From linguistic analysis to public engagement: Improving community access to Tunica language documentation

Andrew Abdalian

Tunica is a language isolate from the southeastern United States whose last known native speaker died in 1948. Before that time, three Tunica speakers had worked with three linguists who documented their speech over three periods (1886, 1907–1910, 1933– 1939). Since 2010, the Tunica-Biloxi Tribe's Language & Culture Revitalization Program (LCRP) and Tulane University's Interdisciplinary Program in Linguistics have collaborated on a project to revitalize Tunica based on this documentation. This collaborative Tunica Language Working Group, or *Kuhpani Yoyani Luhchi Yoroni* (KYLY), is one of community-engaged research, "a framework that seeks and nurtures community involvement, leverages community knowledge, and is led by community need" (Baldwin et al. 2022:176).

From its inception, KYLY has included tribal language workers who decide on the direction and priorities of the language project. Involving the wider community is more challenging. The community must have multiple ways to engage with the language. One of the methods of engagement for which the community has expressed great interest is looking at and interacting with existing language documentation materials.

This presentation explores how KYLY has responded to this community desire, and the successes and challenges of presenting the language documentation transcribed, parsed, and analyzed in SIL's Fieldworks Language Explorer (FLEx) publicly. It examines how software designed for linguistic analysis often lacks built-in tools for easily publishing language documentation in an accessible manner, and the third-party software solutions and bespoke methods that KYLY has developed and is developing in its effort to take the Tunica language documentation in the FLEx database and make it accessible to the community.

Works cited

Baldwin, Daryl, G. Susan Mosley-Howard, George Ironstack, and Haley Shea. 2022. "Community-Engaged Scholarship as a Restorative Action." In Replanting Cultures: Community-Engaged Scholarship in Indian Country, edited by Chief Benjamin J. Barnes and

Stephen Warren. Tribal Worlds : Critical Studies in American Indian Nation Building. Albany, New York: State University of New York Press.

Designing elicitation tools for Asante Twi verbs of separation: issues of ecological validity.

Dorothy Agyepong

In this talk, I report on three data collection methods I have used in documenting verbs of separation in Asante Twi, a Kwa language spoken in the Ashanti and Eastern Regions of Ghana. The methods are video-stimuli elicitation, procedural narratives and textual examples. For the video-stimuli elicitations, I tried to balance the universality with the locality by supplementing an MPI set of stimuli (Bohnemeyer et al. 2001) with comparable stimuli set (Agyepong 2015) for eliciting local cultural verbs for describing separations within the Asante Twi context. Whileboth stimuli portrayed various types of separations involving objects, the Agyepong (2015) videos depicted the separation of objects local to the Ghanaian context, for example, cassava, plantain, coco yam, palm fruit etc. For each of the set, participants were shown the videos 2-3 times and asked follow up questions in Asante Twi. All the elicitations and interviews were both video and audio recorded. For the second set of data, participants provided procedural narratives on events such the killing and cutting up of a chicken for food, the preparation of palm wine from a felled palm tree, the distillation of gin from palm wine as well as the harvesting of crops such as plantain and cocoa. The third set of data was compiled from written texts such as Asante Twi novels, an Akan book of proverbs, and Christaller's (1933) Akan dictionary. These written sources, provided a range of contextualised uses of each of the verbs.

From a methodological point of view, I address issues of ecological validity of the video stimuli; triangulation and how data generated by different methods either converge or diverge. I also attempt to answer the question "How adequate are the various methods in capturing the culture specific semantics of the verbs?"

References

Agyepong, Dorothy Pokua. 2015. "Culture specific Cut and Break videos". Unpublished video clips.

Bohnemeyer, Jürgen, Melissa Bowerman and Penelope Brown. 2001. Cut and break clips. In Levinson, Stephen C., and N.J. Enfield (eds.), *Field Manual 2001, Language and Cognition Group, Max Planck Institute for Psycholinguistics.* Nijmegen: MPI, 90-96.

Christaller, John Gottlieb. 1933. *A Dictionary of the Asante and Fante language called Twi*. Basel: Basel Evangelical Missionary Society.

Addressing the challenges of remote collaboration for language documentation in Nigeria

Chika Kennedy Ajede, Olga Olina, Miracle Oppong Peprah, Nlabephee Kefas Othaniel, Lora Litvinova, Jakob Lesage

Nigeria is home to over 500 languages, nearly half of which are undocumented and lack even a basic description (Hammarström 2018: 6-7). For Nigerian and non-Nigerian researchers alike, a larger emphasis on remote collaboration could offer a possibility to document and describe languages more efficiently and sustainably. This includes a focus on capacity building with part of the funds otherwise reserved for field travel and allows year-round documentation activities. Remote collaboration faces many challenges, however (e.g. Bolaños et al. 2021), and in Nigeria, specific circumstances add to these challenges. Building trust relationships among project members requires creating an atmosphere of transparency, to avoid potential doubts and suspicions. Compensation is also a sensitive issue as many Nigerians struggle to earn a subsistence minimum. This implies that projects cannot expect consultants to be properly engaged in independent documentation activities without offering a stable income, even in communities where speakers feel reluctant to accept monetary compensation. Electricity and internet coverage in Nigeria are notoriously unreliable. This affects the availability of consultants and requires projects to invest in minimal local infrastructure for transcription and electricity-dependent technology.

Using examples from our projects in north-eastern Nigeria, we propose concrete solutions to these obstacles. We offer these in the form of a model work plan which includes:

- Strategies for recruiting consultants and setting up well-informed compensation agreements for independent work.
- Workshops within and across communities on data collection and processing.
- Social media based team management that encourages cooperation, team ownership,

and shared ambition.
Our diverse positionalities, from within and outside Nigeria, as community members and as outside linguists, offer an optimistic yet realistic outlook of what we think fieldwork best-practices can look like in a country where remote collaboration may be perceived as a challenging proposition.

References

Bolaños, Katherine, Jakob Lesage & Sheena Shah. 2021. Planning remote field work for language documentation. *ELDP Remote Fieldwork workshop series*, 15 September, 2021, Online.

Hammarström, Harald. 2018. A survey of African languages. In Tom Güldemann (ed.), *The Languages and Linguistics of Africa*, 1–57. Berlin, Boston: De Gruyter.

Practices of the (Un)sayable: Language Trauma in Language Documentation
Section: Building Relationships

Ioana Aminian Jazi

The erosion of minority languages within communities often results from various forms of language suppression, leading to deep-seated intergenerational trauma among speakers. The dynamic between trauma and language is intricate and reciprocal, involving injuries and disruptions that affect both language and speaking processes. While earlier research predominantly focused on language's role in the discursive construction of trauma as an object of knowledge, recent studies (Busch 2016; Busch & Mcnamara 2020; Busch & Reddemann 2013) have examined language's involvement in the lived experience of trauma. Busch (2015; 2016: 92) introduced the concept of *Spracherleben*, highlighting the pivotal role of language experiences in shaping individuals' perceptions of themselves within verbal interactions across three different axes of recognition/non-recognition, belonging/exclusion, and power/powerlessness.

In this presentation, I will explore how the physical and emotional dimensions of *Spracherleben* imprint themselves on individual memory, shaping deeply ingrained intergenerational dispositions, behaviours, and attitudes towards the heritage language (Abtahian & Quinn 2017). Moreover, language trauma can significantly impact individuals' linguistic repertoire, involvement and participation in language documentation. Additionally, language trauma can influence the "liminal zone of the sayable/unsayable" (Busch 2020: 425) as well as the fragmentation of narratives (Rosenthal 2024 [1995]), conditions for language learning, use, maintenance and abandonment (Anthonissen 2020) as well as agency in language revitalization.

Drawing on insights gained from field research experiences within the Judeo-Spanish language community in Turkey (2013-present), I will explore current frameworks on language and trauma and their relevance to language documentation. Additionally, I will propose strategies for identifying and addressing language trauma within the context of language documentation efforts. Specifically, I will highlight the importance of an intergenerational approach in language documentation as a fundamental element for uncovering and addressing language trauma.

References

Abtahian, Maya Ravindranath and Quinn, Conor Mcdonough (2017): Language shift and linguistic insecurity. In: Language Documentation and Conservation, Documenting Variation in Endangered Languages (13), 137-151.

Anthonissen, Christine (2020): Autobiographical narrative of traumatic experience: Disruption and resilience in South African Truth Commission testimonies. In: Applied linguistics, 41(3), 370-388.

Busch, Brigitta (2015): Zwischen Fremd-und Selbstwahrnehmung. Zum Konzept des Spracherlebens. In: *Mehrsprachigkeit und (Un-) Gesagtes. Sprache als soziale Praxis in der Migrationsgesellschaft. Weinheim/Basel*, 49-66.

Ibid. (2016): Sprachliche Verletzung, verletzte Sprache: Über den Zusammenhang von traumatischem Erleben und Spracherleben. In: *Osnabrücker Beiträge zur Sprachtheorie (OBST)*, 89, 85-108.

Ibid. (2020): Message in a bottle: Scenic presentation of the unsayable. In: *Applied linguistics*, 41(3), 408-427.

Busch, Brigitta and Mcnamara, Tim (2020): Language and trauma: An introduction. In: *Applied linguistics*, 41(3), 323-333.

Busch, Brigitta and Reddemann, Luise (2013): Mehrsprachigkeit, Trauma und resilienz. In: *Zeitschrift für Psychotraumatologie, Psychotherapiewissenschaft, Psychologische Medizin*, 11(3), 23-33.

Rosenthal, Gabriele (2024 [1995]): *Erlebte und erzählte Lebensgeschichte: Gestalt und Struktur biographischer Selbstbeschreibungen*. Frankfurt am Main: Campus Verlag.

Linguistic Fieldwork in Bangladesh with Mru Community: Challenges and Opportunities

Mithun Banerjee

In the presentation, I will describe my recent fieldwork experience in Mru village, located in Bandarban, Bangladesh. Mru is an understudied Sino-Tibetan language spoken in the South-Eastern part of Bangladesh. Mru speakers are known as Mru, who are experiencing severe language shifts due to several socio-cultural reasons and are at a moderate level of endangerment.

In my talk, I will highlight how I fostered an amicable relationship with the Mru community and the experience of my short stay with them. The relationship-building process between the researchers and ethnic minorities is the first step in collecting field data. The foundation of this process focused on respect, mutual understanding, and shared goals. Every speech community is unique regarding social structure, moral foundation, cultural-specific knowledge, and geopolitical surroundings. Therefore, ethical consideration, cultural sensitivity and empowerment of indigenous voices are crucial factors for field linguists in their data collection process. Citizen science in linguistics is a unique approach to documenting, researching, and revitalizing languages. This strategy utilizes the participation of non-specialist volunteers or community people in linguistic projects. By incorporating citizen science into linguistics, there is the possibility of significantly broadening the range and thoroughness of linguistic study, particularly regarding endangered languages.

The Mru people are a peaceful and independent community living in the border regions of Myanmar, India, and Bangladesh. However, security concerns can make accessing this area difficult for researchers. Other unexpected factors can also impact fieldwork, including health matters, political issues, bureaucracies, communication barriers, transportation challenges, and cultural differences. Therefore, it is essential to be well-prepared and knowledgeable in these areas when conducting linguistic fieldwork. Linguistic fieldwork involves building relationships and collaborating with an indigenous community.

Buffalo soup two ways: Showcasing a Dakota/Lakota-led archiving workflow

Elliot Bannister M.Ed.
Tasha Hauff Ph.D., Nacole Walker M.Ed.

Wóoyake (meaning “stories”) is a digital, searchable, user-friendly archive comprised of recordings made by fluent speakers of Dakȟóta/Lakȟóta (a north american Siouan language). It was created by the Standing Rock Sioux Tribe in collaboration with other communities of the Očhéthi Šakówiŋ (Seven Council Fires). This ever-expanding community-led project began in response to the need for access to authentic Dakȟóta/Lakȟóta language – that is, language used by fluent speakers in real life rather than pedagogical materials. Accessible at wooyake.org, it contains digitized audiovisual recordings and written texts going back nearly 200 years.

This presentation demonstrates the workflow we use for getting a written document or audiovisual file onto Wóoyake and ready for use back in the community. This workflow involves a variety of software including Transkribus, ELAN, FLEx, and AdaptIt in order to conduct cataloging, transcription, phonetic transliteration, word glossing, translation, and more. The resulting files are then converted into XML transcripts using a unified schema, and uploaded to the website, along with its media file and rich metadata, to produce searchable, interactive learning experiences. The metadata goes beyond the basics to include real-life community access protocols, rights statements, and interrelational links to other people and places throughout the archive.

Just as importantly as what our workflow is, we explain why we designed it this way and who powers it. Decisions were made to maintain our own protocols, to build community capacity, and to prioritize practical use of the archive over academic pursuits. The process was led at every step of the way by our own Dakȟóta/Lakȟóta speakers – fluent Elders and learners alike. In an era of Indigenous language rematriation, reclamation, and revitalization, Wóoyake’s community-led workflow is a powerful exercise in Očhéthi Šakówiŋ self-determination that serves as a model for decolonizing language documentation.

Musical documentation as language documentation: the Mỹky jakuli

Bernat Bardagil

As an orally-transmitted means of knowledge and communication, traditional music is very adjacent to language. This presentation discusses the connection between the two in the context of language documentation and revitalization initiatives. The focus is on one case study among the Mỹky nation in Brazilian southern Amazonia, that of the documentation of traditional jakuli music and the community's efforts to preserve and revitalize it.

Jakuli, as a musical genre, combines vocal and instrumental music with dance. Typically, men play the reed pipes while women sing the jakuli tunes. The author of this presentation worked together with the three last fluent performers of jakuli in the village of Irupjata, with a heavy involvement of members of the community throughout the process.

The first stages of the revitalization of the jakuli involved the following steps:
- Recording and transcribing the text of some of the tunes.
- Searching for the specific reed used to make katẽntiri pipes, harder and harder to find due to the destruction of the forest by land invaders and ranchers.
- Building a prototype of practice pipes made with PVC piping.
- Developing an easy notation for the music.
- Developing printed and multimedia training materials.

Besides providing an answer to the community's desire to not let jakuli vanish, the ripple effect of our work on its revitalization also encouraged the use of the Mỹky language, by (re)creating a communication space where the Mỹky language is omnipresent, and also by stimulating an interest in the lexicon, myths and oral history connected to the jakuli and to Mỹky-language song more broadly.

Data sovereignty and community engagement at the intersection of language and ecological knowledge: Lessons from two Wixárika (Uto-Aztecan) communities

**Stefanie Ramos Bierge, Alex C. McAlvay,
Gabriel Pacheco Salvador, Tutupika Carrillo De la Cruz**

Indigenous languages and ecological knowledge are tightly interwoven and have been eroded in tandem in many parts of the world. To bolster or revitalize this cultural heritage, many Indigenous Peoples have worked with linguists and/or ethnobiologists, either from within or outside of their communities. Despite an increase in interdisciplinary projects at the intersection of these fields, there has been limited exchange of best practices, ideas, and resources related to ethical and effective work with communities. Two areas where linguists and ethnobiologists have innovated in parallel are data sovereignty and community engagement. While linguists often deposit data in large centralized archives with different levels of access to respect community preferences (Thieberger and Musgrave 2007; Czaykowska-Higgins, 2018; Seyfeddinipur et al., 2019), equivalent archives and graded-access capabilities for ethnobiology are less common. At the same time, ethnobiologists have been experimenting with Traditional Knowledge Labels and Biocultural Labels (Anderson and Hudson, 2020) to ensure that community preferences for information use accompany the data. For community engagement, tools such as "linguistic landscapes" (Shohamy and Gorter, 2009; Blommaert, 2013; Van Mensel et al., 2016) would be easily translated to ethnobiological projects, and community natural history collections (Martin, 2010; Dierig et al., 2014; Balick and Hillman-Kitalong, 2020) could be synergistic with linguistic projects. Linguistics and ethnobiology share significant overlaps not only related to the classification and encoding of biological knowledge, but also in the need for methodologies that foster ethical engagement with local communities. Cross-communication between the two disciplines may lead to mutual benefit. In this presentation, we outline approaches taken to address these issues by both fields, examine opportunities for mutual enrichment, and share experiences from our project with two Wixárika communities in West-Central Mexico.

References

Anderson, J. and Hudson, M. (2020) The Biocultural Labels Initiative: Supporting Indigenous Rights in Data Derived from Genetic Resources. Biodiversity Information Science and Standards 4: e59230. DOI:10.3897/ biss.4.59230.

Balick, M. J. and Hillmann-Kitalong, A. (2020) Ethnobotany of Palau: Plants, People and Island Culture 240. Volume I. Belau National Museum/The New York Botanical Garden, Koror, Palau.

Blommaert, J. (2013) Ethnography, Superdiversity and Linguistic Landscapes: Chronicles of Complexity. Bristol: Multilingual Matters.

Czaykowska-Higgins, E. (2018) Reflections on ethics: Re-humanizing linguistics, building relationships across difference. In B. McDonnell, A. L. Berez-Kroeker, and G. Holton. (eds)

Reflections on Language Documentation 20 Years after Himmelmann 1998. Language Documentation & Conservation Special Publication 15: 110–121 Honolulu: University of Hawai‘i Press.

Dierig, D., Blackburn, H., Ellis, D., and Nesbitt, M. A. R. K. (2014) Curating seeds and other genetic resources for ethnobiology. Curating Biocultural Collections: A Handbook 107– 125. Richmond: Kew Publishing.

Martin, G.J. (2010) Ethnobotany: A methods manual. Routledge.

Seyfeddinipur, M., Ameka, F., Bolton, L., Blumtritt, J., Carpenter, B. Cruz H., Drude, S., Epps, P. L., Ferreira, V., Vilacy Galucio, A., Hellwig, B., Hinte, O., Holton, G., Jung, D., Kasinskaite Buddeberg, I., Krifka, M., Kung, S., Monroig, M., Ngwabe Neba, A., Nordhoff, S., Pakendorf, B., von Prince, K., Rau, F., Rice, K., Riessler, M., Szoelloesi Brenig, V., Thieberger, N., Trilsbeek, P., van der Voort, H., Woodbury, T. (2019) Public access to research data in language documentation: Challenges and possible strategies. Language Documentation and Conservation 13: 545–563. University of Hawaii Press.

Shohamy, E. and Gorter, D. (2009) Linguistic landscape: expanding the scenery. New York: Routledge.

Thieberger, N. and Musgrave, S. (2007) Documentary linguistics and ethical issues. In P. K. Austin (ed.) Language documentation and description 4: 26–37. London: SOAS. http://www.elpublishing.org/PID/048.

Van Mensel, L., Vandenbroucke, M., and Blackwood, R. (2016) Linguistic Landscapes. In O. Garcia, N. Flores, and M. Spotti (eds) Oxford Handbook of Language and Society 423– 449. Oxford University Press.

Building a program for a capacity building center in Colombia: Documentation for community purposes

Katherine Bolaños, Ricardo Palacio Hernandez

In this talk we aim to report on the experience of the first training and support center for language documentation directed to members of indigenous communities in Colombia.

This training and support center, that we have denominated the 'Colombian Hub', was first set as a pilot project in cooperation of two institutions (a local NGO and an international institution), seeking to test a concept of a capacity building center for language support that goes beyond linguistics aims (here meaning that we aim to chase goals other than creating data for academic purposes, linguistic analysis, or formal education purposes) and support, instead, community led initiatives and goals with regard to language documentation and language related actions. As one of the activities of this initiative, we created a one-year-long training program directed to members of indigenous groups in Colombia. For this program we developed a curriculum to be implemented through three, one week-long, training meetings, followed by constant support to participants through virtual meetings and some short visits to their communities. The curriculum included, among others, the following goals: training on equipment and tools for documentation, grant writing, outreach, social appropriation of knowledge, grant management.

In this program participated, at first, members from 6 language groups (Nasa, Wounaan, Kamentsa, Inga, Cofan, and Cabiyarí), from which 4 finalized the one-year long program. In this talk we report on partial results from the first year of the program, emphasizing on the challenges of this initiative (including ethical approaches, funding, curriculum design, connectivity, outreach), as well as on what we regard as successful results of this type of experience (capacity building, leadership, empowerment, language promotion, relationships between different agencies, people and goals, curriculum and universality of application of methods, among others).

Botanical Naming Strategies in Woleaian: A Remote Methods Analysis

Emma Breslow

This study examines the strategies used to name plants in Woleaian, a Micronesian language of the Chuukic subfamily (Sohn 1975). It is spoken in the Federated States of Micronesia, most notably on Eauripik Atoll, where the present study's consultants are from, and Woleai. This work was conducted during the Covid-19 pandemic without visiting Micronesia. Instead, elicitation sessions were held both virtually over Zoom and in-person on Oʻahu, Hawaiʻi. The aim of this documentation was threefold: to act as a proof of concept for remote methodology documenting language about plants, to contribute to interdisciplinary literature on language documentation and ethnobotany, and to create a botanical field guide for the Woleaian community.

Although Eauripik is less than a tenth of a square mile, this project analyzes the ethnobotanical knowledge encoded in over 120 Woleaian botanical terms and how naming strategies pattern with the uses of these plants. Names and knowledge were elicited through a combination of methods including free listing plants of a given theme (Martin 1995), recording and transcribing Woleaian descriptions of the various landscapes on the island, and two sorting tasks based on Berlin (1992), one using a virtual whiteboard and pre-established categories and the other in-person without specifying categories beforehand. The knowledge documented includes Woleaian names, English and scientific names when identifiable, descriptions, uses, categorizations, and images of each plant.

Analysis compares naming strategies, sorting task results both with and without pre-established categories, and morphological breakdown of the botanical terms. While many results are in line with predictions from prior research, some differences may be due in part to the small size of Eauripik, and therefore the nature of how people interact with plants in that community. Ultimately, this work demonstrates that long-distance ethnobotanical linguistic fieldwork is a feasible alternative in situations where travel is not a good option.

References

Berlin, Brent. 1992. *Ethnobiological classification: principles of categorization of plants and animals in traditional societies*. Princeton, N.J: Princeton University Press

Martin, Gary J. 1995. *Ethnobotany*. Boston, MA: Springer US. (doi:10.1007/978-1-4615-2496-0) (http://link.springer.com/10.1007/978-1-4615-2496-0)

Sohn, Ho-Min. 1975. *Woleaian Reference Grammar*. The University Press of Hawaii. (http://hdl.handle.net/10125/62919)

How to document an emerging contact language? The strategies for documenting and archiving the language contact of Warao refugees in Brazil

Dalmo Buzato, Átila Vital

The study of contact languages has faced, since the beginning of scientific investigations in the area, the instability that most of them have. However, with the awakening of several migration flows around the globe, new contact languages have emerged. Among these situations, there is the migratory flow of Venezuelans to Brazil, with the presence of indigenous migrants from the Warao folk, an ethnic group that originally inhabited the Orinoco river delta, in northeastern Venezuela, but also in territories of Guyana and Suriname. Most migrants are Warao L1 speakers and Spanish L2 speakers of some degree of proficiency. Recent studies (Buzato & Vital, 2024, 2023; Buzato, 2023) have indicated the possibility of the emergence of a mixed language between Warao, Spanish and Brazilian Portuguese in this migratory flow. The objective of this presentation is to report the strategies used to date to describe this contact. The first of the activities reported is the creation of a dataset with signs written by refugees with requests for help aimed at the Brazilian population. The analysis of the signs demonstrated a high degree of language contact between the three languages, in addition to the impact of migrants' low literacy on the available data. The second activity to document the languages consisted of collecting spoken and written data available on digital social networks, mainly Facebook and YouTube. The choice of social networks as a data source was motivated by the large amount of potential data, in addition to open access. The third strategy has currently been implemented, and consists of recording spontaneous speech records of Warao refugees in communicative situations in Brazilian society. It is expected that the actions described above can constitute a dataset that describe and preserve this emerging language, and that contribute methodologically to the research field of documenting emerging contact languages.

BUZATO, Dalmo; VITAL, Átila. O contato linguístico em placas de refugiados venezuelanos em Belo Horizonte e região metropolitana: observações preliminares. In: Anais do Congresso Nacional Universidade, EAD e Software Livre. 2023.

BUZATO, Dalmo; VITAL, Átila. Creating datasets for emergent contact languages preservation. In: 3rd Workshop on Digital Humanities and Natural Language Processing Proceedings. Universidade de Santiago de Compostela, 2024.

BUZATO, Dalmo. Universal Dependencies and Language Contact Annotation: Experience from Warao refugees signs in Brazil. In: Proceedings of the 2nd Edition of the Universal Dependencies Brazilian Festival. 2023. p. 509-519.

Language documentation through oral history archives

Silvia Calamai, Ardolino Fabio, Rosalba Nodari

Karl Popper's assertion that disciplines do not exist, but problems do perfectly fit to the field of oral archives, a sphere in which a multitude of issues arise on various levels. It is noteworthy that employing audio or video recording for fieldwork is not confined to linguistics; rather, it is a widely adopted practice across numerous social sciences and humanities disciplines. In the domain of contemporary history, the subfield of oral history stands out as one of the biggest producers of oral data, gathered in oral archives. It has evolved over time into an independent research field, dealing with the systematic gathering of living people's testimony about their experiences. With its dedicated academic journal and associations, oral history has become a thriving domain within historical research (Perks & Thomson, 2015; Calamai, 2023).

In Italy, each discipline approaches oral archives differently, developing its own taxonomies, specific metadata sets, and diverse methods of philologic curation. In addition, analog-born oral archives demand diversified curation, particularly regarding the correlation between the original carrier and its contents (interviews, participant observations, etc.). From an archival perspective, paper materials and field diaries closely interrelate with audio documents, necessitating their inclusion in any digitization process (Calamai et al., 2022).

The contribution presented here intends to discuss the digitization and revitalization efforts undertaken for an oral archive collected in the 1980s by sociologist and anthropologist Vittorio Dini (1925-2018) in a marginalized region of Central Italy (Sestino and Val Marecchia, eastern Tuscany). This area, currently experiencing significant depopulation and economic decline (Dini, 1990), was previously overlooked by dialectologists. The re-analysis of such archival data promises to enhance our comprehension of sound shifts and linguistic repertoires within the broader context of local landscapes, shedding light on how speakers implicitly or explicitly perceive linguistic changes and their connections to social changes.

References

Calamai, S. 2023. "Sociophonetics and oral history". In C. Strelluf (Ed.), The Routledge handbook of

sociophonetics. London: Routledge, pp. 365-82.

Calamai, S., Piccardi, D., Pretto, N. Candeo, G., Stamuli M.F. & Monachini, M. 2022. "Not just paper: enhancement of archive cultural heritage". In D. Fišer & A. Witt (Eds.), CLARIN. The Infrastructure for Language Resources. Berlin: de Gruyter, pp. 647-65.

Dini, V. (Ed.). 1990. Luoghi e voci della memoria collettiva: per un archivio dei saperi e dei vissuti della cultura valtiberina toscana. Documentazione raccolta nei comuni di Sestino e Monterchi dall'anno 1978. [Eng. Places and Voices of Collective Memory: Towards an Archive of Knowledge and Experiences of Tuscan Valtiberina Culture. Documentation collected in the municipalities of Sestino and Monterchi since 1978]. Sestino: Istituto interregionale di studi e ricerche della civiltà appenninica.

Perks, R. & Thomson, A. (Eds.). 2015. The oral history reader. London: Routledge.

Mobilizing Tlingit Oral History Archives to Support Place Names Reclamation Initiatives

Emily Comeau, Tamis Cochrane, Christine Schreyer

Conserving and maintaining community archives can seem an insurmountable task, especially when this involves converting records that were made using obsolete or obsolescent technology into more modern formats. However, archives can contain vital knowledge that may otherwise be lost or forgotten, including information about history, governance, social norms and values, language, and knowledge of the land – such as place names. Audio recordings of Elders speaking their languages are especially important in communities where there are no longer fluent speakers.

Through the TRTFN Oral Histories Project, which has focused on transcribing and translating oral histories recorded from the last several decades, we have uncovered a number of traditional Tlingit place names that, while held in the community's collective memory, previously had not been documented. Traditional place names often provide detailed information about the relationship between the territory's original inhabitants, and the resources available in that area. Place names are also important because they often describe the history of ongoing land use, continuing stewardship of the land, and the cultural significance of specific locations. They can also describe migration events, contact and conflict between groups, and the physical features of the landscape – including landforms, flora and fauna, minerals, climate, and how the landscape has changed over time.

In this presentation, we will share findings from the TRTFN Oral Histories Project, particularly as they relate to place names and language, and we will discuss how this knowledge is being made accessible to community members through a variety of initiatives, including a digital archive and dictionary, a mobile place names app that can be used out on the land, and an updated web-based map, among other projects. This work illustrates how oral history archives and legacy collections can be vital resources for both language and cultural revitalization and place names reclamation.

Training, Archivists, and Language Collections

Sergio I. Coronado, Hugh J. Paterson Ⅲ,
Oksana L. Zavalina, Shobhana L. Chelliah

Language archives play a vital role stewarding language resources—the evidence of language diversity and linguistic cultural heritage. Training archivists to appropriately and effectively manage language collections and engage with their diverse audiences is important to the re-use and curation lifecycle of language resources. Existing training materials related to language archives are often oriented towards depositors, both language scholars (Kung et al., 2020; Miller, 2023) and language-community members (CORSAL, 2024). However, university curriculum specifically addressing language resources and their management is rarely seen in archival-science and library-science training programs. We report on the development of an open-access curriculum to fill this gap and how language collections have been integrated into several courses in the training of information professionals within American Library Association certified graduate degree programs and linguists.

Our work products include a four-module semester-long course and dual deployment of project-developed content in courses including: Corpus Linguistics, Field Methods, Advanced Metadata, Cultural Heritage Stewardship, etc. Some of the important themes addressed include: *Fiduciary responsibility*, *Language identification*, *Intra- and inter-resource relationships*, *Resource's of-ness*, *Interactive modality and materiality*, and *Evaluation of community language archives*.The selection of these areas of emphasis is informed by assessments of students' and practitioners' knowledge gaps (e.g., Aljalahmah & Zavalina, 2023a, 2023b; Zavalin & Zavalina, 2023; Zavalina & Burke, 2021).

As language research becomes more prolific and the need for language resource preservation is growing among both scholars and language communities, a variety of roles within library and archival praxis require training to support these diverse audiences. Our training gives exposure to information professionals who will:

- *Review prospective deposits including data management plans*,
- *Produce ephemeral online interactive experiences containing source materials and the derivative scholarly record*,
- *Conduct curation, collection maintenance and migration tasks*, as well as
- *Plan and evaluate digital infrastructure systems for the long-term stewardship of language-resource collections*.

References

Aljalahmah, S., & Zavalina, O. L. (2023a). Exploration of Metadata Practices in Digital Collections of Archives with Arabian Language Materials. In O. L. Zavalina & S. L. Chelliah (Eds.), *Proceedings of the International Workshop on Digital Language Archives: LangArc-2023* (pp. 11–14). University of North Texas. https://doi.org/10.12794/langarc2114295

Aljalahmah, S., & Zavalina, O. L. (2023b). Student-Created Dublin Core Metadata Representing Arabic Language eBooks: Comparison of Individual and Group Work

Outcomes. *Journal of Education for Library and Information Science*, e20230016. https://doi.org/10.3138/jelis-2023-0016

Computational Resource for South Asian Languages. (2024). *Computational Resource for South Asian Languages*. University of North Texas. Retrieved 25 March, 2024, from https://corsal.unt.edu/curriculum

Kung, S. S., Sullivant, R., Pojman, E., & Niwagaba, A. (2020). *Archiving for the Future: Simple Steps for Archiving Language Documentation Collections*. Teach Online with Teach:able. https://archivingforthefuture.teachable.com

Miller, J. C. (2023). *PARADISEC Workflows*. PARADISEC. https://paradisec-archive.github.io/PARADISEC_workflows

Zavalin, V., & Zavalina, O. L. (2023). Exploration of Accuracy, Completeness and Consistency in Metadata for Physical Objects in Museum Collections. In I. Sserwanga, A. Goulding, H. Moulaison-Sandy, J. T. Du, A. L. Soares, V. Hessami, & R. D. Frank (Eds.), *Information for a Better World: Normality, Virtuality, Physicality, Inclusivity* (pp. 83–90). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-28032-0_7

Zavalina, O. L., & Burke, M. (2021). Assessing Skill Building in Metadata Instruction: Quality Evaluation of Dublin Core Metadata Records Created by Graduate Students. *Journal of Education for Library and Information Science*, *62*(4), 423–442. https://doi.org/10.3138/jelis.62-4-2020-0083

A search engine and a generator of glosses for nheengatu

André Wesley Dantas de Amorim

In this talk I present two projects on which I collaborate that are devoted to the language nheengatu (Tupi-Guarani, yrl), also known as yẽgatu, *ñeengatú, nhengatu, etc.,* spoken today by about 7000 people in Brazil, Colombia and Venezuela. The on-going projects are two applications that are open source, free, and easily accessible through a desktop or smartphone browser. The first app functions as a search engine, while the second one segments words and generates their morphological glosses. They were developed from the data of the project *Romania Amerindia* (Reich & Lessa, 2023), which comprise experiments in the form of language games, collected in 2019 in Brazil. The dataset has 1313 unique words (or entries) and 1441 unique sentences. The search engine takes a given word and returns the searched word, a sentence with the word, and the sentence's segmentation, glosses, translation and audio. The generator of glosses segments words and generates glosses on a given tier of a TextGrid. Both of them were developed in R, but a new version of the search engine has been developed using Python. In relation to the limitations of the apps, the search engine is currently facing a few problems with the alignment, due to the absence of some words in the segmentation and glossing of the original data. Regarding the generator of glosses, the app can currently be summarized as a replacement of an already annotated word to its gloss or segmentation in an already annotated database in a one-to-one relationship, which means that it is lacking a parser. Therefore, any kind of contribution to both projects is encouraged by us and very welcome.

Keywords: nheengatu, search engine, glosses, apps.

References

Reich, Uli & Antônio Lessa. 2023. Corpora of American languages: Interactive language games from multilingual Latin America (Nheengatú). Berlin: Freie Universität.

Proximity and definiteness in complex demonstratives of Otomi

Gabriela De La Cruz

This presentation derives from a larger project that consisted on the documentation of local geography and life memories in Otomi (ISO 639-3 ott), a native Mexican language. The anecdotes took place in the early 70's and were collected during the Winter 2022 in the municipality of Temoaya, Mexico. The methods used to collect the data were (1) interviews, to get to know more about the speakers' everyday life. (2) Hiking (Cruz, E., 2017) in places that are important for the community members such as mountains, spring water, wells, and corn plantations. (3) Description of different areas through pictures and recordings of non-easy access locations. This was a very effective technique to collect data from old or sick speakers who cannot move easily but whose memories are vivid.

The anecdotes contain elders' wisdom related to geography, such as Otomi toponyms along with the stories behind the indigenous local names. The narratives were recorded from 14 adult Temoaya Otomi speakers, then transcribed, and analyzed in Otomi. Finally, they were translated to English and Spanish resulting in a digital collection (See Image 1), accessible to anyone interested in the language and culture.

Image 1. Digital collection of Living memories

The material collected contains meaningful data that allows a detailed analysis of the demonstratives and their role in the definiteness system of the language. So, I will discuss in detail the definite demonstratives shown in table 1.

Table 1 Definite demonstratives

	+PROX	+DIST
	nu̲	**ka**
SG	To xiphi nu̲ Xua to ʃiphi nɤ ʃua 1SG.PST say DEM.SG.PROX John *I told John*	To käjti ka Xua mande to-kãhti ka ʃua mande 1SG.PST-see DEM.SG.DIST John ADV *I saw John yesterday*
PL	**yu̲**	**ku̲**
	Bi tsi yu̲ bahtsi ra mihkwa bi t͡si yɤ baht͡si ɾa-mih-kwa 1SG.PST say DEM.PL.PROX children 3PL.PRS-sit-LOC *The children that are sitting here ate*	Bi maphi ku̲ wene bi-maphi kɤ wene 1SG.PST-scream DEM.PL.DIST baby *The babies screamed*

Then, I will introduce the complex demonstratives that result from definite demonstratives (table 1) and occur in examples like (a) and (b), along with some of the features of the nouns that can take these kinds of demonstratives.

(a)

nu̲-ku̲ khani bi-ntiji bi-ma
DEM.SG-PL. people 3PL.PST-hurry up 3PL.PST-go
and the people hurry up to leave

(b)

nu̲-guegue bi-doi
DEM.SG-3SG 3SG.PST-buy
'he bought it'

This analysis contributes to our understanding of the demonstratives and the morphosemantic features of nouns in Otomi. The goal of this research is to encourage the study of language through documentation of community knowledge, along with the creation of written material that contains elders' wisdom, to promote identity and visibility of the speakers and to invite metis to find their way back to their roots.

REFERENCES

Cruz, E. (2017) *Documenting Landscape Knowledge in Eastern Chatino:Narratives of Fieldwork in San Juan Quiahije.* Anthropological Linguistics, Volume 59, Number 2, Summer 2017, pp. 205-231. University of Nebraska Press.

Developing Keyboards for Digitally Marginalized Languages

Tabea De Wille

This talk presents on the keyboards team at Translation Commons, where we design and develop keyboards for digitally marginalized languages, as well as associated processes and training material. Our goal is to improve access to information and communication channels for the thousands of languages currently left behind in an increasingly digital world. Challenges around supporting languages with large character sets while maintaining ease of learning and use, languages that are not yet represented in Unicode, developing word lists for predictive text and communication with non-linguist community representatives are among the topics that make this work particularly interesting and rewarding.

Schools, Communities, and the Essential Role of Documentation Materials

Connie Dickinson, Alfonso Aguavil, Milton Calazacón, Francisco Aguavil, Milton Callera, Oswando Nenquimo

Documentation has both tangible and intangible benefits for the creation of materials for bi-lingual schools and communities. The tangible benefits include a large body of data as well as the creation of a practical orthography based on hours of transcription. Intangible benefits include a deeper understanding of the language and culture not only for the academic researchers but also for the indigenous researchers. Documentation projects produce competent and confident teams.

Working with documentation materials, we have created a variety of materials including "talking" books, posters, games, spell-check programs, comprehensive dictionaries, and a variety of different QR code cards for Tsachi (Barbacoan), Waorani (Isolate), and Shuar and Achuar (Chicham) bi-lingual schools and communities. The QR codes are used for the "talking" books, various types of flashcards as well as providing a convenient means for presenting important recordings to the community. We use two different kinds of QR codes. One can be used with any phone that has an internet connection. The other is used for specially formatted phones that do not need the internet.

The databases produced by the Tsafiki, Waorani and Shuar/Achuar documentation projects are essential to our work. They allow us to create materials with a breadth and depth that is usually not possible for a previously unwritten language. And some of the materials the communities find most useful would not be possible at all, such as the Tsafiki Word spell-check program which is based on 84,837 written words.

In this presentation we will discuss both the social and technical problems we have encountered in adapting documentation materials for the schools and speakers—everything from addressing different community concerns and needs and solidifying a practical orthography to producing materials that can survive a tropical climate.

“Registering is important, but we also need to reinforce the language use”: community-based documentation for a revitalization program of Djeoromitxi

Andre Djeoromitxi, Vandete Djeoromitxi, Ivan Rocha da Silva, Juliana Solano

This presentation aims to discuss the ongoing documentation of Djeoromitxi, a Brazilian endangered language spoken in Amazonia by the Djeoromitxi people (Macro-Je grouping), with about 35 adults and elderly speakers, funded by the Endangered Languages Documentation Programme. The project focuses on building a lasting, and multipurpose collection hosted by ELAR Archive/ELDP, and ALIM/MPEG. Up to now, the documentation outcomes comprise seventy hours of audio and video, seven hours of transcription and translation with ELAN, one and a half hours of annotation, a partial preliminary grammar, lexicon with eight-hundreds entries, and a preliminary multimedia dictionary. The material recorded was chosen by the community, that is, a range of different types of discourse, since traditional narratives to the daily life of the community as well as word lists, and grammar topics. We have also elaborated pedagogical materials for teaching the languages as L2 for no-speakers. The project was thought to fit into the community's needs of implementing a community-driven program of reinforcement or vitalization of language use. Besides creating material for supporting language re-transmission, the linguist and community are working together to develop a program to raise funds to allow the vitalization activities. The community is setting up a space, a traditional house (maloca), where they can meet regularly to practice the language through its associated knowledge. Paralely, they plan to establish a local in the village as a center of memory (or reference) of the people, showing aspects of their language and culture. The idea is to allow the community to access the Djeoromitxi collection, which originated from this collaborative project, such as video, audio, pictures, texts, etcetera. The community understands the urgency and importance of registering the maximum possible of their language, and its associated knowledge, but they have also claimed that the initiation of the reinforcement of the language use is urgent as well since their language is severely threatened.

Mobilizing archival collections: The Open Text Collections project

Christian Doehler, Sebastian Nordhoff, Mandana Seyfeddinipur

keywords: text collections, interlinear glossed text, open access, FAIR principles

Franz Boas established the “Boasian Trilogy” in language documentation and description (Himmelmann 1998), consisting of a grammatical description, a dictionary, and a text collection. All three are necessary to get a comprehensive overview of a language, a language portrait, or snapshot in time. More im- portantly, the three components complement each other. While we have good publishing outlets for grammars (e.g. Comprehensive Grammar Library) and dictionaries (e.g. Dictionaria), such is not the case for text collections. This means that only few of them are published, and even fewer follow the FAIR principles of findability, accessibility, interoperability, and reusability (Wilkin- son 2016).

The project Open Text Collections (http://opentextcollections.org) will remedy this by making high quality text collections from endangered lan- guages available in an open interoperable format. Next to providing pdfs or printed books to the communities themselves, this setup will also provide the data in Cross-Linguistic Data Format (Forkel et al. 2018) for downstream use in NLP applications.

Most archive collections are the result of a language documentation project, often part of the collectors’ dissertation projects. While countless hours are in- vested into the structuring and glossing of texts, in many cases, however, these texts are not made available in a reusable way. Linguists tend to have them somewhere on their hard drive, or uploaded to an archive but there is no gen- erally established way of publishing them, at least not in a format which would feed further research downstream (e.g. linguistic typology, corpus-based lan- guage description, or NLP). This means that these valuable results of language documentation often fail to be discovered.

Open Text Collections will provide a quality venue for publishing text collec- tions, following the setup established by Language Science Press. The platform is community-driven and aims at being attractive to both data producers (i.e. language documenters) as well as data users (i.e. language communities, ty- pologists, NLP practitioners). For data producers, the platform will provide rigorous peer review, quality control, and top-notch publishing (pdf and print- on-demand), making sure that the time invested in a text collection will not harm job prospects. For data consumers, different outlets will be available to suit different needs: printed books will be available for communities; a search interface (prototype available at https://imtvault.org) will be available for typologists, and all data will be available as CLDF dump for NLP practitioners. By making reuse easy, the research will spread more widely, which in turn is very attractive for the data producers.

As of today, there are 5 regional boards and 40+ proposed text collections. This presentation will showcase the platform, its motivations, and its benefits for data producers and consumers.

References

Forkel, Robert, Johann-Mattis List, Simon Greenhill, Christoph Rzymski, Se- bastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping & Russell D. Gray. 2018. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. Sci Data 5. DOI: 10.1038/sdata.2018.205.

Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. Lin- guistics 36. 161–195.

Wilkinson, M. et al. 2016. The FAIR guiding principles for scientific data man- agement and stewardship. Sci Data 3. 160018. DOI: 10.1038/sdata.2016. 18.

The Digital Archive of Indigenous Languages of the Amazon

**Juan Alvaro Echeverri, Alejandro Prieto,
Abel Antonio Santos, Elio Miraña**

The Digital Archive of Indigenous Languages of the Amazon (Archivo Digital de Lenguas Indígenas de la Amazonia —ARDILIA) is not just a storage space for unpublished audiovisual materials in indigenous languages of the Amazon region. Its purpose is to support local processes of living memory and to make these materials available to the communities they originate from, as well as to other peoples and the wider world. The archive rejects the notion of reducing its contents to mere "digital objects" and instead emphasizes the immanent power of the words contained in it. Archival curation, therefore, goes beyond documenting and preserving objects; it also involves a spiritual healing process. Unlike the largest digital archives of native languages in Europe and the United States, which focus on safeguarding endangered languages from a formal- textualist perspective, ARDILIA takes a different approach. It is a language archive created in collaboration with the communities of the Upper Amazon, located in Leticia, Colombia. The archive gives these communities complete control over the materials and defines its own policies of use and access, serving as a technological instrument for the living memory of the peoples.

Archive URL: https://repositorio.unal.edu.co/handle/unal/82579

Webpage: https://ardiliaorg.wordpress.com

ELAR regional portals: Step by step towards global accessibility

Jonas Engelmann, Vera Ferreira, Hanna Hedeland, Nils Hempel, Mandana Seyfeddinipur

The Endangered Languages Archive (ELAR) is a global digital archive that safeguards intangible cultural heritage from all over the world. The archive's holdings comprise audio-visual language (documentation) collections in over 600 languages, with a majority of the languages being highly endangered. Due to the global focus of the archive, its designated communities – researchers, language communities, the general public – are equally distributed across the globe and partly only literate in local languages. With the current interface only available in English, ELAR is definitely not as accessible as it should be. Since the number of languages that could be considered relevant for a global archive make the localization of the entire archive including the collections' metadata insurmountable, ELAR has opted to proceed step-by-step, creating different regional portals where collections from a particular geographical area are presented in a widely used language of the region (Ferreira et al 2023). In our talk, we will present ELAR's first regional portal, the Latin American portal.

With the implementation of the portal, we are developing an additional discovery layer based on the data and metadata of collections from the Latin American region hosted in the ELAR digital repository. For the portal, comprehensive curation work on the existing metadata was required, and some adaptations to the metadata schema became necessary to model the collections' features properly and to allow for multilingual metadata. In our presentation, we will provide insights into the actual portal development, which include both the transformation of the IMDI metadata into RDF and its enrichment according to linked open data standards and also the development of the new flexible web interface that relies on a SPARQL database to present the materials for browsing and searching. Broader questions related to localization, usability and accessibility of language archives will also be addressed during the presentation.

Wikipedia en wayuunaiki, una lengua indígena en documentación
Wikipedia en wayuunaiki y las nuevas epistemologías

Leonardi Fernández

El periodo colonial fue un episodio ‘’bárbaro’’ para muchas de las comunidades indígenas de nuestra región, ya que este periodo se caracterizó por destruir e invisibilizar la memoria cultural de las comunidades étnicas que habitan el territorio americano. Dicha estrategia respondía (o responde) al plan que buscaba desmoralizar a las comunidades y lograr la homogeneización de la población.

Sin embargo, mucha de la información indígena prevaleció y se afianzó a través de procesos y fenómenos socioculturales que respondieron a las necesidades internas de los grupos, a partir del control de sus elementos culturales y del proceso de negociación de la identidad. Según Levi Strauss (1962), el conocimiento indígena forma parte del acervo cultural del grupo, pero también de cada individuo, dicho conocimiento subyace en la estructura del pensamiento y nace a partir de un proceso de ‘’abstracción’’, en el cual el grupo indígena se relaciona con su mundo a partir del pensamiento.

En el caso particular de la comunidad wayuu, este no es ajeno al mismo proceso que emprendieron otros grupos indígenas, en todo caso, el pueblo wayuu, acumulo un gran número de elementos que le dan lógica a su mundo cultural y que, volviendo al caso de la Wikipedia, esta herramienta le permitirá sistematizar ese conocimiento a partir de un proceso de negociación y validación que la misma población debe experimentar.

En ese caso, Wikimedistas Wayuu está pasando a convertirse en un referente para retratar y ejemplificar el proceso antes descrito, por tal motivo, varias instituciones académicas ligadas a la antropología, la tecnología, la educación, el activismo y los grupos wikimedias han estado interesados en conocer y saber de las ideas que impulsaron a este grupo de personas.

Proyectos locales y empoderamiento

Wikimedistas Wayuu está impulsando el desarrollo de nuevos liderazgos bajo la premisa de la educación y bajo un enfoque local, para ello, se llevaron a cabo talleres de capacitación en los tres primeros meses del segundo semestre de 2023, en dichos talleres participaron 20 líderes y lideresas, quien se comprometieron en replicar y extender la red de wikimedistas para la Guajira colombo-venezolana.

A partir de un taller de prototipado fue posible evidenciar el talento y la creatividad de los docentes participantes del proyecto, quienes se juntaron para crear iniciativas locales con miras a su aplicación en sus propios espacios. De tal experiencia obtuvimos los títulos de 4 propuestas:

- Wanee ekirajawaa jeketü (Una nueva enseñanza). Grupo de la sede Piyushipana.
- Fases de la lengua wayuunaiki. Grupo de docentes y estudiantes de la Universidad Pedagógica Experimental Libertador sede Paraguaipoa.

- Lectura y escritura creativa en wayuunaiki. Grupo de docentes de la Escuela Fe y Alegría Paraguaipoa.
- Sukujalaa Wayuu (Lo que cuentan los Wayuu). Grupo de docentes de la Escuela Francisco Babbini de Guarero.

Un trabajo colectivo

Vale decir que la Wikipedia en wayuunaiki está siendo construida por la comunidad, en ese caso, el grupo de activistas cohesionados en Wikimedistas Wayuu está liderando este proceso a través de talleres locales, editatones, charlas y conferencias en los que se habla principalmente de la importancia de sistematizar el conocimiento indígena, habitar el espacio digital, el control de datos y la gobernanza sobre la información de la comunidad.

Igualmente, la wikipedia en wayuunaiki busca convertirse en una herramienta para la enseñanza de la escritura del idioma, un referente para la estandarización de la escritura y un espacio de consulta para las nuevas generaciones wayuu escolarizadas.

Enlaces de interés

- Página en MetaWiki:

  https://meta.wikimedia.org/wiki/WikiMedistas_Way%C3%BAu/en

- Wikipedia en wayuunaiki: guc.wikipedia.org
- Artículo en wikipedia español:

  https://es.wikipedia.org/wiki/Wikipedia_en_way%C3%BA

- Reseña de Global Voices:

  https://rising.globalvoices.org/lenguas/2024/02/18/wikipedia-en-wayuu/

Community Creativity with New Corpus Creation Tools

Darren Flavelle, Jordan Lachler

The field of Language Documentation has innovated quickly in recent years to facilitate more work being done remotely. There has also been a focus in the last decade for this work to be led by the language communities themselves. For the past two years, we have used newly developed tools made with the express purpose of enabling language communities to create a corpus with less input from outsider linguists. These tools are based on traditional documentary activities, including translation, picture description, conversation prompts, and free-form narratives.

This Presentation will focus on the surprises and challenges we have encountered in training two communities – one in North America, one in Melanesia – to use the tools the way that we intended.

Despite the user-friendly design of the tools, low levels of technological literacy among certain community members created an initial barrier to their participation in the project. This was mitigated through a two-tiered training model which relied on more tech savvy community members to provide hands-on training in the local language.

Once they began using the tools, members of the different communities interpreted the activities from their own perspectives – sometimes prefering to carry out tasks collaboratively, striving for uniformity in their responses, and at other times using the activities to express their linguistic creativity and individuality in ways we could not have anticipated. These new emic perspectives on how the tools should best be used enriched the documentation; though that did lead to a loss of uniformity in the data collection phase and additional unforeseen work for the linguists and communities on the back end.

We aim to showcase these new tools and discuss how we are improving them with the input from the communities in order to prepare them for wider distribution.

Treasures from the archive: The new life of traditional language-learning material in modern technology

Vivien Fröhlich

In the 1990s and early 2000s, speakers of several Alaska Native languages created language courses in collaboration with the Yukon Native Language Centre. These can be found as digital files on the Alaska Native Language Archive's website today.

A current collaboration between Doyon Foundation, an Alaska Native organization whose mission is to provide educational, career and cultural opportunities to enhance the identity and quality of life for Doyon shareholders, and 7000 Languages, a non-profit organization that supports Indigenous and endangered languages communities to create language-learning materials, uses this data to create online language-learning materials that can be used in the classroom or as standalone online courses. The written material and recordings of the past are used to build these courses on the Transparent Language Online platform that are then tailored to the needs of the learners in the Doyon region using the most appropriate of the around 40 different language-learning activities the software offers, including: video players, conversation-based, matching, and slideshow activities. Using this approach, no new content had to be created, instead the voices from the past, of whom some have already passed away, can still be used to teach the ancestral languages.

In 2023 we successfully created five language-learning courses for Alaska Native languages, namely for Upper Tanana, Tanacross, Benhti Kenaga', Denaakk'e, and Inupiaq, which can now be used by teachers in the classroom, as standalone courses for asynchronous language learners, or as a resource for non-Native teachers who teach Native languages giving them the possibility to reuse the audio files in lessons which are tailored to their students' needs without having to either record themselves or find a Native speaker to record – which is, in fact, not possible anymore for some of these languages as there are no first language speakers left.

Developing a community-led documentation project for Bugkalot/Ilongot

Maria Kristina Gallego, Frederick Barcelo

Bugkalot (also known as Ilongot) is a language used by around 5,710 speakers, mainly in the provinces of Nueva Vizcaya, Quirino, Aurora, and Nueva Ecija, Philippines. It is rated as an endangered language by Ethnologue, which means that the younger generations are not learning or using the language (Eberhard, Simons, and Fennig 2024).

This paper reflects on the learnings and insights coming out of a community-led documentation project for Bugkalot. The collaboration between the community and the Department of Linguistics, University of the Philippines Diliman (UP Lingg) began in the 1960s, when UP Lingg conducted field research in a Bugkalot community in Nueva Vizcaya, which focused on gathering basic linguistic data and various oral traditions. In 2009, an undergraduate field methods class of UP Lingg visited the same province to conduct another (unrelated) fieldwork, which produced a sociolinguistic survey and a grammar sketch of the language.

In 2023, Frederick Barcelo, a Bugkalot culture bearer who worked with UP Lingg during the mentioned 2009 fieldwork, reached out to the Department to propose a language documentation project for Bugkalot. The current project is thus a product of a decades-long relationship with the community. While the previous partnerships saw linguists taking the primary role, this current project places the community at the helm of the documentation project. Other significant aspects of the project include the repatriation of legacy materials collected since the 1960s, and capacity-building activities focused on language documentation, orthography development, lexicography, and linguistic analysis.

This paper highlights the importance of maintaining a good and lasting relationship with the community to ensure the sustainability of language documentation. We also emphasize the need for employing best practices in language documentation, such as maintaining metadata and ensuring proper data storage through archiving, in order to be more accountable in our engagements with the community.

Reference:

Eberhard, D., Gary, S. and Fennig, C. (eds). 2024. Ethnologue: Languages of the World. Twenty-seventh edition. SIL International. Online version: http://www.ethnologue.com.

Training native speakers in a community with low degree of contact

David Ginebra

Training speakers of indigenous languages in language documentation skills is one of the first steps for implementing a collaborative documentation project (Benedicto et al. 2007; Czaykowska-Higgins 2009; Yamada 2010). However, the way these trainings are designed and implemented depend, to a great extent, on the degree of contact that the community had had with the national society. In this presentation, I will show how I trained a group of Yamalero speakers from Únuma (Colombia), a community with still quite limited contact with the national society. One year after these trainings, they have already been able to transcribe and translate 8 hours of time-aligned naturalistic speech.

The trainings were divided into two periods (March 2023 and June 2023). In the first one, community members (i) acquired minimal computer literacy skills, and (ii) learnt the existing community orthography, in order to transcribe and translate on ELAN previously segmented recordings. Five community members took part in it, but only one kept working once I left the field. In the second training, they learnt (i) how to video-record a session, (ii) how to collect metadata, (iii) how to process these files, and (iv) how to segment a recording, so that they could become independent documenters. Six community members took part in this training, and three of them have been actively documenting their language since then.

In this presentation, I will discuss the content and organization of these trainings and I will reflect upon some of the problems encountered and the solutions adopted. Additionally, I will discuss the benefits of a hands-on approach, following similar experiences (Harvey & Griscom 2020; Bolaños, Palacios & Seyfeddinipur 2022). Finally, the methods and tools adopted might also be useful to other researchers seeking to implement collaborative work with members of communities with low degree of contact with the national society.

References

Benedicto, Elena, Demetrio Antolín, Modesta Dolores, M.Cristina Feliciano, Gloria Fendly, Tomasa Gómez, Baudilio Miguel, Elizabeth Salomón. 2007. A Model of Participatory Action Research: The Mayangna Linguists Team of Nicaragua. In Proceedings of the XI FEL Conference on 'Working Together for Endangered Languages - Research Challenges and Social Impacts,' pp. 29-35. Kuala Lumpur, Malaysia: SKET, University of Malaya and Foundation for Endangered Languages.

Bolaños, Katherine, Ricardo Palacio and Mandana Seyfeddinipur. 2022. A hub for documentation of languages in Colombia: model, concept building, and tools. Paper presented at the Language Documentation and Archiving Conference. Berlin/online: ELAR/Paradisec

Czaykowska-Higgins, Ewa. 2009. “Research Models, Community Engagement, and Linguistic Fieldwork: Reflections on Working Within Canadian Indigenous communities.” Language Documentation & Conservation 3(1): 15–50. http://hdl.handle.net/10125/4423

Harvey, Andrew, and Richard Griscom. 2020. Haydom language documentation training workshop - January 2020: a report. DOI: 10.5281/zenodo.3971733

Yamada, Racquel-María. 2010. “Speech Community-Based Documentation, Description, and Revitalization: Kari’nja in Konomerume.” PhD diss., University of Oregon, Eugene.

As Told by the Local Voices: Archiving a Corpus of Attitudes on Hawaiian Language Policy

Armando Molina Gómez

The present study reports on the project of cataloging a legacy text-based corpus on the subject of Hawaiian language policy and related topics, including the process of description, tagging, scanning and archiving in the Kaipuleohone University of Hawai‘i Digital Language Archive. The corpus was collected by the late Dr. Albert Schütz over several decades, and comprises a total of 306 entries, predominantly in English, consisting of local press clippings in its majority, alongside other documents such as personal communications. This collection, mostly covering the period between the 1980s and the 2010s, provides insight into the changing language policy on the Hawaiian language over the decades following its recognition as co-official language in 1978. It offers a window into local attitudes and ideologies towards said changes in multiple aspects of public life, including the development of immersion schools, orthographic practices, and public signage, among others. The report will additionally feature a preliminary analysis of a subset of the corpus from 1995-1996, focusing on the inclusion of letters in Hawaiian in Ka Leo O Hawai‘i –the UH Mānoa campus newspaper– and the debate surrounding their translation. By shedding light on the public sentiment regarding language revitalization and normalization in Hawai‘i, this project contributes to the larger conversation on its future, looking ahead into the International Decade of Indigenous Languages.

Indigenous sign language documentation in Africa in the mirror of volunteer work at S-DELI

Judit Hollos

In the past couple of years, numerous milestones have been reached in the field of documenting and revitalizing indigenous languages of the deaf community throughout Africa.

In 2022, SIL LEAD, a non-profit organization released a collection of twenty sign language books for deaf and hard of hearing children in Mali – the first ever digital sign language storybooks in Sign Language for Malian Schools (MsSL). These are now available on Bloom Library where they can be read, shared and adapted. A similar success story is the publication of first-ever children storybooks in Somali Sign Language, thanks to a group of innovators funded through the All Children Reading: Grand Challenge for Development (ACR GCD) Begin with Books prize. AdaSL is an indigenous village sign language used in Adamorobe community in the Eastern Region of Ghana that is believed to have existed as far back as 1733 as a language used by both hearing and deaf people in Adamorobe.
In July 2023, the South African President has officially signed the law recognising South African Sign Language as the 12th official national language.

Through its Indigenous Sign Language Project, the Igbo Literacy Project and Beauty Beyond Speech project, Indigenous Hands and Voices and S-DELI (Save the Deaf and Endangered Languages Initiative) are also continuously working on documenting and preserving minority and endangered indigenous signed and spoken languages, by which they are promoting linguistic diversity and highlighting discrimination people with speech and hearing impairment face on a daily basis.

During the International Week of Deaf People (19-25th September), IHAV used the opportunity to protect and support the lingusitic and cultural identity of sign language users in Africa. S-DELI also organized courses designed for African sign linguists, interpreters and scholars. The Igbo Literacy Project in which I had the opportunity to participate, saw the creation of the Omenka app for the Igbo people in March 2024.

In my presentation, I aim to examine the various ways by which S-DELI strives to promote, preserve and document local indigenous African sign languages and their use in deaf education, through the lens of the activities of its unique volunteer programme in which I have been taking part since June 2023. With special emphasis on the Igbo Literacy Project and the Sign Language Documentation Project, but also drawing parallels with similar programmes in Africa, I wish to explore the possible solutions S-DELI may offer, in an effort to promote early literacy, access to sign language and ensure the survival of these languages and communities.

Key words: indigenous sign languages, Africa, endangerment, survival of languages, preservation, documentation.

What are knowledge graphs and how can they support under-resourced languages?

Elwin Huaman

Knowledge graphs are large semantic nets that integrate and represent knowledge from var- ious domains [1], and have proven to be very important for supporting the development of applications in high-resourced languages. For instance, major technology companies such as Amazon[1], Google[2], and Microsoft[3] and others have been collecting all the data and knowledge from the web with the aim of building knowledge graphs that can support their applications and services [2] in English, Spanish, and a few other languages. However, there are about 7000 languages in the world[4] and only a few of them have the necessary resources to be further implemented on language technologies [3]. Therefore, the creation of interoperable linguistic resources is nowadays more urgent in order to save and help under-resourced languages, and their communities. In this paper, we present our approach for building knowledge graphs that can support under-resourced languages. We adopted a task-based approach to knowledge graph generation, covering the creation, hosting, curation, and deployment phases. We conclude that the methodology and tools adopted, on building an under-resourced language knowledge graph, could enhance the presence of under-resourced languages on the web through the representa- tion of data, text, and media in these languages. Finally, we provide insights, directions, and best practices toward a better architecture for building under-resourced language knowledge graphs.

[1] https://www.aboutamazon.com/news/innovation-at-amazon/making-search-easier, accessed 15 Apr 2024

[2] https://blog.google/products/search/introducing-knowledge-graph-things-not, accessed 15 Apr 2024

[3] https://learn.microsoft.com/en-gb/graph/overview, accessed 15 Apr 2024

[4] https://www.ethnologue.com/

The Cahuilla Pedagogical Grammar Project: An Indigenous Language Research Model (ILRM) Approach to Writing Pedagogical Grammars for Indigenous Communities

Ray Huaute
University of California, Riverside

The Community-Based Language Research Model is defined as "Research that is ***on*** a language, and that is conducted ***for***, ***with***, and ***by*** the language-speaking community within which the research takes place and which it affects" (Czaykowska-Higgins 2009) but where "***by***" usually refers to Indigenous community members who are working "in collaboration" with a non-Indigenous researcher rather than leading the research themselves. In this talk, I present an alternative framework, the Indigenous Language Research Model (ILRM), where research that is ***on*** a language is conducted ***for***, ***with***, and ***by*** the language-speaking community (but where "***by***" can include linguistic research led *by* and Indigenous community researcher), and in a manner that is ***relationally accountable*** to the language-speaking community. This model centers linguistic research around community-specific language reclamation goals, while also creating multiple opportunities for engagement, allowing researchers to identify what type(s) of linguistic research might best support language revitalization efforts in a given community. Specific examples of this approach will be illustrated from my postdoctoral research project, "The Cahuilla Pedagogical Grammar Project".

Working with the last speaker of a language (fieldwork in the situation of language shift almost completed)

Olga Kazakevich

Itelmen is spoken in the South-West of Kamchatka. Some researchers regard it as a part of the Chukot-Kamchatka language family. It is represented by two dialects – Northern and Southern. In the 2010s there were several competent speakers of both dialects, but during the covid epidemic and right after it most of them left this world.

In the paper I present my experience of the fieldwork with the last speaker of the Northern dialect of Itelmen (2022, 2023). This is a lady of 85, Liudmila Egorovna Pravdoshina. Almost all her life she worked as a milkmaid on a collective farm. She is not only most competent in the language, she is also willing to share her knowledge with all those who are interested in it. It so happened that Liudmila Egorovna appeared to be the last competent speaker not only of her native dialect, but also of the Itelmen language in the whole. There are only several “new speakers” of the Southern dialect now.

Liudmila Egorovna is a great language consultant. She is ready to tell life stories and fairy-tales and then to help to transcribe and translate the recorded texts, to pronounce lexemes for a sounding dictionary, to elicitate grammar questionnaires. Most valuable are her life stories: they represent the history of the Itelmen community of Sedanka during all the XXth century. By now we audio and video recorded 15 hours of Itelmen texts, 3 hours are transcribed and translated. This is what might be called the work in progress, next field is fixed for June 2024.

Actually, Itelmen might be called properly documented, but it concerns only its Southern dialect (eg. Stebnitskiy 1934; Volodin 1976; Georg, Volodin 1999; Krasheninnikov 1775; Worth 1961; Bobalik 1996). As for the Northern dialect, there exist only a brief grammar sketch of Tatiana Moll (1960) and two publications by Chikako Ono (2003; 2020).

References

Bobalik, J. 1996. Мəзи’н Амӈə’л. Fairy-tales, legends and stories of the Kamchatka Itelmens.

Georg R.-S. Volodin A. P. 1999. Die itelmenische Sprache. Wiesbaden: Harrassowitz.

Krasheninnikov S.P. 1775. Opisaniye zemli Kamchatki. SPb, Vol. 1, 2.

Moll T.A. 1960. A sketch of phonetics and morphology of the Sedanka dialect of Itelmen // Proceedings of the Leningrad Teacher Training A.I. Hertsen Institute. 1960. Vol. 167.

Ono Chikko. 2003. A Lexicon of Words and Conversation Phrases for the Itelmen Northern Dialect. Kyoto.

Ono Chikko. 2020. Эсхӆэны’н эмӈэ’л и лано’н. Itelmen fairy-tales and stories from Sedanka Osedlaya. Sapporo,.

Stebnitskiy S.N. 1934. The Itelmen (Kamchadal) Language // Languages and wrighting systems of the people of the North. Leningrad, 1934. Part. III.

Volodin A.P. 1976. Itelmen. Leningrad.

Worth, Dean S. 1961. Kamchadal Texts collected by W. Jochelson. ‘s Gravenhage: Mouton.

AILLA 3.0: Intuitive, Equitable & Engaging

Susan Kung

The Archive of the Indigenous Languages of Latin America (AILLA) is about to launch the third version of its repository software, featuring a completely redesigned technology stack and graphical user interface (GUI) that together provide a more intuitive and equitable user experience. Major goals for the rebuild included visual and functional improvements to the user experience and the addition of a Portuguese GUI.

Users can find their way into the collections by using pre-coded filters for languages, countries, persons, and organizations; and facets and filters have been added to the general SOLR search to narrow down results. Files within a single dataset are grouped together in a media carousel, eliminating the need to click on each filename to open its player. The opaque access level numbers have been replaced with easily understood labels (public, login, embargo, restricted); and a built-in form facilitates communication between depositors and people wishing to access their collections. Administrative tools allow depositors to assign specific registered users to be viewers or editors of objects as small as single a media file or as large as an entire collection.

The most equitable component of this rebuild is the addition of a Portuguese GUI. Though AILLA has had English and Spanish interfaces since its launch in 2001, lack of a Portuguese interface has made it inaccessible to an entire country in its target demographic. Portuguese- specific fields were added to every level of the metadata (collections, folders, files, and taxonomies), and existing metadata was translated into Portuguese to fill these fields. This monumental translation effort was handled via both AI and professional translation.

In this presentation, I demonstrate these improvements and discuss how they make AILLA more functional and engaging for its Latin American user base, including Indigenous Peoples whose languages are represented in the repository.

Language and resource agnostic applications

Cat Kutay

When developing online and mobile support for language learning there is a wide range of tools to support this process. Combining these into pipelines for language resource collection (eg Nyingarn.net) are crucial.

This talk is about a web system that evolved out of work with Australian Aboriginal languages in various stages of reclamation or use and explains the tools that have been used, noting that the requirements were so diverse that a pipeline has not been developed form resource to learner, but some success have been achieved in developing resources for teachers.

The languages that are covered are Wiradjuri, Dharug, Yugambeh-Bundjalung, and Kariyarra. The material that we started with ranged from no recordings (Dharug) to transcriptions and sample exercises from teachers. It also included a dialect dictionary extraction using colour codes only.

The tools that are combined on https://dalang.com.au sites are: a wiki blog to provide historical material and worksheets; tools to upload time-aligned text that can be used for pronunciation of words (prioritised over recent recordings); an interactive dictionary from wordlists and sentences provided, with links to recordings; and place to upload sound recordings and images. Finally there is a section for teachers to develop worksheets, which can provide: search interface for word translation, auto linkage to audio files and drag and drop images. Also a question and answer tool was developed using natural Language Toolkit to provide for some variation in answer, but this implements very basic grammar rules so far.

These support the teachers who are often still learning their own language and want to share their work amongst the small group of speakers.

The next stage has been to develop an app that is based on an uploaded word/sentence list from teachers, which then provides an interface for them to collect recordings of these sentences. The same app will then provide resources for their students to practice these sentences.

The emphasis is to build the tools around the material that can be uploaded, as well as facilitating rapid upload of the language material to start the process. Also an emphasis on recording to support these oral languages and communication through sentences, rather than worklists, has been the main factor in the design.

Manchu language learning and socio-linguistic identity in 19th-century China

Dr Lars Peter Laamann (SOAS, University of London)

This paper will introduce a project which has evolved out of a recent cooperation agreement between researchers at the University of Leiden and at SOAS, University of London, which has borne concrete and scientifically satisfying results. Libraries across eastern Asia and Europe contain significant amounts of written materials in the script of the Qing China's (1644–1911) dynastic and second official language: Manchu. However, due to a shortage of scholars who have mastered Manchu, but also because of lacking expertise in our academic libraries, most Manchu collections have remained invisible.

Towards the end of the Qing era, most Manchus had assimilated into a Chinese cultural and linguistic setting, which however did not mean that they had forgotten their roots and distinct socio-linguistic identity. Determined to foster a sense of collective destiny, the Qing court encouraged the members of the military garrison families (*gūsa* ᡤᡡᠰᠠ /*qi* 旗 /banners) to study Manchu by means of taught classes and printed teaching aids (primers). This project aims to compile and make digitally available a corpus which will eventually contain all known versions of Manchu primers and grammars, teaching materials copied or reprinted in other works, as well as hand-written sources such as school book manuscripts and correspondence. This will also include source materials produced by Westerners, e.g. Herbert A. Giles, Paul G. von Möllendorff and Ivan I. Zakharov. Our working hypothesis is that these language aids were intended to link the late imperial Manchu youth to their historical roots, and thus to foster a sense of ethnic belonging. The Manchu primers can hence be seen as veritable "nation-builders", at a time when Manchu as a colloquially spoken language had already become endangered.

We now want to take this project one substantial step further by inviting institutions across Europe and East Asia (China, Taiwan, Japan, Korea) to participate, with the aim of locating, analysing and digitally preserving all original source materials in Manchu, printed or hand-written, as well as applied art objects such as seals or scrolls. The Manchu documents will be made available to the scholarly public by means of a single database, interconnecting textual fragments in our individual libraries so that the viewer is presented with a complete digital text – not unlike the International Dunhuang Project. We can already count on the support of a genuinely global network of Manchu experts, as well as an enthusiastic community of young Manchu learners in China and the West. We look forward to introducing the full scope and concrete examples of our project to the Language Documentation and Archiving conference in Berlin later this year.

Keywords: Manchu language books, Qing dynasty, Chinese minority languages, China modern history

Insights from 25 Years of Co-creative Training in Language Documentation and Revitalization

Jordan Lachler, Nicholas Bunderson-Toler, Timothy Mills, Lex Giesbrecht, Darren Flavelle, Craig Kopris, Marianne Huijsmans

The role of community members in language documentation projects has evolved rapidly over the past 50 years, from informants to consultants to collaborators to project leaders. A key element in this evolution is the training that community members receive, equipping them with the skills necessary to actively guide and shape the work carried out for their languages.

In this presentation, we will provide insights from 25 years of university-based training of Indigenous community members in North America in the core skills of language documentation and revitalization, including phonetics/phonology, morphosyntax, community language planning and technology for language documentation.

We will describe the content and structure of the training in each of these main areas and how they have evolved over the years through a collaborative process involving a combination of trial and error, student feedback, and community guidance.

We will also discuss the benefits and challenges of the different training models we have used. These include mixed-language training sessions, where students from a wide range of language communities learn together, as well as single-language training sessions, where all the students are from the same community.

Our experience has shown the necessity of combining outsider, technical expertise with more local, cultural knowledge and lived experience. The result is a co-creative environment where everyone's contributions are not simply valued, but essential to the shared, reciprocal learning within that space.

We will conclude by highlighting the need for further theorization of both (a) the full skill set needed by community members for documentation and revitalization projects, and (b) the formats and methods used to carry out training in various contexts.

Ticha: leveraging academy-based digital scholarship for community agendas and undergraduate pedagogy

Workshop on Historical language texts, methods for re-use

Brook Lillehaugen, George Aaron Broadwell,
Felipe Lopez, Xóchitl Flores-Marcial

Ticha: a digital text explorer for Colonial Zapotec (Lillehaugen et al. 2016, Broadwell et al. 2020) makes a corpus of alphabetic texts written in Zapotec during the Mexican colonial period available to broad audiences, with high resolution images, plain text transcription, and linguistic and cultural annotations. Co-directed by two linguists, the full Ticha team is interdisciplinary and heavily engaged with contemporary Zapotec-speaking communities, including through the formalization of the Ticha Zapotec Advisory Board. The infrastructure for Ticha's website is based at a small liberal arts college that serves only undergraduate students. Given the nature of support for digital scholarship at this institution, undergraduate students are an important part of creating, maintaining, and growing the website. In addition, the contents of Ticha are used in undergraduate classroom at Haverford and beyond (e.g. Lillehaugen & Flores-Marcial 2022).

In this contribution to the workshop, members of the Ticha team reflect on the challenges and opportunities of leveraging an academy based digital scholarship project for both Zapotec community language reclamation agendas (e.g. Lopez 2021) as well as for undergraduate pedagogical purposes. In many ways, reflective and collaborative conditions that are required for Zapotec community work lend themselves well to responsive and engaged pedagogy. That is to say, placing the "Zapotec agenda" (Plumb et al. 2024) at the center of the work helps foster a meaningful pedagogical experience. Unsurprisingly, there are also areas of tension that must be mitigated or navigated with care, such as heavy handed-ness of the academic construction of time through e.g. academic terms and funding cycles, which do not align with most people's lived experience of time outside of academia. We conclude that articulating and centering community values and priorities in our digital scholarship work is the anchor that allows us to not only serve community agendas but also helps create dynamic, meaningful pedagogical opportunities in and out of the classroom. Moreover, the moments of tension where the team chooses to try to change administrative structure to better support the possibilities of the work, while not always successful, are efforts to decolonize higher education, and when successful will likely benefit a range of projects across the campus for years to come.

Works Cited

Broadwell, George Aaron, Moisés García Guzmán, Brook Danielle Lillehaugen, Felipe H. Lopez, May Helena Plumb, & Mike Zarafonetis. 2020. Ticha: Collaboration with Indigenous communities to build digital resources on Zapotec language and history. *Digital Humanities Quarterly* 14(4). Online: http://digitalhumanities.org/dhq/vol/14/4/000529/000529.html.

Lillehaugen, Brook Danielle, George Aaron Broadwell, Michel R. Oudijk, Laurie Allen, May Helena Plumb, & Mike Zarafonetis. 2016. Ticha: a digital text explorer for Colonial Zapotec, first edition. Online: http://ticha.haverford.edu/.

Lillehaugen, Brook Danielle & Xóchitl Flores-Marcial. 2022. Extending pedagogy through social media: Zapotec language in and beyond classrooms. *Journal of the Native American and Indigenous Studies Association* 9(1): 62—101.

Lopez, Felipe H. 2021. Reclaiming our Languages. In Flores-Marcial et al. (eds), *Caseidyneën Saën–Learning Together: Colonial Valley Zapotec Teaching Materials.* Online: http://ds-wordpress.haverford.edu/ticha-resources/modules/chapter/reclaiming-our-languages/.

Plumb, May Helena, Alejandra Dubcovsky, Moisés García Guzmán, Brook Danielle Lillehaugen & Felipe H. Lopez. 2024. *Growing a bigger linguistics through a Zapotec agenda: a Ticha Project case report*. In Anne Charity Hudley, Christine Mallinson & Mary Bucholtz (eds.), *Decolonizing Linguistics,* 359—374. Oxford: Oxford University Press.

Role and Impact of Community Outreach Programmes for Dimasa Digital Preservation

Monali Longmailai, Monimala Rajkumari Sinha, Christina Wasson

We present an innovative approach to working with communities to collect Indigenous heritage material for a community-based digital archive. The community is Dimasa, primarily based in Assam, India (137,184 people as per 2011 Census of India). Longmailai led the development of "community outreach programmes" as a way for her team of Dimasa linguists to engage with Dimasa living in rural villages. This project responds to several points under the conference theme of Building Relationships, as well as providing a report from an Indigenous-led project (Planning and Design for the Decade theme).

Large set of historical and cultural narratives from Dimasa has been thereby documented since 2021 and uploaded in the "Bodo and Dimasa Heritage Digital Archive" platform which uses Mukurtu, a content management system. The current digital archiving project on Dimasa is a part of the research grant supported by the Indian Council of Social Sciences Research (ICSSR) from 2023 till date.

Community outreach programme, in general words, is a process where an organisation collaborates with the groups and individuals and creates shared experiences and actions for the benefit of a community. The paper will discuss the impact the community outreach programme has made on the Dimasa community from the region for their digital preservation, role of community engagement and financing, and the effectiveness of such programmes on the part of the archive team in collecting bulk data for the documentation and preservation purposes.

It will further highlight the materials the archiving has produced as a result, and how Mukurtu serves as a medium in the promotion of their cultural heritages in the digital world. The paper ultimately will lead to understanding the outcomes for the Dimasa community, which the collaborative Dimasa project aims to serve as such.

Keywords: Dimasa, community outreach, digital preservation

References

Census of India. 2011. Downloaded from https://censusindia.gov.in/census.website/ on March 24, 2024.

Bodo and Dimasa Heritage Digital Archive. 2024. Downloaded from http://bododimasaarchive.org/ on March 21, 2024.

Lingualibre.org : audio documenting 250 languages for Wikimedia

Hugo Lopez

LinguaLibre is an innovative crowdsourced audio documentation project dedicated to preserving vocabularies, housed within the global Wikimedian community. With over six years of continuous activity, our presentation will provide an in-depth overview of the project's current state, encompassing both its technical advancements and community dynamics. We will delve into the extensive quantitative progress achieved, showcasing our documentation efforts across 250+ languages, driven by contributions from a diverse community of over 1500 volunteers. Audio reuses are first done within hundred thousands Wikimedia pages, dataset download is also available for open educational resources.

In addition to highlighting our quantitative achievements, we will offer insights into the qualitative aspects of our work, shedding light on the inherent challenges and limitations associated with crowd-sourced language documentation. Our discussion will explore the varied profiles of our contributors, including their diverse backgrounds, ages, genders, and geographic locations, underscoring the inclusive nature of our collaborative endeavor.

Furthermore, we will present exciting exploratory projects within LinguaLibre, including initiatives focused on documenting regional languages, sign languages, and even whistled languages. Through these explorations, we aim to demonstrate the versatility and potential impact of our platform in facilitating the preservation and celebration of linguistic diversity worldwide.

Building Relationships:
Grammatical Documentation of Mada,
an endangered Niger Congo language of North-Central Nigeria

Shammah Daniel Makpu
(s.makpu@sussex.ac.uk)
University of Sussex

Field linguists acknowledge that native speakers are the most important resource in any fieldwork project, and how they are engaged defines both short-term and long-term outcomes of any language documentation or revitalisation efforts. Consequently, the dynamics of interaction between the researcher and the language community in the field remains an integral aspect of linguistic fieldwork enquiry. The present study is a language documentation project focused on the Mada language, a Niger-Congo language of North-Central Nigeria, which is highly endangered due to insufficient documentation, limited use in critical domains and the prevalence of linguistic exogamy among its native speakers. The dataset is a 100,000-word corpus of public and private monologues and dialogues compiled by the researcher and analysed using a descriptive-typological approach. Critical aspects of data collection and transcription expose the conflict between socio-political expediency on the one hand, and research/personal ethics and other best practices in linguistic fieldwork on the other. The researcher is faced with a wide range of peculiar contextual realities in speech communities, which often challenge one's professionalism and idealistically crafted principles of research, thereby potentially impeding research progress, and stretching the researcher beyond academic competencies, to intercultural competencies. Building meaningful relationships with the language community proves invaluable even for a researcher connected by ethnicity and heritage. A robust morphosyntactic description is the foundation for the longer-term objective of developing primers for language teaching and learning. This research demonstrates the centrality of native speakers in investigating, revitalising and reclaiming endangered languages, and the ways in which the researcher becomes the catalyst for engaging and supporting indigenous communities in collaborative and community-driven documentation efforts. By extension, it exemplifies the interconnectedness of human relations and computational, digital and data-driven approaches in language documentation outcomes.

Key terms: language documentation, language community, building relationships, corpus, grammatical description

Documentation of Numeral Systems

Dr. Kumari Mamta, Postdoctoral Researcher
Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

In the current globalising world, the maintenance of numeral systems that are typologically unusual has been profoundly jeopardised. The homogenisation of numeral systems by a variety of social, economic, and education–related processes obscure the diversity of these systems and it may also lead to erroneous conclusions in regards to the diversity of human conceptualisations of mathematics.

This paper discusses few rare numeral systems with primary focus on South-Asian Languages like Ao, Lotha, Khoibu, Palula, Luro and many more. The paper also provides a global overview of numeral systems. In particular, I discuss the rare phenomenaon of"overcounting" that is rapidly vanishing. Because borrowing triggers the change very fast,it is necessary to study sociolinguistic factors too.

This paper also discusses the type of questionnaire that should be involved while studying any numeral systems to correctly identify and document the structure of them. Proper documentation of numeral systems is urgently needed as Comrie (2005) says that, "Numeral systems are even more endangered than languages". I will introduce the "Numeralbank" database and repository on numeral systems in the world's languages, and also SAND (South Asian Numeral Database) which I am developing keeping in mind the urgency of it.

Diversity of mathematical conceptualizations is severely under threat, particularly because the dominant (*often decimal) numeral systems are encouraged by education systems, language contact and dominance. Furthermore, rare systems are viewed as a deterrent to socio–economic development. Most South Asian languages evidence a shift towards the decimal system by borrowing either the lexicon or structure of the dominant language of the particular region. A discontinuity with the past and the gradual erosion of folk knowledge has also triggered the endangerment.

Key Words: Numeral systems, rarities, numeralbank, documentation

References

Abbi, Anvita & Vysakh R.(2020). Aspects of Word Formation Processes in Luro: The endangered language of the Nicobar Islands. Asian Languages and Linguistics 1, No.1: 9-33

Chan, Eugene. (2015). Numeral systems of the world languages. Retrieved from https://mpilingweb.shh.mpg.de/numeral/

Comrie, Bernard.(2005). Endangered numeral systems. In Bedrohte vielfalt: Aspekte des sprach (en) tods (pp. 203-230). Weißensee-Verl.

Comrie, Bernard.(2011). Typology of numeral systems. Numeral types and changes worldwide.Trends in Linguistics. Studies and monographs, 118. De Gruyter Mouton.

Hammarström, Harald.(2010). Rarities in Numeral Systems. Pp. 11–60 in Rethinking Universals: How Rarities Affect Linguistic Theory, edited by J. Wohlgemuth & M. Cysouw. Berlin: Mouton de Gruyter.

Hanke, Thomas.(2010). Additional rarities in the typology of numerals. (pp. 61-90). De Gruyter Mouton.

Documenting with a smartphone

Bradley McDonnell, Jillian Breithaupt, Nathan Adamson, Kelsey Bialo, Lauren Cornwell, Tyler Demmon, Stephanie Dossett, Orlyn Esquivel, Irina Kolenskaia, Yihan Li, Tracy Preslar, Gillian Sawyer

Documentary linguistics has long recognized the usefulness of smartphones as their use has become almost ubiquitous the world over. This recognition has led to the development of apps that allow anyone with a smartphone to document their language (Bird et al. 2014). At the same time, smartphones as a primary means of documenting the range of linguistic and cultural practices has been discouraged (e.g. Seyfeddinipur & Rau 2020). In a rapidly changing marketplace where tech companies like Apple and Samsung are prioritizing cameras for capturing high-quality images and audiovisual recordings, we aim to determine how well smartphones fare against cameras and camcorders in three different use cases: audio recording, video recording, and document (e.g. fieldnotes) scanning.

Size limitations imposed on smartphone cameras mandates smaller sensors and lenses than those used in standalone cameras, which results in images with higher levels of noise as each pixel sees much less light (Abdelhamed, Lin & Brown 2018). Compared to DSLRs, smartphones are subject to very large fields of depth, small pixel sizes, and limitations in optical zoom capabilities (Blahnik & Schindelbeck 2021). However, recent studies claim that despite such limitations, image/video captured on current smartphone cameras are hardly distinguishable from those of professional cameras in many everyday situations with plenty of ambient lighting (Pattanayak, Malik & Verma 2021). In fact, within the last decade, many academic fields have leveraged smartphones for research, including linguistics (Leemann et al. 2020; Hilton & Leemann 2021).

In an effort to understand the limitations of smartphones, we compare the outputs of several recent models of smartphones to those of well regarded professional/prosumer cameras, camcorders, and audio recorders in the contexts in which language documenters record speech events as well as scan paper documents. Based on these results, we provide recommendations for the use of smartphones in language documentation projects.

References

Abdelhamed, Abdelrahman, Stephen Lin & Michael S. Brown. 2018. A High-Quality Denoising Dataset for Smartphone Cameras. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1692–1700. Salt Lake City, UT: IEEE. https://doi.org/10.1109/CVPR.2018.00182.

Bird, Steven, Florian R. Hanke, Oliver Adams & Haejoong Lee. 2014. Aikuma: A Mobile App for Collaborative Language Documentation. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, 1–5. Baltimore, Maryland: Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-2201.

Blahnik, Vladan & Oliver Schindelbeck. 2021. Smartphone imaging technology and its applications. *Advanced Optical Technologies*. De Gruyter 10(3). 145–232. https://doi.org/10.1515/aot-2021-0023.

Hilton, Nanna Haug & Adrian Leemann. 2021. Editorial: using smartphones to collect linguistic data. *Linguistics Vanguard*. De Gruyter Mouton 7(s1). https://doi.org/10.1515/lingvan-2020-0132.

Leemann, Adrian, Péter Jeszenszky, Carina Steiner, Melanie Studerus & Jan Messerli. 2020. Linguistic fieldwork in a pandemic: Supervised data collection combining smartphone recordings and videoconferencing. *Linguistics Vanguard*. De Gruyter Mouton 6(s3). https://doi.org/10.1515/lingvan-2020-0061.

Pattanayak, Sambhram, Fazal Malik & Manish Verma. 2021. Viability of Mobile phone cameras in professional broadcasting: A case study of camera Efficiency of Apple iPhone 11. In *2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, 452–456. Dubai, United Arab Emirates: IEEE. https://doi.org/10.1109/ICCIKE51210.2021.9410774.

Seyfeddinipur, Mandana & Felix Rau. 2020. Keeping it real: Video data in language documentation and language archiving. *Language Documentation & Conservation* 14. 503–519.

Archiving auto-documentation recordings: challenges and proposals

Amina Mettouchi

Academic-led or collaborative documentation have resulted in several hundreds of documented languages being archived in ELAR or PARADISEC. However, thousands of languages remain virtually undocumented in video format (mostly in Africa, Asia and South America). Scaling-up is needed, and indigenous-led models make this possible. They also are ethically central to how the threat to languages and cultures is tackled.

Auto-Documentation is such a model, of entirely indigenous-led documentation. It is defined as:

The constitution, by speakers/communities, of a body of recordings of their own language and culture, using methods and tools that do not involve academic institutions or funding, NGOs or other external support, but are accessible to any speaker or community in their local environment. (Mettouchi 2021)

Auto-Documentation is currently being implemented in North Africa by us Amazigh speakers, but as a model that does not require outside support, it is suited to many other areas in the world where collaboration with academic institutions or NGOs is extremely difficult or impossible, and/or where national governments are not trusted.

Auto-Documentation has resulted, in North Africa, in a number of culturally significant videos recorded entirely independently of national, academic or NGO networks. However, the status of the recordings that result from this documentation is fragile: they are "saved" locally (on smartphones, computers, on social media apps and other servers) but not "archived" (in a way that guarantees their preservation for future generations — crucially, our communities' descendants).

After presenting the principles, methods and tools of auto-documentation, along with resulting video documentation from indigenous North Africa (by Amazigh speakers and communities), I suggest achievable and sustainable solutions for the integration of such contemporary indigenous recordings into archives in the Global North, in a way that facilitates deposit and consultation by the communities themselves. These solutions, involving social media as well as semi-automated harvesting, form the last building block of the auto-documentation model, and the only one that actually requires collaboration with technologically-equipped institutions in trusted democratic countries. This is where we need you.

References

Mettouchi, Amina. 2021. "Auto-Documentation: Recognizing new relationships", 7th International Conference on Language Documentation and Conservation (ICLDC), University of Hawai‘i at Mānoa, March 4-7, 2021, https://www.youtube.com/watch?v=bijKRVH22-g

Themes of the conference centrally addressed by this presentation

- Building relationships (Language documentation practice as a medium for indigenous agency and revitalization; Models for engaging and supporting indigenous communities
  in collaborative and community-driven documentation efforts)
- Planning and design for the decade (Reports from indigenous-led documentation, training, and revitalization projects; Language archiving: current assessment and future prospects)

Cathedral, bazaar, data garden: the Pangloss Collection

Alexis Michaud, Séverine Guillaume

Eric Raymond's classic *The Cathedral and the Bazaar* (Raymond 1999) contrasts top-down and bottom-up design. How does the Pangloss Collection (pangloss.cnrs.fr), a member of the Digital Endangered Languages and Musics Archives Network, pattern in these terms? There is a *cathedral* aspect to the format used for display and distribution: hierarchical markup (XML) following a fixed structure that has remained essentially stable over the years (Jacobson, Michailovsky & Lowe 2001; Michailovsky et al. 2014). Yet the collection is on the *bazaar* side in terms of the corpora hosted. Any speech dataset can in principle be deposited in the archive, provided that it was collected in connection to linguistic/anthropological research (and belongs within the national scope, since the institution taking care of long-term archiving has national scope). Corpora are archived *as is*: they are documented to a varying extent, they have vastly different sizes, some are annotated with an excruciating amount of detail whereas others only have a transcription and some are untranscribed, or even left untranscribed. The diversity of the archive, and the evolution of the hosted resources over time, suggest a third metaphor: that of the *data garden*, which not only undergoes overall growth over the years, but also allows (and fosters) gradual improvements to the resources. The *gardening* tasks can favour emulation and cross-fertilization in terms of annotation practices, without hard regulations and guidelines. The presentation of the archive will place emphasis (i) on the functions that it plays as part of a broader Open Science environment, (ii) on current plans for improvements to workflows, to interfaces, and to resources and metadata, and (iii) on opportunities, challenges and threats related to Natural Language Processing (such as automatic transcription: Guillaume et al. 2022).

References

Guillaume, Séverine, Guillaume Wisniewski, Benjamin Galliot, Minh-Châu Nguyễn, Maxime Fily, Guillaume Jacques & Alexis Michaud. 2022. Plugging a neural phoneme recognizer into a simple language model: a workflow for low-resource settings. In *Proceedings of Interspeech 2022*. Incheon, Korea. https://halshs.archives-ouvertes.fr/halshs-03625581.

Jacobson, Michel, Boyd Michailovsky & John B. Lowe. 2001. Linguistic documents synchronizing sound and text. *Speech Communication* 33 [special issue: "Speech Annotation and Corpus Tools"]. 79–96.

Michailovsky, Boyd, Martine Mazaudon, Alexis Michaud, Séverine Guillaume, Alexandre François & Evangelia Adamou. 2014. Documenting and researching endangered languages: the Pangloss Collection. *Language Documentation and Conservation* 8. 119–135.

Raymond, Eric. 1999. The cathedral and the bazaar. *Knowledge, Technology & Policy*. Springer 12(3). 23–49.

Integrating domain-specific archiving practice into tertiary education: PARADISEC develops data management, archiving, and digital preservation curricula for linguistics, anthropology, and ethnomusicology

Julia Colleen Miller, Nick Ward

Established in 2003, the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) is a collaborative project involving three Australian universities: The University of Sydney, The University of Melbourne, and the Australian National University. Serving as a digital archive and access platform, PARADISEC focuses on preserving endangered materials from the Pacific region, encompassing Oceania, East, and Southeast Asia. These materials are primarily contributed by researchers specializing in Linguistics, Anthropology, Ethnomusicology, and related fields.

Operating within academic research institutions provides PARADISEC with direct engagement opportunities with students and researchers. Recognizing the importance of integrating archival practices into tertiary education, PARADISEC has developed tailored curricula for linguistics, anthropology, and ethnomusicology. The curriculum emphasizes data management, archiving, and digital preservation, with a focus on safeguarding linguistic and cultural heritage.

Through practical training sessions, students acquire essential skills for managing and preserving diverse data types, including linguistic recordings and ethnographic materials. They are introduced to industry-standard tools and techniques for digitization, metadata creation, and ensuring long-term preservation.

The integration of domain-specific archiving practices into tertiary education aligns with PARADISEC's mission to empower students and researchers in responsibly managing cultural and linguistic resources. Originally designed for credited work placement programs and tertiary classrooms, the curriculum has been successfully implemented at institutions like Charles Sturt University and the University of Manchester. Its effectiveness has led to its adoption for training staff, volunteers, and interested individuals, with growing interest in expanding its reach beyond occasional guest lectures in university classrooms.

Supplementary data uploaded along with this abstract: example 10-week syllabus

(Re)Listening to the Voices of the Ancestors. Kamëntšá Language Documentation and Revitalization Through Meaningful Relationships

Marcelo Miranda, Jully Acuna Suarez, Milena Aguillón Chindoy, Silvia Jamioy Juajibioy

This presentation explores the critical importance of establishing meaningful relationships in the documentation and revitalization of the Kamëntšá language, highlighting the ethical considerations and community-and-family-driven approaches employed at our Tsebionán Curatorial Center in Sibundoy, Colombia.

Our research model is built on *Indigenous Methodologies* (Battiste, 2016; Chilisa, 2020; Smith, 2021), which prioritize collaborative engagement with Indigenous peoples, ensuring that their voices, practices and needs are central to the documentation process. Therefore, our research and documentation process directly involves and benefits the Kamëntšá nation. We identify and respond to support needs while navigating ethical considerations to enhance the accessibility of language materials, thereby fostering linguistic empowerment within the Kamëntšá community.

Furthermore, we present how archived materials, particularly the J. H. McDowell Archives on Kamëntšá language housed in the *Archive of Indigenous Languages of Latin America* (AILLA), are being utilized by our research community. Through a process of re-listening and reinterpreting these archives, we also navigate the complexities of colonial and neocolonial research practices, critiquing methodologies employed by different anthropologists and linguists, who do not take into account community needs, ethical considerations, and are grounded on a poor and often erroneous understanding of Kamëntšá culture, ontology and socio-political context. Our analysis underscores the importance of engaging with communities in critiquing language documentation practices, ensuring that they reflect Indigenous agency and aspirations for revitalization.

In elucidating our research and "family-oriented" methods, developed also by other Indigenous communities (see Custer & Daniels, 2022; Olko & Sallabank, 2021) we advocate for language documentation practices directed at revitalization, creation of pedagogic materials and research centred on a clear and decolonial understanding of Kamëntšá culture developed by elders, educators, artists and scholars, including the works of philosopher Juan Alejandro Chindoy (2020, 2021) and anthropologist Willian Mavisoy (2014, 2018).

Bibliography

Battiste, M. (2016). Research Ethics for Protecting Indigenous Knowledge and Heritage: Institutional and Researcher Responsibilities. In Norman K. Denzin & M. D.

Giardina (Eds.), *Ethical Futures in Qualitative Research. Decolonizing the Politics of Knowledge*. London & New York: Routledge. Pp. 111-132.

Chindoy Chindoy, J. A. (2020). *A Decolonial Philosophy of Indigenous Colombia. Time, Beauty, and Spirit in Kamëntšá Culture*. Lanham: Rowman & Littlefield.

Chindoy Chindoy, A. J. (2021). Dancing as Environmental Aesthetics From Indigenous Colombia. Roczniki Kuluroznawcze, XII (1): 65-73.

Chilisa, B. (2020). *Indigenous Research Methodologies*. New York: SAGE Publications.

Custer, A. & Daniels, B. (2022). *Nehiyawetan Kikinahk? Speaking Cree in the Home. A Begginer's Guide fo Families*. Regina: University of Regina Press.

Mavisoy Muchavisoy, W. J. (2014). Etnografía sobre el quehacer antropológico y las manifestaciones de un antropólogo por su origen. El ejercicio del jtsenojuabnayá "existir reflexionando" entre los Kamëntšá. *Tabula Rasa*, 20: 197-221.

Mavisoy Muchavisoy, W. J. (2018). El conocimiento indígena para descolonizar el territorio. La experiencia Kamëntšá (Colombia). *NÓMADAS*, 48: 240-248.

Olko, J. & Sallabank, J. (2021). *Revitalizing Endangered Languages. A Practical Guide*. Cambridge: Cambridge University Press.

Smith, L. T. (2021). *Decolonizing Methodologies. Research and Indigenous Peoples*. London: Zed Books.

Collaborative Work on Repatriation of Indigenous Folklore Collections in Russia: Perspectives of Indigenous Communities, Archives, Folklore Collectors and Researchers

Maria Momzikova

The paper presents the collaborative work-in-progress on the project [BLINDED] on digitization and repatriation of the Arctic Siberian Indigenous audio folklore recordings collected in the Soviet times in different regions and several Indigenous languages and Russian and kept in the archive [BLINDED] in St Petersburg. Among the project's outputs is the repatriation of digitized audio and other archival materials through collaborative meetings with Indigenous communities and the website. The project started before the Russia-Ukraine war as an international collaboration, and organizers decided to continue the work. I will reflect on building relationships between archival employees, Indigenous communities, scholars, and folklore collectors. Specifically, I will focus on the process of elaborating guidelines on ethical aspects of collaborative work on sharing digitized archival recordings of Indigenous knowledge with communities and how we can use principles of relationality, responsibility, and respect (Liew and Lipscombe 2024) in politically complicated contexts.

Preserving cultural knowledge and doing linguistic research in the collaborative Teop Language Documentation Project

Ulrike Mosel

This presentation shows how the documentation of cultural, historical, and ecological knowledge of the Teop people and corpus-based linguistic research has successfully been combined in a collaborative language documentation project. From the very beginning of the project in 1994, a team of Teop speakers in Bougainville (Papua New Guinea) decided on the content and the form of the documentation. They did the recordings and transcriptions, wrote texts without previous recordings, edited the texts they regarded as suitable for publication, and translated all texts. Two Teop artists contributed drawings for the folk tales and the texts about plants, animals, and the material culture.

My task was to organize the finances and the workflow in Bougainville and in Germany. For security reasons the Teop people did not want me to bring a computer until 2009, so that the typing of Teop texts in Word and ELAN and the compilation of the Teop-English dictionary in Toolbox had to be done by myself and research assistants in Kiel.

The total size of the in ELAN annotated Teop texts is 254 693 words; the size of the transcribed oral texts is 132 396 words and the size of edited texts 122 657 words. Both texts collections comprise folktales, texts about the history and customs of the Teop people, and descriptions of plants, animals, and the material culture. 40 edited folk tales were published in a monolingual schoolbook (2007) and 15 historical texts in a bilingual book (2014). The texts on the material culture, plants and animals have not been published yet in book form, but text excerpts and their translations are found in the encyclopedic articles of the digital Multifunctional Teop-English Dictionary (2019).

The various genres and registers of the texts are considered in my corpus-based grammatical research for which I systematically search the in ELAN annotated texts for grammatical constructions by using Regular Expressions. Such searches do not only find data for a

particular research question, but also unexpected data that contradict hypotheses and lead to new research questions. The results of my recent research include:

- Nouns, verbs, and adjectives are only distinguished by their optional modifiers.
- The ellipsis of arguments is context dependent and not determined by grammatical rules.
- Teop does not have tenses; how the described situations in a text are temporally related to the moment of speech is understood from the context.
- Teop belongs to the rare languages that have a propositional negator.

References

2007 The Teop Language Corpus https://dobes.mpi.nl/projects/teop/team/

2019. A multifunctional Teop-English dictionary.
Dictionaria 4. 1-6488. DOI: 10.5281/zenodo.5526528
https://dictionaria.clld.org/contributions/teop#tabout Accessed on 2024-04-14.)

2022. Teop DoReCo dataset. In Seifart, Frank, Ludger Paschen and Matthew Stave (eds.). Language Documentation Reference Corpus (DoReCo) 1.2. Berlin & Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). https://doreco.huma-num.fr/languages/teop1238 (Accessed on 10/02/2023). DOI:10.34847/nkl.9322sdf2
https://doreco.huma-num.fr/languages/teop1238 (accessed 14 April 2024)

Women´s collaborative cartophonies: how sound shapes indigenous language pedagogies

Deyanira Moya-Chaves, Myriam N. Lamprea-Abril, Carlos A. Barreneche-Jurado, Offray V. Luna-Cardenas, Paula F. Martinez-Pulido

This paper presents the results of a collaborative interdisciplinary research-creation project (Chapman and Sawchuk, 2012), which aims at contributing to the revitalization of four (4) languages spoken in the Amazon, Colombia: *uitoto* (*mɬnɬka*)*, okaina* (*ɬbuza*)*, bora* and *muinane* (*gaigomɬjo*) through the creation of a *cartophony*. Cartophonys are understood here as sound maps that bring together visual and sonic epistemologies and practices which offer rich ground for exploring how representations of time and space are performed between and across the senses (Thulin 2018). Our cartophony is created collaboratively through a wiki and it is comprised of chants, lullabies and songs in the four languages along with soundscapes that represent their meaning and their relationship with the territory. The cartophony also includes the corresponding explanation of each song in Spanish along with key words and expressions in the indigenous languages selected by the women to create a multilingual dictionary. We included illustrations and spectrograms to represent each song visually as well as the necessary metadata to inform the indigenous users. Thus, the cartophony is meant to be used by indigenous teachers and educators, offline and through their cellphones, as pedagogical material for indigenous children in the Amazon. It encodes local, situated indigenous women's ecological knowledge and shows how language is used to heal and to connect with nature. The collaborative creation of this cartophony also shows how revitalization efforts are strongly linked to memory, and to community archiving practices and rights. The paper will also present how the community decided to use the cartophony and how it might impact women's empowerment towards the teaching and use of the different languages that coexist in their territory.

Link to the cartophony: https://cartofonias.tiddlyhost.com/

Safeguarding Sonic Heritage at the Fringe of the Pacific Ocean: A Contextual Approach to Preserving Soundscapes

Ahmad Faudzi Musib and Gisa Jähnichen

Sound preservation goes beyond intentional sounds, encompassing contextual sounds independent of the collector's initial intention. While intentional sounds are often the primary focus during collection, contextual sounds play an equally vital role. These sounds, surrounding the intended sound, may remain passive and unfocused due to their soft amplitude. Nevertheless, their existence is undeniable, as the ear perceives sound spatially through depth and distance.

The conventional method of relying on single sound recordings for preserving sonic heritage imposes limitations, hindering the capture of diverse events. This constraint obstructs the creation of a comprehensive preservation material that vividly portrays specific locations of indigenous groups, especially their social and environmental aspects. The study concentrates on two distinct communities: the Bidayuh community in the village of Annah Rais belonging to Padawan, Sarawak, residing in a longhouse, and the Zhuang community in Napo County, Baise, Guangxi, China, where singing Zhuang songs at a specific park is integral to their daily social interactions. Additionally, literature on this phenomenon is analysed and presented.

By employing contextual sound preservation methods, this joint paper aims to expand the scope of sonic heritage preservation to definite practices. It underscores a multi-dimensional approach, condensing the chosen locations' social and environmental dimensions. The collection process prioritizes social dynamics and daily life interactions; thus offering a deeper understanding of soundscapes and incorporating dynamic social activities within different proximities. Additionally, the research endeavours to document the changes and evolution within these soundscapes over time, considering the broader context and environmental factors. This contributes to a more holistic understanding of preserving audio elements associated with cultural heritage.

Dr. Ahmad Faudzi Bin Hj. Musib is working at Faculty of Human Ecology, Universiti Putra Malaysia. His main working fields are sound synthesis, audio engineering, and audio preservation. He has numerous students at the Music Department. Also, he is a T&E Committee member of IASA and member of the International Council for the Traditions of Music and Dance (ICTMD). He can be best contacted via email or phone: 6003 8946 7120.
Prof. Dr. Gisa Jähnichen is working and researching at Shanghai Conservatory of Music. She did manifold fieldwork in both areas and has many students around the globe. Also, she is the secretary of the IASA-T&E Committee and ambassador of this organisation in Malaysia and China and an active member of the International Council for the Traditions of Music and Dance (ICTMD). She published widely about various aspects of sound preservation. She is best contacted per email: gisajaehnichen@web.de.

Voices of Hope in Waṭa Language Documentation: Prospects for a Sustainable Partnership and Revitalization

Nancy Ngowa and Micheal Maua, Pwani University

nanceyaa@yahoo.com m.maua@pu.ac.ke

There has been more awareness of the rapid pace of language endangerment and death (Ngowa 2021.1). In this case, comprehensive language documentation is considered most urgent. Linguists often embark on such documentation projects based on their perceptions and needs. In this way, the projects normally reflect the interests of researchers other than the speech community. This paper will focus on the new experiences, hopes and impedements encountered in the field during the Waṭa (ssn) language documentation project.

The Waṭa is an Afro-Cushitic language with a speech community estimated at 20,100 speakers (Eberhard, Simons & Fennig, 2021). This language is endangered and most of the youth speak other languages, leaving a very small population of elders speaking the language. This made the elders push for an MoU (Memorandum of understanding) with an academic institution to have their language documented by the linguists. This collaboration resulted in a grant from the ELDP to document this language which was like a gift to the community.

This paper will share the benefits of an academic instituition collaborating with a speech community with mutually benefit from each other and meet a common goal in the language documentation process. Some of the benefits are; Information is all shared without reservation because they rely on institutions more than individuals while the university also gets data that is reliable.

The Waṭa language documentation project has raised voices of hope, enthusiasm and excitement for the prospects of future and more sustainable language documentation and possibility of revitalization. Different stakeholders in the Waṭa language documentation process will share new experiences and perspectives. The stakeholders include the young, the old, the reserach team, the local government and the neighbours of the Waṭa.

On the other hand, this paper will also examine gaps and lapses that hinder the full success of a language documentation project that were encountered in the field. These challenges will inform any future projects of a similar nature.

References

Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). 2024. *Ethnologue: Languages of the World*. Twenty-seventh edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com, Accessed on April 13, 2024.

Ngowa, N. (2021) Documentation of the Waata (ssn): Prospects for a sustainable partnership with the community. Presented in a language Documentation conference in SOAS, 2021.

“to teach another generation”: engaging Wayoro and Makurap communities (Brazil) in collaborative documentation efforts

Antônia Fernanda Nogueira

In this paper, we present a methodology for engaging Wayoro and Makurap communities (Brazil) in collaborative documentation efforts. The goal of our ELDP project is to provide systematic documentation for Makurap and Wayoro (two severely endangered languages, spoken in the Brazilian state of Rondônia) (Galucio, Nogueira, Costa 2023; Nogueira, Galucio, Costa, 2023). At the project’s launch, a meeting was held at the Ricardo Franco village school. The leaders and other participants expressed their concerns about maintaining and documenting their languages. Thereafter, a documentation workshop was held for members of the community. Following the workshop, Makurap and Wajuru community members were encouraged to organize and carry out mini projects, based on Moore (2018), with specific topics they wanted to document. We consider that this methodology was very effective in raising awareness for linguistic and cultural documentation, and also in building documentary capacity among the community members. In the case of the Wayoro language, we would like to highlight the comprehensive community-based approach. By getting in touch with dormant practices and knowledge in the community, the young people strengthen their indigenous identity. In a mini project description (about the production of a sling), Jaqueline Wajuru states her concern for the Wajuru community: “[the aim is] to teach another generation step by step how to produce a sling”. The elders and some young people from both groups (Makurap and Wajuru) have been particularly motivated for carrying on language documentation. The elders want to document their knowledge for their children and grandchildren. The young people are learning about their culture as they document it, as Jociclei Makurap says his interest is both on documenting for the future and for his kids to know, and also to learn more about his own language and culture, as he does the documentation.

References

Galucio, Ana Vilacy, Antonia Fernanda Nogueira, and Carla Daniele Costa. 2023. *Makurap: Documentation of Language and Culture*. Endangered Languages Archive. Handle: http://hdl.handle.net/2196/73b9o01s-1c78-2246-5187-0fm978700a80. Accessed on [08/04/2024].

Moore, Dennis. *Language documentation with a focus on traditional culture among the Gavião and Suruí of Rondônia*. Berlim: Endangered Languages Archive, 2018. Disponível em:<http://hdl.handle.net/2196/00-0000-0000-000F-EE8F-E>. Acesso em : 31/03/2023.

Nogueira, Antonia Fernanda, Ana Vilacy Galucio, and Carla Daniele Costa. 2023. *Wayoro: Documentation of language and culture*. Endangered Languages Archive. Handle: http://hdl.handle.net/2196/5004e53b-79f6-440d-81e6-266a64579366. Accessed on [08-04-2024].

Language documentation on-the-move: a multipurpose endeavour

Karolin Obert

One request to language documenters is the creation of multipurpose records for di5erent uses and users including the community, academic audiences, and local policymakers (e.g., Woodbury 2003; Austin 2006; Himmelmann 2006). While in theory, this is a very reasonable and necessary demand, holding up to this standard in practice can be a di5icult enterprise. Here, I discuss how documentary records produced with GPS-equipped action cameras as people are walking can bridge this gap. When documenting territory with Dâw and Nadëb community members (Naduhup, Brazil), mobile recordings have proven themselves to be highly effective for capturing linguistic categories activated by naturally occurring stimuli in the environment (e.g., landscape, place name, traditional ecological knowledge etc.). Instantiations of such categories that surface spontaneously during routes can be e5iciently selected and transcribed in ELAN, and video and geo-referential data o5er yet another layer for their description and linguistic analysis. For example, fauna terminology mentioned in passing by community members can be captured seamlessly; the video record of the entire route provides crucial information on the habitat of the species, and finally, the GIS record enables their mapping in relation to further aspects of the local landscape. Hence, such contextualized records allow for more accurate representations of Indigenous categories in the communities' own terms. Moreover, integrating geo data on the ELAN timeline enables precise mapping of any category in space, which is a crucial tool for these communities to negotiate land rights with local authorities and for educational purposes. From a linguistic perspective, mobile data collections allow for novel insights into language as uttered in an underexplored realm, a dynamic one. Moreover, integrating geo-data and linguistic data adds a new layer of analysis to unravel the possible footprints of the environment in grammars and lexicons.

In sum, this novel documentation method not only proves to be efficient and engaging but also holds the potential to accommodate the diverse needs and interests of various stakeholders, ranging from communities themselves to academic researchers and policymakers alike.

Legacy Materials in a Digital Environment:
the case of Enggano, a language of Indonesia

Sarah Ogilvie, Gede Rajeg and Daniel Krausse

Language communities, activists, and researchers who want to document, maintain, and revitalize heritage languages are often working with complex archival collections. Trying to piece together a complete picture of a language and culture from incomplete records is never easy, but it is made even more difficult when those records are biased, puzzling, derogatory, or plain wrong. And yet every record plays a part in telling the story and history of a people, their language, and their culture. This paper looks at the challenges facing a team of linguists and community members working with legacy materials relating to Enggano, a vulnerable language spoken southwest of Sumatra, Indonesia, and the tools and techniques used to resolve, or at least mediate, those challenges.

Archival materials mentioning the island and peoples of Enggano date from the late-sixteenth century to the mid-twentieth century. The bulk of materials referring to the Enggano language were recorded from the mid-nineteenth century to the mid-twentieth century. In this paper, we will discuss what happens in a digital environment - especially when creating an online dictionary and educational resources - when the linguistic content of legacy materials differs from contemporary language; when legacy materials are in formats and structures that are difficult to digitize; when they are in languages other than those spoken by the project team; and when they contain material considered culturally insensitive or inaccurate. It is hoped that the tools and techniques discussed will be of value to other researchers working with legacy materials and archives.

Proactive Acquisition: Four-year Progress Report from the California Language Archive

Zachary O'Hagan

The California Language Archive (CLA) at the University of California, Berkeley is a physical and digital archive founded in 1953 for linguistic heritage materials from the Americas. As of March 2024, it holds 432 accessioned collections on 519 named language varieties (18,823 items includ- ing 51,590 digital files). Since January 2021, the CLA has acquired more than 20 collections by outreach to senior (retired) scholars, primarily linguists, anthropologists, and musicologists with research interests in South America. (See Austin 2017, Dobrin and Schwartz 2021, and O'Meara and Good 2010 for issues inherent in legacy materials.) Examples include relatively large collec- tions, such as Gerald Weiss's (1932-2021; O'Hagan 2021) papers, tape recordings, and photographs (majority 1961-1980) from ethnographic fieldwork in Ashaninka communities (Peru), and small ones, such as Alfred Pietroforte's (b. 1925) single tape recording of songs in Mono, Northern Paiute, Tachi, and Wukchumni (California), from 1959. These proactive acquisitions complement deposits by active researchers and instructors (and Indigenous efforts to locate at-risk materials; Tamez 2021), and can circumvent models by which an intermediary must facilitate deposits by these scholars. They respond to the precariousness of collections in private hands, but they also risk creating processing backlogs for understaffed repositories.

This presentation discusses these acquisitions in order to frame a general progress report on CLA accessions since this date, together with activities including coordination with California tribal groups (e.g., Kawaiisu, Modoc) as part of cataloging, community research visits supported by multiple organizations, collaborations with museums, and preliminary digital materials re- turn events in Kukamiria and Urarina communities (Peru). The result is a variety of collections stemming from classical research projects conducted by faculty members, graduate students, and others, longstanding collaborations with community members, linguistic field methods courses, linguistics lectures, discoveries in departmental spaces, and the work of missionaries.

References

Austin, Peter. 2017. Language Documentation and Legacy Text Materials. Asian and African Languages and Linguistics 11.

Dobrin, Lise M., and Saul Schwartz. 2021. The Social Lives of Linguistic Legacy Materials. Language Documentation and Description 21:1–36.

O'Hagan, Zachary. 2021. Obituario, Gerald Weiss (1932-2021). Amazoń ıa peruana 34:279–286. O'Meara, Carolyn, and Jeff Good. 2010. Ethical Issues in Legacy Language Resources. Language and Communication 30:162–170.

Tamez, Sonia. 2021. Saving Our Stories: Sustaining California Indigenous Knowledge. California Institute for Community, Art, and Nature.

Languages in Berlin: Adapting Existing Endangerment Models to Urban Settings

Olina, Olga; Behling, Bruno; Ilgner, Bastian.

Several models have been proposed to assess language endangerment worldwide (AES, EGIDS, LEI & UNESCO Atlas). While all of them recognise globalisation and urban migration as key factors contributing to language decline, their applicability to urban contexts is limited. Our research indicates that while many factors included in these models are helpful, others are irrelevant, and some are left unconsidered.

We conducted interviews with thirty speakers of thirty-five primarily non-European heritage languages spoken in Berlin. These interviews focussed on the participants' linguistic backgrounds, their daily language use and attitudes towards their heritage languages, German and other global languages. Our results show that some of these languages are more likely to be maintained and transmitted to the next generation, while others are more likely to fall into disuse.

Our presentation will critically examine existing models for assessing language endangerment and their applicability to urban settings, specifically Berlin. We propose that factors such as the presence of diaspora neighbourhoods, community size and cohesion, institutional support, and the political and migratory contexts of the home country influence language vitality. Additionally, we will discuss several challenges of conducting urban fieldwork, including participant recruitment and ethical considerations when engaging with minority populations in city settings.

References

AES – Hammarström, Harald; Forkel, Robert; Haspelmath, Martin; Bank, Sebastian (eds.). *Glottolog 5.0*. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://glottolog.org/langdoc/status, last access 26 March 2024.

EGIDS – Lewis, M. Paul; Simons, Gary F. 2010. *Expanded Graded Intergenerational Disruption Scale*. https://www.ethnologue.com/about/language-status, last access 26 March 2024.

LEI – Catalogue of Endangered Languages. 2022. *The Language Endangerment Index.* University of Hawaii at Manoa. https://www.endangeredlanguages.com/about_catalogue/, last access 26 March 2024.

SBEE – Amt für Statistik Berlin-Brandenburg. 2021. *Statistischer Bericht: Einwohnerinnen und Einwohner im Land Berlin am 31. Dezember 2020*. Potsdam. https://download.statistik-berlin- brandenburg.de/fa93e3bd19a2e885/a5ecfb2fff6a/SB_A01-05-00_2020h02_BE.pdf, last access 26 March 2024.

UNESCO Atlas – Mosely, Christopher; Nicolas, Alexandre. 2010. *Atlas of the World's Languages in Danger*. https://unesdoc.unesco.org/ark:/48223/pf0000187026, last access 26 March 2024.

Building Foundational Language Data with A Community-First Approach

Subhashish Panigrahi

The historical language data collection, archiving, and publishing practices have been critiqued within a post-colonial context. Although professional documenters collect data ethically, issues including prioritisation mismatches between communities and documenters, as well as paywalled and less accessible archived materials, persist, leading to extractive documentation. The extensive language data collection for building generative artificial intelligence (AI) and machine learning (ML) demands decolonising language documentation practices to prevent individual and community rights infringements.

For-profit entities have already displayed questionable language data collection models, particularly for medium- and large-resourced languages. Low-resourced languages remain vulnerable, particularly as the AI/ML industry promotes the idea of swift, error-free content generation. The ongoing political campaigns in the 64 countries with elections in 2024, which use emerging technology to spread political messages, including misinformation, in multiple languages and mediums, highlight this potential risk. Indigenous language data could soon be targeted for training language models. Current models are often trained using data scraped from publicly hosted sites without permission or compensating the data owners.

While emerging technology could aid in documenting, processing, and analysing Indigenous and endangered languages, mainly when led by speaker communities, careful and conscious practices are critical. The socio-economic understanding of low and limited-resourced language speakers is crucial to the future of language data collection. Te Hiku Media, a Māori-language community media organisation, provides arguably the most effective methods for empowering communities in a language documentation process, both financially and socio-politically. While standardising data collection practices is challenging, local- and community-first approaches should be thoroughly documented and taught to replace current practices. Similarly, low- and medium-resourced language activists are also under-resourced. However, there is potential for more well-resourced language archives and networks and the startup ecosystem to develop data collection and archiving models where the language community is actively involved and remunerated.

Engaging communities through training in language documentation

Leah Pappas, Khairunnisa Khairunnisa

In recent decades, community members have played an increasingly larger role in documentary projects; often, they are not just consultants but rather project collaborators trained in documentary tools, methods, and theories. Such community engagement leads to more sustainable documentation that aligns with both the linguist's and the community's goals (Leonard & Haynes, 2010; Czaykowska-Higgins, 2009) and also results in data that more accurately reflects the community's knowledge and language use (Olko, 2018). We discuss our strategies for working towards collaborative documentation of Saparua (ISO 639-3 spr), an endangered Austronesian language of Central Maluku, Indonesia.

Specifically, we share the details of a workshop that we hosted in the community during the first week of fieldwork. The workshop was a simplified, yet applicable, training package in language documentation that included modules on language endangerment in Indonesia, ethics, techniques in language documentation (e.g., recording, ELAN, metadata), and a field practicum. Although there were challenges to adapting the training for the community, the workshop greatly benefited our project. Four community members joined the project in a full time capacity. They collect, transcribe, and translate data, and they also advise on what and who to record by facilitating communication with the rest of the community. The workshop was also a space for discussion. As a post-conflict society, the Saparua community had reservations about the project, but through the workshop we were able to plan an approach for documenting Saparua that aligned with their needs and expectations. We suggest that this model was a crucial way to encourage collaboration at all stages of our project, and we believe that it is applicable to other fieldwork contexts, particularly in Indonesia.

References:

Czaykowska-Higgins, E. (2009). Research Models, Community Engagement, and Linguistic Fieldwork: Reflections on Working within Canadian Indigenous Communities.

Language Documentation and Conservation, 3, 15–50.

Leonard, W. Y., & Haynes, E. (2010). Making "collaboration" collaborative An examination of perspectives that frame linguistic field research. Language Documentation &

Conservation, 4, 268–293. http://hdl.handle.net/10125/4482

Olko, J. (2018). Spaces for participatory research, decolonization and community empowerment: Working with speakers of Nahuatl in Mexico. Language Documentation and Description, 16, 1–34.

The Importance of Agent Paradigms in Role Usage

Hugh Paterson III

The pioneering work (Bird and Simons 2001; Huang 2002; Hughes 2004; Johnson 2002) which resulted in the OLAC Role Vocabulary (OLAC-RV) has been under-discussed since its promotion to candidate status (Johnson 2006). However, OLAC roles are considered important for language archives as well as citation and referencing practices (Andreassen et al. 2019). Roles as part of the complete metadata record play a part not only in acknowledging the intellectual merit and contribution of contributors but also the character, type, and nature of the resource. For example, "speaker" may be an OLAC role which is used to indicate a language proficient voice in a resource, but "interviewee" may actually more accurately characterize the resource, not just the contributor.

Two important aspects of the OLAC-RV deserve attention: first, the interdependent nature of the OLAC-RV and the MARC relator roles vocabulary (LOC 2024) and, second, the unfinished nature of the OLAC-RV where its authors state: "To do: Elicit roles from language researchers other than documentary linguists." By addressing agent paradigms this paper lays a foundation for principled continuous development of OLAC-RV.

The OLAC-RV usefully presents 24 roles, but provides no guidance when these roles should be co-used with other roles. We rely on the undocumented but implicit paradigm invoked by a role. We further suggest that the documentation of these paradigms are paramount for clear and consistent application of roles across stewardship organizations. The documentation of language-resource agent/creation-paradigms is also important for successful communication with "non-documentary linguists" for the efficient elicitation of roles. Finally, since the OLAC-RV is interdependently built upon the MARC relator role vocabulary, it is important to understand the paradigms present in both vocabularies. We present 10 paradigms based on roles in the OLAC-RV so that conversations with "other language researchers" can commence.

References

Andreassen, Helene N., Andrea L. Berez-Kroeker, Lauren Collister, Philipp Conzett, Christopher Cox, Koenraad De Smedt, Bradley McDonnell, et al. 2019. Tromsø Recommendations for Citation of Research Data in Linguistics. Etterbeek, Belgium: Research Data Alliance Linguistic Data Interest Group. https://doi.org/10.15497/RDA00040.

Bird, Steven, and Gary F. Simons. 2001. "The OLAC Metadata Set and Controlled Vocabularies." In Proceedings of ACL/EACL Workshop on Sharing Tools and Resources for Research and Education, edited by Thierry DeClerck, Steven Krauwer, and Mike Rosner, 7–18. Université de Toulouse, France: EACL-ACL; elsnet. https://www.aclweb.org/anthology/W01-1506.

Huang, Chu-Ren. 2002. "Suggestion for Adding a Proofreader Role," November 12, 2002. https://listserv.linguistlist.org/pipermail/olac-implementers/2002-November/000209.html.

Hughes, Baden. 2004. “OLAC Metadata,” June 21, 2004. https://listserv.linguistlist.org/pipermail/olac-implementers/2004-June/000245.html.

Network Development and MARC Standards Office, Library of Congress. 2024. “Relator Code and Term List -- Term Sequence: MARC 21 Source Codes.” Washington, D.C.: Library of Congress. https://www.loc.gov/marc/relators/relaterm.html. Johnson, Heidi. 2002. “OLAC Role Vocabulary.” Slides presented at the IRCS Workshop on Open Language Archives, Second

OLAC Workshop, University of Pennsylvania. https://slideplayer.com/slide/717312/. Johnson, Heidi. 2006. “OLAC Role Vocabulary.” Recommendation. Dallas, TX: Open Language Archive Community.

http://www.language-archives.org/REC/role.html.

Zur Sprache kommen: Bringing endangered languages to the wider public in a museum context

Naomi Peck, Uta Reinöhl

Exhibitions are a valuable opportunity to engage the public with issues surrounding language documentation and archiving. In this talk, we introduce the exhibition "Zur Sprache kommen: Forschung zu bedrohten Sprachen sichtbar gemacht" (Coming to language: Visualising research on endangered languages), held at the Uniseum Freiburg in summer 2024. We share in this presentation lessons learnt during preparation and execution of the exhibition.

The purpose of the exhibition was twofold. The first goal was to inform the Freiburg public on ongoing research with endangered languages (e.g. *Non-Hierarchicality in Grammar*, DFG project no. 406074683; *Russinisch als eine Staatsgrenzen überschreitende Minderheitensprache*, DFG project no. 263457843). The second -- perhaps more important -- goal was to encourage visitors to draw connections between the experiences that speakers of endangered languages have to their own lived experiences.

These goals informed how we approached the design of the exhibition and its accompanying educational programme. Dedicated sections focussed on the process of fieldwork, its outcomes, and its ethical repercussions, as well as how speakers of endangered languages relate to their language(s). An interactive display allowed visitors to experience a wide range of language data, including archival data. In the accompanying programme, we reached a more varied audience with public lectures, guided tours in different languages, and participation in the yearly Museumsnacht.

Several challenges arose in the framing and presenting of the exhibition contents. We aimed to display fieldwork in a way that acknowledges its roots in coloniality, while also highlighting movements within decolonial and indigenous frameworks. Another challenge was how to best portray language diversity while not feeding into exoticisation and othering. This exhibition represents a case study in how we can engage a broader public with important issues in language documentation and archiving through a critical lens.

Tangible and Intangible Heritage and Culture: among endangered languages, archives, and museum collections.

Alessandra Praun

When considering the urgency of addressing themes related to endangered or dormant languages—as highlighted by Indigenous Kokama linguist and researcher Altaci Rubim—it becomes essential to reflect on the interconnections or interdependencies between tangible and intangible heritage and culture. In 2019, UNESCO proposed the International Decade of Indigenous Languages (2022-2032) to encourage the preservation, dissemination, revitalization, and integration of these languages. Subsequently, some cultural institutions have incorporated these themes into their discussions. For example, the Museu da Língua Portuguesa in São Paulo inaugurated celebrations in line with UNESCO's proposal, and the Humboldt Forum in Berlin with a series of events in 2024.

The disappearance or risk of extinction of these languages is directly linked to the enduring remnants of colonialism, practices that have influenced museum collections and continue to shape narratives today. Museums and cultural institutions, particularly those with ethnological collections, play a crucial role in advocating for these languages. After all, many endangered languages resonate with the groups and cultures represented in museum collections, telling their stories and capturing the essence of their communities. Furthermore, archives document these languages, transforming the living language—which is fluid and accompanies its speakers—into static, yet essential, documentation. Therefore, this paper aims to explore the interdependencies and relationships between tangible and intangible heritage in the context of museum collections, archives, and endangered languages.

Evolution of linguistic archiving in India

Chaithra Puttaswamy, Mrunmayee Amshekar

Archiving documented material is a crucial segment in the process of documentation. As in the methodology of documentation, there have been technological advances in the archiving process too. Over the last few decades, archiving practices have moved from paper to digital formats: audio-visual data stored in internet archives. In this paper we take an overview of archiving practices for Indian languages at the local (national) level.

Language data is no longer necessarily just archived in libraries in physical format like Emeneau's documentation of the languages Kota (Emeneau, 1944-46) and Toda (Emeneau, 1971, 1984) or Grierson's LSI data (Grierson, 1903-28). With emerging scholarship on theory of archiving and better technological tools, a shift in evolution of archiving goals and methods can be observed. Archiving language documentation has three kinds of options in India:

- Using services provided by the local library to archive documentation of projects hosted
- Creating a specialized archive to host all the documentation projects funded by a specific institution
- Utilizing the services provided by remote archives such as CORSAL

This paper reviews the merits of the three options mentioned above.
Increasing awareness about identity and cultural heritage has also encouraged entities other than linguists and anthropologists to create archives. Woodbury's (2014) emphasis on making archives accessible to communities is acknowledged in practice (e.g. SDML (*SDML*, n.d.) made separate sections for narrations – one for the linguists and one for non-linguists). Meta documentary data, encouraging transparency in the methods of data collection is also made available with the archived data (Narayanan, 2020; *SDML*, n.d.). The archival attempts are not limited to efforts by linguists or institutions, community initiatives to archive their language and culture have increased in number.

References

Emeneau, M. B. (1944). *Kota texts* (Vols. 2–3). University of California Press.

Emeneau, M. B. (1971). *Toda songs*. Clarendon Press.

Emeneau, M. B. (1984). *Toda grammar and texts*. American Philosophical Society.

Grierson, G. A. (Ed.). (1903). *Linguistic survey of India*. Office of the Superintendent of the Government Printing.

Narayanan, R. K. (2020). Made in India SiDHELA Indias First Endangered Language Archive. *DESIDOC Journal of Library & Information Technology*, *40*(05), 292–299. https://doi.org/10.14429/djlit.40.05.16349

Survey of Dialects of Marathi Language Online. (n.d.). Retrieved April 12, 2024, from https://sdml.ac.in/en

Woodbury, A. C. (2014). Archives and audiences: Toward making endangered language documentations people can read, use, understand, and admire. *Language Documentation and Description*, 19-36 Pages. https://doi.org/10.25894/LDD161

Sustaining Practices: A Multivocal Approach to Archiving

Karthick Narayanan Ramakrishnan, Sumitra Ranganathan

Community engagement in archiving practices is vital for safeguarding endangered linguistic and cultural practices, yet current approaches often lack inclusivity and perpetuate power imbalances. We report upon a prototype for community-engaged archiving in which a linguist and an ethnomusicologist build upon a growing body of work[1] on strategies for mediated and participatory archives to address this critical gap.

The prototype archive draws records from field recordings created over the last decade by the authors. The linguist focuses on creating an audio-visual repository of Toda verbal arts, renowned in the field of ethnopoetics (Emeneau 1971; 1984). Similarly, the ethnomusicologist works on compiling a comprehensive archive of a critically endangered tradition of Dhrupad, the oldest genre of north Indian classical music. The initiative's main aims were not simply preservation of these cultural heritages but sustaining their continued practice.

To make the archive integral to community efforts to sustain their practices, the authors worked in close collaboration with community members to develop a conceptual model for digital archiving that reconfigured some basic archival principles and practices. This reconfiguration prioritises the co-custody of records, respectful representation, and decentralised decision-making.

Custody in traditional archiving practices is centralised, with records ownership bestowed on the archiving institution. However, the prototype entered an ongoing partnership between the repository and the community to develop a co-custody model similar to those reported by Mary Stevens and Shepherd 2010, Flinn 2011, and Rodrigues 2016. In this model, records are treated as cultural assets. The archive foregrounds ownership and participation of the community members as equal partners. Secondly, the collection development and appraisal policies were developed with criteria drawn from community knowledge and traditional practices instead of the techno-quantitative criteria decided by the archives. Thirdly, the arrangement schemas were developed to mirror traditional knowledge in the community. While traditional archives approached archival arrangement from the principle of original orders and provenance, the prototype relies heavily on communities' meta-knowledge about their linguistic and cultural heritage to create meaningful order in the archival arrangement. Finally, the descriptive schema of the records is designed to expose the rich context within which the materials were created and to provide for findability, access, and (re)use of the record along with respectful, trusting, and sustained collaboration.

[1] Mary Stevens and Shepherd 2010; Flinn 2011; Rodrigues 2016; Dale 2021; Holton 2013; Garrett 2014; Newman 2012; Wilbur 2014

References

Dale, M. 2022. "Creating workflow for mediated archiving in CoRSAL".*The Electronic Library*, Vol. 40 No. 5, pp. 568-578. https://doi.org/10.1108/EL-02-2022-0027

Emeneau, M. B. 1971. *Toda Songs*. Oxford: Clarendon Press.

———. 1984. *Toda Grammar and Texts*. Memoirs of the American Philosophical Society; v. 155. Philadelphia: American Philosophical Society.

Flinn, Andrew. 2011. “Archival Activism: Independent and Community-Led Archives, Radical Public History and the Heritage Professions.” *InterActions: UCLA Journal of Education and Information Studies* 7 (2). https://doi.org/10.5070/D472000699.

Garrett, Edward. 2014. “Participant-Driven Language Archiving.” *Language Documentation and Description* 12 (: Special Issue on Language Documentation and Archiving): 68–84.

Holton, G. 2014. “Mediating language documentation”. Language Documentation and Description 12, 37-52. doi: https://doi.org/10.25894/ldd163

Stevens, Mary, Andrew Flinn, and Elizabeth Shepherd. 2010. “New Frameworks for Community Engagement in the Archive Sector: From Handing over to Handing On.” *International Journal of Heritage Studies* 16 (1–2): 59–76. https://doi.org/10.1080/13527250903441770.

Newman, Joanna M A. 2012. “Sustaining Community Archives: Where Practice Meets Theory.” *Australasian Public Libraries and Information Services,* 25 (1).

Rodrigues, Antonio. 2016. “Introducing an Archival Collecting Model for the Records Created by South African Portuguese Community Organisations.” *Archives and Manuscripts* 44 (3): 141–54. https://doi.org/10.1080/01576895.2016.1258582.

Wilbur, J. 2014. “Archiving for the community: Engaging local archives in language documentation projects”. Language Documentation and Description 12, 85-102. doi: https://doi.org/10.25894/ldd166

Exploring Chinantec of Comaltepec using published legacy materials in a diaspora online micro-community of learning

Cynthia Montaño Ramírez

Using published legacy materials for language revitalization efforts is challenging due to the nature of such materials, since they are generally not created for that purpose. The challenges are: language community members not being aware of the existence of these materials; the terminological accessibility of the language resources due to the interests and values of researchers who create them rather than the speech communities. (Austin, 2017; O'Meara & Good, 2010).

To explore published legacy materials, I work with the diaspora community of Santiago Comaltepec Oaxaca, Mexico, in California, USA, using the resources made by the Summer Institute of Linguistics (SIL) for Chinantec (ISO-639: cco): a grammar, a dictionary, an alphabet and a writing handbook. Our main goals are filling the gap between language published materials and their access by community members, exploring the language resources of Comaltepec Chinantec published by the SIL along with the speakers, so they can get familiar with them, and evaluating how to adapt them to future revitalization efforts.

I work mainly with 2 people: R.G., a native speaker of Comaltepec Chinantec based in Los Angeles, and Diana, a heritage Chinantec language learner based in Boston. They are uncle and niece, and since we are mainly three people, I have called us an "online micro-community of learning" to offer another model in language revitalization in diaspora with a multi-sited approach in contrast to efforts made by larger groups of people, like language classes, or revitalization community projects *in situ*. R. G., Diana, and I first met in June 2023 to discuss our project, and we have met one hour weekly since then until now (March 2024), and our project is still going.

Our experience has been fruitful. We have learned and practiced the sounds of Chinantec, the tones, and understanding the writing system, as well as how animacy agreement works. Engaging native speakers with the legacy materials can also stimulate the discussion about the data documented in the collection: variation in the vocabulary, different meanings of a given word, narrative styles, old stories (Spencer, 2018).

Bibliography:

Austin, P. (2017). Language documentation and legacy text materials. Asian and African Languages and Linguistics, 11, 23–44.

O'Meara, C., & Good, J. (2010). Ethical issues in legacy language resources. Language & Communication, 30(3), 162–170. https://doi.org/10.1016/j.langcom.2009.11.008

Spence, J. (2018). Learning Languages Through Archives. En The Routledge Handbook of Language Revitalization. Routledge.

A workflow for independent time-aligned transcription in low-resource documentation contexts

Eleanor Ridge

Transcription is an often undervalued component of language documentation, viewed as a time-consuming 'bottleneck' holding up more interesting stages of analysis (Himmelmann 2018), or as mechanical task ripe for delegation to computational tools. Independent transcription, where speakers transcribe audio or video recordings without the presence of a linguist researcher (Jung & Himmelmann 2011), can help to make transcription more efficient, without losing the valuable reflective experience that the process of transcription can be for community members. This paper presents a workflow for independent time-aligned transcription using free and widely available software tools (SayMore (Moeller 2014), ELAN (MPI Nijmegen 2023) and Audacity (Audacity Team 2023)) to generate labelled and numbered mp3 files of short segments, which transcribers can listen to and transcribe by hand, noting codes and segment numbers so that they can be easily typed up into a time-aligned annotation file. This workflow is especially appropriate in contexts where there is access to mobile phones or mp3 players and means to power them, as well writing materials and literacy; but where use of computers is not practical due to availability, restricted power, or limited digital literacy. As well as outlining the workflow, the paper will reflect on two projects where this workflow has been used in Vatlongos-speaking communities in Vanuatu: 1) a PhD project focussing on corpus analysis of variation in morphosyntax of verbs, involving three transcribers; and 2) an ongoing project working with young Vatlongos community members, involving dozens of inexperienced transcribers across four rural and urban communities, transcribing archival recordings from the 1950s-2000s, as well as young people's own recordings of monologues and discussions. Beyond the technical workflow, this paper will discuss training, communication, logistical considerations, as well as the transformative potential of transcription as a reflective linguistic process in contexts of language endangerment and reclamation.

Audacity Team. 2023. Audacity: Free, open source, cross-platform audio software for multi-track recording and editing. https://www.audacityteam.org/.

Himmelmann, Nikolaus P. 2018. Meeting the transcription challenge. University of Hawai'i Press.

Jung, Dagmar & Nikolaus P. Himmelmann. 2011. Retelling data: working on transcription. In Geoffrey L. J. Haig, Nicole Nau, Stefan Schnell & Claudia Wegener (eds.), *Documenting Languages: Achievements and Perspectives.*, 201–20. Berlin: Mouton De Gruyter.

Moeller, Sarah Ruth. 2014. SayMore, a tool for language documentation productivity. *Language Documentation and Conservation* 8. 66–74.

MPI Nijmegen. 2023. ELAN. Nijmegen, The Netherlands: Max Planck Institute for Psycholinguistics, The Language Archive. https://archive.mpi.nl/tla/elan.

Exploring user-friendly functionalities to access audiovisual materials: A web-based interface for the Tsuut'ina Language Archive.

Lorena Martín Rodríguez

In recent years, language archiving in the field of language documentation has undergone considerable changes in its format. The creation of digitized databases has explored some of the advantages of digital methods, such as the possibility of indexing large textual collections for user-friendly searches (Enge et al., 2015). Moreover, social changes have required us to redefine established recommended practices. Traditional boundaries and roles between the participants of documentary linguistics projects have been challenged through the empowerment of Indigenous communities in the documentation and revitalization of their languages (Henke and Berez-Kroeker, 2016), as well as in the creation of newly developed methods of participatory models and community-language research (Czaykowska-Higgins, 2009). These conditions have also impacted the design choices of language archives. Recent literature has emphasized the need of tailoring their design and functionalities for their speech communities (Wasson et al., 2016). This is especially relevant when language archives are conceived not only as a tool for language documentation, but also as a key element to support language revitalization (Perez Báez et al., 2019).

This presentation introduces a web-based interface for the Tsuut'ina Language Archive, a collection of language materials in Tsuut'ina (ISO 639-3: srs, Glottocode: sars1236). This interface was prototyped along with members of Tsuut'ina Nation following user-centered design principles to tailor its materials and functionalities to the needs of the language community. This presentation focuses on the implementation of user-friendly methods of discovery and delivery tailored to audiovisual files annotated using ELAN (Brugman & Russel, 2004): full-text search and the visualization of time- aligned annotations online. Examining the collaborative design of these approaches facilitates the analysis of how digital methods in language archiving can improve accessibility to archived materials while supporting the documentation and revitalization goals of the language community.

References

Blokland, R., Chuprov, V., Levchenko, D., Fedina, M., Fedina, M., Partanen, N., & Rießler, M. (2016). Komi media collection. Syktyvkar: FU-Lab. http://videocorpora.ru

Brugman, H., & Russel, A. (2004). Annotating multi-media/multi-modal resources with ELAN. Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04), 2065–2068.

Czaykowska-Higgins, E. (2009). Research models, community engagement, and linguistic fieldwork: Reflections on working within Canadian Indigenous communities. Language Documentation and Conservation 3(1).15–50. http://hdl.handle.net/10125/4423.

Enge, J., Głowacz, A., Grega, M., Leszczuk, M., Papir, Z., Romaniak, P., & Simko, W. (2009). OASIS Archive – Open Archiving System with Internet Sharing. In: Mauthe, A., Zeadally, S., Cerqueira, E., Curado, M. (eds) Future Multimedia Networking. FMN 2009. Lecture Notes in

Computer Science, vol 5630. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-02472-6_28

Henke, R., & Berez-Kroeker, A. L. (2016). A brief history of archiving in language documentation, with an annotated bibliography. Language Documentation & Conservation, 10, 411-457.

Himmelmann, N. P. (2006). Language documentation: What is it and what is it good for? In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), Essentials of Language Documentation, 1–30. Berlin: Mouton de Gruyter.

Northern Pomo Language Tools. (2023). https://northernpomolanguagetools.com/

Pérez Báez, G., Vogel, R., Patolo, U., (2019). Global Survey of Revitalization Efforts: A mixed methods approach to understanding language revitalization practices. Language Documentation & Conservation, 13, 446–513.

Spence, J. (n.d.). Hupa Online Dictionary and Texts. http://nalc.ucdavis.edu/dictionaries/hupa-lexicon.php

Wasson, C., Holton, G., & Roth, H. S. (2016). Bringing user-centered design to the field of language archives. Language Documentation & Conservation, 10, 641-681.

Agile frameworks for holistic data stewardship: A relational approach to developing tools within community-university partnerships

Daisy Rosenblum, Dante Cerron, David Gaertner, Olivia Chen, Julia Chu

Many Indigenous communities are embracing new media technologies to support their goals of maintaining linguistic and cultural continuity. Community researchers are seeking innovative ways to mobilize collective knowledge through audio, video, and other 'content' deployed on web- and mobile-based platforms. These tools generate and draw on extensive reserves of born-digital and digitized legacy data of diverse formats, types, and origins, all of which needs to be managed, labeled, described, and organized. The resulting processes and practices of data governance and stewardship are critical sites of collaboration for community- and university-based researchers.

This paper presents a Content Management System (CMS) built in Drupal and React, created for the stewardship of digital data related to prototyping new media storytelling projects. We describe our attention to principles such as CARE (Collective Benefit, Authority to Control, Responsibility, and Ethics), OCAP (Ownership, Control, Access, and Possession), and FAIR (Findable, Accessible, Interoperable, and Reusable), alongside the ways these principles sometimes contradict each other. Considering the role of data in relation to sovereignty and resurgence, relational ethics, and decolonization guides our application of these frameworks.

Our work highlights the value of adapting Agile methodologies from industry contexts (cf. Beck et al. 2001) to community-engaged research models. Agile development's incremental and evolutionary approach prioritizes individuals and interaction, collaboration with community stakeholders, and a flexible, responsive stance. Agile's iterative processes frequent opportunities for feedback are particularly well-suited to community-engaged research partnerships, offering a flexible and responsive development model that supports the ethical, participatory, and dynamic needs of community-engaged research. In contrast to the "waterfall" lifecycle of traditional data management which follows a linear lifecycle, moving from planning to storage in a predictable sequence and echoing the legacy of extractive approaches to research 'on' communities' languages, rather than *with* or *for* them, Agile development offers the potential for a more nuanced, equitable approach to data lifecycle management that recognizes and rectifies the imbalance between community contributions and researcher gains.

Beck, Kent; James Grenning; Robert C. Martin; Mike Beedle; Jim Highsmith; Steve Mellor; Arie van Bennekum; Andrew Hunt; Ken Schwaber; Alistair Cockburn; Ron Jeffries; Jeff Sutherland; Ward Cunningham; Jon Kern; Dave Thomas; Martin Fowler; Brian Marick (2001). "Principles behind the Agile Manifesto". Agile Alliance. Archived from the original on 14 June 2010. Retrieved 6 June 2010.

Documenting minority languages in politically sensitive contexts: ethical considerations based on the case of Romeyka in Turkey

Laurentia Schreiber

Building relationships and supporting speech communities is one of the central pillars of language documentation. This presentation is concerned with settings of language endangerment where the speakers are not favorable towards language maintenance and documentation as can be the case, for example, in settings of political sensitivity and/or historical trauma. The presentation raises questions as to (I) whether it is acceptable to document an endangered language against the preference of the (majority of) speakers, (II) how to pre-empt potential negative effects of language documentation for the community, and (III) how to support communities to facilitate language maintenance or linguistic self-determination. These questions are addressed based on the case study of Romeyka, a variety of Pontic Greek spoken in Turkey which has been classified earlier as 'definitely endangered' whereby conflicting language identities and negative language attitudes are central factors in the advanced language shift towards the national language Turkish (Schreiber & Sitaridou 2016). From the general and Greek linguistic perspective, Romeyka is highly interesting showing, among other traits, retention of several archaic features, which are otherwise absent from modern Greek varieties (Bortone 2009; Sitaridou 2013, 2014; Schreiber 2023). The dichotomy between research interests and the interests of the speech community evokes ethical questions such as 'Who "owns" the language?' and '"Who can/should be the agents in language documentation?'. On the example of the use of Romeyka on social media platforms such as YouTube and Facebook, a bottom-up approach to assist language maintenance is presented that exemplifies how complex sociolinguistic factors are interwoven and how they affect language vitality.

References

Bortone, Pietro. 2009. Greek with no history, no standard, no models: Muslim Pontic Greek. In Alexandra Georgakopoulou and Silk, Michael (eds.), *Standard Languages and Language Standards: Greek, Past and Present*, 67-89 London: Ashgate.

Schreiber, Laurentia & Ioanna Sitaridou. 2017. Assessing the sociolinguistic vitality of Istanbulite Romeyka: An attitudinal study. *Journal of Multilingual and Multicultural Development* 39(1), 1-16. DOI: 10.1080/01434632.2017.1301944.

Schreiber, Laurentia. 2023. *A (contact-)grammar of Romeyka*. University of Bamberg/University of Ghent Dissertation.

Sitaridou, Ioanna. 2014. The Romeyka Infinitive: Continuity, Contact and Change in the Hellenic varieties of Pontus. *Diachronica* 31(1). 23-73.

Sitaridou, Ioanna. 2013. Greek-speaking enclaves in Pontus today: The documentation and revitalization of Romeyka. In Mari Jones and Sarah Ogilvie (eds.), *Keeping Languages Alive. Language Endangerment: Documentation, Pedagogy and Revitalization*, 98-112. Cambridge: Cambridge University Press.

A new archival data structure: PARADISEC's move to RO-Crate

Peter Sefton, Amanda Harris and Nick Thieberger

Over the life of the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) we have used a few different tools for describing items in the collection. Each had its benefits, but each was superseded and allowed its contents to be re-used in the subsequent system, but each separated the catalog from the items in the collection.

In 2024, we have adopted an emerging specification, the Research Object Crate [1]. The advantage of using RO-Crate as the underlying storage format is that it allows the collection to be self-describing, not relying on a database for contextual information. This makes the data more sustainable and transportable.

A major advantage of using RO-Crate for PARADISEC is that we can now index the text files in RO-Crate and search within items, allowing users to search text files, and also to search transcripts of media (in ELAN's .eaf format), with transcripts linked to playable media. This result then resolves to just the selected segment in a media file, a major advance over the view provided by the previous catalog.

RO-Crate provides Linked Data metadata for data resources using JSON-LD, and the main metadata vocabulary is Schema.org. Linked data as used in RO-Crate also allows for interrelationships between files to be described, for example showing that an ELAN file is an annotation of a video file, and potentially which tiers in the file are transcriptions, translations etc.

The advantages of using RO-Crate metadata also gives the project access to multiple tool implementations such as (software libraries, metadata editors and data-discovery services. etc)

In this presentation we will outline the new data structure of our collection and demonstrate the improved access to items in the collection that it provides.

[1] Stian Soiland-Reyes, Peter Sefton, Mercè Crosas, Leyla Jael Castro, Frederik Coppens, José M. Fernández, Daniel Garijo, Björn Grüning, Marco La Rosa, Simone Leo, Eoghan Ó Carragáin, Marc Portier, Ana Trisovic, RO-Crate Community, Paul Groth, Carole Goble (2022): Packaging research artefacts with RO-Crate. Data Science 5(2) https://doi.org/10.3233/DS-210053

The Open Gazetteer of Open Maps for Europe and Minority Place Names

Roman Stani-Fertl

Open Maps for Europe (OME) is a collaborative European project focused on open geographic data. It is managed by EuroGeographics, the membership association for National Mapping and Cadastre Agencies (NMCAs). Among the six datasets within the project is the Open Gazetteer, which offers authoritative geographical names as maintained in the source data of the NMCAs. It includes standardized endonyms in all the national languages of the participating countries. However, certain types of toponyms are initially excluded from the gazetteer due to the NMCAs' limited jurisdiction to their national territories and officially recognized languages.
The sub-project "Exonyms Management," funded by the German BKG, aims to expand the Open Gazetteer by incorporating toponyms that are absent from the national names databases. These missing toponyms primarily fall into two categories:

- Names of geographical features in national languages that lie beyond the NMCAs' jurisdiction (exonyms) (e.g. Prag, Prága, Prague, etc., for Praha)

- Toponyms in endonymic languages not officially recognized in a respective country (endonyms and exonyms) (e.g. Perpinya [cat] Perpignan; Roazhon [bre] Rennes FR; Gduńsk [csb] Gdańsk PL; Podstupim [hsb] Potsdam DE)

According to INSPIRE, all toponyms are classified as *endonyms* or *exonyms*, and their name status is determined to be *official*, *standardized*, *historical*, or *other*. Currently, the Open Gazetteer includes toponyms from more than 50 languages categorized based on the mentioned name status.

By incorporating these name forms into the Open Gazetteer, minority toponyms will be directly linked to their endonyms along with essential attributes such as coordinates, feature type, or other related name forms. This expansion of the Open Gazetteer aims to enhance the visibility of minority toponyms within an open Pan-European dataset built on authoritative data sources.

Contributor Roles in a Mixed Archive

Hannah Tarver, Mark Phillips, Hugh Paterson Ⅲ

Many archives (PARADISEC, ELAR, Pangloss, AILLA) operate as standalone units. This affords them the opportunity to make decisions about controlled vocabularies used within their information systems. Assumptions about the independent nature of preservation and memory organizations have characterized much of the scholarly discussion regarding language documentation metadata. Less discussed is the growing number of language resource collections which are embedded within digital repositories at memory institutions with a broader preservation focus. If institutions with broadly scoped missions are to equally participate in supporting language revitalization and language documentation efforts, then metadata for describing language resources needs to be able to be drawn from controlled vocabularies used across the broader cultural heritage industry. The OLAC Roles (Johnson 2006) and the MARC relator terms (MARC 2024) are two such vocabularies.

We present on contributor roles as they are implemented in an archival collection at an institution which has language resource and non-language resource collections and requires metadata to be conformat within the larger context across all collections. We have compared three sets of contributor roles (OLAC, 24 terms, MARC, 299 terms, and our institutional terms, 74 terms). When comparing OLAC terms with MARC terms, we show that 10 of the 24 terms have compatibility issues. Six are in the context of their semantics, and four have no equivalence. We then analyzed our institutional terms for congruence with MARC and found that 3 of 74 terms were locally created and not part of the MARC relator vocabulary. Finally, we analyzed which of the roles in our local systems were compatible with the OLAC role vocabulary. There are only 14 comparable roles between our local metadata vocabularies and the OLAC role vocabulary. Of these roles, three have semantic issues between the terms used and the definitions expected between the metadata namespaces.

References

Network Development and MARC Standards Office, Library of Congress. 2024. “Relator Code and Term List -- Term Sequence: MARC 21 Source Codes.” Washington, D.C.: Library of Congress. https://www.loc.gov/marc/relators/relaterm.html.

Johnson, Heidi. 2006. “OLAC Role Vocabulary.” Recommendation. Dallas, TX: Open Language Archive Community. http://www.language-archives.org/REC/role.html.

Documenting Endangered Languages: Revitalizing Judeo-Georgian through Archival Materials[1]

Maka Tetradze (PhD student) Ivane Javakhishvili Tbilisi State University
Tsira Janjghava (PhD) Ivane Javakhishvili Tbilisi State University

The Georgian-speaking Jewish communities scattered throughout the Republic of Georgia face the threat of disappearance due to ongoing immigration, predominantly to Israel, which began in the 1970s. With approximately 120,000 members repatriated, only around 3,000 Jews remain in Georgia today. Their unique speech, known as Judeo-Georgian, represents a variety of the Georgian language. While not a separate dialect, Judeo-Georgian exhibits linguistic features influenced by the diverse origins and environments of these communities.

Our project, "Documenting Endangered Languages: The Case of Judeo-Georgian," aims to preserve and revitalize Judeo-Georgian by documenting it according to modern standards of language documentation. Utilizing video recording and linguistic annotation, we are compiling a comprehensive dataset primarily sourced from the Israeli community. Additionally, our project seeks to breathe new life into archival texts collected nearly 70 years ago in the Kutaisi region by Roza Tavdidishvili. These manuscripts, recently digitized and integrated into the Georgian dialect corpus, lack glossing and annotation[2].

Through the annotation process, we confront various challenges, particularly in determining the appropriate theoretical framework for grammatical analysis, especially concerning the complex morphology of Georgian verbs. Nevertheless, by leveraging tools like Flex, we navigate these challenges, addressing issues of grammatical categorization and finding solutions.

Our efforts culminate in the creation of fully annotated Judeo-Georgian language data, presented in both Georgian script and Latin transliteration (IPA), ensuring accessibility to international academia and language preservation communities. By providing this rich linguistic resource, we contribute to the broader understanding and conservation of Judeo-Georgian for future generations.

[1] The work is supported by the Shota Rustaveli National Foundation of Georgia (grant number FR-21-20266).
[2] corpora.co

Access to Australian Indigenous language manuscripts

Nick Thieberger

Libraries, archives and museums hold many manuscript sources representing Australian Indigenous languages. On paper or microfilm, they are difficult to access, impossible to search, and their content remains obscure to the people most concerned to read them, the descendants of the people whose language is recorded in them. In a collaboration with the National Library of Australia (2011-2017), I tested an access method for a set of 26,000 pages of manuscripts created in 1904, representing a number of Australian Indigenous languages in the Daisy Bates collection [1]. Microfilms of the originals were digitised and then key parts of this data set (some 4,000 pages) were typed, and encoded using the Text Encoding Initiative guidelines, allowing all pages to be linked and presented online. With standard metadata descriptions for each vocabulary, this allowed geographic mapping of words. A fuzzy search mechanism that includes known variations of graphemes maximises findability of items in the data set, that, being on paper, was previously unsearchable.

This work has been universally well received, especially by Aboriginal people, but also by other researchers who can now search this mass of information. However, it is a fixed set of records, not designed to grow or to have user feedback added. Motivated by the success of this project, in 2021 I led a team to successfully apply for Australian Research Council funding to build a platform that would allow users to upload manuscript images, to transcribe them, and to then download them in various textual formats to build a corpus of materials in these languages. We call this platform Nyingarn [2], an open source project [3], and, after nearly three years of development, it provides a workspace for existing transcripts in various formats (e.g., MS Word, csv, text, Transkribus XML, and others) or it can use Amazon Textract for Optical Character Recognition (OCR) of text in images. We also successfully use an existing crowdsourcing platform to have documents transcribed. Initial preparation of transcripts is done in a secure workspace in which a user can nominate other users to collaborate on a manuscript, but it is otherwise not visible to anyone else.

Nyingarn currently has 190 users who have transcribed some 980 manuscripts, a testament to its usefulness. The original design of Nyingarn includes both the workspace and a repository in which finished manuscripts can be housed and accessed. For each manuscript to enter the repository, we require it to have permission both from the copyright holder or source institution, and a language authority (an Indigenous person associated with the language who approves the use of the manuscript). The project has a range of State libraries and the National Library of Australia as partners, and is running on servers provided by the Australian Institute of Aboriginal and Torres Strait Islander Studies who will take responsibility for Nyingarn into the future.This presentation will outline these features and show recent developments in Nyingarn.

[1] Nick Thieberger. 2017. Digital Daisy Bates. Web resource. http://bates.org.au
[2] https://nyingarn.net
[3] https://github.com/CoEDL/nyingarn-workspace

Do we need digital language and music archives?

Nick Thieberger, Amanda Harris, and Mandana Seyfeddinipur

Over the past 20 years few digital language archives have started, and existing archives have not developed new interfaces and methods. Considering the number of languages and the amount of documentation occurring it would seem that there is a need for much more infrastructure to curate and make this material available. There has been discussion about what is needed in language archives, and critiques of existing archives, but none of this has resulted in the development of a new archive.

In part, we suggest this lack is due to the low uptake by linguists and musicologists of the possibilities offered by documentation methods, which is itself a reflection of the lack of incentive provided by academia for the creation of citable primary records, and of engagement with the communities whose language are represented in the research. A further reason for this is also a lack in training in documentation methods and archiving and the use of archival collections in teaching and research.

As the OLAC website is at the end of its life, decisions need to be made about whether to revive it or not. We explore the arguments for and against the ongoing development of language and music archives, first, because they are an accessible platform for curaton of the specific kind of records created by ethnographic fieldwork and second, because they participate in OLAC's directory of language resources. Bird and Simons [3] suggest mainstreaming "language archives by replacing OLAC's parochial metadata format with a generic application profile, steering OLAC and the cataloguing of language resources into the library and information systems mainstream".

We will discuss possible ways forward for the community of language archives and report on developments in the OLAC harvester.

The Torwali Revitalization Program

Zubair Torwali
Idara Baraye Taleem wa Taraqi (IBT)
Bahrain Swat

A Dardic language of the Indo-Aryan family within the Indo-European, Torwali is a small language spoken in the Bahrain and Chail areas of District Swat in Northern Pakistan. According to some estimates the Torwali people count themselves more than 120,000 (Manan, Channa, Tul-Kubra, & David, 2021) while recent research counts the number of speakers of the Torwali language around 130,000 (Lunsford, Sagar, Ahmad, & Haider, 2021). 'Possibly half of them live in the heartland, which is located in northern Pakistan, in the Swat River Valley in the Khyber-Pakhtunkhwa province (Lunsford, Sagar, Ahmad, & Haider, 2021)'.

A Swat based organization in Pakistan, Idara Baraye Taleem wa Taraqi (IBT), that is constituted of researchers, writers, teachers and activists has been implementing a rigorous revitalization program for the Torwali language. This program is bottom-up and is immensely holistic for it has not only focused the children but also the youth and adults are targeted via this program. On the other hand, this program has focused many domains where Torwali is used. These domains include informal school education in Torwali for children, enhancing literacy of Torwali in the state-owned schools, using social media for writing Torwali, using audio visual media for Torwali; and by fostering the Torwali identity among the youths. The program also includes developing and promoting the Torwali folk poetry, folk music, and cultural events.

Rigorous work on the revitalization of Torwali language has been carried out since 2004 and under this program a mother-tongue-based 'MLE program that was established in 2005 by IBT, a registered community-based organization, which includes a large volume of locally produced curriculum and resources, two glossaries (not full dictionaries) have been produced and published by Torwali speakers' (Lunsford, Sagar, Ahmad, & Haider, 2021).

So far considerable written material has been produced in Torwali which include books on the folk Torwali poetry, folktales, daily usage trilingual book, small dictionaries for students and 'also several biographies about well-known historical figures which were written by respected scholars in Urdu have been translated into the Torwali language' (Manan, Channa, Tul-Kubra, & David, 2021). The researchers associated with the organization, Idara Baraye Taleem wa Taraqi (IBT) has also produced six music videos that portrayed traditional Torwali musical songs that had been adapted to incorporate some contemporary sounds. These videos are watched and liked widely, even among the Torwali diaspora, and greatly appreciated by Torwali speakers everywhere. A local cable TV station, which includes some Torwali programming, was established with help and support from IBT. Most of the activities described above have happened since the year 2006.

I intend to present a report on this model of language revitalization wherein the indigenous community-led organization has greatly revitalized not only the language but also the ethnic identity of the people.

Works Cited

Lunsford, W. A., Sagar, M. Z., Ahmad, E., & Haider, A. (2021). The Guide" in Six Speech Communities of Northern Pakistan. In D. M. Eberhard, & S. A. Smith (Eds.), *Planning Language Use Case Studies in Community-Based Language Development* (1st ed., pp. 76-121). SIL International.

Manan, S., Channa, L. A., Tul-Kubra, K., & David, M. K. (2021, April 2). Ecological planning towards language revitalization: The Torwali minority language in Pakistan. *International Journal of Applied Linguistics, 31*(3), 438-457. doi:https://doi.org/10.1111/ijal.12340

Collection replication for enhanced re-usability and preservation planning

Paul Trilsbeek, Kavon Hooshiar, Antonina Werthmann, Andreas Witt

The Language Archive (TLA) is one of the first Language Archives created after Himmelmann (1998), housing the DOBES collections of endangered language data. The development of the metadata schema and access restrictions for TLA represent innovations that have now been adopted and streamlined by newer DELAMAN archives. In this paper, we present an ongoing collaboration between TLA and the Leibniz Institute for the German Language (IDS) to improve the security, accessibility, and discoverability of DOBES collections.

TLA will replicate collections to IDS to create a novel layer of institutional redundancy to the data, which goes beyond redundancy of location and media. We review details and challenges of planning this infrastructure, including: maintaining the integrity of the data, metadata, and access restrictions of the original collections; renegotiating relationships with depositors; legal constraints; and automating digital infrastructure across institutions.

We aim to improve the accessibility and discoverability of the data by:
a) Asking depositors to review access restrictions, and to provide: secondary contacts; criteria for who should have access; and criteria for who should be allowed to moderate access requests in their absence.
a) Enhancing and standardising documentation for each collection, encompassing annotation formats and conventions.
c) Converting structured text transcriptions into the TEI-based ISO standard format (2462:2016), enabling easier automatic migration to diverse formats and usability across different disciplines.
d) Integrating the data into Germany's National Research Data Infrastructure (NFDI) via Text+.

In this context, the project focuses on questions like: What happens to restricted data once depositors are no longer available to moderate requests? How can we insulate the data from unforeseen institutional changes? This project seeks to plan for the future of these data by improving their accessibility, discoverability, and security, thereby benefiting scholars across various disciplines and contributing to the long-term preservation of global linguistic diversity.

Converting Historical Language Texts into Structured Data Using Multimodal Models

Daan Van Esch

Recent multimodal models like GPT-4[1] and Gemini 1.5 Pro[2] can accept a mixture of text and image inputs, and can be given instructions in natural language to describe what actions should be taken with these inputs. Among other use cases, these kinds of multimodal models are starting to be used in various settings to extract structured data from images. In a language documentation and revitalization context, this capability may also enable direct conversion of e.g. images/scans of historical dictionaries into structured XML or JSON files – with an implicit OCR step happening in between.

For example, in initial experiments, Gemini 1.5 Pro can create a JSON file based on a photo of a dictionary page, with different types of linguistic content correctly extracted into different fields, yielding a machine-readable structured format. Such conversion may be helpful in settings where no digital text version of the historical dictionary or linguistic resource is available, which is a common occurrence with legacy data[3].

This brief overview will discuss a few initial experiments using multimodal models that are publicly available today, and discuss pros and cons of this approach compared to cascaded approaches like (1) OCR + post-editing + structured data extraction, or (2) manual data entry. Preliminary experiments so far look promising, but some issues remain challenging, e.g. handling diacritics and successfully getting models running on a laptop/workstation locally.

---

[1] OpenAI 2023, *GPT-4 Technical Report* (arXiv:2303.08774 [cs.CL]).
[2] Gemini Team 2024, *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context* (arXiv:2403.05530 [cs.CL]).
[3] See e.g. Nick Thieberger and Conal Tuohy 2017, *From Small to Big Data: paper manuscripts to RDF triples of Australian Indigenous Vocabularies* (ComputEL 2017) and Nick Thieberger 2016, *Daisy Bates in the digital world* (in *Language, land & song: Studies in honour of Luise Hercus, edited by Peter K. Austin, Harold Koch & Jane Simpson*. London: EL Publishing. pp. 102-114).

Multi-sited dictionary compilation for Sonsorolese

Vasiliki Vita, Lucy Pedro, Tiffany Pedro

This paper focuses on a multi-sited dictionary compilation process for Sonsorolese, a language spoken in the Republic of Palau, Micronesia. From wider Micronesia to Taiwan and Japan, to Hawaii and the USA (Walda-Mandel 2016), speakers of Sonsorolese journey with their languages and cultures as a compass to navigate life. In a Participatory Action Research theoretical framework (Kemmis & McTaggart 2005), following Sonsorolese values of respect, relationships and flexibility (Walda-Mandel 2016) we engaged in a process for multi-sited dictionary compilation. This occurred in a hyper-collaborative model of working within an expanded network (Vita et al. 2022) with the Young Historians of Sonsorol (YH), a youth group focusing on the preservation of Sonsorolese languages and cultures, and collaborators. Multi-sited does not refer only to a) different geographical locations (Marcus 1995), in our case, Echang and Dongosaro, but also b) modalities (online and in person), c) individuals (older and younger, local and external researchers), and d) practices for data compilation (collaborative and individual efforts). Thus, by describing the process of the Sonsorolese-English dictionary initiated by YH and funded by the Endangered Languages Documentation Programme (ELDP) (Vita et al. 2023), we aim to show how shifting the perspective of dictionary compilation during language documentation from data input to a teaching and learning process can a) be inclusive, b) validate speakers' desires and c) produce materials that are accessible to a variety of users by potentially tapping into networks of diaspora speakers.

Keywords: language documentation, multi-sited, dictionary, Sonsorolese, Micronesia

References

Kemmis, Stephen & Robin McTaggart. 2005. Participatory action research: Communicative action and the public sphere. In *The Sage handbook of qualitative research , 3rd edition*, 559–604. Thousand Oaks: SAGE Publications Ltd.

Marcus, George E. 1995. Ethnography in/of the World System: The Emergence of Multi-Sited Ethnography. *Annual Review of Anthropology*. Annual Reviews 24. 95–117.

Vita, Vasiliki, Chelsea Pedro, Lincy Lee Marino, Daphne Nestor & Young Historians of Sonsorol. 2023. Collaborative corpus building for Sonsorolese. Endangered Languages Archive. http://hdl.handle.net/2196/2de60f8b-676c-4b6f-c40c-417242964d7h.

Vita, Vasiliki, Sydney Rey, Leonore Lukschy & Pierpaolo Di Carlo. 2022. Hyper-collaboration: Creating networks beyond usual suspects. In *Where Do We Need to Go from Here? Language Documentation and Archiving during the International Decade of Indigenous Languages*. Berlin and online.

Walda-Mandel, Stephanie. 2016. *"There Is No Place Like Home" : Migration and Cultural Identity of the Sonsorolese, Micronesia* (Heidelberg Studies in Pacific Anthropology). Heidelberg, GERMANY: Universitätsverlag Winter. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=nlebk&AN=2041180&site=ehost-live. (19 January, 2024).

A story of stories: how Standing Rock created its digital language archive

Nacole Walker M.Ed.
Elliot Bannister M.Ed., Tasha Hauff Ph.D.

Wóoyake (meaning “stories”) is a digital, searchable, user-friendly archive comprised of recordings made by fluent speakers of Dakȟóta/Lakȟóta (a north american Siouan language). It was created by the Standing Rock Sioux Tribe in collaboration with other communities of the Očhéthi Šakówiŋ (Seven Council Fires). This ever-expanding community-led project began in response to the need for access to authentic Dakȟóta/Lakȟóta language – that is, language used by fluent speakers in real life rather than pedagogical materials. Accessible at wooyake.org, it contains digitized audiovisual recordings and written texts going back nearly 200 years. These are being cataloged, transcribed, and translated so that the Očhéthi Šakówiŋ community can search for them and explore the historical and social connections between them.

The project is based on our unique constellation of needs and has been community-led in all aspects: visioning, planning, grant writing, project management, web development, and community implementation. Wóoyake has become a powerful tool for research of all kinds, including genealogical, historical, cultural, linguistic, and more. It has been applied in classrooms where instructors use its content to better facilitate language transmission. Wóoyake diversifies resources for learners by removing the barriers preventing engagement with Dakȟóta/Lakȟóta language material while, at the same time, upholding Indigenous data sovereignty by prioritizing traditional access protocols over academic pursuit.

Wóoyake represents a new standard for preserving endangered languages and cultures through innovative technology. This presentation will narrate Wóoyake’s inception and showcase ways it is being used to develop Dakȟóta/Lakȟóta language, culture, and linguistic understanding among the Očhéthi Šakówiŋ and their allies.

Empowering Community-based Language Documentation and Revitalization Efforts: Approaches by 7000 Languages

Stephanie Witkowski

In the landscape of language revitalization efforts, communities facing language endangerment often encounter challenges in accessing essential resources and expertise. Academic institutions, while repositories of valuable knowledge, may be inaccessible due to historical and socioeconomic factors, leaving many Indigenous communities without vital support for their revitalization endeavors. Consequently, external assistance, albeit sometimes effective, can lead to dependency and limited community input in program development.

Our initiative, 7000 Languages, a US-based nonprofit serving communities internationally, aims to address these gaps by prioritizing community-led efforts in language revitalization. Through direct engagement with communities, we provide expertise, resources, and support, particularly in the realm of language learning technology. Our approach emphasizes the training and development of community leadership, ensuring sustainable, internally driven revitalization efforts.

In this presentation, we will showcase our partnership models, highlighting successful collaborations between our initiative and community-based organizations as well as discuss Indigenous data rights and access in this space. These partnerships include our Fellowship program, which empowers individuals to create resources for their language communities within weeks of initiation, and ongoing community collaborations. By sharing our experiences, we aim to demonstrate pathways to partnership that foster long-term success without perpetuating dependency on external support or extractive processes.

Through our participation, we contribute to discussions on the impacts of documentation and archiving, showcasing how archived materials are utilized by Indigenous communities to create online language learning materials. Additionally, we engage with the planning and design aspects for the future decade, focusing on indigenous-led projects, accessibility solutions, and broader public engagement initiatives.

We hope our presentation will contribute to the conference's goal of sharing knowledge, developing capacity, and supporting communities in their efforts to record and revitalize Indigenous languages.

Records for language and history:
Western Nahuatl Documents from 1557—1750

Rosa H. Yanez Rosales

During the colonial period of Mexico (1521-1821), language policies promoted by Spain, included writing in several of the Mesoamerican languages. Scribes used the alphabetic system to write wills, petitions, land measurements, accusations, etc.

In recent decades, more and more texts have been found in parochial, state, and national archives. The languages recorded are mostly Nahuatl, Zapotec, Mixtec, Caqchiquel, Maya, Purhépecha, and others. There are websites where one can find the manuscript alone, or even the transcription and translation. The philological analysis of such texts has allowed us to become aware not only of language change, but also of facts that bring us closer to the daily life of the communities, facts that are rarely narrated in the sources written by the Spanish crown or church representatives.

Regarding the western part of New Spain, so far, only documents in Nahuatl have been found. There are two dialects clearly identified in the documents where Tlajomulco-Guadalajara, stands as a boundary: a northern one that runs from Tlajomulco-Guadalajara towards Huaxicori, Nayarit, and a southern one from Tlajomulco-Guadalajara towards Ostula, Michoacán.

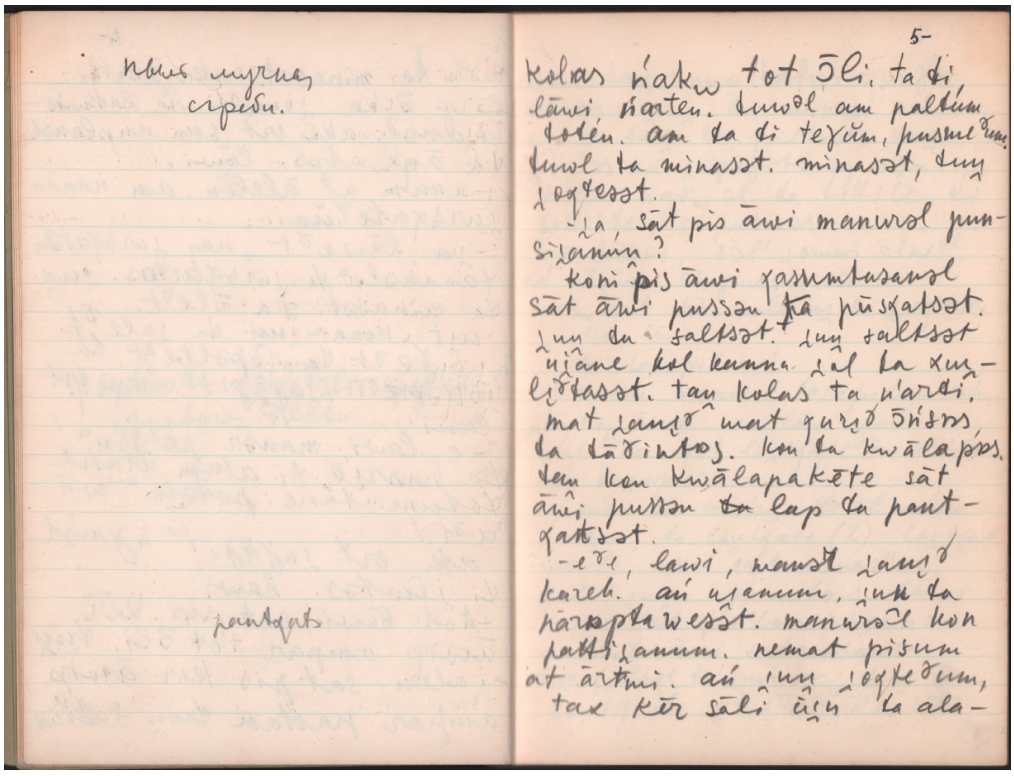
This paper is about Nahuatl texts written between 1557-1750, within the western region of Mexico. It shows how such documents recorded language variation and change, and the concerns and reclamations of small Indigenous communities. The documents are a valuable tool for reconstructing both: the similarities and differences between western and other Nahuatl dialects, and the course of events that took place in the communities. The paper highlights the importance of archival work, the training of linguists and historians to deal with this type of records, to develop data bases that will allow communities to learn about their history, scholars (sociolinguists, dialectologists, historians) to propose hypothesis about reclamation and agency within the communities, and about language variation and change.

Mansi archival data: processing workflow and community involvement

Daria Zhornik

The Mansi languages are an Ob-Ugric (< Finno-Ugric < Uralic) language subgroup in Western Siberia, of which Northern Mansi is the only existing variety with approximately 1200 speakers according to the 2020/2021 national census. In this talk, I will present the workflow and results of a 2019-2021 project on processing of Valery Chernetsov's archive (1925-1938) with data from Northern and Eastern Mansi varieties. The linguistic part of the archive consists of 16 notebooks, approximately 1.000 pages in total, with untranslated texts in Northern Mansi, written in pencil by hand with very few commentaries. 15 notebooks contain Northern Mansi data, while one notebook has materials in Eastern Mansi.

Fig. 1. A Northern Mansi text from Valery Chernetsov's notebook No. 50 (1938)



In my talk, I will describe the difficulties the project team encountered in gaining access to the archive's materials. Then, I will present the workflow of various stages of data processing: digitization, translation, annotation. I will describe in detail the translation stage, in which three Northern Mansi native speakers were most actively involved. For one of the speakers, this work became a full-time job for two years and it was carried out in close interaction between linguists and community members. I will present the current state of the processed data and plans for its further usage. For example, in Khanty-Mansi Autonomous Okrug, which is one of the main administrative areas of Russia where Northern Mansi is spoken, a project of a Northern Mansi online-translator is being carried out. There are plans to expand it into a project of a national Northern Mansi corpus, and the native speakers have great interest in including the data from Valery Chernetsov's archive into that work, as well as other materials processed by the same team.

TermLog: Bridging Linguistic and Technical Divides Through English-Nigerian Indigenous Language Glossaries

Adesina Ayeni

The ongoing TermLog metalanguage project aims to fill a critical gap by developing a collaborative and community-led glossary of technical terms spanning various domains, disciplines, and subject matters in both English and Nigerian indigenous languages. With a focus on supporting Nigerian language development and archival efforts, TermLog endeavors to bridge linguistic gaps and foster inclusivity within the digital landscape.

TermLog serves as a dynamic repository of terminology, facilitating seamless communication and knowledge dissemination across linguistic boundaries. By collating and standardizing technical terms in English and Nigerian indigenous languages, the project strives to empower native-language communities.

Key features of the TermLog project include:

Linguistic Diversity: Embracing the rich linguistic diversity of Nigeria, TermLog encompasses a wide array of indigenous languages, including Hausa, Igbo, Yoruba, and more. Through meticulous research and collaboration with linguistic experts, the project ensures the accurate translation and contextualization of technical terms in each language.

Domain Specificity: Recognizing the diverse fields of knowledge and expertise, TermLog caters to multiple domains, including science, technology, engineering, mathematics, the humanities, and beyond. By providing domain-specific glossaries, the project facilitates language development and archival efforts across various disciplines.

Community Engagement: TermLog actively engages with language enthusiasts, translators and interpreters, educators, and researchers to solicit feedback and contributions. Through participatory approaches, the project fosters a sense of ownership and collective responsibility towards preserving indigenous languages.

Technological Integration: Leveraging digital technologies, TermLog employs user- friendly interfaces to enhance accessibility and usability. At present, TermLog is a web-based platform, but it has the hope of migrating to mobile application technology to democratize access to linguistic resources and promote language revitalization efforts.

In conclusion, the TermLog metalanguage project represents a significant step towards advancing Nigerian language development and archival initiatives.

Sharing archives: A test case with Zotero

Prof. Claire Bowern, Seth Katenkamp

Language documentation collections are often heterogeneous, with many genres, collectors, time periods, communities and formats. This creates challenges for archiving and the larger the collection, the more difficult organization can be. Such collections must also be accessible to community members who also use materials in many different ways (Burke et al 2021, 2022). Additional challenges are created when collections are under continued development and in active use for research and education. Bird & Simons' (2003) dimensions of "portability" include adequate citation, sharing, and searchability, in addition to long-term storage. In this talk, we explore Zotero (zotero.org) as a partial solution to these issues.

Zotero is open source bibliographic software for storing, cataloguing, and sharing bibliographic items. It has a prespecified and user-entered metadata options. Reference collections can be public or restricted, and the digital primary sources can be stored and shared in multiple formats. It is fully searchable and uses a combination of folders, tags, relationship definitions, and free-text for organization and searchability. Users can annotate records or items and collaboratively upload materials; these can be hosted through WebDAV or elsewhere. Zotero is not recommended for long-term preservation (that is, it does not replace digital archives) but for working with collections under active development from multiple sources, it is a stable and promising resource (cf. LOCKSS; Bird & Simons 2003). We illustrate this talk with examples from a collection under active development (language redacted for anonymity). Items are organized in primary folders by genre (stories, fieldnotes, songs, school-materials, publications, dictionary) and secondarily by researcher. Items can be filtered by speaker and are tagged for attributes. Materials from the same session (e.g. audio, transcript) are linked. Zotero's ability to export annotations on pdfs has also been a welcome way to link photographs of handwritten fieldnotes with digital materials in other formats.

Family networks and maintaining the Indigenous language in Paiwan communities

Chun-Mei Chen

This paper examines how strong family networks contribute to the maintenance of the Indigenous language in the Paiwan families, the aboriginal people of Taiwan. The Paiwan language is one of the 16 government-recognized Austronesian languages spoken in southern Taiwan. The Indigenous communities use Mandarin, Paiwan, Taiwanese (Minnan), and Hakka. Elementary schools in the communities teach the Paiwan language, while schools introduce the Indigenous culture in Mandarin. The local epistemic ecology shows the dominance of Mandarin. Family networks (Velázquez 2008, 2013), language use in family activities, and Paiwan language proficiency were assessed based on language documentation and collaboration with 12 Paiwan family members to understand the retention of the Indigenous language in the communities. Specifically, Paiwan epistemological conversations in Indigenous families require further attention because of the social implications of family language maintenance in the communities.

Through micro-analysis of the linguistic practices of Paiwan family members in the communities, this study examines how Paiwan people locate and realize the knowledge status of family networks and family members. The intertwining of the Paiwan family networks and language policies is studied from the perspective of interactive talks. In addition, local schools have established sequential structures in the communities of the nature of language practices. The findings suggest that strong family networks and Paiwan-Mandarin translanguaging interpreting activities ensure children's well-being and Indigenous identity, and that Indigenous languages are passed on from generation to generation. On the other hand, networks and spaces of family activities are crucial in interpreting the language practices and socialization of Paiwan family members. The findings also reveal the family networks of language socialization of multilingual Indigenous children, as well as the situational nature of different values and resources in the Indigenous communities.

The Atchan Song and Story Corpus: Towards a sustainable, adaptable online resource

Yao Maxime Dido, Siddharth Ganapathy, Lindsay Hatch,
Rebecca Jarvis, Katherine R. Russell, Marie-Anne Xu

This presentation introduces the Atchan Song and Story Corpus (Recueil de contes et chants Atchan/ÁCAN NANMƐ LÊ ÁLɛ́ BHƆ́), a community-oriented website presenting glossed and translated songs and narrative texts in Atchan (Kwa; spoken on the Ébrié Lagoon, Côte d'Ivoire). This project is a collaboration between researchers at UC Berkeley (USA) and University of Alassane Ouattara (Côte d'Ivoire), including one native speaker of Atchan. The website, currently under development, has two broad aims: to host recordings and transcriptions of songs and narrative texts, and to promote the recently-developed Atchan orthography.

This presentation focuses on the workflow involved in managing and adding to the website. The primary data is hosted in two places: recordings are hosted on YouTube (with storytellers'/ singers' consent), and transcriptions and translations are hosted in the online database Twisted Tongues (Ewert 2015). To support the generation and population of the website, we have written a Python script that converts text from TSV to JSON format, and a JavaScript program that generates the actual website. The script is adapted from the Moro Database script (Sande et al. 2016), with additional support for a bilingual interface (FR/EN), both IPA and orthographic representations, and video embedding. The website will be compiled via GitHub Actions, and the source code will be made publicly available. We expect the website to be live by summer 2024.

The workflow is designed to support community involvement in the website, minimize technical knowledge needed to contribute to it, and promote longer-term sustainability: Twisted Tongues can be used online, and any appropriately-formatted TSV can be an input to the website. Because the source code will be public, the workflow could also be straightforwardly adapted to support use in other communities.

Translating typologically unusual structures in linguistics examples: A survey and evaluation of translation strategies used in the case of western Austronesian voice

Holly Drayton

Within Language Documentation, annotation of data (e.g. glossing and translation) is seen as peripheral to data collection and analysis (Beier 2015). Many documentary texts have been noted to be unsuitable for use beyond the original research project, due to poor-quality translation (Evans & Sasse 2007; Penfield & Tucker 2011; Austin 2017: 35).

This paper explores the question of how syntactic structures are translated when there is no equivalent in the target language. *Western Austronesian Symmetrical Voice* provides an ideal testing ground since analysis of these structures is controversial, particularly for the Undergoer or Patient voice, which has been proposed to be a Passive, Ergative and Symmetrical construction, each giving rise to a different translation.

Examples of English translations of two varieties of Malay/Indonesian: Desa and Standard Indonesian are shown below, presenting translations where similar UV constructions are translated as both English active (1a), passive (1b) constructions and active sentences with fronted theme arguments (2a-b).

Example 1 UV translations from Desa

a. Meja Yetn Aku Teipel
table DEM 1sg Touch
'I touch the table.'
(Erlewine & Sommerlot 2023: 9)

b. Buku Itu Ku-beli
Book DET 1sg-buy
'The book was bought by me.'
(Erlewine & Sommerlot 2023: 10)

Example 2 UV Translations from Standard Indonesian

a. Buku Itu Saya baca
book DEM 1sg read
'The book, I read.'
(Arka & Manning 2008: 2)

b. Buku Itu Ku-baca
Book DET 1sg-read
'The book, I read.'
(Arka & Manning 2008: 2)

This paper draws on concepts from Translation Studies: Pedersen's taxonomy of strategies for cross-cultural translation (Pedersen 2011) and The Agency of the Translator (Tymoczko 2007; Venuti 2018) to closely examine 30 texts (10 archival deposits, 10 descriptive grammars, 10 journal articles) to determine:

(i) The translation strategies used for translating Western Austronesian voice constructions into English;
(ii) To what extent these strategies are transparent.

References

Arka, I. Wayan & Manning, Christopher. 2008. Voice and grammatical relations in Indonesian: A new perspective. In Peter Austin & Simon Musgrave (eds.), *Voice and grammatical relations in Austronesian languages*, 45–69. Stanford, Calif: Center for the Study of Language and Information. Retrieved from http://hdl.handle.net/1885/26154

Austin, Peter K. 2017. Language documentation and legacy text materials. *Asian and African Languages and Linguistics* 11. 23–44.

Beier, Christine. 2015, 04. *Text translation in the context of endangered language documentation: The case of Iquito*. Conference presentation, Fieldwork Forum, Department of Linguistics, University of California, Berkeley. Retrieved from http://www.cabeceras.org/beier_translation_fforum_20150423.pdf

Erlewine, Michael Yoshitaka & Sommerlot, Carly. 2023. *Voice and extraction in Malayic*.

Evans, Nicholas & Sasse, Hans-Jürgen. 2007. Searching for meaning in the Library of Babel: Field semantics and problems of digital archiving. *Language Documentation & Description* 4. 58–99.

Pedersen, Jan. 2011. *Subtitling Norms for Television: An exploration focussing on extralinguistic cultural references* Vol. 98. Amsterdam: John Benjamins Publishing Company. DOI: https://doi.org/10.1075/btl.98

Penfield, Susan D. & Tucker, Benjamin V. 2011. From documenting to revitalizing an endangered language: where do applied linguists fit? *Language and Education* 25(4). 291–305. DOI: https://doi.org/10.1080/09500782.2011.577219

Tymoczko, Maria. 2007. *Enlarging translation, empowering translators*. Manchester, UK ; Kinderhook, NY: St. Jerome Pub.

Venuti, Lawrence. 2018. *The translator's invisibility: a history of translation* Third edition. New York: Routledge.

Collaborative efforts for indigenous languages documentation and revitalization in Paraguay

Celeste Escobar

Multimedia resources were developed for Maká (Mataguaya), Paĩ Tavyterã Guaraní (Tupí-Guaraní) and Ayorean (Zamucoan) in indigenous communities in Paraguay (lower Amazonian Basin) from 2021 until the present. This presentation focus on the collaborative work to propose the development of contents for learning and revitalization materials in multimedia formats in and with the involved communities. Documenting the language with different levels of grammar content for communicative use was one of the main focus. The approach for the elaboration of resources for revitalization purposes was applying "multimodality" (Kress 1997; Stein 2007; Boch 2016), from which, the way the new generations of indigenous languages native speakers learn, communicate and express themselves using different means. At the same time, these materials are meant to be used in hybrid learning environments in the face of the rapid displacement and jeopardy of losing the richness of their cultural diversity embedded in the languages of this work. These experiences also are meant to show the contextualized approach according to each case taking into consideration that though they are all endangered languages, their sociolinguistic contexts may affect the type of documentation and revitalization process. However, what they all share in common is that native speakers were actively involved and constantly consulted in the process of documenting and generating the cited multimedia resources.

Key words: Indigenous languages. Endangered languages. Learning resources. Revitalization. Collaborative work.

Empowering under-resourced communities with AI-illustrated picture books: An initiative in Palauan

Orlyn Esquivel

Quality children's literature, as an art form, combines carefully crafted language, expressive images, and sensitive design (Kiefer, 2008). Many examples of quality children's literature are picture books. Through these books, children develop their understanding of the world and its people (Albers, 2009) and an understanding of their place within it (Driggs Wolfenbarger & Sipe, 2007). Particular features of picture books, such as illustrations, hold importance in illuminating texts and meaning. These visual cues not only serve to capture children's attention but also aid in interpreting complex concepts and understanding narrative themes (Greenhoot et al., 2014). Therefore, producing quality illustrations for children's literature is an enormous responsibility. However, high-quality picture book art is primarily associated with commercially produced books. Commercial publishers operate on a for-profit basis, prioritizing picture books that have the potential to generate revenue. Hence, commercially available picture books are often published in dominant languages, causing a scarcity of quality literature for smaller communities.

Artificial intelligence (AI) has recently made its mark in the landscape of creating children's literature. The once-labor-intensive task of illustrating a children's book has now become significantly easier with the assistance of AI technology. Quality picture book art and illustrations, commonly found in commercially produced picture books, can now be generated through natural language descriptions, called prompts. This advancement from AI models offers transformative possibilities for creating and distributing quality picture books for under-resourced communities.

Thus, this project showcases five bilingual picture books written in Palauan and English for preschool children. This includes translating existing English materials and adapting local Palauan stories and songs with the assistance of a native Palauan speaker. I used Midjourney, a generative AI program, to create appropriate and culturally relevant illustrations for the books. I also compared three leading AI models, DALL-E, Midjourney, and Leonardo Ai, to explain why I chose Midjourney for this project.

Generally, the goal is to facilitate the preservation of Palauan language and culture and provide Palauan children with access to quality picture books. With high hopes, this project will serve as a model to motivate the creation of AI-powered picture books for Indigenous communities grappling with educational disparities and underrepresentation in literature.

Keywords: artificial intelligence; under-resourced communities; Palauan; picture books

References

Albers, P. (2009). Reading students' visual texts created in Language Arts classrooms. *Language Arts Journal of Michigan, 25*(1), 6-16.

Driggs Wolfenbarger, C., & Sipe, L. R. (2007). A unique visual and literary art form: Recent research on picture books. *Language Arts, 84*(3), 273-280.
Greenhoot, A. F., Beyer, A. M., & Curtis, J. (2014). More than pretty pictures? How illustrations affect parent-child story reading and children's story recall. *Frontiers in psychology, 5*, 76510.

Kiefer, B. (2008). What is a picture book, anyway? The evolution of form and substance through the postmodern era and beyond. In L. R. Sipe & S. Panteleo (Eds.), *Postmodern picture books: Play, parody and self-referentiality* (pp. 921). New York, Routledge.

Change Over Time: Ethnobotanical Data in Linguistic Documentation Dictionaries from 1960-2020

Lauren Hall

Linguistic documentation practices have undeniably evolved over the years. Such change has largely been in response to developments like advancement of technologies and the evolution of academic and cultural values which then get reflected in fieldwork methods employed. It stands to reason that typically non-linguistic topics of cultural value can be reflected in linguistic documentary work. This study focuses on the socio-political movement of environmentalism which gained popularity in the 1960s and turned greater attention to preservation of natural resources and biodiversity. The effects of such a long-lasting and globally recognized movement are frequently analyzed in relation to preservation of biodiversity and the climate crisis, but the effect of this movement on research like linguistics is not widely known. There is, however, ample literature on what linguists should include in documentations, such as the inclusion of the binomial nomenclature of plant species when possible. However, more recent guidelines emphasize including use(s) of the plant, taxonomy, and cultural significance as well. An assessment of the application of such documentary practices has not been widely observed. This research explores linguistic documentation dictionary works from 1960 to 2020 to assess efforts made in the past six decades to include data deemed relevant and important for scientific identification and cultural value. The primary method of analysis consists of assessing trends of what types of information are typically included or excluded in linguistic dictionary works as well as if factors like year of publication or type of publication affect inclusion of plant-related words and information. Being aware of any change over the past six decades will help to identify any room for improvement to implement in future linguistic documentation efforts and to thus help preserve linguistic diversity in the long term.

Mobilising the Archive: Training Modern Speech Technology Models with Digitalised Fieldwork Recordings

William Havard, Emmanuel Schang, Benjamin Lecouteux

Over the years, community members and linguists have recorded speakers and peers in the field to formally study their languages and write grammars, and to preserve cultural knowledge. Up to now, most of the gathered recordings are archived and remain untranscribed. They are therefore impossible to index and navigate, as indexing and navigation rely on the existence of transcriptions, and remain unsearchable (and potentially unusable) for both community members and linguists.

In our work, we leverage the power of modern self-supervised speech-processing tools (wav2vec, Baevski et al. 2020) and the existence of archival material. We pre-trained self-supervised models of speech processing on digitalised fieldwork recordings (350h) in Haitian Creole, collected 40 years ago in Haiti and digitalised by the French National Library. We further train the models on a speech recognition task, and obtain competitive results on fieldwork material (24.1% character error rate, CER) and read speech (15.2% CER), with models requiring only 40 minutes of transcribed speech to be trained.

To the best of our knowledge, our work is the first that only uses fieldwork recordings to train state- of-the-art speech processing models at every step of the training process. We show that old fieldwork recordings, that were not collected for computational applications, can be repurposed and used to train speech recognition models. We conclude that the 'mobilising the archive'-approach advocated by (Bird, 2020) is a promising way forward to design speech technologies for new languages, and make archival material accessible both for community members and linguists. In future works, we would like to explore query-by-example approaches that would leverage the need for transcriptions altogether and allow users to query and navigate the archive by simply pronouncing a key word.

References

A. Baevski, H. Zhou, A. Mohamed, and M. Auli. Wav2vec 2.0: A framework for self-supervised learning of speech representations. In Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

S. Bird. Decolonising speech and language technology. In D. Scott, N. Bel, and C. Zong, editors, Proceedings of the 28th International Conference on Computational Linguistics, pages 3504–3519, Barcelona, Spain (Online), Dec. 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.313. URL https://aclanthology.org/2020.coling-main.313.

Digitally Preserving Dialectal Expressions: Techniques for Sustaining Linguistic Diversity

Pu Meng

This study addresses the urgent need for preserving the unique expressions found within dialects, which are an integral part of linguistic diversity and cultural heritage. With the rapid decline of many dialects due to globalization and language standardization, specialized expressions that carry cultural, historical, and social significance are at risk of disappearing. The research aims to explore and optimize digital archiving and documentation techniques to ensure the sustainable preservation and organization of these expressions for future linguistic research and cultural heritage conservation.

The primary data source for this study includes a collection of dialectal expressions from various regions, identified through fieldwork and literature review. Employing digital humanities methodologies, the study integrates linguistic annotation tools, audio-visual recordings, and metadata schemata to document and archive these expressions in an accessible and interpretable format.

Through qualitative analysis, the study identifies key challenges in dialectal expression preservation, such as regional variations, the lack of standardized orthographies for many dialects, and the need for context in understanding expressions. The research then evaluates different digital archiving solutions, including open-access repositories and interactive databases that incorporate multimedia elements to convey the full richness of expressions.

Results indicate that a multimodal documentation approach, combining text, audio, and visual data, significantly enhances the comprehensibility and accessibility of dialectal expressions. Moreover, community involvement in the documentation process emerges as crucial for ensuring the authenticity and contextual accuracy of the archived expressions.

The study contributes to the field of linguistic preservation by proposing a model for digital archiving that prioritizes ease of access, the inclusion of diverse data types, and community participation. This model aims to serve as a blueprint for future initiatives focused on safeguarding dialectal diversity.

References:

- Auer, P., & Schmidt, J. E. (2010). Language and Space: An International Handbook of Linguistic Variation. De Gruyter Mouton.
- Bird, S., & Simons, G. (2003). Seven dimensions of portability for language documentation and description. Language, 79(3), 557-582.
- Crystal, D. (2000). Language Death. Cambridge University Press.
- Himmelmann, N. P. (1998). Documentary and descriptive linguistics. Linguistics, 36(1), 161-195.
- Thieberger, N., & Barwick, L. (2012). Sustainable data from digital fieldwork. Sydney University Press.

A Corpus with Latin Poetry have you: The Lost OLAC Discourse Type

Bret Mulligan, Hugh Paterson Ⅲ

The Open Language Archiving Community (OLAC) and its DC-aligned metadata application profile are widely used by langage archives (Bird and Simons 2001; 2003; 2021; Simons and Bird 2003a; 2003b). Genre representation in digital libraries is not a well defined practice (Dragon 2020). Groundbreaking work in linguistic genre identification led by Johnson and Aristar-Dry (2012) resulted in the OLAC Discourse Types Vocabulary (OLAC-DTV) for cross-disciplinary language resource preservation and discovery. OLAC-DTV contrasts with both the concept of genre in literature studies (e.g., epic, tragedy, comedy, etc.) and the genre-and-form vocabularies often used within bibliographic records (e.g., MARC Genre Terms, Library of Congress Genre/Form Terms, etc.). OLAC-DTV is especially useful in applications requiring description of corpora or textual units within corpora.

OLAC-DTV has undergone several revisions (2002-11-21, 2002-12-17, 2003-01-27, 2006-04-06, 2012-02-04 → 2002-11-21), the latest being a reversion to the original proposal due to the accompanying XML/XSD not being maintained in step with the approved text. We maintain that: the management processes should have brought the XML/XSD file into alignment with the approved textual representation; and after intentional inclusion (Aristar-Dry and Sriram 2002) the term for 'poetry' was removed from OLAC-DTV between the 2002-12-17 and 2003-01-27 versions.

Using a digital library (archive) of Latin texts arranged for language-teaching (Author et al 2023) we show that poetry is important as a discourse genre and is relevant in language teaching as well as corpus based analysis. Wide consensus exists that Latin poetry and prose have distinct syntactic and other linguistic attributes (Pinkster 2021; Chaudhuri et al. 2019; Ferri 2011; Sciarrino 2011; Gale 2004) and should be treated appropriately when making corpora based claims about the language (Egbert, Biber, and Gray 2022; Biber 1993b; 1993a). Therefore, we argue that the term 'poetry' should be reinstated in OLAC-DTV.

References

Acquisitions and Bibliographic Access Directorate of the Library of Congress. 2024. "Library of Congress Genre/Form Terms."
https://www.loc.gov/aba/publications/FreeLCGFT/freelcgft.html

Author et al. 2023. Work discussing latin digital library. ACL.
Aristar-Dry, Helen, and Gayathri Sriram. 2002. "Linguistic Data Types & Discourse Types & Linguistic Fields." Slides

presented at the IRCS Workshop on Open Language Archives, Second OLAC Workshop, University of

Pennsylvania. http://www.language-archives.org/events/olac02/proceedings.pdf.
Biber, Douglas. 1993a. "Representativeness in Corpus Design." Literary and Linguistic Computing 8 (4): 243–57.

https://doi.org/10.1093/llc/8.4.243.
———. 1993b. “Using Register-Diversified Corpora for General Language Studies.” Computational Linguistics 19 (2):

219–41. https://www.aclweb.org/anthology/J93-2001.
Bird, Steven, and Gary F. Simons. 2001. “The OLAC Metadata Set and Controlled Vocabularies.” In *Proceedings of*

ACL/EACL Workshop on Sharing Tools and Resources for Research and Education, edited by Thierry DeClerck, Steven Krauwer, and Mike Rosner, 7–18. Université de Toulouse, France: EACL-ACL; elsnet. https://www.aclweb.org/anthology/W01-1506.

———. 2003. “Extending Dublin Core Metadata to Support the Description and Discovery of Language Resources.” *Computers and the Humanities* 37 (4): 375–88. https://doi.org/10.1023/A:1025720518994.

———. 2021. “Towards an Agenda for Open Language Archiving.” In *Proceedings of the International Workshop on Digital Language Archives: LangArc 2021*, edited by Oksana Zavalina and Shobhana Lakshmi Chelliah, 25–28. Denton, Texas: University of North Texas. https://doi.org/10.12794/langarc1851171.

Chaudhuri, Pramit, Tathagata Dasgupta, Joseph P Dexter, and Krithika Iyer. 2019. “A Small Set of Stylometric Features Differentiates Latin Prose and Verse.” Digital Scholarship in the Humanities 34 (4): 716–29. https://doi.org/10.1093/llc/fqy070.

Dragon, Patricia M. 2020. “Form and Genre Access to Academic Library Digital Collections.” Journal of Library Metadata 20 (1): 29–49. https://doi.org/10.1080/19386389.2020.1723203.

Egbert, Jesse, Douglas Biber, and Bethany Gray. 2022. Designing and Evaluating Language Corpora: A Practical Framework for Corpus Representativeness. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316584880.

Ferri, Rolando. 2011. “The Language of Latin Epic and Lyric Poetry.” In A Companion to the Latin Language, edited by James Clackson, 1st ed., 344–66. Wiley. https://doi.org/10.1002/9781444343397.ch20.

Gale, Monica, ed. 2004. Latin Epic and Didactic Poetry: Genre, Tradition and Individuality. Swansea: The Classical Press of Wales. https://doi.org/10.2307/j.ctv1n357vk.

Johnson, Heidi, and Helen Aristar-Dry. 2012. “OLAC Discourse Type Vocabulary.” Recommendation. Dallas, TX: Open Language Archive Community. http://www.language-archives.org/REC/discourse.html.

Network Development and MARC Standards Office. 2017. “MARC Genre Terms List”. https://www.loc.gov/standards/valuelist/marcgt.html

Pinkster, Harm. 2021. The Oxford Latin Syntax: Volume II: The Complex Sentence and Discourse. Oxford, New York: Oxford University Press.

Sciarrino, Enrica. 2011. Cato the Censor and the Beginnings of Latin Prose: From Poetic Translation to Elite Transcription. Columbus: Ohio State University Press.

Simons, Gary F., and Steven Bird. 2003a. “Building an Open Language Archives Community on the OAI Foundation.” *Library Hi Tech* 21 (2): 210–18. https://doi.org/10.1108/07378830310479848.

———. 2003b. “The Open Language Archives Community: An Infrastructure for Distributed Archiving of Language Resources.” *Literary and Linguistic Computing* 18 (2): 117–28. https://doi.org/10.1093/llc/18.2.117.

Indexing language data between libraries and archives

Heike Renner-Westermann, Tobias Weber, Ivana Vrdoljak

A deposit of language data has several purposes and meanings: it documents and represents language use for researchers and language communities alike, captures a snapshot of the global linguistic diversity, offers an empirical basis for linguistic work, and serves as a documentation of (academic) activity (e.g. Himmelmann 2006; Holton 2012; Chelliah 2021). For these reasons, depositors aim for their data to be as findable, accessible, interoperable, and reusable as possible (Wilkinson et al. 2016), while also paying attention to (indigenous) community needs (Carroll et al. 2020), with the ultimate goal of supporting discovery, citation, and reuse of their data sets.

This goal links archiving to practices in publication and academic discourse, which prominently takes place in book and journal publications where data sets are cited and analysed. Linguistic descriptions and underlying data are often kept in different infrastructures – archives and repositories for data and libraries for publications. Even if a deposit has elaborate technical and descriptive metadata, a librarian can only establish a link to a publication in a catalogue if they are able to locate and assess the data. This paper presents challenges, possible solutions, and opportunities from the perspective of a specialised library service for indexing linguistic publications and linking them with their data sets.

To our knowledge, no automatic approaches to linking data citations exist, links are mostly established and checked by a human. This is a work-intensive task, partially due to diverging metadata schemas and institutional requirements. At the same time, authority control can add benefits for depositors and users, which also covers controlled language nametags and classifications. These steps shall lead to increased searchability through semantic web technologies across library catalogues and repositories, as a goal of the current project.

Carroll, Stephanie Russo et al. 2020. The CARE Principles for Indigenous Data Governance. *Data Science Journal* 19 (1): 43

Chelliah, Shobhana L. 2021. *Why Language Documentation Matters*. Cham: Springer.

Himmelmann, Nikolaus P. 2006. Language documentation: What is it and what is it good for? In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of Language Documentation*. Berlin – New York: De Gruyter Mouton. 1-30.

Holton, Gary. 2012. Language archives: They're not just for linguists any more. In Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann, Dagmar Jung, Anna Margetts, and Paul Trilsbeek (eds). 2012. *Potentials of Language Documentation: Methods, Analyses, and Utilization*. Honolulu: University of Hawai'i Press. 111-117.

Wilkinson, Mark D. et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3: 160018.

Preserving Murut Bookan through Collaborative Documentation Initiatives in Sabah

Siak Bie Soh

Sabah, known for its rich cultural tapestry and linguistic diversity, hosts numerous ethnic groups and indigenous communities. However, many minority indigenous oral languages remain under-documented, particularly those of the Murut tribe. The Murut Bookan language is verbal and lacks sufficient documentation. The Murut Bookan community, numbering around 2,100 to 2,400 individuals, represents one such group. Interviews with over 50 Murut Bookan community members revealed that fluency in Murut Bookan is primarily preserved among the elderly, aged 70 to 86 years, with some influence of the Malay language observed among individuals aged 50 to 69. Concerningly, individuals below 50 years old, as well as younger parental generations in their 30s to 40s, exhibit minimal to no fluency in Murut Bookan. This demographic shift raises serious concerns about the vitality of the language and the sustainability of language documentation efforts within the community across various age groups. It highlights the importance of engaging and supporting indigenous communities in documentation initiatives and a pressing need to establish and nurture relationships among community members, local authorities, and researchers/academicians, refine collaboration approaches, and actively engage and support indigenous communities in collaborative and community-driven documentation projects. The significance of this presentation lies in advocating for the adoption of adaptive and flexible methodologies that can evolve in response to community input and evolving circumstances. Central to these efforts is the collaborative development approach, wherein community members are recognized as equal partners in shaping the direction and implementation of documentation projects, such as co-creating project objectives, methodologies, and outcomes that resonate with cultural values and address the unique needs and context of the community. The cultivation of relationships from the outset ensures that community voices are prioritized in decision-making processes and that indigenous knowledge systems are honoured and safeguarded against exploitation or appropriation.

Assessing linguistic legacy materials – socially embedded meaning-making beyond philology and hermeneutics

Tobias Weber

All acts of documentation are of intrinsic communicative nature, whereby the documenters (and their consultants) put observations on record that they consider important in subsequent communicative events. This can involve archiving, discussing, analysing, disseminating and proliferating these observations and making them accessible to other interested parties. Language documentation, thus, involves different layers of communication and meta-communication embedded in the artefacts as outputs of the documentation efforts. This paper deals discusses how we can engage with these different layers of communication, as well as the relevant literacies and competencies to access and interpret language documentation outputs, and linguistic legacy materials as a particular example.

Legacy materials are all artefacts of language documentation – language samples, stimuli, documents, metadata – where the documented events are too far from the present day for them to be recreated from memory. In many cases, documenters and/or consultants cannot be asked directly and the artefactually contained observations take primacy over memory and renarrations. This poses challenges in the reconstruction and contextualisation of the observation, rendering the re- assessment of legacy materials an interpretation that goes beyond the reading and deciphering of texts (hermeneutics) and a comparative contextualisation as in philology.

These challenges are not exclusive to legacy materials since learning from past documentation can inform present-day practices and guide reflections on meta-documentary linguistics, as all artefacts from current documentation projects will turn into legacy materials within the next 100 years. Current documentary practice places emphasis on thorough descriptive and technical metadata, as well as the contextualisation amidst (indigenous) processes of knowledge generation. Yet, to fully understand (legacy) materials, we would need access to the mental processes of the involved parties and socio-cultural (and historical) embeddings in the negotiation of meaning. Conceptualising language artefacts as discursive, communicative entities in social contexts keeps them as active parts in processes of meaning-making in the past, present, and future.

The Recent Shadow Puppet Theater of Pingjiang, Hunan, China

Shengkai Zhang, Xilin Li

In Pingjiang, shadow puppet theater is a famous traditional opera. Many people know it. It was popular until the end of the last century.

The Pingjiang shadow puppet theater has unique dialogue and vocal styles, which makes its lyrics difficult to understand, even for people over the age of 50 who are familiar with the local dialect and customs, and of course, for young people. Therefore, in recent years, shadow puppet theater has gradually declined.

Pingjiang shadow puppetry is very complicated. It employs three speech styles and two singing styles.

The first speech style mirrors the ordinary dialect spoken by the common person. The second speech style comes from written language, a speech style used by people of high class. The third speech style is a mixture of the first two styles. It is one that a common person might use when trying to talk to a high-class person.

The first singing style uses the written language of the Pingjiang dialect. The singing is accompanied by a *suona*, a local instrument. This style of singing is called the *suona* style. It is unique to the Pingjiang Theater.

The second singing style uses the written language of the Yueyang dialect and is accompanied by the *erhu*. This style is commonly known as *qin* opera.

To understand the contents of the story, you must understand the historical background.

Nowadays, most people invite Pingjiang shadow puppetry artists to perform in order to fulfill the wishes of the gods, and shadow puppetry shows are most commonly performed for the gods in temples with no human audience. People who put on the theater are usually only wanting to satisfy the gods. Their goal is not usually to enjoy the artistic nature of the performances. They believe the gods can help them become healthy, get into a university, get richer, etc.

Currently, there is almost no one who wants to become a shadow puppet performer. The youngest performer was born in 1979, and is now 45 years old. Some females have been studying the art recently, but they cannot perform very well. as they began to learn

shadow puppetry in adulthood. A good shadow puppet artist must begin learning puppetry skill from childhood. Traditionally, a novice had to eat and live together with the teacher. An adult cannot master the skill very well. So, shadow puppetry is becoming simplified, and as a result may vanish in the not-too-distant future.

I started recording and archiving shadow puppetry theater in 2017. Now, I have a large library of recordings and scripts, as below. Each theater performance is comprised of three parts.

 Hexi (congratulations drama, short story) 40 (8 hours);

 Zhengxi (long story, main drama)40 (80 hours);

 Shuaxi(final drama)40 (7 hours);

 scripts (the abstracts of each complete performance) 100,000 character.

I will continue to archive the shadow puppet theater and will publish the materials on the internet. I also want to build a shadow puppet museum to display shadow puppet props, with subtitles to videos and audios, and I want to help train young artists.

Some shadow puppetry artists have agreed to work with me. They will join me in my archiving efforts. We started working together in Aug. 2023.

The artists are currently recording the dialogue, sound, and video for me, and after they do, I transcribe the spoken word and make the scripts available for the artists and other interested people. In this way, ordinary people can have an opportunity to understand and enjoy this art form.

As a team, our goal is to increase the influence of Pingjiang shadow puppetry and increase the size of the audience, in the hopes of encouraging young people to want to take up the art form.

Close cooperation between researchers and new artists can promote and protect Pingjiang shadow puppetry, and our work and cooperation can become a model for localities across China, localities which have their own culture to promote and protect.