Building capacity for community-led documentation in Erakor, Vanuatu

Ana Krajinović[1,2,3], Rosey Billington[1,2], Lionel Emil[1,4], Gray Kaltapau[1,4], Nick Thieberger[1,2]
[1]Centre of Excellence for the Dynamics of Language, Australia, [2]University of Melbourne,
[3]Humboldt-Universität zu Berlin, [4]Nafsan Language Team, Erakor Village, Efate, Shefa Province, Vanuatu

Contact: ana.krajinovic.rodrigues@gmail.com, rbil@unimelb.edu.au, gkkaltkpau@gmail.com, thien@unimelb.edu.au

Language Technologies for All, 4-6 Dec 2019, UNESCO, Paris

Paper download: https://bit.ly/37ePmxa

Introduction

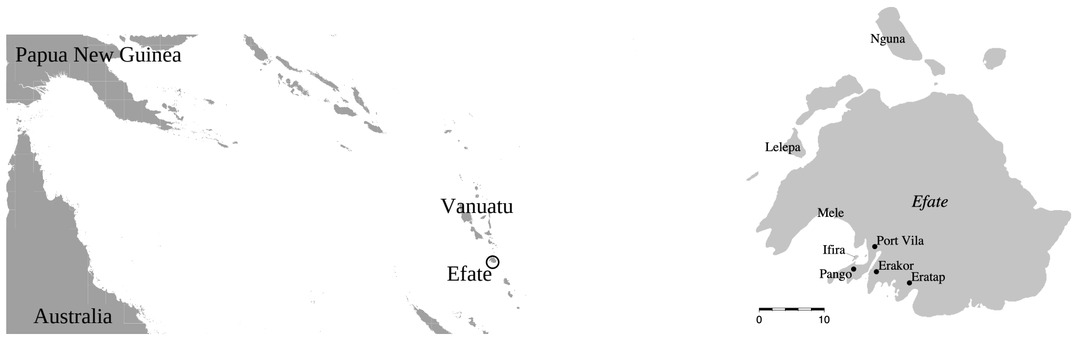

Figure 1: Location of Vanuatu on the left and the island of Efate on the right

- we focus on collaborative work between researchers and community members on Nafsan (Oceanic)
- we argue that benefits of collaborative work outweigh the challenges
- 6,000 speakers (Lynch et al., 2002) of Nafsan in Erakor, Pango, and Eratap
- missionary translations from the 19th century, word list data (e.g. Tryon, 1976)
- comprehensive reference grammar of Nafsan (Thieberger, 2006), corpus data, a book of stories (Thieberger, 2011b), and a dictionary (Thieberger, 2011a)
- the previous research and the practice of returning materials to the community by Nick Thieberger laid the groundwork for the more recent fieldwork

Sharing technical and procedural skills

Figure 2: Nafsan Language Team in Erakor Village

How it started

- following the 2017 dictionary workshop, there was community interest in collecting more stories, continuing to update the dictionary
- a recorder and a computer were made available, later a camera
- need for training in data collection and management
- help with the transcription and audio and video recordings

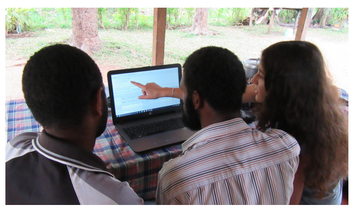

Figure 3: Training in ELAN transcription

Training: audio, video, data management

- using Zoom H1N and camera, choosing the right environment
- discussing consent, spoken metadata
- time-aligned transcription with ELAN: template (Gaved & Salffner, 2014), step-by-step documentation of the process
- file-naming conventions and metadata in preparation for archiving in PARADISEC
- file storage and backup
- understood as a part of the workflow of making a recording

Outcomes of the community-led project

Results

- 17 audio files totaling 05:26:37 (custom and life stories)
- 25 video files totaling 04:25:34 (weaving instructions)
- 7 recordings transcribed fully and 2 partially
- all the recordings archived in PARADISEC: Gray Kaltapau (collector), 2017; Nafsan recordings (GKLE), Digital collection managed by PARADISEC. [Open Access] DOI: 10.26278/5c8fb27b3a40a. `http://catalog.paradisec.org.au/repository/GKLE`

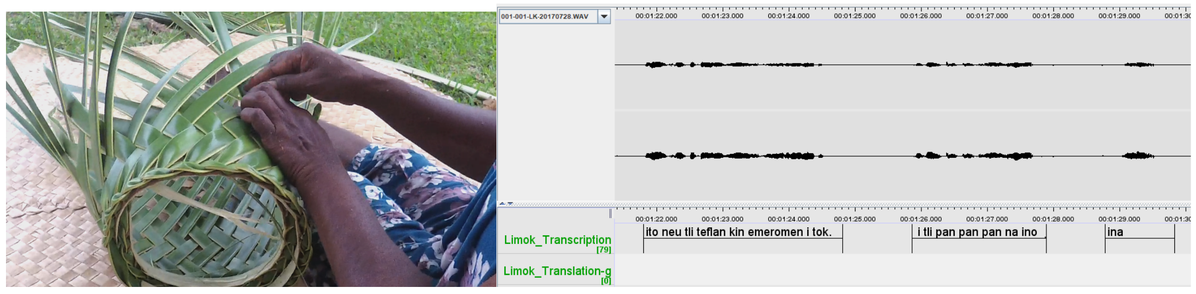

Figure 4: Left: Marian Kalmary weaving *naal pool* (GKLE-013), right: Orthographic transcription of *Napre nig Taler* (a story about a demon) told by Limok Kaltapau (GKLE-001)

Benefits

Community perspective:

- native knowledge of Nafsan facilitates the recording and transcription process
- specific activities are better documented with video than audio
- using knowledge of activities to plan the shooting
- can decide which activities are most important to document for preservation of language and culture

Linguist perspective:

- scale and quality of documentation
- materials representative of community's needs and more useful for supporting language and cultural maintenance
- data management, metadata collection, and discussing access conditions was built into the initial training

Challenges

- sharing one laptop and one recorder and finding time among other commitments
- availability of participants: people are afraid to show up in a recording or video
- stabilizing the camera: particularly hard with some kinds of activities like weaving, partially solved by providing a tripod
- long-term sustainability of this type of collaboration
- difficulty for linguists to allocate time to produce useful outputs for the community

Potential for applications of language technology to less-resourced languages

'Transcription bottleneck'

- more data is recorded than can feasibly be transcribed and added to a corpus (e.g. Brinckmann, 2009)
- Do community-led projects create even more data that *cannot* be easily used by communities or researchers?

Current problem with automatic speech recognition (ASR) in less-resourced languages:

- ASR usually requires very large speech corpora
- noisy fieldwork conditions

Our insight

Community researchers may be better placed than visiting linguists to collect high-quality audio recordings (given appropriate training)!

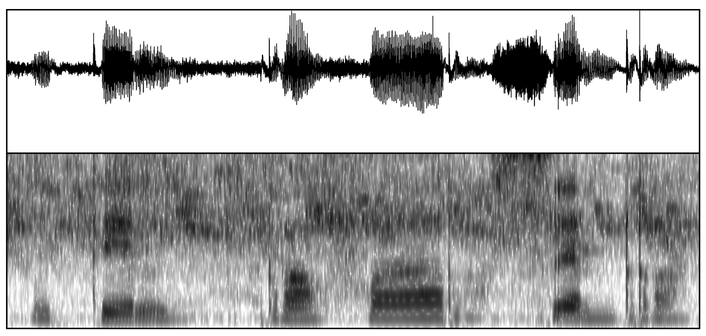

Figure 5: Recording made by the linguist in noisy conditions

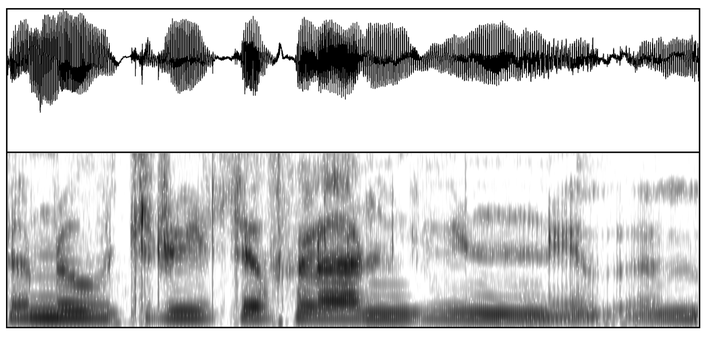

Figure 6: Recording made by the community member in quiet conditions

ASR for Nafsan

- There are now several projects aiming to develop and adapt ASR tools to be applied to less-resourced languages, especially for language documentation (e.g. Persephone)
- first test for ASR for Nafsan done by using Kaldi, via the in-development Elpis pipeline `https://github.com/coedl/elpis` (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% (Foley et al., 2018)
- the case of Nafsan shows the potential for the increased amount of data emerging from collaborative projects to be more easily processed and made useful
- it is not out of reach to develop language technology for less-resourced languages

Community-led documentation projects have clear benefits:

- results in data useful for the community:
  - using knowledge of activities to decide what to document, and how (audio, video, under which conditions)
- data useful for linguists and language technologies:
  - larger collections of high-quality data, transcribed by native speakers
  - might have even more potential to be used in training automatic speech recognition (ASR) than data collected by linguists

References

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

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 & 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*, 200–204.

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

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

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. `https://minerva-access.unimelb.edu.au/handle/11343/28968`.

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

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