Cite as: Krajinović, Ana, Rosey Billington, Lionel Emil, Gray Kaltapau & Nick Thieberger. 2019. Building capacity for community-led documentation in Erakor, Vanuatu. In Zygmunt Vetulani & Patrick Paroubek (eds.), Human language technologies as a challenge for computer science and linguistics — 2019, 185–189. Poznan: Wydawnictwo Nauka i Innowacje.

Building capacity for community-led documentation in Erakor, Vanuatu

Ana Krajinović[12], Rosey Billington[13], Lionel Emil[4], Gray Kaltapau[4], Nick Thieberger[13]

[1]Centre of Excellence for the Dynamics of Language, Australia
[2]Humboldt-Universitt zu Berlin, Unter den Linden 6, 10099 Berlin, Germany; krajinoa@hu-berlin.de
[3]University of Melbourne, Royal Parade, Parkville VIC 3052, Australia; rbil@unimelb.edu.au, thien@unimelb.edu.au
[4]Nafsan Language Team, Erakor Village, Efate, Shefa Province, Vanuatu; gkkaltkpau@gmail.com

Abstract

Close collaboration between community members and visiting researchers offers mutual benefits, including opportunities for new research insights and an expanded scope for supporting language maintenance and developing practical materials. We discuss a collaboration in Erakor, Vanuatu aiming to build the capacity of community-based researchers to undertake and sustain language and cultural documentation projects. We focus on the technical and procedural skills required to collect, manage, and work with audio and video data, and give an overview of the outcomes of a community-led project after initial training. We discuss the benefits and challenges of this type of project from the perspective of the community researchers and the external linguists. We show that the community-led project in Erakor, in which data management and archiving are incorporated into the documentation process, has crucial benefits for both the community and the linguists. Two most salient benefits are: a) long-term documentation of linguistic and cultural practices calibrated towards community's needs, and b) collections of large quantities of data of good phonetic quality, which, besides being readily available for research, have a great potential for training and testing emerging language technologies based on machine learning.

1. Introduction

There has been increasing recognition that greater collaboration between external linguists and language communities can be mutually beneficial, and aid language maintenance efforts (e.g. Czaykowska-Higgins, 2009; Rice, 2011; Bowern and Warner, 2015). Approaches incorporating technical training and empowering people to undertake community-led projects are noted to be vital for inclusive collaboration (e.g. Yamada, 2007; Yamada, 2014). In Vanuatu, there are many examples of productive collaborations on language and cultural documentation projects (e.g. Regenvanu, 1999; Tryon, 1999; Barbour, 2010; Guérin and Lacrampe, 2010; Taylor and Thieberger, 2011). In this paper we describe one process of building community capacity to engage in language maintenance and corpus building through linguist-community collaboration. We focus on the community of Erakor, on the island of Efate, Vanuatu (Fig. 1), near the capital, Port Vila.[1]

The language of the community in Erakor, as well as nearby Eratap and Pango, is Nafsan (also known as South Efate), a Southern Oceanic language with an estimated 5,000-6,000 speakers (Lynch et al., 2002). Nafsan is one of 130+ languages in Vanuatu, and is spoken alongside Bislama, one of three official languages and a lingua franca across the archipelago. Education is carried out in English and French. Vanuatu is undergoing an information and communications technology revolution (Cave, 2012; Finau et al., 2014), and around 86% of households now have home access to mobile networks (Vanuatu National Statistics Office, 2017). Access to technologies other than mobile phones is still limited, though increasing, and both mobile and internet use is claimed to be linked to changing patterns of language use (Vandeputte-Tavo, 2013).

Figure 1: Location of Vanuatu and the island of Efate

Records of Nafsan extend back to the mid-1800s, in materials produced by missionaries (see Thieberger, 2019). Modern linguistic research began with a focus on the phonology and genetic classification of Nafsan (e.g. Tryon, 1976; Clark, 1985; Lynch, 2000). A comprehensive reference grammar of Nafsan has been produced by Thieberger, 2006, accompanied by corpus data, a book of stories (Thieberger, 2011b), and a dictionary (Thieberger, 2011a), which is regularly updated at community workshops and will soon have a new edition. All of this previous research laid the groundwork for the more recent activity, first by creating a corpus that new researchers could use to begin work on the language, and, second, by demonstrating a quid pro quo of returning materials to the village in forms that could be used there. The main aim of this paper is to demonstrate ways in which linguists can support community efforts in language documentation and maintenance through building capacity, and how these collaborations can result in larger quantities of quality data. We describe the process of training community members in using technology for recording, transcribing and building a corpus (§2), and discuss the outcomes, benefits and challenges of

[1]We wish to thank all the speakers of Nafsan that participated in this documentation project and we are also grateful for the feedback we received at Vanuatu Languages Workshop, 25-27 July 2018 in Port Vila, Vanuatu. This work has been funded by the ARC Centre of Excellence for the Dynamics of Language (Australia) and the German Research Foundation DFG (MelaTAMP project with number 273640553).

a documentation project undertaken by the third and fourth authors (§3). We also identify ways that both the language community and the wider linguistics community can benefit if there is greater consideration of the extent to which data will be conserved, accessible, and amenable for use with current software and tools as well as emerging language technologies (§4), and conclude in §5.

2. Sharing technical and procedural skills

To build on the previous documentation and description of Nafsan, the first two authors (AK & RB) began fieldwork in 2017 in Erakor, aiming to collect new Nafsan data for targeted semantic and phonetic analyses (e.g. Krajinović, 2018; Billington et al., 2018). In the beginning of their fieldtrip, they participated in a dictionary workshop in Erakor led by the fifth author (NT). During the workshop sessions, it became clear that besides the work on the dictionary, there was community interest in collecting more narratives in Nafsan. NT gave a Zoom H1N recorder to a community member, GK, the fourth author, who partnered with the third author, LE, to develop ideas for a recording project. Given that there was an intention of data collection in the absence of linguists, AK & RB realized that there was a need for training in data collection and management. During their semantic and phonetic experiments, they started familiarizing GK and LE with the process of making a recording, transcribing it, and managing the data. GK & LE assisted AK & RB in different types of fieldwork tasks, such as transcription and video recording, and a computer was made available for them to use for independent transcription, using ELAN (The Language Archive, 2018). As GK & LE became more comfortable with transcribing pre-segmented audio files in ELAN, AK & RB organized more formal training of linguistic tools.

The training focused on four indispensable activities in a language documentation workflow: planning and discussing a recording with participants (including archival access conditions), making a recording, data management, and transcription. For the recording process, GK & LE practiced using the Zoom H1N and including basic spoken metadata at the beginning of each audio recording, and we discussed some basic principles of video recording. The data management component was slightly more challenging as it involved familiarizing the community members with the use of spreadsheets and file-naming practices. We practiced the workflow as a routine of making a recording, transferring it to a computer, entering metadata in a spreadsheet, and backing up the data. This process was easily followed as each activity was understood as an essential part of the workflow. The last step was learning how to use ELAN (see Fig. 2). Until this point, GK & LE were already familiar with transcribing spoken Nafsan in a single pre-segmented tier. These skills were extended to creating a new file and importing audio files together with a template (Gaved and Salffner, 2014) that facilitates exporting into FieldWorks (SIL, 2018), in which it can be semi-automatically glossed. The use of a more complex template required some explanation of the hierarchical organization of tiers, e.g. that the translation tier depends on the tier of the original text. The focus of transcription efforts was filling in the first tier with orthographic Nafsan, as in Fig. 3. In this training we focused on highlighting the structure of the workflow, and making sure that the community members understood the importance of data management that follows the creation of each recording. Understanding the technical aspects of using different types of software proved to be relatively easy. However, documenting instructions in a simple text was also helpful.

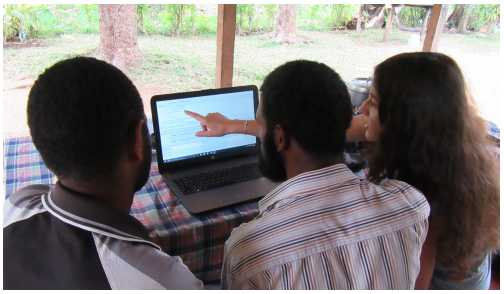

Figure 2: Training in ELAN transcription

3. Outcomes of community-led project

3.1. Summary of collected materials

Between July 2017–June 2018, GK and LE, as community researchers, collected audio and video data relating to 21 recording sessions. Some sessions were recorded only using either video or audio, and others were recorded with simultaneous video and audio, for later synchronization. In total, the collected data comprised 17 audio files totalling 05:26:37, and 25 video files totalling 04:25:34. Recording sessions took place primarily in Erakor, but some took place in Eton, a village further to the north on the coast of Efate, with strong ties to Erakor. The recordings were all of natural speech and related to diverse topics, driven by the interests of the community researchers and the community members they engaged with for their project. Among the recordings which were primarily audio, two were 'kastom' (traditional) stories, four were personal life histories, and three were stories about people and events in Erakor and Eton. Among the recordings which were primarily video, there was a story about the first permanent house in Erakor, and many videos demonstrating techniques for weaving baskets, fans and mats using coconut and pandanus leaves. The community researchers chose weaving as a focal topic because of concern that traditional weaving skills are not being passed on to younger generations, and a desire to document these skills and develop educational resources. All of the recordings have been archived in PARADISEC[2] with accompanying metadata, and apart from one, all are open-access (Kaltaƀau and Emil, 2017). Good progress has also been made on transcribing these recordings in ELAN; seven recordings have been fully transcribed, one partially transcribed, and one long recording has been fully segmented and made ready for transcription. The project is ongoing, and future plans include engaging more community members as participants, recording material for a documentary about Nafsan, and identifying ways to use collected videos for educational purposes.

[2] http://www.paradisec.org.au

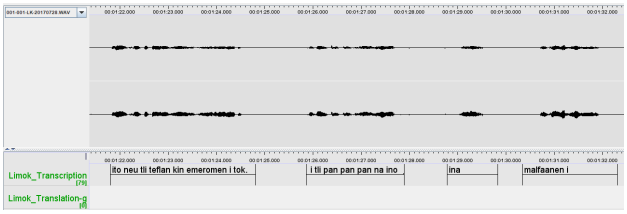

Figure 3: Orthographic transcription of *Naɲre nig Taler* (a story about a demon) told by Limok Kaltaɲau (GKLE-001)

Figure 4: Marian Kalmary weaving *naal pool* (GKLE-013)

3.2. Benefits and challenges

From the perspective of the community researchers, there are a number of advantages to language and cultural documentation projects led by community members. One clear advantage is first-hand knowledge of the language. In most documentation projects, linguists are visitors, and while they may acquire the language of study to varying extents, in most cases they are unlikely to acquire competence approaching that of native speakers. Native knowledge of Nafsan facilitates more accurate and efficient transcription, and also facilitates the process of undertaking recording sessions with different community members. Community researchers also have a significant advantage in that they have better knowledge of the linguistic and cultural practices which may feature in documentation recordings. They are well-placed to decide which activities are better documented with video rather than audio, based on the type of activity and also what participants are most comfortable with, and are also able to use their knowledge of particular activities to more effectively plan and capture these using video. For example, if the goal of a recording is to document the process for weaving a particular type of basket (e.g. Fig. 4), and the community researchers are familiar with what this entails, they can choose the most appropriate framing and zoom level at different stages, so that viewers can identify exactly what the participant is doing. In comparison, an external researcher may focus on capturing the whole scene in every frame, perhaps to include gestures or background interlocutors, but this will be less useful to someone wanting to watch the recording to study the weaving technique. Community researchers are also better able to identify which activities are most important to document, and of the greatest interest to the community, particularly in contexts where a project aims to support language and cultural maintenance.

Challenges noted by GK & LE relate to both the practicalities of using equipment and technology as well as the logistics of managing a project. While the actual transcription process in ELAN was relatively manageable, making a new .eaf file could be difficult. The template provided by AK & RB was helpful, and consistently used, but the main issue was remembering how to navigate the ELAN interface and access the template when starting a new transcription. Sharing one laptop also limited the ability of the community researchers to undertake transcription and data management tasks at the times most convenient to them. Similarly, it was often difficult to find time to spend on recording and transcription among other family and community commitments. It was also not always easy to find people who were willing and available to participate. In some cases people were interested but had limited time, and in other cases people were intimidated by the prospect of being in an audio or video recording. A particular challenge when recording video was shakiness caused by camera movement. Activities such as weaving required GK & LE to be able to move around in order to best capture different parts of the process, and this proved to be difficult to do without excessive movement caused by using a handheld video camera. Some of the challenges noted here have since been addressed, for example by acquiring a tripod to reduce camera shakiness even if carrying by hand, and an additional laptop, allowing an easier division of tasks between the two community researchers.

From the perspective of the visiting researchers, there is no doubt that building local capacity to undertake language and cultural documentation offers benefits in terms of both the scale and quality of documentation. The community-led project contributes to a more comprehensive record of Nafsan, and allows for new research questions to be explored and existing research questions to be addressed more thoroughly. Importantly, the resulting materials are more representative of community priorities and interests, and more useful for developing materials supporting language and cultural maintenance. These and many other ways that collaborative and community-led projects benefit both the specific goals of a community, and the scientific endeavor of linguistic research, have been discussed in detail elsewhere (e.g. Czaykowska-Higgins, 2009; Rice, 2011; Bowern and Warner, 2015). An additional benefit of the particular approach taken in the current project is that data management and metadata collection was built into the initial training, as were strategies for discussing archiving and access conditions with community participants. While data and metadata management has required some ongoing support, and can be difficult when internet access is limited, the result is that not only is there a rich set of materials collected by the community researchers, but that these materials have been easily archived along with details of their content, and are accessible and therefore usable by others, including community members who have some previous experience accessing Nafsan materials collected by NT via PARADISEC. Other researchers discussing collaborative language documentation acknowledge that there can be logistical, institutional, and interpersonal challenges to the sustainability of community-led projects, but we find, as they do, that the benefits of community-led documentation far outweigh the challenges.

4. Potential for applications of language technology to less-resourced languages

One problem arising, which may not seem like a problem at first, is too much data. Scaling up documentation in the way described here leads to more audio and video recordings than would otherwise have been collected thus far, but not all have been transcribed. While engaging community members in transcription is often seen as a way to speed up the process, and to transcribe a higher percentage of recordings than a solo linguist (with less fluency in the language) could manage, community members are generally not able to work on these tasks to the exclusion of other responsibilities, or other interests within a project. As long as transcription is fully reliant on human effort, there remains an issue of the 'transcription bottleneck', whereby more data is recorded than can feasibly be transcribed and added to a corpus within time and resource limitations. (e.g. Brinckmann, 2009). However, there are several types of machine-learning technologies currently under development which focus on ways to apply automatic speech recognition (ASR) and transcription to less-resourced languages. Training a language model for adequate speech recognition generally requires very large speech corpora, but these are not typically available for languages which are relatively under-described. Recent discussions argue that field linguists should modify their practice to assist the task of machine learning, for example by making high-quality recordings using head-mounted microphones (Seifart et al., 2018), given that recordings made in often noisy fieldwork conditions can be challenging for ASR technologies (van Esch et al., 2019).

To add to this discussion, we note that community researchers may be better placed than visiting linguists to collect high-quality audio recordings, given appropriate training opportunities. External researchers typically visit for a set time frame, and generally have specific goals, for example related to collecting a certain number of hours of particular data types, with a range of participants. This means that recordings are often undertaken opportunistically, where and when community members are available, and it is not always possible to have a great deal of control over factors such as environmental noise. Fig. 5 shows a sample waveform and spectrogram of a recording made by the second author in one such opportunistic setting. The recording was made with a hypercardioid head-mounted microphone in a location with as much sound attenuation as possible within the available options, but unfortunately took place exactly at dusk, which meant substantial noise from a flock of birds settling in to roost in a tree nearby. As can be seen, the signal-to-noise ratio is not ideal; there is a lot of additional noise in the higher frequency range. While this recording would still be fairly usable for phonetic analyses of fundamental frequency or duration, it would be less useful for analyses of fricative energy or formant transitions, and would also present more of a challenge to ASR.

In comparison, community researchers are able to be more flexible in their project schedules, and can choose to make audio recordings in a quiet environment at a preferred time of day, and to make video recordings under optimal weather and lighting conditions. They may also be better able to negotiate a recording situation which prioritizes both the comfort of the participant and the quality of the recording (in ways that visitors are not always equipped to do appropriately). Fig. 6 shows a sample waveform and spectrogram[3] of a recording collected by GK. He chose to record this late at night, after the noise of people, birds and vehicles and generators had stopped, in a small room with closed windows. He also sat close to the speaker in order to hold the recorder at a constant and appropriate distance from her mouth. This recording was made with the inbuilt stereo microphone of the Zoom H1N, which, being less directional, would pick up more background noise than the microphone used for Fig. 5, but as can be seen this is clearly the cleaner recording. Recordings like this are much better suited to training of ASR models. Preliminary tests of developing a speech recognition model for Nafsan have been undertaken using Kaldi, via the in-development Elpis pipeline, and show promising results (Foley et al., 2018). A model based on just 3 hours of audio as training data was applied to untranscribed data and returned a word error rate of 42.7%; a 'reasonably decent' result for a first pass using sample data with limited coverage and limited tuning of parameters in the pronunciation model. The potential offered by these kinds of technologies, as they continue to be refined for use in documentation contexts, is clear. In addition, there are various natural extensions of these speech and language technology toolkits which would not only further aid data processing and analysis, but also better support the use of less-resourced languages in digital domains (van Esch et al., 2019).

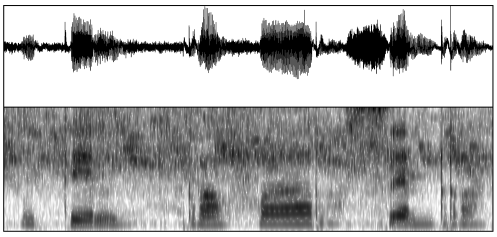

Figure 5: Recording made in noisy conditions

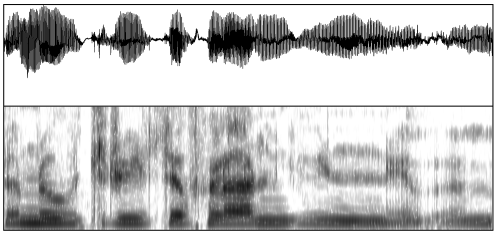

Figure 6: Recording made in quiet conditions

5. Conclusion

In this paper we described the process and outcomes of building capacity for community-led documentation in Erakor, Vanuatu. We highlighted the benefits of direct community involvement in language documentation and maintenance efforts for both the community and the external linguists. We showed that the community researchers are

[3]Fig. 5 and 6 correspond to samples of 200ms; spectrograms show frequencies up to 5000Hz with a 60dB dynamic range.

able to contribute to overall larger quantities of linguistic data than that collected only by visiting linguists during fieldwork. Moreover, in some cases the data gathered by community researchers is better than that collected by external linguists, in terms of either content or audio quality. This happens mainly for two reasons: a) the community members are best placed to decide what linguistic and cultural practices to document, and how, thus making the resulting materials more useful for the community, and b) they may have greater choice in and control over recording conditions, resulting in better acoustic quality of audio recordings (and image quality in video recordings). The former aspect is crucial for supporting language maintenance efforts and the latter aspect allows for favorable results from applications of ASR technologies to less-resourced languages. The potential scope for language technology applications is expanded when data of good technical quality is combined with well-maintained corpus materials. More generally, both linguists and the community benefit greatly from archival collection of the materials, which become available for linguistic research and to the community now and in the future.

6. References

Barbour, Julie, 2010. Neverver: A study of language vitality and community initiatives. In Margaret Florey (ed.), *Endangered languages of Austronesia*. Pp. 225–244.

Billington, Rosey, Janet Fletcher, Nick Thieberger, and Ben Volchok, 2018. Acoustic correlates of prominence in Nafsan. In *Proc. 17th Australasian International Speech Science & Technology Conference*. Pp. 137–140.

Bowern, Claire and Natasha Warner, 2015. Lone Wolves and collaboration: A reply to Crippen & Robinson (2013). *Language Documentation & Conservation*, 9:59–85.

Brinckmann, Caren, 2009. Transcription bottleneck of speech corpus exploitation. In Verena Lyding (ed.), *Proc. Second Colloquium on Lesser Used Languages and Computer Linguistics*. Pp. 165–179.

Cave, Danielle, 2012. Digital islands: How the Pacific's ICT revolution is transforming the region. Technical report, Lowy Institute for International Policy.

Clark, Ross, 1985. The Efate dialects. *Te Reo*, (28):3–35.

Czaykowska-Higgins, Ewa, 2009. Research models, community engagement, and linguistic fieldwork: Reflections on working within Canadian Indigenous communities. *Language Documentation & Conservation*, 3:15–50.

Finau, Glen, Romitesh Kant, Sarah Logan, Acklesh Prasad, Jope Tarai, John Cox, et al., 2014. Social media and e-democracy in Fiji, Solomon Islands and Vanuatu. Association for Information Systems.

Foley, Ben, Josh Arnold, Rolando Coto-Solano, Gautier Durantin, T Mark Ellison, Daan van Esch, Scott Heath, František Kratochvíl, Zara Maxwell-Smith, David Nash, Ola Olsson, Mark Richards, Nay San, Hywel Stoakes, Nick Thieberger, and Janet Wiles, 2018. Building speech recognition systems for language documentation: The CoEDL Endangered Language Pipeline and Inference System (ELPIS). In *6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*. Pp. 200–204.

Gaved, Tim and Sophie Salffner, 2014. Working with ELAN and FLEx together: An ELAN-FLEx-ELAN teaching set. Retrieved from `https://www.soas.ac.uk/elar/file122785`.

Guérin, Valérie and Sébastien Lacrampe, 2010. Trust me, I am a linguist! Building partnership in the field. *Language Documentation & Conservation*, 4:22–33.

Kaltaƀau, Gray and Lionel Emil, 2017. Nafsan recordings (GKLE), Digital collection managed by PARADISEC. `http://catalog.paradisec.org.au/collections/GKLE`.

Krajinović, Ana, 2018. Comparative study of conditional clauses in Nafsan. In *SIL Language and Culture Documentation and Description 41 (Proc. COOL 10)*. Pp. 1–18.

Lynch, John, 2000. South Efate phonological history. *Oceanic Linguistics*, 39(2):320–338.

Lynch, John, Malcom Ross, and Terry Crowley, 2002. *The Oceanic languages*. London: Routledge.

Regenvanu, Ralph, 1999. Afterword: Vanuatu perspectives on research. *Oceania*, 70(1):98–100.

Rice, Keren, 2011. Documentary linguistics and community relations. *Language Documentation & Conservation*, 5:187–207.

Seifart, Frank, Nicholas Evans, Harald Hammarström, and Stephen C Levinson, 2018. Language documentation twenty-five years on. *Language*, 94(4):e324–e345.

SIL, 2018. Fieldworks Language Explorer (FLEx) 8.3. Retrieved from `https://software.sil.org/fieldworks/`,.

Taylor, John and Nick Thieberger (eds.), 2011. *Working together in Vanuatu: Research histories, collaborations, projects and reflections*. Canberra: ANU Press.

The Language Archive, 2018. ELAN (Version 5.2) [Computer software]. Nijmegen: Max Planck Institute for Psycholinguistics. Retrieved from `https://tla.mpi.nl/tools/tla-tools/elan/`.

Thieberger, Nicholas, 2006. *A grammar of South Efate: An Oceanic language of Vanuatu*. Honolulu: University of Hawai'i Press.

Thieberger, Nick, 2011a. *A South Efate dictionary*. Parkville: University of Melbourne.

Thieberger, Nick, 2011b. *Natrauswen nig Efat: Stories from South Efate*. Parkville: University of Melbourne.

Thieberger, Nick, 2019. Guide to the Nafsan, South Efate collection. Retrieved from `https://www.nthieberger.net/sefate.html`.

Tryon, Darrell, 1999. Ni-Vanuatu research and researchers. *Oceania*, 70(1):9–15.

Tryon, Darrell T., 1976. *New Hebrides languages: An internal classification*. Canberra: Pacific Linguistics.

van Esch, Daan, Ben Foley, and Nay San, 2019. Future directions in technological support for language documentation. In *Proc. 3rd Workshop on Computational Methods for Endangered Languages (vol. 1)*. Pp. 14–22.

Vandeputte-Tavo, Leslie, 2013. New technologies and language shifting in Vanuatu. *Pragmatics*, 23(1):169–179.

Vanuatu National Statistics Office, 2017. 2016 Post-TC Pam Mini Census Report. Technical report, Ministry of Finance & Economic Management, Port Vila, Vanuatu.

Yamada, Racquel-María, 2007. Collaborative linguistic fieldwork: Practical application of the empowerment model. *Language Documentation & Conservation*, 1:257–282.

Yamada, Racquel-María, 2014. Training in the community-collaborative context: A case study. *Language Documentation & Conservation*, 8:326–344.