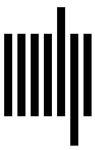
The Open Handbook of Linguistic Data Management

Edited By: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister

Citation:

The Open Handbook of Linguistic Data Management
Edited By: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister
DOI: 10.7551/mitpress/12200.001.0001
ISBN (electronic): 9780262366076
Publisher: The MIT Press
Published: 2022



The MIT Press

35 Managing Data Workflows for Untrained Forced Alignment: Examples from Costa Rica, Mexico, the Cook Islands, and Vanuatu

Rolando Coto-Solano, Sally Akevai Nicholas, Brittany Hoback, and Gregorio Tiburcio Cano

1 Introduction: Why did we use untrained forced alignment?

Forced alignment is a technique to align the audio signal of a spoken utterance with its transcription, so that the boundaries between words, and even between phones,[1] can be determined automatically (Wightman & Talkin 1997). Figure 35.1 shows an example of this in Spanish, where the spectrogram of an utterance is accompanied by lines indicating the approximate temporal limits of words and individual phones within the recording. The main advantage of this technique is that it accelerates phonetic research: the algorithm can tag these boundaries approximately thirty times faster than expert humans can (Labov, Rosenfelder, & Fruehwald 2013).

Forced alignment depends on a language model that describes the spectral features that characterize each of the phones in the sample provided to the algorithm, as well as the probabilities for the occurrence of different phones and words within the language sample. These probabilities are derived from a supervised training set: annotated examples are provided to the algorithm, so that it can learn to connect spectral features to phones. Training these requires large amounts of data (e.g., twenty-five hours for the American English model in the Forced Alignment and Vowel Extraction [FAVE]-align aligner [Rosenfelder et al. 2011]). Because such large data sets are not available for most Indigenous and minority languages, a technique called *untrained* forced alignment has been devised (DiCanio et al. 2013; Strunk, Schiel, & Seifart 2014). In untrained forced alignment, the model for one language (e.g., English) is used to process the phones of a different language (e.g., Cook Islands Māori). This is possible because, even though the words for both languages are completely different, many of the phones are acoustically similar, so that bootstrapping is possible. For example, the 'm' phones in both English and Cook Islands Māori have similar spectral cues, so the English model's idea of an 'm' can also find 'm's in the Cook Islands Māori data. Many of the phones are not similar. For example, the glottal stop of Cook Islands Māori /ʔ/ has no direct equivalent in American English, French, Spanish, or other European languages with available models. However, the phones that are not available in English can be approximated. For example, the /ʔ/ stops the air flow like /t/ and /k/ do, and these similarities have been exploited to detect /ʔ/ in languages such as Triqui (DiCanio et al. 2013). These transformations allow for the use of an existing model with audio from another language, and because the model has not been explicitly trained on data from the Indigenous language, we say that this method is untrained forced alignment.

This untrained method has been fruitfully applied to languages such as Triqui from Mexico (DiCanio et al. 2013), Nikyob from Nigeria (Kempton 2017), Nambo and Matukar Panau from Papua New Guinea (Kashima et al. 2016; González et al. 2018), Australian Kriol (Jones et al. 2017, 2019), and Tongan (Johnson, Di Paolo, & Bell 2018). Research teams have also created pan-language models to take advantage of the similarities across an entire language family and pooled together the resources from these various underresourced languages to increase the size of the training set, such as the Australian language aligners in WebMAUS (Strunk, Schiel, & Seifart 2014).

On first encountering this technique, we tested and confirmed its potential to accelerate phonetic research in underresourced languages, as well as to contribute to the larger program of natural language processing (NLP). We investigated the potential of using forced alignment to create new data sets for the training of other NLP techniques, including speech recognition, automatic part-of-speech tagging, and parsing. Finally, the data

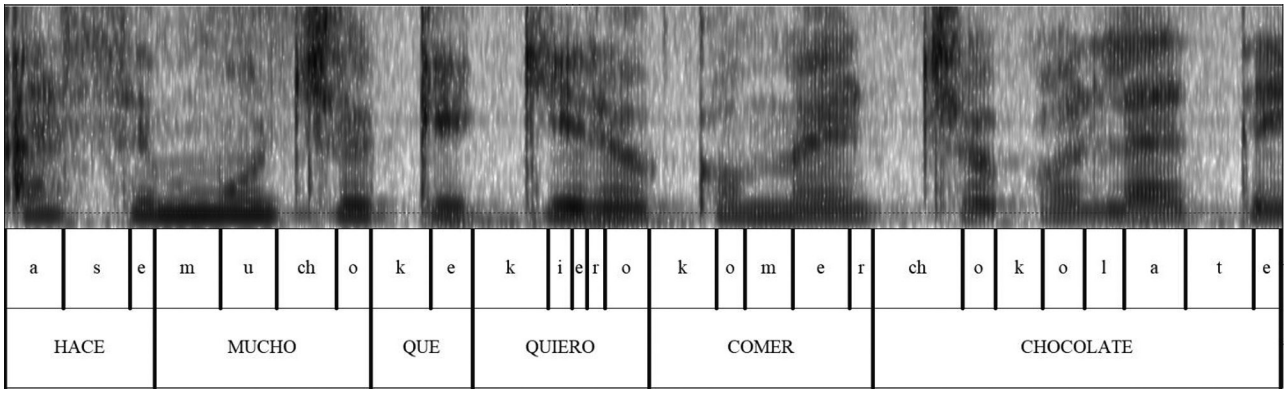


Figure 35.1
Praat (Boersma 2001) spectrogram and TextGrid with the Spanish phrase 'I've been wanting to eat chocolate for a long time,' with words and phones aligned.

sets we generated can themselves be corpora that can be used by linguists and community members for their research. As we did this, we found numerous challenges in data management given the many intermediate steps between transcriptions and completed TextGrids, and the many intermediate files generated along the way. We also faced issues with sharing the data in accordance with community principles of access and sharing.

The following sections report on these challenges. Section 2 focuses on our initial attempts and shows a 'typical' attempt to combine NLP and linguistics: disorderly and with little replicability. Think of it as a guide to how not to perform forced alignment. Section 3 reports our attempts to improve the data management flow, from collection to sharing, as well as the incorporation of consultation for the purpose of public archiving and reuse of data for language revitalization and to train other NLP applications. Section 4 focuses on software maintenance and replicability of the algorithms themselves across different platforms and users. Sections 5 and 6 present our ideas about a functional workflow for data management in forced alignment, and our recommendations for those who wish to replicate this technique.

While doing this, each section will also present six use cases of adaptation of forced alignment to six underresourced languages: Bribri, Cabécar and Malecu (Chibchan, Costa Rica), Me̲'pha̲a Váthá́á (Otomanguean, Mexico), Cook Islands Māori (Austronesian, Cook Islands), and Denggan (Austronesian, Vanuatu).

Before we begin, it is useful to clarify that this is just one way of performing forced alignment, and there are in fact numerous other aligners and ways of doing alignment. In addition to the P2FA (University of Pennsylvania Phonetics Lab Forced Aligner) algorithm (Yuan & Liberman 2008), which is used in FAVE-align (Rosenfelder et al. 2011), there are other algorithms for forced alignment, such as the Montreal Forced Aligner (McAuliffe et al. 2017), which is used in the Dartmouth Linguistic Automation, or DARLA, website (Reddy & Stanford 2015). There are also EasyAlign (Goldman 2011), Prosodylab-Aligner (Gorman, Howell, & Wagner 2011), LaBB-CAT (Fromont & Hay 2012), and the Munich Automatic Segmentation System, or MAUS (Strunk, Schiel, & Seifart 2014). Despite this wealth of options, we believe the workflow presented here represents a valid case for data management in forced alignment studies.

2 Good alignment, bad data management: Forced alignment and languages in the Americas

Our first attempt to run these algorithms on linguistic data was with languages of the Americas. We tested the feasibility of the technique by using previously existing recordings from three Chibchan languages from Costa Rica: Bribri, Cabécar, and Malecu (ISO 639-3: bzd, cjp, gut) (Coto-Solano & Flores Solórzano 2016, 2017). After that, we ran the algorithms using fieldwork-collected data from the Otomanguean language Me̲'pha̲a Váthá́á (no specific ISO 639-9 code, Glottocode: zila1239) in order to study its tonal phonetics (Coto-Solano 2017).

The phonological systems of these languages have numerous elements that are different from those in European languages that could potentially pose challenges for English language models or other models based on

well-resourced languages. For example, three of the languages are tonal (Bribri, Cabécar, and Me̲'phaa Váthá), while Malecu has phonemic vowel length. The Chibchan languages all have liquids that don't appear in English, French, or Spanish, such as /l, r, ɾ/ for Bribri and Cabécar and /ɬ, r, ɾ/ for Malecu. Bribri, Cabécar, and Me̲'phaa Váthá have contrastive vowel nasality, and Me̲'phaa Váthá has aspirated and prenasalized stops: /p, p^h, m, ^{m}b/, /t, t^h, d, ^{n}d/. Given all of these differences, we tested the performance of English and French language models to align Chibchan data (Coto-Solano & Flores Solórzano 2016), and after determining that the English model had better performance, we used it to align Me̲'phaa data (Coto-Solano 2017). We used a web-based interface for the aligner (Rosenfelder et al. 2011), and the performance of the algorithm was satisfactory: 8% error when aligning the center of words (i.e., the center of the word was off by 8% of the duration of the word), and 23% error when aligning the center of vowels compared to alignment made by a human expert.[2] However, after our initial publications, we found that we had a mess of files and experiments strewn all over our personal computers, with little hope of being able to replicate it without starting from scratch.

Figure 35.2 shows the workflow for generating and processing information from forced alignment algorithms. In particular, it shows fourteen steps where files are generated, modified, or susceptible to updates and changes. These files include the source files for the recording and its transcription (steps 1 and 2), Python code to generate intermediate transformations of the data (steps 5, 8, and 9), and text files that contain those intermediate transformations (steps 3, 4, and 6). Using all of these, we generate Praat TextGrids (Boersma 2001) with the alignment of the audio and transcription (steps 7 and 11), as well as comma-separated values data sets (steps 10 and 13) that are then processed using statistical software such as R (step 14) (R Core Team 2017).

The main challenge is to transform the data into a form that can be processed by an English aligner (the *data preparation* section of figure 35.2), and then to undo those transformations so that the data are not 'deformed' by having fit it into an English language mold. Critical to this is the *glyph management*. This entails three processes: First, we have to study the way in which the alignment system expresses the phonology of English. For example, FAVE-align uses the Arpabet transcription system (Zue & Seneff 1996). *Arpabet* is an English-based phonetic alphabet that expresses a word such as 'read' using the following glyph sequence: R IY1 D. The glyphs R and D represent the phones 'r' and 'd', while the glyph IY1 represents the vowel. The number 1 represents the fact that the vowel

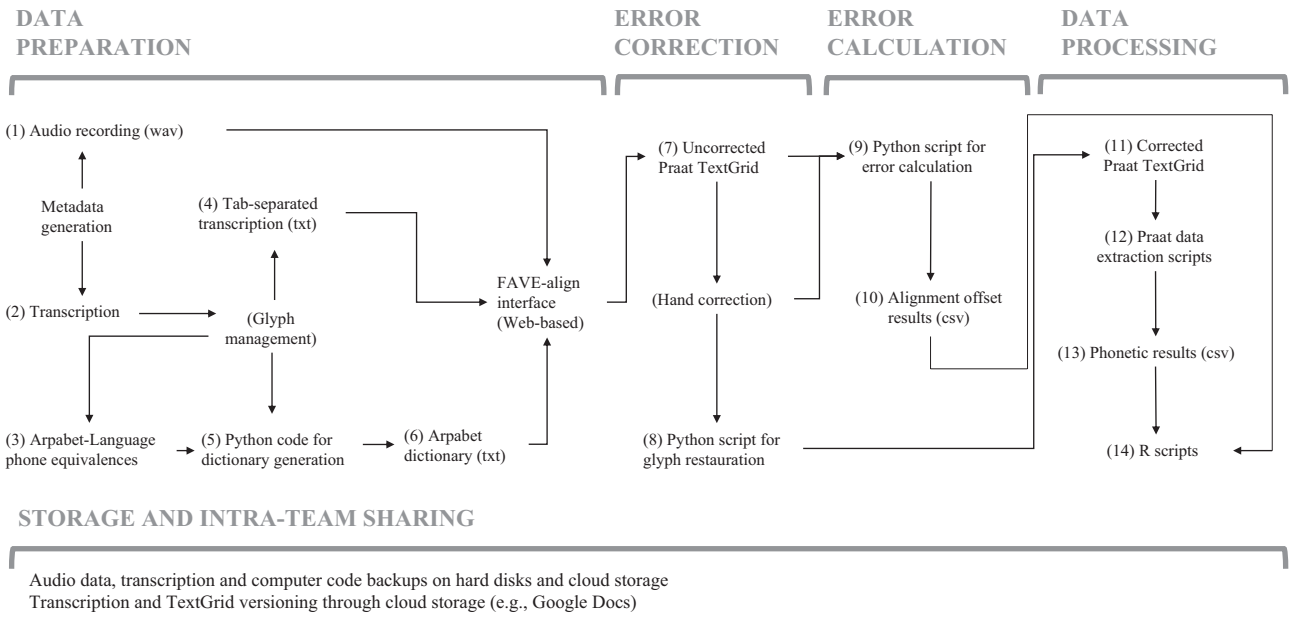


Figure 35.2
Workflow for untrained forced alignment with FAVE-align. The numbers indicate steps where data files are generated during the alignment process.

is in the stressed syllable of the word. (Arpabet also uses 0 for unstressed syllables and 2 for syllables with secondary stress). Second, in the glyph management process, we have to decide which phones of the original language correspond to which glyphs in English Arpabet, so we can express the Chibchan words in the recording using the Arpabet transcription system. For example, the Bribri phones /a/ /aɭ/, /à/ /aɬ/, and /á/ /aꞯ/ only have one equivalence in English Arpabet: AE1. This means that both the words *alà* /aɬˈɭaɬ/ 'child' and *alá* /aɬˈɭaꞯ/ 'thunder' will have the English Arpabet AE1 L AE1. (The tonal contrast between these two words is lost at this point, but it will be recovered on step 8 of figure 35.1, *glyph restoration*.) The third issue in glyph management has to do with encoding compatibility. Depending on its configuration, the aligner might be unable to process Unicode characters that appear in ELAN transcriptions (Wittenburg et al. 2006). This means that a mark such as that for nasalization, which is written with a line under the vowel (e.g., ù /ũɬ/ 'pot'),[3] had to be expressed in some other way in the tab-separated transcription that will become the input for the aligner. For example, the glyph {h}, which is not used in the Bribri orthography, was chosen to represent nasality, so that nasal /u/ with a high tone (ù) becomes ùh in the tab-separated transcription. Python scripts deal with both problems of glyph management, creating a dictionary (steps 5 and 6) and then transforming the uncorrected TextGrids from Arpabet (AE L AE1) back to the original representation of the language (*alà*) (step 7). If the reader is interested in the specifics of how these dictionaries are structured, section 2.1 of Coto-Solano and Flores Solórzano (2017) provides a detailed explanation.

These dictionaries are used not only to instruct the aligner on the phones of our languages, but they are also used to reconstruct phonemic contrasts that were lost during the alignment process. Figure 35.3 shows TextGrid tiers for the Bribri phrase *ì kuéki wìm òr darèrè* 'why does the monkey shout so loud' (Jara Murillo & Segura 2009). Figure 35.3a shows the output of the aligner, with the glyphs still in Arpabet. Some of the phonemic distinctions of Bribri are not represented. For example, the word *ì* 'what' has a high tone oral vowel /iɬ/, whereas the final phoneme of the word *kuéki* /kwẽꞯkĩɭ/ 'because' has a low tone nasal vowel /ĩɭ/. Despite these two being different, they are both represented by the Arpabet IY1. Figure 35.3b shows the TextGrid where the Arpabet has been replaced with the original labels for each glyph (i.e., {ì} and {ih}). These changes are effected by Python code (available in the GitHub repository https://github.com/rolandocoto/cim-aligned). This code goes through the dictionary and looks for the phones that were originally in the word. This restores any lost phonemic distinctions and leaves the TextGrid in a form that is usable for phonetic research.

A second issue in data management came with the correction of the raw results from the forced alignment (the uncorrected TextGrid in step 7 of figure 35.2). After performing the forced alignment, the resulting Praat TextGrids need to be checked and hand-corrected so that the boundaries detected by the system correspond to the actual boundaries of the phones.

This is an exacting process and is best achieved by a multiperson team with specialized knowledge in phonetics, where the team is jointly deciding on the

(a)

IY1	K	W	EHI	K	IY1	W	IY1	M	OW1	R	D	AE1	R	IH1	R	IH1
Ì	KUHÉHKIH					WÌM			ÒHR		DARË‘RË‘					

(b)

ì	k	uh	éh	k	ih	w	ì	m	òh	r	d	a	r	ë	r	ë
Ì	KUHÉHKIH					WÌM			ÒHR		DARË‘RË‘					

Figure 35.3
Praat TextGrid tiers with phones and words from Bribri: *ì kuéki wìm òr darèrè* 'why does the monkey shout so loud?' (a) The phone tier has Arpabet glyphs. (b) The phone tier now has the phones in a form closer to the standard spelling of Bribri, and phonemic distinctions such as tones have been reinserted into the TextGrid.

numerous issues that will arise during boundary marking. For example, phones can be phonetically combined or reduced to the point where it's difficult to neatly separate them from others (Ernestus & Warner 2011), and the correction team will need to make decisions about how to manage these reductions. This marking will have an impact on the type of data that can be extracted for research down the line.[4]

A third issue in data management was the sharing of transcriptions, recordings, and TextGrids, so that they could be accessed by all team members and backups could be kept. At this stage we used consumer-level password-controlled cloud storage (e.g., Google Docs). We used relatively simple versioning,[5] with file names indicating the latest edition of each file, and regular backups limited only to the main inputs of the alignment algorithm (source recordings and transcription: steps 1 and 2), the main outputs (the TextGrids: steps 7 and 11) and some of the code (the R scripts: step 14).

The management of the other files did not take sharing or future use into account. Much of the code that generates the intermediate data transformations was only accessible to one member of the team (the person who ran the NLP algorithm), and it was not accessible at all by other scholars. As for the dictionary, no effort was made to generate one 'large' dictionary that contained all the words from all the transcriptions. Instead of creating one large file that could serve as a main dictionary and progressively populating it with the words from each recording (so that in the future we have a dictionary like the English Carnegie Mellon University Pronouncing Dictionary), we used a minidictionary that we had to generate for each recording. Finally, the code for the aligner itself was being used from a third-party interface (Rosenfelder et al. 2011), which meant that we didn't have to access the code itself to run the program. Having a prebuilt interface made it easy to use the aligner at the time, but as we discuss in section 4, it proved to be a vulnerability once that server was taken offline.

In summary, at this point of our work, the data management of the project was haphazard. This is probably typical of other linguists using NLP tools: files stored only in the researchers' devices, code that isn't shared, and complex software that the researchers use but ultimately don't control.

3 Sharing our results: Alignment in Cook Islands Māori

The real overhaul came with the work on Southern Cook Islands Māori (CIM; ISO 639-3: rar). Here, we attempted to involve the community members in our decisions for data management and dissemination, and we also improved the versioning of the files in the project.

CIM is an East Polynesian language originating from the Southern Cook Islands and spoken by approximately 1,500 people in the Cook Islands and in diaspora populations in New Zealand and Australia. Within the language community, the majority of whom reside in a diaspora, there are very low rates of intergenerational transmission and, as such, there is a very advanced shift toward English as the main language of communication (Nicholas 2018:46). There have been recent efforts to increase the documentation for the language (Nicholas 2017) that have resulted in a corpus of spoken CIM of approximately eighty hours. The manual orthographic transcription of this collection is in progress; at the time of writing, about 30% of it has been transcribed using ELAN. The existence of these transcriptions has allowed the possibility of using untrained forced alignment to study the phonetics of the language. For example, we used this method to describe the distribution of glottal stop realizations throughout the Southern Cook Islands. These realizations range from a full glottal stop (e.g., *ta'i* ['ta.ʔi] 'one') on some islands, to a realization without glottal closure but with creaky voice on the vowels on other islands (e.g., *ta'i* ['ta.ḭ] 'one') (Nicholas & Coto-Solano 2019).

The input for the CIM alignment came from a documentation project that had been ongoing for multiple years, with many speakers and sources for the recordings, as well as different file formats. This meant that the data required more metadata and curation to manage them. For example, the recordings were captured with both audio and video, significantly increasing both the number and size of the files to be managed. In addition to this, many of the recordings involve naturalistic conversation, which is linguistically rich, but difficult to prepare for NLP processing. The audiovisual material of the corpus (Nicholas 2012) is stored and made publicly available at the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC; http://www.paradisec.org.au), but when we first began

the forced alignment for CIM, the file transcriptions were stored in the main researcher's personal computer, which was different from the computer where the data were being aligned. To add to these issues of file management, early on we found a divergence in transcription styles: sometimes the needs for 'linguistically oriented' documentation were not the same as those of the alignment algorithm. The following are three examples of these differences: First, linguistically oriented transcriptions need to retain punctuation and case sensitivity, whereas transcriptions for forced alignment usually need to be as simple as possible, eliminating punctuation and reducing the number of glyphs in the text to those that correspond to sounds in the recording. Second, linguistically oriented transcriptions also have multiple tiers, including tiers for translations, annotations of linguistic features (e.g., morphemic glossing, discourse level tags), while forced alignment transcriptions need to be separated by speakers and almost completely free of data that are not directly related to the signal. Third, linguistically oriented transcriptions need to include instances of code-switching (e.g., between Spanish and Bribri or between Bislama and Denggan) to present a complete picture of the linguistic patterns of the speakers. Processing multilingual data, however, would further complicate the use of untrained forced alignment, as well as the generation of training sets for further NLP tools.

One final difference between linguistically oriented and alignment-oriented transcription styles is that transcriptions for forced alignment need to have a more precise temporal delimitation of each separate utterance. In linguistically oriented transcriptions, the temporal limits for each utterance can include noise or even phrases in other languages without degrading its usability. Noise, on the other hand, can engage the alignment at the wrong points of the recording. 'Noise,' by the way, includes roosters, children, pigs, dogs, wind, and cicadas, elements that are common to all the fieldwork settings presented in this chapter. Because those are constant, their influence has to be managed and minimized. In the future, we hope to bring noise-robust speech recognition techniques from large languages to help with this in fieldwork settings (Liao & Gales 2008; Seltzer, Yu, & Wang 2013). The interruptions of other people and other background conversations can also result in errors in alignment. When eliciting word lists, for example, false starts of words or half-utterances can fool the alignment program into thinking that they're the beginning of the word, and then an indiscriminate vowel sound or hesitation might extend a real target's ending. All of these details of the transformation from transcription to the 'tab-separated transcription' intended as input to the aligner demanded attention and time when generating the input for the forced alignment algorithm.

As described in section 2, TextGrids generated from forced alignment need to be corrected to make sure that the boundaries are aligned with the phones correctly, a time-consuming process that benefits from a multiperson workforce. These corrections distributed across a larger team, combined with the many intermediate files generated during the alignment process, quickly caused an organizational chaos that forced us to rethink our file management process. We started pushing critical files (e.g., uncorrected, partially corrected, and completely corrected TextGrids) onto a GitHub repository (in this case https://github.com/rolandocoto/cim-aligned). This has been shown to provide effective versioning for code and for linguistic data (Partanen 2016), so we could have files at different versions of progress and work on them asynchronously. Very quickly it was obvious that other files, such as ELAN files, the Arpabet dictionaries, and the R and Python code would also benefit from stable backups and versioning.

This process brought to the fore questions about the public availability of our data. Complete access to data is not always desirable in the context of Indigenous language work because the data might contain data that are private to a family, private for religious or cultural reasons, or that constitutes an intangible cultural or intellectual property of the community, and are therefore susceptible to appropriation or theft (cf. Dwyer 2006; Kukutai & Taylor 2016; Keegan 2019). Great care must be taken to make sure the needs of the language community are being met in this regard. In the Cook Islands context, there is widespread support for open access to most types of linguistic materials due to the perception that this approach provides the best chance for successful language revitalization. During the documentation process, the contributors are always given the option to exclude their recording, or part of it, from the open access corpus or to ask for it to be excluded at a later date. However, we must always remain responsive to the wishes of the contributors, and we must continue to make sure contributors understand what might happen to and

result from their recordings and keep them informed as these possibilities change with technological advances.

In addition to public access for the data sets, the goals of the documentation project included contributing to the revitalization efforts of CIM. Therefore, alongside our data management needs for supporting the scholarly use of the results of forced alignment, we also needed to overhaul the mechanisms to report our findings to make sure that they would produce materials accessible to the language community that would benefit their revitalization efforts. As it stands, the community has several ways to access the raw data. The audiovisual recordings themselves, as well as the aligned and corrected TextGrids can be accessed through PARADISEC (Nicholas 2012). There is also a text-based corpus in the Annotation of Information Structure search and visualization platform (Centre of Excellence for the Dynamics of Language & Nicholas 2019). Finally, the code, TextGrids, and ELAN files needed to replicate the alignment are available on a GitHub repository. However, these formats are not layperson friendly, so we've had to think of a number of ways for community members to use our materials so that they can use them when teaching the language to themselves or others. These community-oriented resources need to occupy less bandwidth, be topically focused, and be easily accessible through a wide range of devices. The most well-populated of these resource channels are the YouTube channel *Araara Māori Kuki Airani* (Nicholas 2019a) and the GERLINGO collection of CIM narrations (Nicholas 2019b). We are also currently engaged in an ongoing educational relationship with the language community. Specifically, we work with primary and secondary school language teachers, coordinating classes on how to use all these community-facing resources as well as how to contribute to creating more of them.

Finally, we have taken the wealth of well-annotated data that the forced alignment workflow has generated and started using it to train NLP algorithms, in the hopes of accelerating the documentation process. We have begun training automatic speech recognition algorithms so that the transcription can be automated and turned into a process of mere correction and verification. As we produce more transcribed data, these could be fed into the algorithm again, increasing the accuracy of the automated transcription (Foley et al. 2018; Foley, Van Esch, & San, chapter 36, this volume). We have also used the data to train part-of-speech tagging (Coto-Solano, Nicholas, & Wray 2018), which is currently working with 92% accuracy (currently at http://cimpos.appspot.com). We hope to create a 'virtuous circle' where more documentation can be used to train more NLP tools. This will not only help with the documentation process but also help us bring the language closer to the computer devices that permeate people's lives, thereby creating a 'symbolic impact' in favor of language reclamation (Aguilar Gil 2014; Jones & Ogilvie 2013; Terrill 2002) and expanding the domains of usage of CIM.

Figure 35.4 shows a summary of the workflow for the untrained forced alignment of CIM data, including the data transformations, and the process for storage, archiving, and sharing research outputs with the community and with academia.

As we learned more about managing our files, new challenges emerged. One of them led us in a new direction: How can we make sure that other researchers can replicate the workflow, and that they can run the code and use untrained forced alignment to work on their languages?

4 Maintaining reproducibility: Vanuatu and the future

After early successes in using untrained forced alignment to generate research on Me̲'pha̲a Váthá́á and CIM phonetics, we decided to expand this technique to a new language. Denggan, formerly known as Banam Bay Language or Burmbar (ISO 639-3: vrt), is a language spoken in the Pacific archipelago of Vanuatu. Denggan is an Austronesian language spoken by an estimated nine hundred language users and has recorded word lists of 300–1,500 words (Tryon 1976; Charpentier 1982), but it is otherwise undocumented. While it still maintains high vitality and intergenerational transmission, it is considered endangered based on the small speaker population and the prominence of the national language, Bislama, in official domains such as church and government and in communication with people from different language backgrounds throughout Vanuatu. Our immediate goal was to use fieldwork recordings of both casual and elicited speech (i.e., Swadesh word lists) to plot F1 and F2 formants, so that we could start the documentation of the vowels. This is part of a larger project by Hoback to document the grammar of the language and aid in creating language maintenance resources.

The lessons learned from the archiving of CIM data and community sharing were useful for the Vanuatu

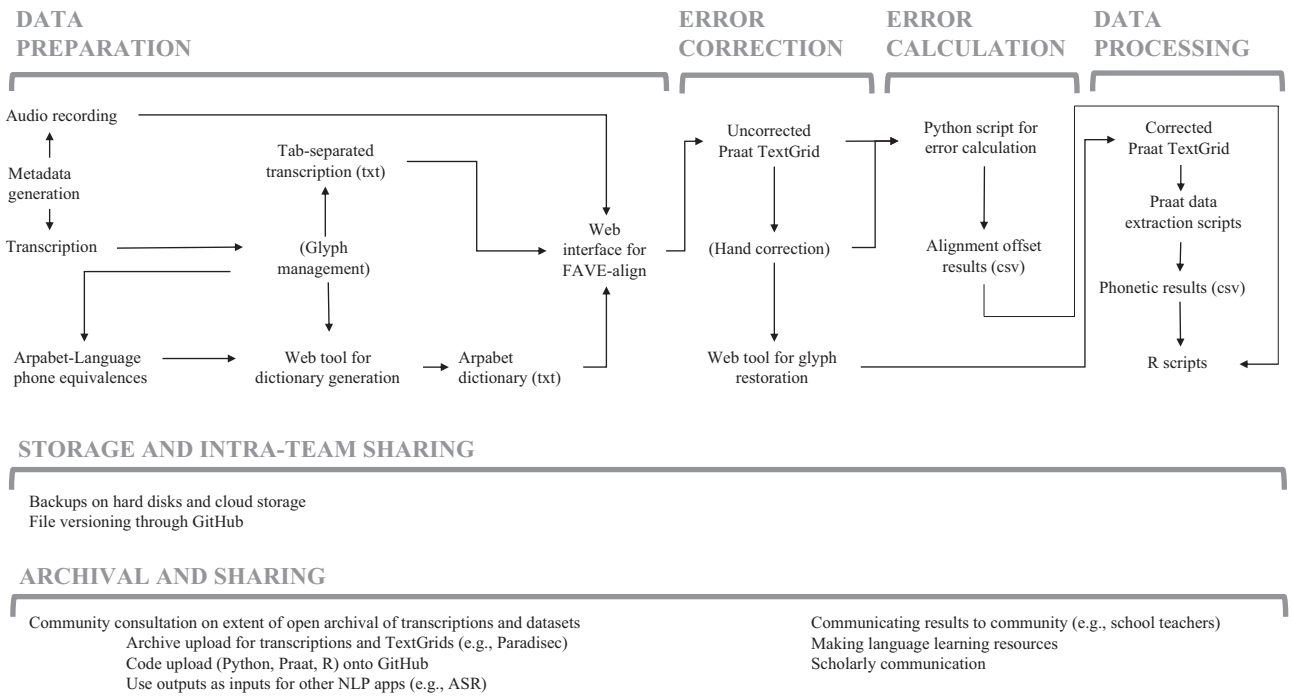


Figure 35.4
Workflow for CIM data. ASR = automatic speech recognition.

case, given that this had a formal requirement for its related files to be archived in the Endangered Languages Archive (ELAR; https://www.soas.ac.uk/elar/), and that it belongs to a much larger language documentation project. Like with the CIM archival materials, the resources for Denggan were archived both as sound files, ELAN transcriptions, and corresponding TextGrid files (Hoback 2019; ELAR deposit identifier: 0549). These were bundled together through CMDI Maker (Zimmer 2014) and can be accessed separately for use of the community (of which audio files or transcriptions would be the most useful). It can also be accessed as a full bundle, including aligned TextGrids, which can be accessed to provide accountability and replicability of the phonetic analysis. The majority of these resources are open access to registered users of ELAR. For the community, this means that there is one registered account for the whole community that individuals can use to access the resources in ELAR. As found in the CIM project, accessing the files of interest can be difficult for speakers through this interface. It also requires downloading the files (e.g., video files of several gigabytes), which is not the most economical way of accessing materials through mobile data, which is the main Internet access for the majority of community members. We are therefore trying to use other means to distribute language resource files, particularly through social media sites such as YouTube or Facebook, whose interfaces are familiar to Denggan speakers.

Denggan presented some challenges that were now familiar: When transforming the data, we found numerous phones that did not correspond to any English Arpabet glyphs. For example, Denggan has the prenasalized consonant /^{m}b/, as well as a uvular fricative [χ]~[ʁ] that also affects the acoustic characteristics of the surrounding vowels. More importantly, the recordings from Denggan were more multilingual than the ones we had processed before were. The Chibchan and Me̲’pha̲a̲ recordings were almost exclusively monolingual. The CIM recordings had code-switching among English, CIM, and occasional words in Te Reo Māori from Aotearoa/New Zealand. However, the recordings from Vanuatu constantly used at least three languages: English, Bislama, and Denggan. English and Bislama were used as contact languages and as a way to elicit Denggan data. Because the transcriptions were segmented into continuous conversation chunks, the aligner would pick up the prompt in the contact language if there was a sound that also corresponded to an expected English phone. This caused the alignment to begin at the prompt and extend, usually, to the end of the actual target word. In addition to this, the phonotactics

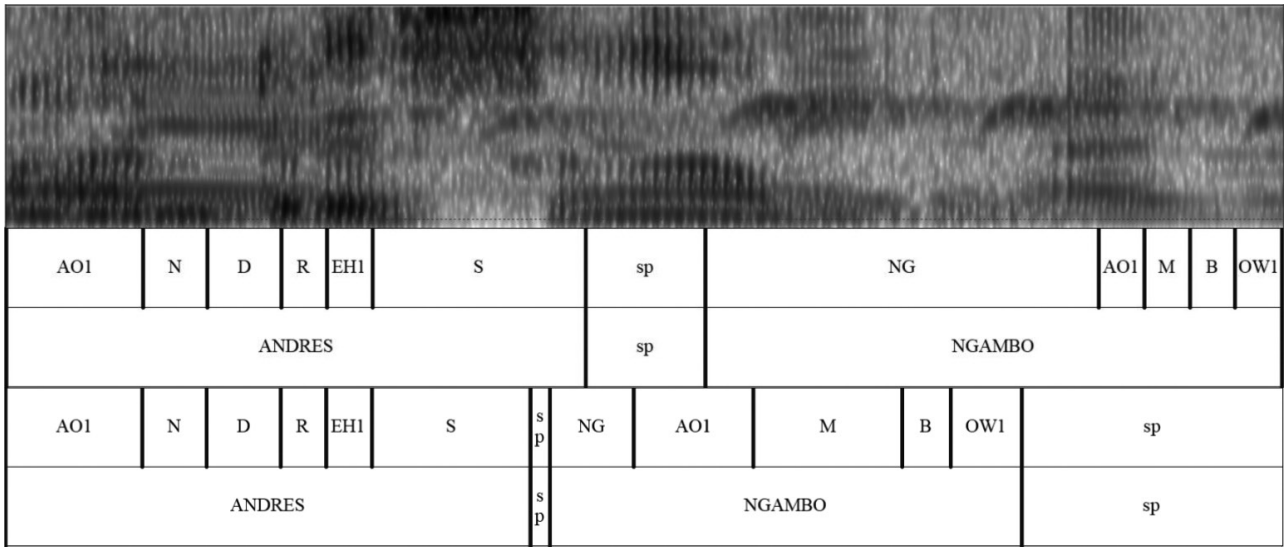


Figure 35.5
Alignment of word-initial /ŋ/ in Denggan data. The first TextGrid tier shows an automatic alignment for /ŋ/ that extends beyond the end of its word. The third tier shows the hand-corrected alignment.

of Denggan would sometimes affect the alignment. For example, the velar nasal /ŋ/ can occur word-initially (as shown in figure 35.5). Because English phonotactics do not allow for this, the English-trained algorithm had problems identifying them in the Denggan recordings and would often misalign word boundaries. This resulted in more manual hand-correction and a workflow similar to that for CIM.

One new challenge, however, had a major impact on the project: While working on CIM, and right before we started working with Denggan, we lost access to the web interface for the aligner software. This website was taken offline, and this left the community of FAVE-align users scrambling to find solutions. The most common fix was to download the FAVE-align code to our own computers and run it via Python. Here, the problems of encoding and code maintenance came back with a vengeance. We were working on multiple operating systems (Windows, Ubuntu, and Mac OSX), and while we installed the software, trained ourselves in running it manually, and made sure our transcriptions were compatible across all platforms, the temporary files piled up in multiple computers with little thought given to versioning or backups.

Our first attempt to deal with this issue was to create a short manual with command line instructions, so that we could run it on our personal computers. This worked for a while, but it was highly dependent on the experience that each user had with command line work (which in this case was the *bash* command language). Our second attempt was to run our code using a remote virtual machine, instantiated using an Amazon Web Service AWS-EC2 CentOS (https://aws.amazon.com/es/ec2/) and accessible through a secure shell protocol.[6] We ran a workshop using this method, and while it was intimidating to new users, they could use this remote text-based interface to align novel data after about four to five hours of training.[7] These solutions worked, but they still presented formidable obstacles to first-time users. Ultimately, a solution where all the team needs programming experience is completely unsustainable.

After this workshop, we decided that a web-based graphical interface was necessary. This new interface would both allow us to customize the code and give the users an easy way to run the scripts and automate the data transformations without resorting to text-based instructions to a Python script. The resulting interface can be used at http://icldc-align.appspot.com. This new interface operated using the AWS storage and virtual machines as the back end for the processing, and a Java-based interface for the front end. Figure 35.6 shows a screenshot of the instructions to align a recording.

We were able to test this new interface at a workshop in the International Conference on Language Documentation and Conservation conference at the University of Hawai'i, in two sessions attended by over a hundred people (Coto-Solano et al. 2019). Each of the

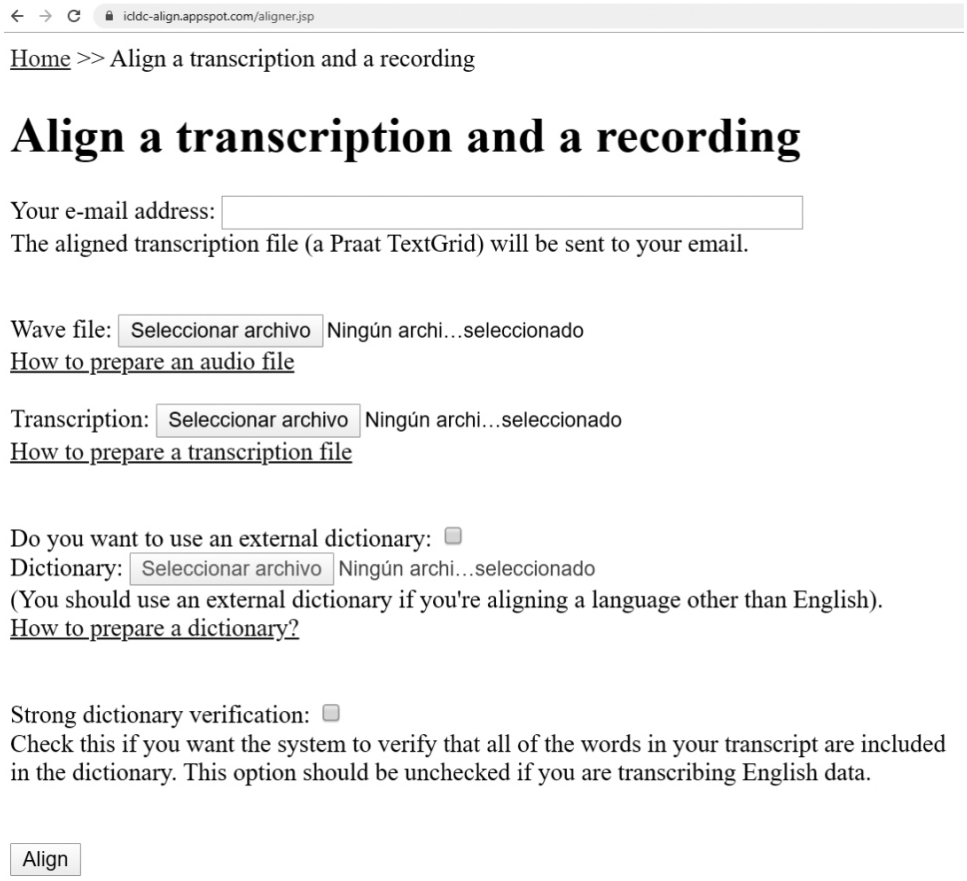


Figure 35.6
Interface to align audio and transcription using FAVE-align.

sessions lasted about ninety minutes, and users were able to generate alignment for previously untrained languages during that time. This new interface also has an important element: we added tutorials for how to download the code and run all the steps of the alignment from a user's personal computer. We also pointed users to the GitHub repository and to the documentation there. By taking these steps we are adding a layer of protection so that users can run the code on their own in case our own website goes offline. We hope that this will increase the reproducibility of this method of linguistic research.

5 Key elements of data management and future work

As we developed the methods of untrained forced alignment, we quickly found that linguistic data management is linked to the ethics of fieldwork and documentation, to Indigenous data sovereignty, and to the intersection of these with ideas from the open access and open science movements. As figure 35.4 shows, the life cycle of the data includes not just the input and the output for the NLP algorithm, but also considerations of (1) how the data and the outputs will benefit the community that speaks the language, (2) how much of it can be shared with researchers are do not belong to the community, and (3) how the publicly shared outputs can serve multiple stakeholders, including other academics, language learners and teachers, and activists involved in language reclamation projects. These ethical considerations are a key part of the data management process, as they will determine which data are archived where, who has access to backups and intermediate files, and how the data will continue to inform research on the language.

The untrained forced alignment workflow also made us examine the reproducibility of computer code environments and the creation of asynchronous workflows for large teams. This made us recapitulate the principles of open source code and thinking of our in-house code as objects that could potentially be useful to other researchers. Also, the concept of data sovereignty directly

affected us. We needed to understand the code and create an environment where we could ensure its future execution so we could continue to use this technique.

One key element is that many of the issues we found were recurrent across these six very disparate languages. The issues in CIM glyph and data management were mirrored in Denggan, for example, and therefore such issues can be addressed and resolved similarly. We are increasingly confident that this technique is replicable and that it can be used by other teams for phonetic research. Exploring these questions has also made us reconsider what has happened to the data from previous projects. For example, could the data from the Chibchan experiments be reused by other researchers? The public archiving of the Me̲'phaa Váthää data has also begun as a result of improvements in the management of CIM and Denggan data.

The workflow presented in figure 35.4 is just one example of how the untrained forced alignment technique can be applied, and there are numerous improvements that we would like to implement. For example, we would like to write scripts that automatically push partially corrected Praat TextGrids into GitHub repositories. We would also like to improve the support of non-Roman systems from any operating system, so that a wider range of languages can be analyzed using this technique. As we explore improvements, we find ourselves coming back to the same questions: How do we keep this complex task coordinated among the team participants, and how do we make sure that our research has an impact in both academia and in language reclamation efforts? These are elements that remain constant even as the NLP algorithms and their environments change.

6 Conclusions

We have presented an example of a workflow for managing the data involved in untrained forced alignment. During this process we faced the task of organizing and keeping coordinated numerous data transformations across teams of linguists and data annotators, and we found that online platforms that provide explicit versioning helped us with this task. We also had to figure out ways to appropriate the computer code for the algorithm so we could ensure the replicability of our results. In doing these, we explored the ethics of community-based language documentation and research (Czaykowska-Higgins 2009), as well as the issues of research replicability and data reusability in linguistics. In light of the need for work on Indigenous languages, we need to bring in tools that can help get more work done, and we hope that these results will provide ideas for how to use existing NLP tools and knowledge to help the speakers of these languages in their goals of language documentation and reclamation.

Acknowledgments

We want to thank Dr. Christian DiCanio (whose work first introduced us to this technique), Dr. Tyler Peterson, Dr. Miriam Meyerhoff, Dr. Samantha Wray, Dr. Bradley McDonnell, Dr. Sofía Flores, and Alí García Segura; software testers at Victoria University of Wellington, Jawaharlal Nehru University in New Delhi, and the International Conference on Language Documentation and Conservation in Honolulu; as well as audiences at University of Costa Rica, Accenture Costa Rica, Dartmouth College, University of the South Pacific in Suva and Rarotonga, University of Auckland, and the National Library of New Zealand. We also thank the anonymous reviewers of this chapter for their revisions and useful suggestions for tool improvement. In addition to them, we also thank the Fulbright Foreign Student Program, the Tinker Foundation, the Center for Latin American Studies of the University of Arizona, and the Graduate and Professional Student Council at the University of Arizona (grant RSRCH-201FY16) for funding the work on Me̲'phaa alignment, as well as the Me̲'phaa Váthää teachers of School Sector 76 in Guerrero, Mexico, for their support with the Me̲'phaa project.

We would also like to acknowledge the many CIM speakers who have contributed to this project and particular thanks must go to the 2018 cohort of the Diploma in Pacific Vernacular Languages (Cook Islands Māori) at the University of the South Pacific Rarotonga. In that respect we would like to acknowledge the recent passing of Mama Kairangi Daniel and Uriaau George, *moe mai ra e te ngā taeake*. We would also like to thank the staff and students of Ma'uke School, our most prolific collaborator Jean Tekura Mason, *ē tō Ake kōpū tangata kātoatoa*. Together we will make our language thrive again.

Finally, we would like to acknowledge the communities of S. E. Malekula and speakers of Denggan for the active interest and participation in the language project and for their patience providing recordings. Also to

Brittany's partner, Jim, who has been a constant help transcribing and explaining phonological features hand-in-hand with our visual analyses of his language. We would like to acknowledge Victoria University of Wellington and the Endangered Languages Documentation Programme SOAS University of London for funding for the Banam Bay Language Documentation Project (grant IGS0329) and one more time Dr. Miriam Meyerhoff for her insight and supervision during this and subsequent phases of the research. *Sipa ran emdro ran naut Banam Bay nge mun matbafi nenggis san enge san sor Denggan. Sipa ran fafu raru, boti fana sangk, boti nating raru, Isaac boti Jeffry*.

Notes

1. We will use the word *phone* for any unit of sound in a word that can be transcribed using a symbol in the International Phonetic Alphabet, regardless of its status as a phoneme or allophone in a language.

2. This amount of error is comparable to that of aligning non-standard English dialects. When performing untrained forced alignment on British dialects such as Sunderland or Westray English (MacKenzie & Turton 2019), approximately 80% of the phones have errors of less than twenty microseconds when marking their onset. When marking the onset of Bribri phones, 80% of the vowels have an error of less than thirty-one microseconds, and 80% of the consonants have an error of less than twenty-nine microseconds (Coto-Solano & Flores Solórzano 2017).

3. Bribri has several spelling conventions. Nasalization can be written either as a line underneath the vowel (*ù* /ũ̀/ 'pot') or as a tilde above the vowel (*ù̃* 'pot') (Jara Murillo & Segura 2009).

4. For example, when a glottal stop loses its closure but you still have vocalic laryngealization, do you keep a 'glottal stop' interval in the Praat script? Our solution in Nicholas and Coto-Solano (2019) was to mark the smallest possible interval to indicate that a glottal stop was expected there, but so small that we could easily filter it out. Other solutions are possible. For example, adding additional tiers with full transcriptions or explanations is also a possibility for retaining phonological information, as is done in the Corpus of Spontaneous Japanese (NINJAL 2006).

5. *Versioning* or 'version control' refers to the practice of keeping track of the small differences between iterative versions of a complex documents as they are developed and edited, commonly computer code, but in our case ELAN files or Praat TextGrids.

6. We also considered deploying a container-like environment on a platform such as Docker (Merkel 2014). However, this presented the same challenge in terms of user interface and usability.

7. This workshop saw some of the biggest challenges in terms of file encoding. The current FAVE-align algorithm (implemented on Python 2) has issues using Indic and other complex writing systems.

References

Aguilar Gil, Yásnaya Elena. 2014. *¿Para qué publicar libros en lenguas indígenas si nadie los lee?* E'px, September 10. https://archivo.estepais.com/site/2014/para-que-publicar-libros-en-lenguas-indigenas-si-nadie-los-lee/.

Boersma, Paul. 2001. Praat: A system for doing phonetics by computer. *Glot International* 5 (9–10): 341–345.

Centre of Excellence for the Dynamics of Language and S. A. Nicholas. 2019. *Cook Islands Māori*. ANNIS. http://www.corpus.dynamicsoflanguage.edu.au/. Accessed September 29, 2019.

Charpentier, J. M. 1982. *Atlas linguistique du Sud-Malakula (Vanuatu)*. Vol. 1. Paris: Société d'Études Linguistiques et Anthropologiques de France, Peeters Publishers.

Coto-Solano, R. 2017. Tonal reduction and literacy in Me'phaa Váthää. PhD dissertation, University of Arizona.

Coto-Solano, R., and S. Flores Solórzano. 2016. Alineación forzada sin entrenamiento para la anotación automática de corpus orales de las lenguas indígenas de Costa Rica. *Káñina* 40 (4): 175–199. https://doi.org/10.15517/rk.v40i4.30234.

Coto-Solano, R., and S. Flores Solórzano. 2017. Comparison of two forced alignment systems for aligning Bribri speech. *CLEI Electronic Journal* 20 (1): 21. https://doi.org/10.19153/cleiej.20.1.2.

Coto-Solano, R., S. A. Nicholas, and S. Wray. 2018. Development of natural language processing tools for Cook Islands Māori. In *Proceedings of the Australasian Language Technology Association Workshop 2018*, ed. S. Mac Kim and X. J. Zhang, 26–33. Dunedin, New Zealand: Australasian Language Technology Association. https://www.aclweb.org/anthology/U18-1003/.

Coto-Solano, R., S. A. Nicholas, S. Wray, and T. Petersen. 2019. Accelerating the analysis of your audio recordings with untrained forced speech alignment. Paper presented at the 6th International Conference on Language Documentation and Conservation (ICLDC), University of Hawai'i at Mānoa, March 3. http://hdl.handle.net/10125/44886.

Czaykowska-Higgins, E. 2009. Research models, community engagement, and linguistic fieldwork: Reflections on working within Canadian indigenous communities. *Language Documentation and Conservation* 3 (1): 182–215.

DiCanio, C., H. Nam, D. H. Whalen, H. Timothy Bunnell, J. D. Amith, and R. C. García. 2013. Using automatic alignment to analyze endangered language data: Testing the viability of untrained alignment. *Journal of the Acoustical Society of America* 134 (3): 2235–2246. https://doi.org/10.1121/1.4816491.

Dwyer, A. M. 2006. Ethics and practicalities of cooperative fieldwork and analysis. In *Essentials of Language Documentation*,

ed. J. Gippert, N. Himmelmann, and U. Mosel, 31–66. Trends in Linguistics: Studies and Monographs 178. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110197730.31.

Ernestus, M., and N. Warner, eds. 2011. An introduction to reduced pronunciation variants. In *Speech Reduction*. Special issue, *Journal of Phonetics* 39 (3): 253–260. https://doi.org/10.1016/s0095-4470(11)00055-6.

Foley, B., J. T. Arnold, R. Coto-Solano, G. Durantin, T. M. Ellison, D. van Esch, D.,S. Heath, et al. 2018. Building speech recognition systems for language documentation: The CoEDL Endangered Language Pipeline and Inference System (ELPIS). In *Proceedings of SLTU*, 205–209. https://doi.org/10.21437/sltu.2018-42.

Fromont, R., and J. Hay. 2012. LaBB-CAT: An annotation store. In *Proceedings of the Australasian Language Technology Association Workshop*, ed. P. Cook and S. Nowson, 113–117. Dunedin, New Zealand: Australasian Language Technology Association. https://www.aclweb.org/anthology/U12-1015/.

Goldman, J. P. 2011. *EasyAlign: An Automatic Phonetic Alignment Tool under Praat*. http://latlcui.unige.ch/phonetique/easyalign.php.

González, S., C. E. Travis, J. Grama, D. Barth, and S. Ananthanarayan. 2018. Recursive forced alignment: A test on a minority language. In *Proceedings of the Seventeenth Australasian International Conference on Speech Science and Technology*, ed. J. Epps, J. Wolfe, J. Smith, and C. Jones, 145–148. Sydney, Australia: Australasian Speech Science and Technology Association. https://assta.org/proceedings/sst/SST-2018/SST_2018_Proceedings_Rev_A_IDX.pdf.

Gorman, K., J. Howell, and M. Wagner. 2011. Prosodylab-aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics* 39 (3): 192–193.

Hoback, B. 2019. *Banam Bay Language: Documentation and Endangered Language Maintenance*. London: SOAS, Endangered Languages Archive. (Deposit ID: 0549). https://elar.soas.ac.uk/Collection/MPI1202117.

Jara Murillo, C., and A. G. Segura. 2009. *Se' e'yawö bribri wa. Aprendemos la lengua bribri*. San José: Editorial de la Universidad de Costa Rica.

Johnson, L. M., M. Di Paolo, and A. Bell. 2018. Forced alignment for understudied language varieties: Testing Prosodylab-Aligner with Tongan data. *Language Documentation and Conservation* 12:80–123.

Jones, C., K. Demuth, W. Li, and A. Almeida. 2017. Vowels in the Barunga variety of North Australian Kriol. In *Proceedings of Interspeech*, 219–223. https://doi.org/10.21437/interspeech.2017-1552.

Jones, C., W. Li, A. Almeida, and A. German. 2019. Evaluating cross-linguistic forced alignment of conversational data in north Australian Kriol, an under-resourced language. *Language Documentation and Conservation* 13:281–299.

Jones, M. C., and S. Ogilvie, eds. 2013. *Keeping Languages Alive: Documentation, Pedagogy and Revitalization*. Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9781139245890.

Kashima, E., D. Williams, T. Mark Ellison, D. Schokkin, and P. Escudero. 2016. Uncovering the acoustic vowel space of a previously undescribed language: The vowels of Nambo. *Journal of the Acoustical Society of America* 139 (6): EL252–EL256. https://doi.org/10.1121/1.4954395.

Keegan, T. 2019. Language normalisation through technology: Te Reo Māori example. In *6th International Conference on Language Documentation and Conservation (ICLDC)*. http://hdl.handle.net/10125/44883.

Kempton, T. 2017. Cross-language forced alignment to assist community-based linguistics for low resource languages. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, 165–169. https://doi.org/10.18653/v1/w17-0122.

Kukutai, T., and J. Taylor. 2016. *Indigenous Data Sovereignty: Toward an Agenda*. Vol. 38. Canberra, Australia: ANU Press.

Labov, W., I. Rosenfelder, and J. Fruehwald. 2013. One hundred years of sound change in Philadelphia: Linear incrementation, reversal, and reanalysis. *Language* 89 (1): 30–65. https://doi.org/10.1353/lan.2013.0015.

Liao, H., and M. J. F. Gales. 2008. Issues with uncertainty decoding for noise robust automatic speech recognition. *Speech Communication* 50 (4): 265–277. https://doi.org/10.1016/j.specom.2007.10.004.

MacKenzie, L., and D. Turton. 2019. Assessing the accuracy of existing forced alignment software on varieties of British English. *Linguistics Vanguard* 6 (s1): 1–14.

McAuliffe, M., M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger. 2017. Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Proceedings of Interspeech*, 498–502. https://doi.org/10.21437/interspeech.2017-1386.

Merkel, D. 2014. Docker: Lightweight Linux containers for consistent development and deployment. *Linux Journal* 2014 (239): 2.

Nicholas, S. A. 2012. *Te Vairanga Tuatua o te Te Reo Māori o te Pae Tonga: Cook Islands Māori* (Southern dialects) (sn1). Digital collection managed by PARADISEC. [Open Access]. doi:10.4225/72/56E9793466307. http://catalog.paradisec.org.au/collections/SN1.

Nicholas, S. A. 2017. *Ko te Karāma o te Reo Māori o te Pae Tonga o Te Kuki Airani*: A grammar of Southern Cook Islands Māori. Unpublished PhD thesis, University of Auckland. http://hdl.handle.net/2292/32929.

Nicholas, S. A. 2018. Language contexts: *Te Reo Māori o te Pae Tonga o te Kuki Airani* also known as Southern Cook Islands Māori. *Language Documentation and Description* 15:36–64.

Nicholas, S. A. 2019a. *Araara Māaori Kuki Airani* (YouTube channel). https://www.youtube.com/channel/UCXow-aTZOg3hEkBYAz7F8Cw. Accessed October 1, 2019.

Nicholas, S. A. 2019b. *Cook Islands Māori*. https://www.gerlingo.com/language_detail.php?langID=26. Accessed September 29, 2019.

Nicholas, S. A., and R. Coto-Solano. 2019. Glottal variation, teacher training and language revitalisation in the Cook Islands. In *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia*, ed. S. Calhoun, P. Escudero, M. Tabain, and P. Warren, 3602–3606. Canberra, Australia: Australasian Speech Science and Technology Association.

NINJAL (National Institute for Japanese Language and Linguistics). 2006. Construction of the Corpus of Spontaneous Japanese. https://pj.ninjal.ac.jp/corpus_center/csj/en. Accessed December 8, 2019

Partanen, Niko. 2016. Using a Git repository for language documentation corpora (web log message). https://langdoc.github.io/2016-05-20-langdoc_with_Git.html.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/.

Reddy, S., and J. N. Stanford. 2015. Toward completely automated vowel extraction: Introducing DARLA. *Linguistics Vanguard* 1 (1): 15–28. https://doi.org/10.1515/lingvan-2015-0002.

Rosenfelder, I., J. Fruehwald, K. Evanini, K. Seyfarth, S. Gorman, K. Prichard, and J. Yuan. 2011. *FAVE (Forced Alignment and Vowel Extraction) Program Suite*. https://doi.org/10.5281/zenodo.9846. Accessed December 2020.

Seltzer, M. L., D. Yu, and Y. Wang. 2013. An investigation of deep neural networks for noise robust speech recognition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 7398–7402. IEEE. https://doi.org/10.1109/icassp.2013.6639100.

Strunk, J., F. Schiel, and F. Seifart. 2014. Untrained forced alignment of transcriptions and audio for language documentation corpora using WebMAUS. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '14)*, ed. N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, et al., 3940–3947. Reykjavik, Iceland: European Language Resources Association. https://www.aclweb.org/anthology/L14-1123/.

Terrill, Angela. 2002. Why make books for people who don't read? A perspective on documentation of an Endangered Language from Solomon Islands. *International Journal of Society and Language* 2002 (155–156): 205–219. https://doi.org/10.1515/ijsl.2002.029.

Tryon, D. T. 1976. *New Hebrides languages: An internal classification*. Canberra: Department of Linguistics, Research School of Pacific Studies, the Australian National University.

Wightman, C. W., and D. T. Talkin. 1997. The Aligner: Text-to-speech alignment using Markov models. In *Progress in Speech Synthesis*, ed. J. P. H. Santen, J. P. Olive, R. W. Sproat, and J. Hirschberg, 313–323. New York: Springer. https://doi.org/10.1007/978-1-4612-1894-4_25.

Wittenburg, P., H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes. 2006. ELAN: A professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC '06)*, ed. N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, and D. Tapias, 1556–1559. Genoa, Italy: European Language Resources Association. https://www.aclweb.org/anthology/L06-1082/.

Yuan, J., and M. Liberman. 2008. Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America* 123 (5): 3878. https://doi.org/10.1121/1.2935783.

Zimmer, S. 2014. *CMDI Maker 2.20*. https://beta.cmdi-maker.uni-koeln.de/. Accessed December 8, 2019.

Zue, V. W., and S. Seneff. 1996. Transcription and alignment of the TIMIT database. In *Recent Research towards Advanced Man-Machine Interface through Spoken Language*, ed. H. Fujisaki, 515–525. Amsterdam: Elsevier Science BV. https://doi.org/10.1016/b978-044481607-8/50088-8.