University Library

MINERVA ACCESS

A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Krajinović, A;Billington, R;Emil, L;Kaltaƥau, G;Thieberger, N

Title:
Community-Led Documentation of Nafsan (Erakor, Vanuatu)

Date:
2022-01-01

Citation:
Krajinović, A., Billington, R., Emil, L., Kaltaƥau, G. & Thieberger, N. (2022). Community-Led Documentation of Nafsan (Erakor, Vanuatu). Vetulani, Z (Ed.) Paroubek, P (Ed.) Kubis, M (Ed.) Lecture Notes in Computer Science Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, 13212 LNAI, pp.112-128. SPRINGER INTERNATIONAL PUBLISHING AG. https://doi.org/10.1007/978-3-031-05328-3_8.

Persistent Link:
https://hdl.handle.net/11343/325204

Author Proof

Check for updates

Community-Led Documentation of Nafsan (Erakor, Vanuatu)

Ana Krajinović[1,2(✉)], Rosey Billington[2,3], Lionel Emil[2,4], Gray Kaltapau[2,4], and Nick Thieberger[2,5]

[1] Heinrich-Heine-Universität Düsseldorf, Universitätsstraße 1, 40225 Düsseldorf, Germany
ana.krajinovic@hhu.de
[2] ARC Center of Excellence for the Dynamics of Language, Canberra, Australia
[3] The Australian National University, Canberra, ACT 0200, Australia
rosey.billington@anu.edu.au
[4] Nafsan Language Team, Erakor Village, Vanuatu
[5] The University of Melbourne, Parkville, VIC 3010, Australia
thien@unimelb.edu.au

Abstract. We focus on a collaboration between community members and visiting linguists in Erakor, Vanuatu, aiming to build the capacity of community-based researchers to undertake and sustain documentation of Nafsan, the local indigenous language. 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 documentation 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 community-led documentation such as this 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. The two most salient benefits are: a) long-term documentation of linguistic and cultural practices calibrated towards community's needs, and b) collection of larger quantities of data by community members, and often of better quality and scope than those collected by visiting linguists, which, besides being readily available for research, have a great potential for training and testing emerging language technologies for less-resourced languages, such as Automatic Speech Recognition (ASR).

We wish to thank all the speakers of Nafsan who participated in this documentation project and we are also grateful for the feedback we received at the Vanuatu Languages Workshop, 25–27 July 2018 in Port Vila, Vanuatu. We also benefited from discussions at the 9th Language & Technology Conference, May 17–19, 2019 in Poznań, Poland, and Language Technologies for All, 4–6 Dec 2019, UNESCO, Paris. This work has been funded by the ARC Centre of Excellence for the Dynamics of Language (Australia) (project ID: CE140100041) and the German Research Foundation DFG (MelaTAMP project with number 273640553).

© Springer Nature Switzerland AG 2022
Z. Vetulani et al. (Eds.): LTC 2019, LNAI 13212, pp. 1–17, 2022.
https://doi.org/10.1007/978-3-031-05328-3_8

Author Proof

Keywords: Community-led language documentation · Technical training · Less-resourced languages · Technology for indigenous languages · Nafsan · Vanuatu · Automatic speech recognition

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. [8,12,33]). Approaches incorporating technical training and empowering people to undertake community-led projects are noted to be vital for inclusive collaboration (e.g. [48,49]). In Vanuatu, there are many examples of productive collaborations on language and cultural documentation projects (e.g. [1,18,32,36,43]). 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 (Figs. 1 and 2),[1] near the capital, Port Vila.



Fig. 1. Location of Vanuatu and the island of Efate.

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 [26]. Nafsan is one of 130+ indigenous languages in Vanuatu, and is spoken alongside Bislama, an English-based creole, which is one of three official languages and a lingua franca across the archipelago. Education is mainly carried out in English and French. Vanuatu is undergoing an information and communications technology revolution [10,14], and around 86% of households now have home access to mobile networks [46]. 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

[1] All the maps in this article were produced by using the Generic Mapping Tools [47].

Author Proof

use, such as greater use of Bislama [45]. However, new modes of communication also offer new environments in which indigenous languages of Vanuatu can be used. For example, social media platforms like Facebook include a number of pages in which Nafsan is the main language. This is a positive development, as speakers are actively using Nafsan in its written form on social media, which is not done in other areas of life, where English, French, and Bislama predominate as written languages. In our collaboration, Facebook is also the main means of communication between the researchers and the community members.

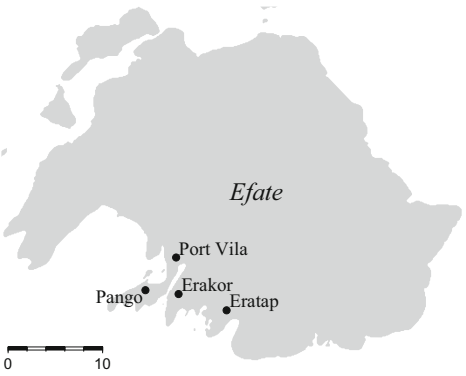


Fig. 2. The island of Efate indicating the places where Nafsan is spoken

Records of Nafsan extend back to the mid-1800s, in materials produced by missionaries (see [41]). Modern linguistic research began with a focus on the phonology and genetic classification of Nafsan (e.g. [11,25,44]). A comprehensive reference grammar of Nafsan has been produced by Thieberger [38], accompanied by corpus data, a book of stories [40], and a dictionary [39], which is regularly updated at community workshops and was published in 2021 [42]. 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 (§3) and the benefits and challenges (§4) of a documentation project undertaken by the third and fourth authors. We also identify ways that both the language community and the wider linguistics community can benefit from community-led documentation, especially if there is greater consideration of how the data will be conserved, archived, and made accessible in different formats (e.g. WAV files, time-aligned transcribed text etc.), which are used in linguistic

Author Proof

research and in the development of language technologies. In §5 we argue that the community-led documentation leads to the collection of high-quality data in larger quantities than those collected by visiting researchers, and is thus of crucial importance for emerging language technologies for indigenous and less-resourced languages, such as Automatic Speech Recognition (ASR). We note some promising results with new speech recognition and forced-alignment technologies applied to Nafsan data. We conclude in §6.

2 Technical and Procedural Training

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. [3,4,21,22,30]). In the beginning of their field trip, they participated in a dictionary workshop in Erakor led by the fifth author (NT), focused on checking, correcting and adding to entries for the Nafsan dictionary through group discussions with community members. 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 [37]. 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. 3). 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 [17] that facilitates exporting into FieldWorks [35], in which it can be semi-automatically glossed. The use of a more

Author Proof

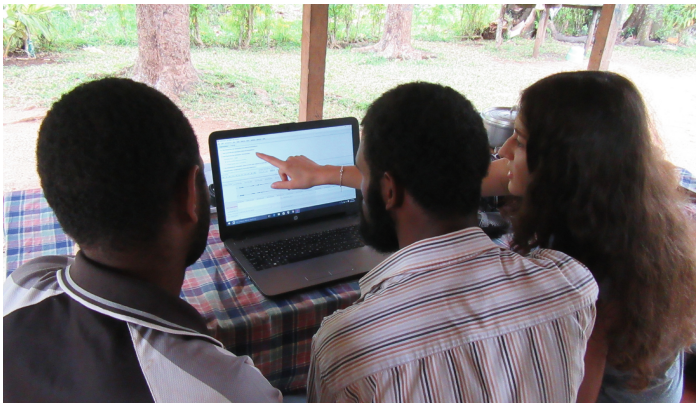


Fig. 3. Training in ELAN transcription

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. 4. 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.

Our training also included a discussion on the importance of explicitly explaining to the speakers who will be recorded and how the recordings will be stored and used, and making sure it is understood that they have a choice to select their own preferences regarding the access rights to the recording. Through their previous collaborations with NT, many community members, including GK & LE, were aware of the concept of an archive and the benefits of archiving the collected language data for posterity, as NT has been providing the community with their own copy of previously collected data, both locally and through online open access to the PARADISEC archive.[2] All the recordings made by AK [24] and RB [6] are also archived in PARADISEC. All the physical language materials created as a result of our research and collaboration, such as the storyboard booklet in Nafsan [23], have also been deposited in the *Vanuatu Kaljoral Senta* (Vanuatu Cultural Centre).[3]

3 Outcomes of the Community-Led Documentation

Between July 2017 and June 2018, GK and LE, as community researchers, collected audio and video data in 21 recording sessions. Some sessions were recorded

[2] Available at https://www.paradisec.org.au.
[3] https://vanuatuculturalcentre.gov.vu.

Author Proof

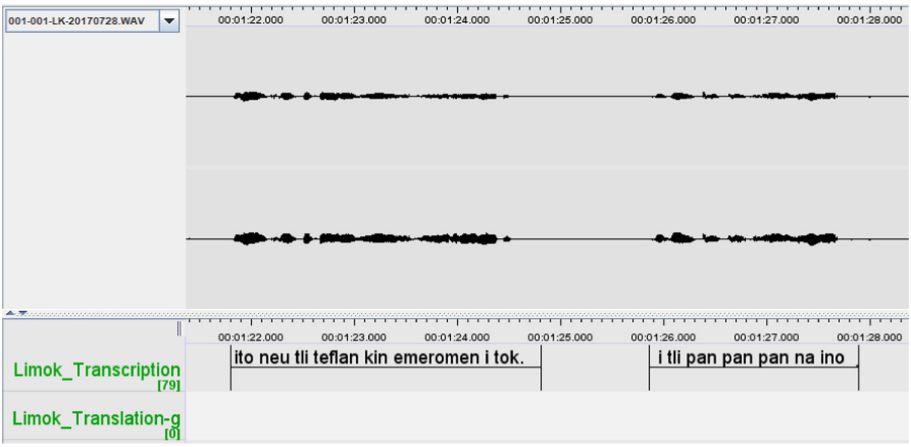


Fig. 4. Orthographic transcription of *Naṽre nig Taler* (a story about a demon) told by Limok Kaltaṽau (GKLE-001)

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 totaling five and a half hours, and 25 video files totalling four and a half hours. 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 (see Fig. 5). 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[4] with accompanying metadata, and apart from one, all are open-access [20].

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. The skills gained by GK and LE have also since been extended to new projects; building on the

[4] https://catalog.paradisec.org.au/collections/GKLE.

Author Proof

initial recordings made by GK in Eton, GK and RB visited Eton to record some traditional narratives in the language of that village[5] [7], for which published information is limited to a wordlist [44] and a comparative study of languages of the Efate region [11]. GK facilitated these recording sessions, and LE has assisted with initial transcription. Longer term, the project team hope that further collaboration with the Eton community will allow for a deeper understanding of the linguistic and cultural connections across southern Efate.

Item Identifier (e.g.	**Item Title** (e.g. Introductory Materials)	**Item Description** (e.g. Four text stories for interviews)
001	Napre nig Taler	Story about Taler's encounter with the erakor demon (natopu)
002	Natrauswen nig Oftau go Tiawi nig tutufur	Story about Oftau and his friend from tutufur
003	Nafsan nfapuswen	Farewell speech
004	Natrauswen nig ati touraan teflaan I patlas	Life story
005	Natrauswen ni limok go kaltpau	Life story
006	Natrauswen ni tesa nmatu ralim iskei go	Custom story
007	Natrauswen namolien ni apu abel naar	History of Abel Naar , an evangelist to Eton village
008	Natrausen ni Linmas kalsilik	Life story
009	Linmas i traus natrauswen nig apu samuel	Story of late chief samuel
010	Nfauwen ni likat	Weaving instruction
011	Natiltaewen ni nafkaworwen	Explanation of the nafsan word "pkawor"
012	Nfauwen ni niif	Fan weaving
013	Nfawen ni naal pool	Ball basket weaving
014	Nfawen ni toofrak	Plate weaving
015	Nafeifeien nig naal	Display of baskets
016	Nfawen ni naal	Basket weaving
017	Nfawen ni likat	Likat basket weaving
018	Napnotien ni likat	Finishing likat
019	Natfagien ni nasum pei ni natkoon Erakor	Building of the first permanent house in Erakor
020	Nfawen ni tefkau	Tefkau weaving
021	Nfawen ni tefkau	Tefkau weaving

Fig. 5. A part of the metadata of the recordings made by GK and LE

4 Benefits and Challenges

4.1 Benefits: The Engagement of Community Researchers Improves the Results of Language Documentation

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

[5] The language is also known as Eton.

Author Proof

better documented with video rather than audio, based on the type of activity and also what participants are most comfortable with. They 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. 6), 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.

Another benefit of community-led documentation is the quantity of collected data. In a relatively short time frame, GK and LE were able to collect large quantities of spoken Nafsan data, without any interference of another language, and with a minimized observer's bias. On the other hand, visiting researchers are outsiders, who maximize the observer's bias, and often communicate in another language, adding to the complexity of linguistic influences in the recorded data. Moreover, linguists often collect language data catered towards their specific research questions, either through experiments or elicitation, without necessarily prioritizing community interests. Usually only certain types of data commonly collected for research, such as telling of traditional stories, can be made immediately useful for the language community. Linguists typically undertake additional activities outside of their research topics in order to produce materials for community-wide use, such as the Nafsan dictionary [42]. In contrast, the Nafsan data collected by the community researchers can serve multiple purposes from the start: among others, education in language or cultural practices, entertainment, promotion of the local culture and products, linguistic research, and use in developing language technologies for indigenous and less-resourced languages (see §4).

4.2 Challenges: Sustainability

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.

Author Proof



Fig. 6. Marian Kalmary weaving *naal pool* (GKLE-013)

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.

The internet in Vanuatu is most readily accessible via mobile data. While this means that the internet is theoretically available even in many remote areas, in our experience, this has been both a benefit and a challenge. While GK & LE can stay in contact with researchers and even transfer ELAN .eaf files, the expense of mobile data and limited connection speed means sharing actual audio and video recordings is hard. Thus, the sharing of the recordings still needs to happen in person. This problem has been especially challenging during the COVID-19 pandemic, which made it impossible for researchers to travel to Vanuatu.

The COVID-19 pandemic has shown that now more than ever we need to improve the sustainability of these types of collaborations. As researchers might be unable to travel to distant fieldwork sites, where communities are especially vulnerable to potential outbreaks, these communities need to able to independently carry out language documentation and associated activities in order to continue making progress towards particular community goals. However, so far it has been hard to ensure the sustainability of such collaborations, primarily because of the lack of a fast and affordable internet connection that would allow for long-distance communication, and the lack of resources to maintain the hardware necessary for language and cultural documentation. Nevertheless, we have been able to transfer some of the discussion regarding our ongoing research on

Author Proof

Nafsan to Facebook chats. The information collected in these chats typically concerns words that can be included in the dictionary [42] and grammaticality judgments for semantic research [22]. Unfortunately, phonetic data cannot be collected this way.

4.3 Benefits Outweigh the Challenges

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. [8,12,33]). 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 (e.g. [49]), but we find, as they do, that the benefits of community-led documentation far outweigh the challenges.

5 Increased 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. Manual transcription, whether it is done by a native speaker or otherwise, is also

Author Proof

extremely time consuming, and depending on the familiarity with the language and the nature and level of detail of the transcription, can take anywhere from 2.5 h [16] to 200 h [9] per hour of recorded speech. As long as transcription is fully reliant on human effort, there remains an issue of the 'transcription bottleneck' [9], whereby more data is recorded than can feasibly be transcribed and added to a corpus within time and resource limitations.

At the same time, corpora are most useful for linguistic research, and many community goals, when they are transcribed. For instance, the Nafsan data that has been transcribed has been used in several cross-linguistic projects focusing on different linguistic domains, e.g. phonetics and morphology [28], inferring grammar from texts [19], and semantics [30,31]. Finding ways to improve transcription workflows is therefore vital to being able to extend the scale and usability of language corpora. Speech and language technologies, such as tools for automatic speech recognition and automatic transcription, can significantly aid the process of transcribing spoken language. However, there is limited availability of usable tools of this sort for many languages (e.g. [2]). In part, this is due to the limited quality and scope of language material available for less-resourced languages.

Regarding quality, one challenge in developing ASR for less-resourced languages is the sometimes variable audio quality of language documentation corpora, where recordings are often made in noisy fieldwork conditions. 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 [34]. 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. Figure 7 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

Author Proof

the quality of the recording (in ways that visitors are not always equipped to do appropriately). Figure 8 shows a sample waveform and spectrogram[6] 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. 7, but as can be seen this is clearly the cleaner recording. Recordings like this are much better suited to training of ASR models.

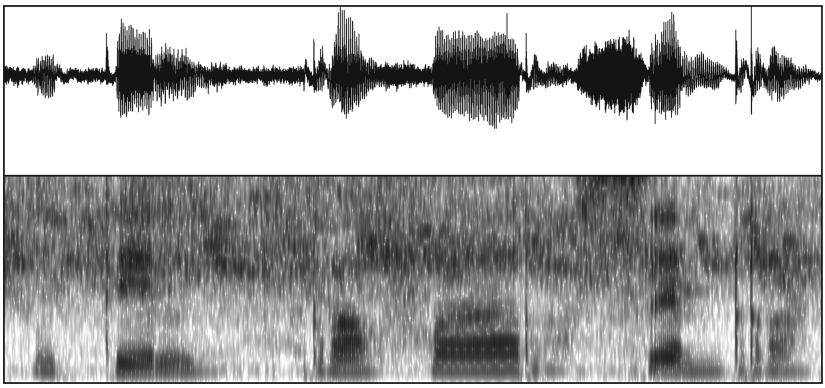


Fig. 7. Recording made in noisy conditions

The amount of data available for developing new speech and language tools is another challenge. 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. In recent years, there has been increasing interest in finding ways to adapt automatic speech recognition and transcription methods to work more effectively with small corpora of the sort typically collected during language documentation. Preliminary tests of developing a speech recognition model and semi-automated transcription for Nafsan have been undertaken using the Kaldi speech recognition engine [29], via the in-development Elpis pipeline, and show promising results [15]. A model based on just 3 h 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.

[6] Figures 7 and 8 correspond to samples of 200 ms; spectrograms show frequencies up to 5000 Hz with a 60 dB dynamic range.

Author Proof

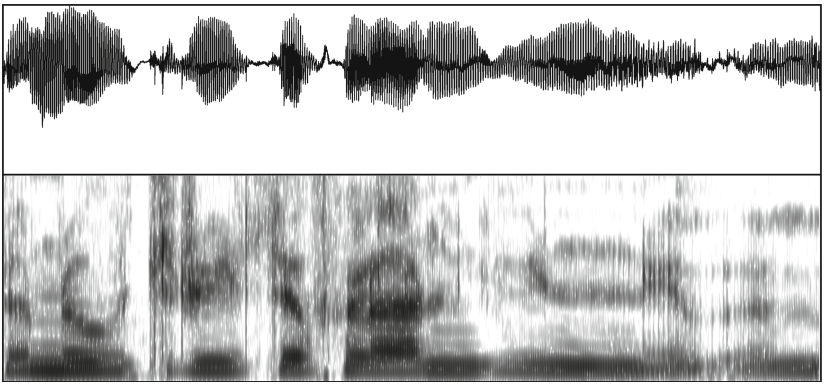


Fig. 8. Recording made in quiet conditions

Language modelling for the purposes of forced alignment has also been tested, focused on using utterance-level orthographic transcriptions to produce phone-level and word-level alignments. These more granular annotations of speech are of particular interest in phonetic research, where it is necessary to be able to take acoustic measurements with reference to individual speech sounds. Using the Montreal Forced Aligner (MFA) [27], which also uses Kaldi, speech modelling and forced alignment was undertaken based on just over 2 h of Nafsan data [5]. The output phone alignment was very accurate, and preliminary analyses of vowel tokens showed comparable acoustic patterns to those obtained in previous experimental datasets. It is important to note that the quality of the output of automatic transcription and forced alignment processes depends on the quality of the data used in the modelling, not just in terms of audio but also the accompanying transcriptions. In cases where the amount of data is limited, language corpora which have been carefully developed in collaboration with community members, including contributions to transcriptions and analyses, will lead to better results, and in turn aid the expansion and enrichment of the corpus.

The potential offered by these kinds of technologies, as they continue to be refined for use in documentation contexts, is clear. There is also scope to draw on language models for a given language to develop models for related or phonetically similar languages which may have even more limited speech material available. This is currently being explored for several languages of the Efate region [5]. 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 [13].

6 Conclusion

In this paper we described the process and outcomes of building capacity for community-led documentation in Erakor, Vanuatu. We showed that archiving

Author Proof

research materials provides a base for reciprocity with the speakers of the language, and then permits further research to be built on existing work in ways that were not previously possible. 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 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 an archival collection of the materials, which become available for linguistic research and to the community now and in the future.

References

1. Barbour, J.: Neverver: A study of language vitality and community initiatives. In: Florey, M. (ed.) Endangered languages of Austronesia, pp. 225–244. Oxford University Press, Oxford (2010)
2. Besacier, L., Barnard, E., Karpov, A., Schultz, T.: Automatic speech recognition for under-resourced languages: a survey. Speech Commun. **56**, 85–100 (2014)
3. Billington, R., Fletcher, J., Thieberger, N., Volchok, B.: Acoustic evidence for right-edge prominence in Nafsan. J. Acoust. Soc. Am. **147**(4), 2829–2844 (2020). https://doi.org/10.1121/10.0000995
4. Billington, R., Thieberger, N., Fletcher, J.: Nafsan. J. Int. Phonetic Assoc. 1–21 (2021). https://doi.org/10.1017/S0025100321000177
5. Billington, R., Stoakes, H., Thieberger, N.: The Pacific Expansion: Optimizing phonetic transcription of archival corpora. In: Proceedings of INTERSPEECH 2021, pp. 2021–2167. International Speech Communication Association, Brno (2021). https://doi.org/10.21437/Interspeech
6. Billington, R.: Rosey Billington Nafsan materials. Collection BR1 at catalog.paradisec.org.au [Open Access] (2017). https://dx.doi.org/10.26278/GXDM-J159
7. Billington, R.: Recordings of the language of Eton. Collection BR2 at catalog.paradisec.org.au [Open Access] (2019). https://dx.doi.org/10.26278/TRS1-XP03
8. Bowern, C., Warner, N.: 'Lone Wolves' and collaboration: a reply to Crippen & Robinson (2013). Lang. Documentation Conserv. **9**, 59–85 (2015)

Author Proof

9. Brinckmann, C.: Transcription bottleneck of speech corpus exploitation. In: Lyding, V. (ed.) Proceedings of the Second Colloquium on Lesser Used Languages and Computer Linguistics (LULCL II) Combining Efforts to Foster Computational Support Of Minority Languages, pp. 165–179. Europäische Akademie (EURAC book, 54) (2009)
10. Cave, D.: Digital islands: how the Pacific's ICT revolution is transforming the region. Tech. rep. Lowy Institute for International Policy (2012)
11. Clark, R.: The Efate dialects. Te Reo **28**, 3–35 (1985)
12. Czaykowska-Higgins, E.: Research models, community engagement, and linguistic fieldwork: reflections on working within Canadian Indigenous communities. Lang. Documentation Conserv. **3**, 15–50 (2009)
13. van Esch, D., Foley, B., San, N.: Future directions in technological support for language documentation. In: Proceedings 3rd Workshop on Computational Methods for Endangered Languages, vol. 1, pp. 14–22 (2019). https://doi.org/10.33011/computel.v1i.341
14. Finau, G., et al.: Social media and e-democracy in Fiji, Solomon Islands and Vanuatu. In: Twentieth Americas Conference on Information Systems. Association for Information Systems, Savannah (2014). http://hdl.handle.net/1885/75381
15. Foley, B., et al.: Building speech recognition systems for language documentation: the CoEDL endangered language pipeline and inference system (ELPIS). In: 6th International Workshop on Spoken Language Technologies for Under-Resourced Languages, pp. 200–204 (2018)
16. Foley, B., Durantin, G., Rakhi, A., Wiles, J.: Transcription survey. In: Paper Presented at the Australian Linguistic Society Annual Conference (2019). Retrieved from 15 October 2021. http://bit.ly/ALS-survey
17. Gaved, T., Salffner, S.: Working with ELAN and FLEx together: an ELAN-FLEx-ELAN teaching set (2014). Retrieved from 15 October 2021. https://www.scribd.com/document/357359102/Working-with-ELAN-and-FLEx-together-pdf
18. Guérin, V., Lacrampe, S.: Trust me, I am a linguist! building partnership in the field. Lang. Documentation Conserv. **4**, 22–33 (2010)
19. Howell, K.: Inferring grammars from interlinear glossed text: extracting typological and lexical properties for the automatic generation of HPSG grammars. Ph.D. thesis, University of Washington (2020)
20. Kaltaƀau, G., Emil, L.: Nafsan recordings (GKLE), Digital collection managed by PARADISEC (2017). http://catalog.paradisec.org.au/collections/GKLE
21. Krajinović, A.: Comparative study of conditional clauses in Nafsan. In: Boerger, B.H., Unger, P. (eds.) SIL Language and Culture Documentation and Description 41, Proceedings of COOL 10, pp. 39–61. SIL International (2018). https://www.sil.org/resources/publications/entry/82335
22. Krajinović, A.: Tense, mood, and aspect expressions in Nafsan (South Efate) from a typological perspective: the perfect aspect and the realis/irrealis mood. Ph.D. thesis, Humboldt-Universität zu Berlin and The University of Melbourne (2019). https://minerva-access.unimelb.edu.au/handle/11343/237469
23. Krajinović, A., et al.: Natrauswen ni tesa nen rumtri ki nafsan ni Erakor. ISBN 978-1721654246 (2018)
24. Krajinović, A., (collector).: Nafsan recordings (AK1). Digital collection managed by PARADISEC. [Open Access] http://catalog.paradisec.org.au/collections/AK1 (2017). https://doi.org/10.4225/72/5b2d1d0a315a2
25. Lynch, J.: South Efate phonological history. Oceanic Linguistics **39**(2), 320–338 (2000)

Author Proof

26. Lynch, J., Ross, M., Crowley, T.: The Oceanic Languages. Routledge, London (2002)
27. McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., Sonderegger, M.: Montreal forced aligner: trainable text-speech alignment using Kaldi. In: Lacerda, F., House, D., Heldner, M., Gustafson, J., Strombergsson, S., Wlodarczak, M. (eds.) Proceedings of Interspeech 2017, pp. 498–502. ISCA, Stockholm (2017)
28. Paschen, L., Delafontaine, F., Draxler, C., Fuchs, S., Stave, M., Seifart, F.: Building a time-aligned cross-linguistic reference corpus from language documentation data (DoReCo). In: Proceedings of the 12th Language Resources and Evaluation Conference, pp. 2657–2666 (2020)
29. Povey, D.: The Kaldi speech recognition toolkit. In: Proceedings of IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society (2011)
30. von Prince, K., Krajinović, A., Krifka, M.: Irrealis is real. Language (in press)
31. von Prince, K., Krajinović, A., Krifka, M., Guérin, V., Franjieh, M.: Mapping Irreality: storyboards for eliciting TAM contexts. In: Gattnar, A., Hörnig, R., Störzer, M., Featherston, S. (eds.) Proceedings of Linguistic Evidence 2018: Experimental Data Drives Linguistic Theory. University of Tübingen, Tübingen (2019). https://publikationen.uni-tuebingen.de/xmlui/handle/10900/87132
32. Regenvanu, R.: Afterword: Vanuatu perspectives on research. Oceania **70**(1), 98–100 (1999)
33. Rice, K.: Documentary linguistics and community relations. Lang. Document. Conserv. **5**, 187–207 (2011)
34. Seifart, F., Evans, N., Hammarström, H., Levinson, S.C.: Language documentation twenty-five years on. Language **94**(4), e324–e345 (2018)
35. SIL: Fieldworks Language Explorer (FLEx) 8.3 (2018). Retrieved from 15 October 2021. https://software.sil.org/fieldworks/
36. Taylor, J., Thieberger, N. (eds.): Working together in Vanuatu: Research Histories, Collaborations, Projects and Reflections. ANU Press, Canberra (2011)
37. The Language Archive: ELAN (Version 5.2) [Computer software]. Nijmegen: Max Planck Institute for Psycholinguistics (2018). Retrieved from 15 October 2021. https://tla.mpi.nl/tools/tla-tools/elan/
38. Thieberger, N.: A grammar of South Efate: An Oceanic Language of Vanuatu. University of Hawai'i Press, Honolulu (2006)
39. Thieberger, N.: A South Efate Dictionary. University of Melbourne, Parkville (2011). https://minerva-access.unimelb.edu.au/handle/11343/28968
40. Thieberger, N.: Natrauswen NIG Efat: Stories from South Efate. University of Melbourne, Parkville (2011)
41. Thieberger, N.: Guide to the Nafsan, South Efate collection (2021). Retrieved from 15 October 2021. https://www.nthieberger.net/sefate.html
42. Thieberger, Nicholas with Members of the Erakor Community: A Dictionary of Nafsan, South Efate, Vanuatu: M̃pet Nafsan ni Erakor. Oceanic Linguistics Special Publications No. 41, University of Hawaii Press, Honolulu (2021)
43. Tryon, D.: Ni-Vanuatu research and researchers. Oceania **70**(1), 9–15 (1999)
44. Tryon, D.T.: New Hebrides Languages: An Internal Classification. Pacific Linguistics, Canberra (1976)
45. Vandeputte-Tavo, L.: New technologies and language shifting in Vanuatu. Pragmatics **23**(1), 169–179 (2013)
46. Vanuatu National Statistics Office: 2016 Post-TC Pam Mini Census Report. Tech. rep., Ministry of Finance & Economic Management, Port Vila, Vanuatu (2017)

Author Proof

47. Wessel, P., et al.: The generic mapping tools version 6. Geochem. Geophy. Geosyst. **20**, 5556–5564 (2019). https://doi.org/10.1029/2019GC008515
48. Yamada, R.M.: Collaborative linguistic fieldwork: practical application of the empowerment model. Lang. Documentation Conserv. **1**, 257–282 (2007)
49. Yamada, R.M.: Training in the community-collaborative context: a case study. Lang. Documentation Conserv. **8**, 326–344 (2014)