bennett & Henderson 2022), at the same time significantly broadening the range of topics and the coverage of speakers by age, gender, and local variety of the language, allowing for a broader comparative approach to spoken Kaqchikel. A subcollection of 2,707 tokens has been glossed in FieldWorks.

The transcriptions and Spanish translations were prepared by a native speaker from Patzún. We further verified the data and added English translations.

Although the shift in goals may appear to mark a return from a community-oriented to research-oriented project, the corpus remains fundamentally shaped by its community-driven origins. The texts are also being made available online through a dedicated community-oriented webpage. This project demonstrates that aligning documentation practices with community priorities does not undermine academic quality but enhances sociolinguistic diversity, discourse richness, and long-term accessibility. By transforming a community-targeted cultural project with local relevance into an open-access corpus, we created a resource that simultaneously serves Patzún Kaqchikel speakers and advances linguistic research.

Bangladesh's Kui in transition: Language ecology and ritual practice

Syed Shahrier Rahman (Professor, Department of Linguistics, University of Dhaka)

Kui is a severely endangered Dravidian language spoken by the Kondo community in northeastern Bangladesh. At present, only two elderly speakers remain fully fluent, corresponding to the most critically endangered levels on language vitality scales (UNESCO, 2003). Unlike the large Kondo population in Odisha, India (Friend-Pereira, 1909), this group in Bangladesh descends from migrant labourers and has lived for generations in close contact with Bangla and its dialects, as well as other local languages, producing a distinct contact setting (Thurston, 1909). Over time, this contact has shaped a distinct Bangladesh variety of Kui in contrast to varieties described for India. To understand its present form, we must look at how and where it is used, how people move between regions, and how ritual life still holds certain speech forms together, in line with an ecological view of language in its social environment (Haugen, 1972; Fill and Mühlhäusler, 2001). Kui emerges as a social resource tied to memory, ancestry in Odisha, tea-garden labour history, and changing relations with the wider Bangladeshi society, echoing broader arguments about endangered languages and cultural heritage.

This study adopts a language-ecology approach, treating languages as part of a living environment rather than as isolated systems (Haugen, 1972; Fill and Mühlhäusler, 2001). The idea is inspired by Einar Haugen's concept of the "ecology of language," where languages shift and mix across social spaces under changing sociopolitical conditions (Haugen 1972). In daily life at home, Kui is rarely passed down from older to younger generations, a pattern characteristic of late-stage shift (UNESCO, 2003). Younger Kondo members mostly speak Bangla or local mixed varieties, while Kui appears mainly in short phrases, kinship terms, and respectful expressions addressed to elders. It no longer serves as the main tool for everyday talk, and in schools, offices, markets, and health centres, Bangla dominates these institutional domains (UNESCO, 2003).

Ritual life forms one of the last strong spaces for Kui, mirroring cross-cultural observations that the ceremonial and religious spheres often preserve endangered languages the longest. Here, the language still has weight and visibility. The research draws on more than five hours of recorded stories, life histories, and ritual speech with the last fluent speakers. Therefore, the chief research query is: How does the Bangladesh variety of Kui function within its changing language ecology, especially in ritual life, during a late stage of language shift? The research design follows a collaborative and participatory approach. Community members help plan recording sessions and also assist in interpreting the meaning of ritual texts. Two interested members of the Kondo community are involved and receive training in digital recording methods and basic metadata management. The data collection procedure includes multi-angle video recording, high-resolution audio capture, and digital annotation platforms to record speech together with gestures, spatial context, and the mixing of languages that often occur during rituals. This study serves as a pilot initiative to document a vanishing language and to encourage local involvement in preserving cultural heritage.

Change and Persistence: The Game of Being a Language Archive

Felix Rau (University of Cologne), Anke Debbeler (University of Cologne), Francisco Mondaca (University of Cologne), Sara Saleh (University of Cologne)

With the upcoming launch of the third and newest version of the Language Archive Cologne (LAC), we have the chance to reflect on past and future strategies for the sustainable operation of the language archive.

All language archives are confronted with the question of how to ensure the long-term preservation of their data. This involves not only technical challenges but also organizational, financial, and social factors. While all language archives face the same overarching problems, the specific problems vary greatly depending on the context of the individual archive. Starting from our experience with LAC, we will discuss these challenges in a general way and suggest approaches for sustainable management.

After completely rebuilding LAC, we will report on our experiences with past technological decisions and reevaluate our strategy for a maintainable and sustainable setup. However, we want to put the technical aspect into a broader context and emphasize that there is no technological fix for sustainability.

In the first part of the talk, we will present the history of LAC, its organisational context and the technological decisions that were made in the past. We will review the key design principle of LAC that the archive can be reconstructed entirely from its digital objects, decoupling the application layer from long-term storage. We will discuss the challenges we faced and how we addressed them. In the second part, we will reflect on our current strategy for sustainability and discuss how we can ensure that LAC continues to be a valuable resource for researchers and communities in the future. In this context, we will touch upon the strategic role of technical topics such as metadata, digital object packaging standards – such as OCFL and Bagit – and the purpose and the maintainability of domain-specific value-added services – such as ELAN file viewers, audio visualization.

Mapping and Reassembling Iskonawa Cultural Heritage for Language Revitalization

Carolina Rodríguez Alzza (University of Texas at Austin)

The Iskonawa are an Indigenous people of the Peruvian Amazon whose language is critically endangered yet currently undergoing revitalization. In this context, archival materials have gained increasing significance as potential resources for community-led initiatives. However, Iskonawa cultural heritage is geographically dispersed across institutions in South America, North America, and Europe, creating substantial barriers to community access.

Mapping these collections reveals how Indigenous knowledge has frequently been divided into “material” and “immaterial” domains (Lewy 2020). Language recordings, for example, have often been archived separately from material artifacts, reflecting the distinct trajectories of collectors, scholars, and institutions. This fragmentation has produced what I call an institutional geography of heritage, a configuration in which Iskonawa cultural knowledge is distributed across distant repositories and organized according to external classificatory logics. As a result, access is constrained not only by physical distance from the Amazon to these institutions, but also by linguistic barriers and by archival systems that produce materials difficult for community members to navigate.

Drawing on my experience facilitating Iskonawa access to parts of these collections, including visits to artifact holdings in a governmental institution in Peru and the circulation of photographs from museums in the United States and Poland through community-based exhibitions, this presentation centers on how Iskonawa people envision engaging with their dispersed and fragmented heritage. I examine both the possibilities these encounters open and the challenges they entail.

Engaging with scholarship on the return of heritage knowledge to Indigenous communities, particularly work focused on sound materials and digital repatriation (Kummels & Cánepa 2023; Lewy 2020; Lewy & Brabec 2023; Seeger 2018; Vapnarsky 2020), I argue that addressing the Iskonawa institutional geography of heritage requires a global reassemblage of Iskonawa collections. Such reassemblage must include not only dispersed artifacts and recordings, but also the relations that have reshaped and reconfigured these collections within their current institutional locations (Vapnarsky and Noûs 2021). In this way, these collections can be more effectively integrated into language revitalization efforts, not as static materials of the past, but as resources reactivated in relation to present-day needs and aspirations.

StivorLing: Digital Documentation and Social Media Revitalization of a Critically Endangered Romance Variety

Marija Runic (University of Banjaluka), Rossella Montibeler (University of Banjaluka)

This paper presents **StivorLing** (<https://stivorling.com/en/home-english/>), the first comprehensive, community-driven documentation and revitalization project dedicated to the heritage Italo-Romance variety of Štivor, a small village in Bosnia and Herzegovina. Currently, the variety is **critically endangered, with only approximately 10 fluent speakers remaining**, all belonging to the oldest generation. Given this extreme urgency, our project moves beyond traditional academic models to center on a deep, active collaboration between a community member and an academic researcher.

We explore three key dimensions of the project's digital and social ecosystem:

1. Collaborative Documentation and the Archive: We detail a process where community members are the primary agents of recording. By prioritizing authentic cultural narratives, we have archived oral histories that reflect a disappearing reality, such as memories of the village school in the 1950s, the history of local wells, and personal stories of selling wine and grapes in Derventa. These recordings provide a vital linguistic record of specialized agricultural and domestic vocabulary, including the production of *rakija* and traditional tools.

2. Social Media Mobilization and Diaspora Outreach: We analyze the strategic evolution of the StivorLing Instagram page as a tool for revitalization. By repurposing archival materials—such as traditional songs like “*La bella bionda*” and “*Marcjar*”—into simple, “catchy” videos and audio stories, we make the variety visible to the younger generation and the global diaspora. This approach moves the language from the private domestic sphere of a few elderly speakers into a curated, engaging digital space that gives newfound value and prestige to the variety.

3. Tailored Educational Outcomes: The StivorLing archive serves as a living foundation for language strengthening, moving beyond static preservation to active revitalization. We showcase a range of educational initiatives, including the creation of tailored pedagogical materials—such as illustrated agricultural vocabularies and village maps—designed for both community descendants and non-speakers. Crucially, the project extends into the public sphere through targeted workshops for high school students. These sessions are designed to raise awareness of the Štivor variety among the dominant population, challenging its invisibility and fostering linguistic prestige. By bridging archival records with digital engagement and classroom outreach, we illustrate how historical documentation is transformed into a dynamic tool for language learning, cultural pride, and broader social awareness.

As the first project of its kind in the region to fully involve community members in every stage, StivorLing demonstrates how technology and grassroots collaboration can effectively document and provide a future for a variety on the brink of extinction.

Mobilizing the ‘beautiful mess’: four decades of documenting Tłıchǫ dè eyıts’ǫ nàowò

Renee Saucier (Tłıchǫ Government)

The Tłıchǫ are a self-governing Dene First Nation in Denendeh (Northwest Territories, Canada).

External researchers began documenting Tłıchǫ Yatıì (language) and nàowò (knowledge) in the mid-twentieth century, depositing their recordings and notes at colonial repositories such as the Canadian Museum of History and the Northwest Territories Archives. The Tłıchǫ began auto-documenting traditional knowledge in the 1980s to pursue land claim and self-government negotiations, and to respond to mining company activities. The local school board (the first Indigenous-run school board in Canada) began creating resources to transmit Tłıchǫ Yatıì and nàowò through schools and ‘Tłıchǫnize’ the education system. Through partnership with academic linguists and missionary organizations, the Tłıchǫ developed standard orthography, fonts and written materials in Tłıchǫ, such as the Bible and prayer books.

These self-documentation activities produced extensive collections of recordings of Tłıchǫ Yatıì in audio, video and cartographic formats. A substantial portion of the audiovisual materials have been digitized and are stored on local servers. Thousands of recordings have been transcribed (into English), and Tłıchǫ place names data has been integrated into a Geographic Information System that supports land management and regulation. While many community members have contributed to the creation and digitization of the materials across several decades, the collection has been moved multiple times (scattering materials) and has only recently gained a central steward (archivist).

The Tłıchǫ Government Department of Culture and Lands Protection now stewards this archival collection, known as Tłıchǫ Nàowò Deghats’etsaa: Tłıchǫ Cultural Commons Digital Archives.

Our challenge is to mobilize this collection —affectionately called “the beautiful mess” — to enable access and re-use of existing documentation by community members, the Tłıchǫ Online Dictionary, and the Language Revitalization Division of the Tłıchǫ Government. While there are thousands of hours of audio in Tłıchǫ Yatıì, these recordings were generally not made specifically for the exclusive purpose of documenting the language itself. While audio recordings have been transcribed, locally there are diverging views about the appropriate use of English transcripts to access knowledge transmitted in Tłıchǫ Yatıì.

Our paper addresses current efforts to activate Tłıchǫ Nàowò Deghats’etsaa: Tłıchǫ Cultural Commons Digital Archives in service of language learning and revitalization, with strategies including processing archival backlog and surfacing materials for use by Language staff; digital repatriation of records held by external institutions; establishing data management planning procedures with researchers; and working towards implementing repository software to enable online access. We consider some of the complexities of facilitating the re-use of a large, rich collection of cultural materials for linguistic purposes.

Storytelling as a living tradition: collecting Wapichan and Taruma traditional narratives in Marurana

Anna Serke (Radboud University), Adrian Gomes (University of Guyana)

Oral narratives are highly culturally significant in many Indigenous languages, but when a language without a long written tradition is endangered, so are the stories of its speakers; for this reason, traditional stories have increasingly been recorded and written down during language documentation and maintenance projects in recent decades (Chelliah & de Reuse 2010; Fitzgerald & Hinson 2016). At the same time, the act of storytelling itself also forms part of the community's linguistic and cultural repertoire, and is therefore explicitly used as a (re-)vitalization strategy in some projects (see, e.g. Tossa 2008, Fitzgerald & Hinson 2016). In this context, we present two story collection projects which involve Wapichan (Arawakan) and Taruma (unclassified), and take place in collaboration with the WWA (Wapichan Wadauniinao Ati'o, a local organization promoting literacy in Wapichan) in Marurana, Guyana. Wapichan is spoken by 6000 speakers in Guyana and is used as the primary language at home in Marurana, but it is transmitted less and less to the younger generations, who grow up as passive bilinguals or as monolingual speakers of Guyanese Creole English (Eberhard et al. 2026, Anonymous 2022). Taruma is an endangered and previously undescribed language of which the only known family of speakers lives in Marurana and is bilingual in Wapichan (Eberhard et al. 2026). There are only a few locally published story booklets in Wapichana, while a handful of Taruma stories summarized in English can be found in Farabee (1917, 1918); in both languages, oral storytelling is on the decline.

The first project we present concerns a book containing a collection of stories in Taruma with corresponding Wapichana translations, requested by the family of speakers. The family is also involved as transcribers and proofreaders in the production of this book. The second project is a pilot community event combining both storytelling and story collection: elders being live-recorded while sharing traditional stories with other community members. These narratives will also be collected in a book, and both publications will feature illustrative artwork by community members. These books, both in printed and digital format, will be resources for the community, to be kept in the library and used as educational materials in the local schools. To sum up, the pilot event has the following two aims. Firstly, to produce some tangible output preserving the stories for the future and increasing the prestige of Wapichan by using it as a written language, and secondly, to create opportunities for community members to come together and experience storytelling as a living tradition. By shifting this practice from the family to the community level, narratives are passed on to speakers who never heard them at home.

We propose a report on and evaluation of these story collection projects, especially regarding the aims of the pilot event and its viability to be expanded into a larger initiative: its potential to become a recurring event in Marurana and to take place in other Wapichan communities as well as in neighboring speech communities of other Indigenous languages who are interested in comparing strategies for maintaining and promoting their language and culture.

From data to dialogue: Indigenous community-collaborative documentation for Murut Bookan language and culture in North Borneo

Siak Bie Soh (University of Malaya)

Language documentation is often framed as an act of preserving endangered linguistic data. However, for many Indigenous communities, documentation is also about regeneration, circulation, and the future transmission of cultural knowledge. Drawing on collaborative documentation with the Murut Bookan community in Sabah, Malaysia (Author, 2023), this talk discusses how digital archives can function not merely as technical repositories but as relational infrastructures through which knowledge, authority, and cultural continuity are shaped through ongoing negotiations between researchers, communities, and knowledge systems.

Based on over 75 hours of audio-visual recordings across diverse genres, including oral histories, ecological narratives, ritual practices, and everyday interactions, the project was developed through collaborative fieldwork and capacity-building initiatives involving local youth and community members. Rather than treating documentation as extractive data collection or privileging externally imposed archival structures and research timelines, the project shifted toward Indigenous epistemologies and community priorities through sustained consultation, youth training, and shared decision-making on recording practices, annotation depth, and access protocols, particularly in relation to documenting traditional ecological knowledge embedded in oral traditions.

I discuss three interrelated dimensions: (1) negotiating differing expectations between academic archiving standards and community priorities; (2) capacity-building and skills transfer to emerging local documenters to ensure long-term sustainability; and (3) the transformation of archived materials into pedagogical and cultural resources that support intergenerational knowledge transmission. Through this case study, I also discuss digital archives function as sites where authority over representation, circulation, and interpretation is actively negotiated, opening possibilities for more community-centred approaches to language documentation. By situating Murut Bookan documentation within broader discussions of Indigenous-led methodologies and contemporary archiving practices, the talk highlights how documentation projects can move beyond preservation toward regenerative, community-anchored futures.

The Eyak (dAXunhyuuga') Dictionary Project: A case study in revitalization lexicography

Jake Swenson (Eyak Cultural Foundation), Guillaume Leduey (Eyak Cultural Foundation), Gary Holton (University of Hawai'i at Mānoa)

Eyak (dAXunhyuuga') is the traditional language of the Gulf Coast of Alaska, situated at the boundary between two much larger Indigenous languages, Tlingit (to which it is distantly related) and Sugpiaq (a member of the unrelated Yupik family). The arrival of the colonial language English more than a century ago further exacerbated the pressures of language shift, with the result that by the middle of the 20th century, only a handful of people identified as Eyak speakers. However, over the past decade or so this trend has reversed, and the Eyak language community is growing. A dedicated team of language champions have led language classes, developed curriculum materials, created online language courses, and organized immersive language camps. As there are no longer any birth speakers of Eyak, these efforts rely crucially on documentary materials collected half a century ago. While these materials have been archived and in most cases digitized, practical access remains challenging, particularly for lexical documentation. In this presentation we describe these challenges and report on our efforts to overcome them in order to produce an accessible dictionary of Eyak which furthers the goals of language learners. Moreover, we emphasize that the dictionary making process itself is an act of language revitalization, serving to unite the language community and facilitate engagement with language—a process that Anderson (2020) has called revitalization lexicography.

Modern documentation of Eyak began in 1961, and a major dictionary was compiled in 1966, distributed as gray literature comprising some 3000 typescript manuscript pages (Krauss 1970). The dictionary adopts an organizational structure common to that found in most technical dictionaries of the related Dene languages (e.g., Young et al. 1992), organizing entries by abstract root. Thus, the entry for *díšičeʔ* 'I am hungry' is found under the headword *če* (Fig. 1). The dictionary employs a technical Americanist orthography with an idiosyncratic ordering based on the syllable structure of the root. Copious examples are given with references to narrative texts, but these are largely opportunistic and do not always serve to illustrate paradigmatic alterations. These factors together make it challenging to use the dictionary in language learning contexts. Additional grammatical information can be found in a recently published reference grammar (Krauss 2024), but this publication is also not aimed at learners.

Beginning in 2012 one of us began working to enter the existing Krauss (1970) dictionary into a Tshwanelex database, updating the orthography and adding new material which had been collected since the original compilation, nearly doubling the number of headwords. While in theory the process of generating a dictionary from this database should be straightforward, in practice it has not been so easy. We discuss some of these challenges and describe the tools and processes we have been using to reformat the database. More recently, a larger group began working to create an English to Eyak index, a process which has necessitated deep engagement with the dictionary definitions, written more than half a century ago. Organization of the index has become a deliberate act of language and culture reclamation, emphasizing themes such as personal names and incorporating new forms. As Hermes et al. (2012) have argued, language is more than mere content. Though originally conceived as a legacy project to digitally transform archival materials, the Eyak Dictionary Project has become a vehicle for engaging with the language and reimagining a new future for Eyak.

Language documentation, ethnography and social networks: a critical review and some recommendations

Adam James Ross Tallman (Structure et Dynamique des Langues)

Some recent discussions of sociolinguistic documentation have emphasized the importance of adopting a variationist perspective in language documentation (Stanford & Preston 2009; Meyerhoff 2019; Stanford & Mansfield 2017). In such a perspective, variation is contextualized and explained with reference to social categories (age, gender, race) and social structural relations, particularly social class (Labov 2001, *inter alia*) and social network position (Milroy 1980, *inter alia*). This talk will discuss the issue of eliciting, coding and archiving dimensions of socially structured relations (social class, social networks, etc.) in the context of sociolinguistic language documentation. Indeed, such a practice seems to be entailed by sociolinguistic documentation. However, I argue that language documentation should not adopt the methods of social networks (e.g. Milroy 1980) from variationist sociolinguistics uncritically. Rather, I suggest that language documentation should integrate network ethnography (White & Johansen 2005) into language documentation. I discuss the application of network ethnography in the context of language documentation projects with Chácobo-Pacahuara (Pano, Bolivia) (Tallman 2022), and Idi-Dhmiri-Agob (Pahoturi, PNG) (Schokkin et al. 2021) based on original fieldwork experience from myself and other colleagues (Diego Villar and Lorena Córdoba), emphasizing network ethnography's power in acquiring more systematic metadata and in developing quantitative metrics for analyzing social structure. Social networks are generally described as social formations, which are complementary to social classes in explaining sociolinguistic variation, based on ties (friendship, brother, father, etc.) between speech community members. It has also been suggested that they do not suffer from the ethnocentrism of (Labovian) social classes or castes (Milroy 1980: 178). In principle, sociolinguistic documentation should attempt to integrate social network position into language documentation and archiving, but developing and importing the concept from variationist sociolinguistics is not straightforward. The reason for this is that so-called social network metrics from variationist sociolinguistics (Milroy 1980 and Meyerhoff 2019 for relevant sources) are based on ad hoc questions and hard simplifying assumptions: they depend on presupposing an *a priori* network centre in the speech community and then select a set of questions, the results of which can be aggregated to position each speaker on a scale in relation to how close they are to this hypothetical centre.

I suggest that language documentation should instead adopt the methods of network ethnography developed by White and Johansen (2005), which involves coding person-to-person relations in a graph database structure, from which social structure can be discovered from the bottom-up in a less *a priori* fashion. White and Johansen's (2005) network data analysis and ethnography develop and complement each other in tandem in a way reminiscent of autotypology. Network coding forces more systematic coding of person metadata, while the resulting in graph databases that can direct researchers to emergent social organization that they may not have considered otherwise – an important issue especially given the outsider status of many language documenters (see Stanford & Preston 2009). I discuss the application of network ethnography with a focus on kin relations (alliance and descent), to language documentation in the context of community driven language documentation with Chácobo-Pacahuara (Pano, Bolivia) and Idi-Agob (Pahoturi, PNG). The results provide rich person-level metadata from which multiple sociolinguistically relevant metrics can be derived. I also suggest that the resulting database structures are better equipped to engage with issues related to social class because clans, lineages, house societies, endogamous blocks, alliance cycles, status groups etc. can be derived and tracked from them when they are coupled with information about village and property (see Mewett 1982: 125; Brudner & White 1997 for relevant arguments). Overall person-to-person-level graph coding is a more useful tool for language documentation, because it is more ethnographically sensitive, cross-culturally portable, and less ideologically saturated than Labovian social class or Milroyian social network metrics.

A Digitized Corpus of Veps Folktales: Thematic Classification and Applications for an Endangered Finno-Ugric Language (vepsian language)

Ruslan Terentev

Veps, an endangered Finno-Ugric language with approximately 4,000 speakers, lacks comprehensive digital resources, limiting its revitalization and computational study. This paper presents a novel corpus of 55 Veps folktales (sourced from the Veps Corpus, ID 767–821, http://vepsian.krc.karelia.ru/text/#subcorpus_Tales), categorized into six thematic clusters to preserve cultural heritage and enable linguistic and NLP applications. As a native Veps speaker and language educator, the author ensured community-aligned development. We propose a baseline algorithm for automated thematic classification, achieving 82% alignment with manual annotations. The corpus, annotated with motifs and translations, supports language revitalization, educational tools, and NLP tasks such as narrative parsing and cross-lingual transfer. It will be shared via LRE Map under a Creative Commons license, contributing to the preservation of Veps and low-resource language research.

Reviving the Open Language Archives Community (OLAC)

Nick Thieberger (PARADISEC), Mandana Seyfeddinipur (Endangered Languages Archive), Paul Trilsbeek (The Language Archive), Alexis Michaud (Pangloss), Thomas Kraemer (GESIS - Leibniz-Institut für Sozialwissenschaften), Felix Rau (Language Archive Cologne), Sylvain Loiseau (Pangloss)

Language archives have benefited from the foundational work done by the Open Language Archives Community (OLAC) in the early 2000s, led by Gary Simons and Steven Bird. Following a consultative process funded by the USA National Science Foundation via the E-MELD (Electronic Metastructure for Endangered Languages Data) conferences, OLAC established a standard language metadata set and a metadata harvesting service. Every day, metadata was harvested by the OLAC database from each participating archive. This metadata was presented on a single page per language, providing a targeted search mechanism for language records.

Since that early work the number of participating archives reached a peak of around 60. In 2018, Gary Simons and Steven Bird asked the Digital Endangered Languages and Musics Archives Network (DELANMAN) to take on the task of updating OLAC. Their thoughts about improvements in such areas as participation, access, and sustainability are set out in a detailed 20-year update on OLAC published in 2022 (Simons and Bird 2022). DELAMAN members sought funds to do this work, and an Australian Research Council project that included the OLAC work began in 2022.

Initially, it was necessary to update the current OLAC site to ensure it was secure and stable. The next phase focused on auditing the existing infrastructure and to plan the upgrade of OLAC.

We have since transitioned into a comprehensive modernization of the harvesting and search infrastructure, established continuous integration and containerized deployment on a new infrastructure.

This includes an enhanced metadata processing pipeline to enriched metadata with language identifiers from Austlang and Glottolog, mapped via ISO 639 codes to improve findability.

One component of this effort is the development of a new search interface, which provides advanced filtering while maintaining access using OLAC IDs through links to legacy metadata.

Beyond serving legacy and current metadata, the platform hosts over 700 content pages, including the formal OLAC metadata specifications and a curated list of community resources, such as events, tools, and services.

Looking ahead, the project will implement automated metrics (metadata quality score) and link validation to ensure data integrity, and to incentivize the participation of associated institutions.

In this presentation, we will demonstrate the functionality of the new OLAC site and the enhanced search mechanism.

Accompanied Fieldwork in Linguistics

Martha Tsutsui “University of California, Santa Barbara”

The stereotypical linguistic fieldworker is often imagined as single, mobile, and unencumbered by caregiving responsibilities (Cassell 1987; Newman 2009). Even when researchers do bring their families to the field, they are frequently represented as “sole actors” studying others who, in contrast to the researchers, are “presented as social beings whose identities and practices are defined by their social relations.” (Korpela et al. 2016:3). Fieldwork manuals rarely acknowledge that researchers may have responsibilities beyond research itself, the notable exception being Everett (2006; later Sakel & Everett 2012). This silence raises a broader question: must parenting and linguistic fieldwork be mutually exclusive? This paper addresses that question through a case study of my own “accompanied fieldwork” (Korpela et al. 2016) alongside accounts and testimonials from other scholars. In fall 2024, I began a postdoctoral fellowship five months postpartum following the birth of my daughter. Having previously completed two solo field trips during my PhD, I prepared to return to the field site following a multi-year hiatus due to the COVID-19 pandemic and my pregnancy- this time with an infant and my partner. While seeking models for conducting linguistic fieldwork with a young child, I found few examples in linguistics. Apart from Tsutsui (2020), most discussions of accompanied fieldwork come from adjacent disciplines such as sociology (Jones 2012), anthropology (Braukmann et al. 2020), and ethnomusicology (Stobart 2021). The relative absence of such narratives in linguistics indicates both a gap in our professional discourse and a barrier to participation. Drawing on four field trips now successfully completed with my family, this paper outlines practical and conceptual considerations for accompanied fieldwork across regional and cultural contexts, such as logistics (e.g., budgeting, health and safety, childcare, and education), as well as local cultural expectations, shifting researcher positionality, and the impact on data collection and community relationships. I argue that accompanied fieldwork does not merely pose logistical challenges but also reshapes field relations in productive ways, foregrounding relationality and care as central to linguistic practice. Building on Macaulay’s (2011:204) call for linguists to speak openly about all aspects of fieldwork experience, I contend that parenting must be included in our methodological conversations. More than a decade later, the realities of fieldworkers who are parents remain under-discussed in linguistics. By centering accompanied fieldwork, this paper seeks to broaden our understanding of who can participate in field research and under what conditions. Ultimately, I hope to initiate a sustained conversation about how institutions, funding structures, and disciplinary norms can better support researchers who are caregivers, ensuring that parenthood does not obstruct access to linguistic fieldwork.

Preserving Traditional Quechua Ways of Being, Knowing, and Doing in Urban and Migratory Settings: A Native Speaker–Led Documentation Initiative

Gaby Vargas (University of Arizona)

In this talk, I will share a native speaker–led documentation initiative aimed at preserving traditional ways of being, knowing, and doing among speakers of South Bolivian Quechua (SBQ) in the Valle Bajo of Cochabamba, Bolivia. Although the region has been in sustained contact with Spanish since the colonial period, monolingual Quechua speakers remained common until relatively recently. In recent decades however, increasing bilingualism, migration, and urbanization have accelerated a shift toward Spanish. As a result, forms of knowledge, cultural practices, and lived experiences that have historically been transmitted through oral traditions are increasingly at risk of disappearing.

I will describe the development of a community-engaged documentation project conducted as part of my graduate research and grounded in my experience as a native speaker from this area and a Quechua Bolivian linguist. Using a case study of an extended family, I will illustrate the sociolinguistic dynamics of language use in the context of urbanization and migration. I will also outline the practical process of carrying out the project, including recruiting participants, following IRB procedures, producing audiovisual recordings, annotating and translating materials in ELAN, and curating documentation that can support both long-term archiving and community access.

I will reflect on the challenges and opportunities of conducting documentation as a young bilingual speaker who was born and raised in these communities but now navigates different social and academic contexts. Finally, I will also discuss efforts to train other native speakers in this project and to engage community members in the process, contributing to broader initiatives of language maintenance and cultural revitalization.

Language documentation as task: Integrating community-based documentation into task-based Sonsorolese learning

Vasiliki Vita (School of Oriental and African Studies), Chelsea Pedro (Young Historians ofonsorol, Palau Community College)

Language documentation and revitalisation are often described as complementary yet remain institutionally distinct, with documentation oriented toward archival preservation and revitalisation toward pedagogy. (Himmelman 1998; Woodbury 2011; Hinton 2011)

This paper argues that community-based language documentation can itself function as the central task in task-based language learning. Drawing on Vita et al.'s (2023) ELAR collection, we analyse Team's task of learning about the process of making a coconut drink (sn_hs_nov_21_22_1 and sn_hs_nov_21_22_2).

Mapping these activities onto Long's (2009) methodological principles, we demonstrate that language documentation as a task in task-based learning is successful for learning Sonsorolese. Documentation activities generated authentic communicative interaction making a coconut drink, interacting with the teacher and other learners. Participants negotiated meaning in real time through asking questions about the seasons of harvesting and equipment used in the process. Archival production and language learning became simultaneous processes. This is demonstrated by asking the teacher questions, preparing their roles for the session and evaluative questions on their learning, such as what went well, what didn't go well, how we could improve.

Rather than repurposing documentation outputs for teaching, this reframing embeds documentation within pedagogy, collapsing the archive/classroom divide. We suggest that framing documentation sessions as pedagogical tasks provides a practical mechanism for aligning documentation with revitalisation goals and for repositioning speakers as active producers of both linguistic knowledge and archival records.

From Colonial Prose to Structured Data: Formalizing Example–Translation Relations in Colonial Mixtec Grammars

Lena Weißmann (Freie Universität Berlin)

Colonial-period grammars of Mixtec (16th–18th century) constitute a key corpus of early linguistic documentation of this Indigenous language of Mesoamerica. However, due to their historical particularities, they remain difficult to use. Recent critical print editions have improved scholarly access by providing essential philological analysis, yet the print format is limited in its ability to render the structural complexity of the sources transparent. In particular, linguistic examples and their Spanish equivalents remain embedded in dense colonial prose. Even when example-translation correspondences are listed in appendices, print editions cannot present them in a way that is both systematically retrievable and context-preserving. Digital relational encoding, by contrast, preserves them in their explanatory context while making them transparent to readers and computationally reusable.

DEKoMix-P develops digital scholarly editions of four colonial Mixtec grammars using TEI-XML and establishes a documented workflow for the future integration of additional texts. With the broader objective of enhancing accessibility, reusability, and long-term preservation of the material, the present pilot phase of the project focuses on establishing structured, machine-readable “basic editions” as a foundation for future linguistic annotation and follow-up research—including potential collaborative work with contemporary Mixtec-speaking communities in the case of one text.

The formalization of example–translation relations is a central methodological issue at this “groundwork” stage. In the original texts, Mixtec forms and their Spanish equivalents are frequently linked through ambiguous syntax, punctuation, and discursive markers. The project therefore implements a reusable workflow for: (1) systematic marking of Mixtec examples, (2) explicit linking to author-provided Spanish translations, and (3) normalization of the Spanish descriptive prose.

By doing so, the project addresses a complementary structural dimension to prior major digitization initiatives, most notably TICHÁ for colonial Zapotec, which has set important standards in community collaboration and in the integration of interlinear glosses generated with FLEx. While foreign-language segments are systematically marked and, in some cases, enriched with linguistic annotations, the example–translation correspondences remain textually implicit rather than formally modeled.

By formalizing these correspondences at the level of the historical documentation itself—prior to the introduction of modern linguistic analysis—DEKoMix-P renders the documentary logic of colonial linguistic description explicit and transparent both to readers of the digital edition and to computational processing. In doing so, it preserves the epistemic layer of the colonial author as a clearly identifiable level of description, distinct from later analytical interpretation. This separation establishes a robust foundation for subsequent linguistic analysis, allowing historical documentation and present-day linguistic knowledge to be related without collapsing the two perspectives.

By transforming colonial linguistic prose into structured, interoperable data, DEKoMix-P establishes the technical conditions for making the linguistic information contained in these sources accessible and methodologically reusable. It thereby creates a foundation on which both scholarly research and community-led engagement can develop new questions and interpretations of the historical material.

Reconstructing multilingual pasts: Field linguistics meets data-science in multilingual Northwest Amazonia

Katarzyna Wojtyla (University of Warsaw), Diego Ruiz (Independent Researcher)

This paper presents a computational approach implemented around the GEDCOM data format—a data format for genealogical data storage—designed to support sociolinguistic documentation and analysis of multilingualism in indigenous Amazonia. Drawing on pilot studies conducted in 2024 among elders from seven indigenous groups (Murui, Minika, Ocaina, Nonuya, Bora, Muinane, and Andoke) of the Caquetá–Putumayo (CP) river basins in Colombian Amazon, the project integrates genealogical reconstruction and ethnolinguistic interviews documenting more than 1,100 individuals and 295 families across five generations (Wojtylak 2025).

The workflow proposed and used in this project integrates three interrelated work streams:

- a) the digitization of an extensive corpus of genealogical field notes and its formalization as structured CSV datasets;
- b) the implementation of a service in Python to parse these datasets into a relational database and map the records as GEDCOM data structures (*Gedcom Creator*, Ruiz 2025), validating the factual consistency of genealogical relationships—a process that proved highly error-prone and required iterative verification; and
- c) the use of Python-based tools to reconstruct genealogies from SVG images and produce quantitative analyses and visualizations (e.g., plots, charts, infographics) illustrating language (transmission) patterns (Wojtylak & Ruiz 2025).

Together, these components constitute a comprehensive workflow linking field data collection, validation, visualization, and analysis. They allow researchers to measure the number of languages spoken per generation, compare multilingualism across groups, and trace correlations between intermarriage and language vitality.

The analysis revealed a shift in the CP region from high multilingualism in the early twentieth century—up to four indigenous languages per person—to near-total Spanish monolingualism among younger speakers today. While the multilingual character of the region was previously known from qualitative studies (e.g., Echeverri 1997, Gasché 2009), no quantitative data had ever been available. The use of the presented workflow made it possible to quantify CP multilingualism for the first time, providing empirical measures of language use across generations. The results also show that CP multilingualism was unevenly distributed: dominant groups such as Murui, Minika, and Bora tended to remain largely monolingual, whereas smaller or socially less dominant groups—such as Nonuya, Ocaina, and Andoke—were markedly multilingual. These findings highlight the value of genealogical data for understanding long-term processes of language contact and shift, demonstrating how historical multilingualism can be reconstructed even in regions with limited archival evidence.

The paper discusses methodological considerations in adapting genealogical data structures for sociolinguistic purposes. It argues that combining Python-based data processing with GEDCOM-compatible visualization provides a scalable and replicable approach to documenting language shift, diffusion, and resilience, as well as reconstructing past multilingualism in contexts where historical documentation is scarce.

Identifying and Acknowledging Language Resource Aggregates in Language Archive Catalogs

Oksana Zavalina (University of North Texas), Hugh Paterson III (Institute for Advanced Study, Princeton), Sergio Coronado (University of North Texas)

Language archives support language scholars and communities by stewarding language resource collections. Discovery services are paramount for maintaining connections with audiences. Digital language archives have deployed a variety of methods for arranging the contents of their collections (Paterson III 2021b, 2021a). Popular across language archives is the ‘bundle’ (Johnson 2002) or an aggregation wherein several resources might be represented by a single descriptive metadata record. Bundles can be internally complex and have external relationships to other resources. Examples include multiple interviews recorded on the same audio tape, or several audio and video recordings of the same interview. Another common bundle arrangement is the inclusion of transcriptions with their associated audio or video resources. Huber (2023) points out that bundles present challenges for rights management, while Burke and Zavalina (2019) note that they complicate API access to individual resources.

Libraries around the world take an alternative approach. They have centered descriptions of resources in their collections around the Library Reference Model IFLA-LRM (Riva, Le Bœuf, and Žumer 2017) which has as its core the Work-Expression-Manifestation-Item (WEMI) ontology. The IFLA-LRM model which facilitates search and discovery based on the WEMI entities’ attributes and relationships is not attested to be in use at any language archive in the DELAMAN community. While the full IFLA-LRM and its closely associated Resource Description and Access (RDA) model provide guidance for representing bibliographic entities beyond WEMI, along with their numerous attributes and relationships, Coyle (2022) points out that a minimal ontology of WEMI entities with model specific attributes and relationships allows for accurate record keeping and the flexible building of discovery services across a variety of industries.

In this paper we present some examples of how WEMI would be applied to metadata records for two types of resources: (1) a legacy audio recording which has been digitized and (2) an audio or video recording with an accompanying transcript. We then contrast this with the metadata records currently observed in digital language archives. We address critical questions about the independent nature of transcriptions and annotations as works (as opposed to expressions) in the WEMI model. We then describe how metadata record independence or conflation impact citation and referencing of resources in bibliographic contexts. Our conclusions suggest that the adoption of WEMI within language archives would provide more consistent discovery services within and across archives, increase compatibility of metadata with that used by libraries which often host language archives, improve citation and referencing capabilities by providing a framework for the assignment of persistent identifiers, and streamline the metadata management by aligning metadata records with unique individual resources as units of description.

POSTER PRESENTATIONS

Making Space for Both the Silent and the Spoken: Building the Tłıchq Archives and Museum from Traditional Values

Rebecca Bourgeois (Western University), Renee Saucier (Tłıchq Government), Danielle Dacanay (Tłıchq Government)

National narratives are cemented in places through sites of pedagogy, such as memorials and museums, wherein material culture is used as a tool to construct and project selective discourses of past life (Anderson 2017; Hooper-Greenhill 1988; Lleras 2017). The power of such narratives is embodied through the language and framing of exhibits, and is furthered in how we write about them (Jackson 1989; Low 2009), constructing national narratives centered on the stories that a nation wishes to project (Anderson 2017; Lleras 2017; Weiser 2009, 2015, 2017).

The Behchokq Community Museum—whaehdog kg, meaning “house of the ancestors” in Tłıchq yatił (formerly Dogrib)—opened in 2021 to celebrate the 100th anniversary of the signing of Treaty 11 (1921). Its inaugural exhibits focused on the history of Treaty 11, and Chief Mohwhi, who signed on behalf of the Tłıchq Nation. The Tłıchq are a self-governing Indigenous Nation with four communities (Behchoko, Whati, Gamèti, Wekweèti) in what is now known as the Northwest Territories, Canada. This space served as a venue through which the Tłıchq could tell their own story to both their citizens and the world. The museum was welcomed warmly by the community, and thus the Tłıchq Government (TG) decided to expand its programming in concert with the development of Tłıchq Nàowoo Deghats’etsaa: the Tłıchq Cultural Commons Digital Archive (TCCDA). The problem, however, was that there were no policies or long-term plans in place to support the museum or the digital archive, especially in terms of access, acquisition, vision, and purpose. A research project was launched in 2021 to address these gaps in policy and redefine principles of curation according to Tłıchq traditions of caring for knowledge, investigating 1, the purpose of Tłıchq cultural heritage collections in the context of the breadth of language, culture, and way of life, and 2, ways to build a museum and archive according Tłıchq traditional values, instead of imposing existing colonial structures.

The Elders we worked with relied on the moral teachings that Tłıchq yati carries across

generations in discussing our shared goals ([Author 1] 2025). Some of these traditional teachings, however, meant that we needed to distance ourselves from an ethnographic process that functionally translates the oral into written, instead adopting methods that allowed for the description of silence (Hirschauer 2006). To do so, we built a strong like two people research methodology, which was based on the Tłıchq value values of interpersonal unity through the meeting of diverse perspectives ([Author 1] 2026). During our research process—wherein we employed learning circles and semi-structured interviews—we decided to predominantly speak in terms of museums, as the Elders we work with are more familiar with that than the rhetoric of archives. When reimagining an archive, the last thing we wanted to do was explain to them what an archive was and how it differed from a museum. Our approach employed ethnographic description not as documentation, but instead, as voiced by Hirschauer (2006, 413), as a means to “solve the problems of the voiceless, the silent, the unspeakable, the pre-linguistic, and the indescribable.” The Elders spoke for the voiceless (e.g., past and future generations) while communicating through silence (e.g., where cultural protocols kept them from sharing sensitive information from outsiders) in pursuit of explaining the indescribable (e.g., the system by which knowledge is organized and shared).

In this presentation, we walk through the example of the Behchoko Community Museum and the TCCDA to highlight how we used ethnographic-, linguistic-, and Indigenous-informed research methodologies to address the voiceless, the silent, and the indescribable to build a system of curation based within Tłıchq traditions of caring for knowledge.

Multilingual Transcription as a Continuous Temporal Field

Tracy Burnett (Independent Author)

Spoken language documentation typically aligns one or more transcriptions and translations to time-indexed audio using tier-based annotation systems, in which annotations are organized as ordered sequences within parallel tiers. This paper presents an alternative representation: treating multilingual transcription as a continuous temporal field in which all linguistic segments are anchored directly to shared time, without requiring linear order within a tier.

In the presented system, each segment is defined only by a start time, end time, language, and textual content. Segments may overlap freely, nest, or occur out of chronological order within a given tier. Cross-language correspondence emerges from partial and fuzzy temporal overlap rather than from predefined tier relationships. Because all segments occupy a shared continuous timeline, segments in one language can be used to trace, assemble, and reinterpret structure in another at varying analytic scales. This permits alternative interpretive sequencing, reassembly of narrative units, and systematic comparison of structural groupings across languages.

By preserving overlapping, heterogeneous, and non-linear segmentations within a unified temporal framework, the model allows multiple structural interpretations to coexist without rewriting underlying data. The approach supports dynamic rendering of multilingual segments at different granularities, enabling interactive transcripts, educational interfaces, and structured comparison of segmentation and interpretation across languages. This presentation describes the conceptual rationale, the interval-based storage model, and its implementation across existing annotation formats. It also discusses implications for documentation workflows and long-term archiving. More broadly, it explores how time-continuous representations may contribute to more flexible, durable, and inclusive preservation of spoken language data.

The Dictionary as Pedagogical Technology: Language Documentation, Schooling, and Linguistic Sovereignty in the Kheul Karipuna and Galibi-Marworno Context

Amanda da Costa Carvalho (Instituto Federal do Amapá, Universidade Federal de São Carlos), Janina da Silva Forte (Associação das Mulheres Indígenas em Mutirão – AMIM), Jaciara Silva dos Santos (Governo do Estado do Amapá), Glauber Romling da Silva (Universidade Federal do Amapá), Mara Santos (Universidade Federal do Amapá)

Language documentation has increasingly moved beyond descriptive goals to support pedagogical, socio-political, and revitalization-oriented practices in Indigenous contexts. In Brazil, Indigenous languages face pressures related to multilingual contact, schooling centered on Portuguese, and shifting patterns of language transmission, making educationally oriented documentation strategies essential. Research on Kheul do Uaçá, a French-based creole shared by the Karipuna and Galibi-Marworno people in the Franco-Brazilian border region, shows that the language emerged in colonial contact settings and later consolidated as a central marker of collective identity and multilingual interaction (Silva, 2021; Jennings & Pfänder, 2018; Lüpke et al., 2020). This paper analyzes the online bilingual and multimedia dictionaries of Kheul Karipuna and Kheul Galibi-Marworno as pedagogical technologies developed through collaborative language documentation. Produced within participatory and formative frameworks that include Indigenous researcher training, lexical database construction, and school-oriented materials, the dictionaries exemplify how documentary outputs can function directly in teaching and learning processes (Silva et al., 2023). Evidence from other Indigenous documentation initiatives in Brazil similarly demonstrates that documentation products (such as grammars, lexical databases, and narrative collections) can strengthen mother-tongue education and preserve ancestral knowledge within school environments (Chagas; Taffarel; Txicão, 2024). Historically, the production of didactic materials in Kheul has operated as both a pedagogical strategy and a form of cultural resistance, contributing to linguistic valorization, differentiated education, and social justice in Indigenous schooling (Aniká; Machado, 2025). These processes align with national curricular principles that define Indigenous education as intercultural, community-based, and grounded in linguistic plurality and autonomy (Brasil, 1998), while recent language-policy advances, such as the co-officialization of Indigenous languages in Amapá, reinforce the institutional and educational legitimacy of Kheul. By positioning the dictionary as a living pedagogical artifact linking documentation, schooling, and language policy, this study argues that documentary lexicography can operate as a key instrument of linguistic sovereignty, intergenerational transmission, and culturally grounded education.

Canastofonias”: local indigenous language archives

Deyanira S. Moya-Chaves (Pontificia Universidad Javeriana)

Four Amazonian languages from Colombia -Uitoto, Bora, Muinane, and Okaina- spoken in La Chorrera by the Peoples of the Center were once silenced and prohibited. Grandmothers and grandfathers resisted this silencing and kept ancestral knowledge encoded in their languages, in their memory and in oral tradition. However, what could be considered a linguistic shame established mainly as a consequence of colonial processes has taken its toll on new generations, and therefore, today we face a progressive decrease in speakers. To reverse this situation, the Peoples of the Center have included in their Plan of Life actions to recover and strengthen their languages: they teach them in schools and preserve them, for example, in verbal art or *rafue uitoto* (word with power) (e.g. Echeverri 2022; Becerra 1998; Urbina Rangel 2010). Even so, grandparents wonder how to ensure that both the languages and the knowledge they convey are not lost, especially as technology booms in this Amazonian territory. Under this framework, *canastofonías* emerge as a feminist linguistic archive co-created with elderly women who have kept in their living memory lullabies and other children's songs full of ecological knowledge that they themselves want to recover and transmit to new generations. This talk will present the local technological tools created to develop such archive as well as the process of reflection on what indigenous language archives themselves entail.

Linguistic Documentation and Archiving of Tjwao: Making Digital Archives More Navigable and Engaging

Shumirai Nyota (Zimbabwe Open University)

The documentation and archiving of Tjwao, a critically endangered Khoisan language spoken by the San community in the western Zimbabwe district of Tsholotsho, has moved from initial efforts that prioritized phonetic and lexical corpora data. Current efforts prioritize strategic planning and policy-driven innovation to move beyond mere storage. The goal is to transform passive data repositories into dynamic, accessible, and community-centric resources that ensure the language's survival and accessibility in the digital age. A primary technical objective involves updating Open Language Archives Community (OLAC) metadata. By aligning with these global standards, Tjwao records become more discoverable across international networks. This structural refinement ensures that the documentation is integrated into the broader ecosystem of global linguistic research, making it easier for scholars, researchers and community members alike to locate specific cultural and linguistic assets. To increase user engagement, the project is integrating culturally relevant interfaces. This includes visual navigation tools, indigenous taxonomies, and multimedia previews that reflect the Tjwao world-view. By making the interface intuitive and aesthetically aligned with San cultural heritage, the archive evolves from a technical database into a digital home for the community's collective memory. In Tsholotsho where Tjwao speakers reside like in rural Zimbabwe, there are challenges to "digital archiving" due to lack of limited infra-structure. To bridge this gap, the project uses offline access solutions. This includes use of portable media kits (such as pre-loaded tablets and durable hard drives) and the optimization of files into low-bandwidth formats.

These measures ensure that the outcomes of documentation remain available to those who provided the data, the Tjwao community regardless of internet connectivity. Sustainability of the Tjwao archive rests on capacity development. Training programs are used to equip community members, local educators, and archivists with essential skills in metadata management and digital literacy. This shifts the role of the Tjwao from being objects of study to participants. Hence participatory methodology is used. By shifting the role of the community from "subjects of study" to "stewards of data," the study fosters deep community ownership. This ensures that the Tjwao community manage their archived data, updating it as and when it becomes necessary to do so for generations to come. Through these integrated strategies, the Tjwao documentation and archiving study serves as a replicable model for other endangered languages. It demonstrates that future-ready language preservation must be as much about inclusive access and human agency as it is about technical precision.

Vibe coding an online ELAN viewer

Paul Trilsbeek (Max Planck Institute for Psycholinguistics)

This paper describes an effort to use Agentic AI for the creation of a web-based viewer for interacting with media that are annotated with ELAN (Wittenburg et al. 2006).

From their first introduction, Large Language Models (LLMs) have been used to assist with writing computer code, initially starting with context-aware code completions but quickly evolving to writing entire designs and complete pieces of code as well as performing testing, debugging and code optimization/refactoring tasks. The relatively strict syntax of programming languages makes them a particularly suitable domain for LLMs. The increased accuracy of subsequent LLM generations as well as the introduction of specialized models for coding made them perform increasingly well.

In the second half of 2025, Agentic AI solutions became widely available from most of the larger commercial LLM creators including Anthropic, Google and OpenAI. In these solutions, one or more “autonomous agents” interact with one or more LLMs to perform an entire workflow that is specified by the user either in natural language or in somewhat more structured Markdown, without the user having to be involved for each step of the workflow. The agents interact with the models to interpret the workflow, to define specific tasks that need to be performed, and to ultimately perform the tasks and report back to the user when the workflow is completed. The software solutions built around the agents can interact with the user’s local file system, a local web browser, as well as services that are running either on the user’s computer or elsewhere, allowing for complete cycles of planning, analysis, design, implementation, deployment and testing to be performed.

In February 2026, we used Google’s Antigravity agentic AI development platform to create a new version of a web-based viewer for media files that are annotated with ELAN. At the time of writing, Antigravity uses two LLMs from Google (Gemini 3.1 Pro and 3 Flash), two from Anthropic (Claude Opus 4.6 and Sonnet 4.6) and one from OpenAI (GPT-OSS). There are hourly and weekly usage limits for these models that depend on the complexity of the model, whether or not it is hosted by Google, and what kind of subscription the user has (free or paid Pro/Ultra tiers).

The starting point for this effort was a very basic existing web-based viewer for the Drupal 7 content management system that included a parser for ELAN EAF files in PHP 7 and a viewer that showed a table of annotations using the Angular JavaScript framework. Within two hours, the existing solution was ported to Drupal 11, the latest version of PHP and the Vue JavaScript framework that is better suited for the type of application. Surprised by how well this worked, we kept specifying additional features to be added to the viewer and after about one week we had a fully functional viewer that includes most of the view modes that the stand-alone ELAN tool offers: Timeline, Grid, Text and Subtitles. Also, a waveform and spectrogram option were included. Developing the same functionality in a traditional manner would have taken at least several months.

Some observations:

- it was remarkable how much knowledge the models had about concepts such as “multi-tiered timeline”. Without giving much further instruction, the tool produced a view that looked very much like a modernized version of the ELAN timeline, see figure 1.
- the implemented solutions did typically need some further refinement or fixing and there was a large variation in how many iterations were needed before any particular issue was resolved.

- specifying a single feature per iteration seemed to work better than combining several features in one prompt.
- at times, the system lost important context information after a number of iterations. This could be mitigated by defining “skills” and “workspace rules” that the agent always has access to.
- the “simpler” models such as Gemini 3 Flash are perfectly adequate for tasks that do not involve higher level reasoning. Given the rate limit differences, it is advisable to use those by default and only use the more advanced models for overall design or complex debugging tasks.
- Antigravity is currently in “public preview” . We sometimes experienced issues where it got stuck and did not respond. It is also unknown whether the usage limits might change when the first stable version is released.

While perhaps being overhyped for general use in people’s personal lives, agentic AI is radically changing the world of software development and it has only just started. It will have a major impact on what can be done with fewer resources in less time. This small exercise has shown that this also applies to software in the domain of language documentation and archiving and it may well extend to other systematic tasks in the domain such as archiving workflows.

An Annotated Searchable Online Corpus of the Juhuri Language

Ksenia Viktorova (European University at St. Petersburg)

Juhuri is a Southwestern Iranian language traditionally spoken by the Mountain Jewish community. Until the late 20th century, Mountain Jews resided in various settlements across the Eastern Caucasus, a region now divided between the Russian Federation and Azerbaijan. Four distinct dialects emerged within this territory: Northern (Chechnya, Kabardino-Balkaria, and Northern Dagestan), Central (Derbent and its environs), Southern (Baku, Krasnaya Sloboda, and surrounding villages), and Vartashen (the town of Vartashen).

In recent decades, the majority of Mountain Jews have left the Eastern Caucasus for major Russian cities, Israel, the United States, and other locations. This migration has accelerated a language shift toward dominant national languages. However, the pace of this shift varies significantly across dialects. Field observations in 2024–2025 [Viktorova in press] indicate that the Central dialect now has the fewest speakers while the Southern dialect remains relatively vital.

Through community-led initiatives, a substantial corpus of Juhuri texts has been digitized and made accessible online. It represents primarily the Central dialect which served as the basis for the literary language during the Soviet era and the Southern dialect which is the most popular today. In contrast, publications in the Northern dialect are scarce and largely remain undigitized. The Vartashen dialect lacks a written tradition entirely; at present, documentation of this variety is limited to field recordings collected from several speakers in 2024, a collection that is expected to expand through ongoing research.

To preserve and study all the varieties of Juhuri, an annotated searchable online corpus is being developed on the Tsakorpus platform. The workflow includes the following stages:

1. **Data Collection and Metadata:** Digitizing Northern publications and transcriptions of Vartashen records. Assigning comprehensive metadata to each entry (dialect, author, date, source, genre, original/translation/elicitation).
2. **Script Unification:** Juhuri texts exist in Hebrew, Latin (1920s–30s and modern Azerbaijani), and various Cyrillic scripts. To establish a consistent searchable layer, all materials are converted into a unified spelling standard using Python transliteration scripts. The *ad hoc* standard is based on the dictionaries [Agarunov and Agarunov 2010] and [Naftaliev 2016]. They also serve as the source for lexical meanings.
3. **Manual Annotation:** A “gold standard” subset of approximately 1,500 word forms for each dialect is annotated manually. The morphological markers and glosses are based on the following descriptions: for the Northern dialect [Miller 1901], Central [Anisimov 1932, Authier 2012], Southern [Miller 1952].
4. **Semi-automated Analysis:** This annotated subset is used to train the UniParser tool, enabling semi-automated morphological glossing for the rest of the data. Morphological homonymy is resolved manually through the Tsakorpus interface.
5. **Indexing and Deployment:** The final dataset, comprising approximately 50,000 annotated word forms, is converted into JSON format and indexed using Elasticsearch.
6. **New materials are added on the regular basis through the automated ingestion pipeline.**

The corpus is deployed on the server of the European University at St. Petersburg. It provides linguists, Jewish studies specialists, and the Mountain Jewish community with an open-access tool for the documentation and preservation of Juhuri linguistic heritage.

From lexicons to life stories: A mixed-methods documentation of Northeastern Neo-Aramaic Senaya

Kathy Zaruba (Leibniz-Zentrum Allgemeine Sprachwissenschaft [Berlin])

This contribution addresses the dialectic between breadth and depth in data collection by introducing a mixed-methods documentation corpus of Northeastern Neo-Aramaic Senaya among heritage speakers in the United States.

Elicited fictional narratives using the Multilingual Assessment Instrument for Narratives (MAIN; Gagarina et al. 2019) are complemented by open-ended personal narratives to capture both structured and naturalistic language use and enable systematic comparison as well as comprehensive coverage of semi-spontaneous production. To enhance data interpretability, the methodological apparatus includes adapted versions of the Heritage Language Experience questionnaire (HeLEx; Tomić et al. 2023) to profile speakers' language backgrounds and sociolinguistic characteristics; the Controlled Oral Word Association Task (COWAT; Benton 1969) and Crosslinguistic Lexical Tasks (CLTs; Haman, Łuniewska & Pomiechowska 2015) to assess productive and receptive lexical range in Senaya; and the Lexical Test for Advanced English Learners (LexTALE; Lemhöfer & Broersma 2012) to measure majority language (English) proficiency and bilingual competence.

By integrating sociolinguistic profiling and lexical competence measures to contextualize semi-spontaneous narrative data, the methodological approach presented in this contribution generates a multilayered documentary infrastructure which balances breadth and depth of its datasets and enhances longevity and versatility of the corpus. Beyond its application to Senaya, the test battery is designed for adaptability: the standardized instruments enable cross-linguistic comparison while the overall framework can be deployed across documentation contexts, establishing a foundation for networked corpora of under-documented (heritage) languages.

Drawing on preliminary findings from fieldwork using these measures, this work demonstrates how competing demands of documentary efficiency, analytical rigor, and community responsiveness may be addressed in documentation projects. It advances ongoing discussions on documentation standards and archival practices by offering a methodological case study for building interoperable, multi-purpose corpora of endangered heritage languages.