Using LARA to rescue a legacy Pitjantjatjara course

Manny Rayner
FTI/TIM
University of Geneva, Switzerland
Emmanuel.Rayner@unige.ch

Sasha Wilmoth
University of Melbourne and
ARC Centre of Excellence for the
Dynamics of Language, Australia
sasha.wilmoth@unimelb.edu.au

Abstract

Learning resources for endangered languages are rare, and good ones may remain relevant for a long time. They will however be much more attractive if they can be repackaged in a modern format. Here, we describe a case study where we used the open source LARA platform to transform a substantial course for the Australian Indigenous language Pitjantjatjara, developed in the 1960s, into a web-compatible multimodal format. The transformation process required about two person-months of effort and produced a resource, freely available on the web, containing about 4.7 hours of Pitjantjatjara audio. The methods are simple and general, and could easily be used to update other legacy resources of this nature. Online documentation is provided.

1 Introduction and background

Endangered languages are almost invariably low-resource languages, and this particularly applies to learning resources. A good learning resource may remain relevant for a long time. However, the format will typically have dated so much that the resource can be hard to use in practice. It is thus interesting to explore methods that can be used to update legacy learning resources into formats appropriate to the twenty-first century.

In this short paper, we present a case study where we used the open source LARA toolkit to transform a substantial course, developed in the 1960s for the Central Australian Indigenous language Pitjantjatjara, into a multimodal web format. We briefly present background on Pitjantjatjara and LARA, describe a simple extension we added to LARA to support incorporation of legacy audio, and outline how we used it to convert the course. Links are given to the result, which is freely available on the web, and to relevant online documentation.

1.1 Pitjantjatjara

Pitjantjatjara is a Pama-Nyungan language spoken in Central Australia. It is one of the most widely spoken Australian Indigenous languages, with over 3,000 speakers as of the 2016 census (Department of Infrastructure, Transport, Regional Development and Communications et al., 2020). It is spoken particularly in the Aṉangu Pitjantjatjara Yankunytjatjara (APY) Lands in the far north of South Australia, but also in other communities in the neighbouring regions of the Northern Territory and Western Australia, and some communities further south in South Australia (such as Yalata). There are also substantial numbers of Pitjantjatjara speakers in urban centres such as Adelaide and Alice Springs. It is still being acquired as a first language by children, and in most Pitjantjatjara-speaking communities, English is learned as an additional language at school, and is mostly used in interactions with government services such as education and health. While there is copious description and documentation of the language by professional linguists, there are fewer publicly available options for learning the language. This is particularly an issue for non-Indigenous people living and working with Aṉangu[1]. The University of South Australia offers a Pitjantjatjara/Yankunytjatjara Language and Culture Program (https://study.unisa.edu.au/short-courses/pitjantjatjara-yankunytjatjara-language/; see Gale et al. 2020 for more information). For those who cannot enrol in these courses, options are more limited. *Wangka Wiṟu* (Eckert and Hudson, 1988) is a detailed grammatical description aimed at non-specialist language learners, however, it is not a language course as such, and does not include audio. The same applies to other resources such as dictionaries (Goddard and Defina, 2020, for example). Two courses with audio have been developed, one in the 1960s at the University of Adelaide (Downing et al. 1967; see details in Amery 2020, p. 492 and §3 immediately below), and *Wangka Kulintjaku* (Kirke, 1985). Neither are

[1] *Aṉangu* literally means "person/people" but is used in the region in both Pitjantjatjara and English to mean "Aboriginal person", particularly an Aboriginal person from the Western Desert. Compare with *Inuit*.

```
@Once upon a time@ there were#be# four little Rabbits#rabbit#,
and their names#name# were#be# Flopsy#Flopsy#, Mopsy#Mopsy#,
Cotton-tail#Cottontail#, and Peter#Peter#.||

They lived#live# with their Mother in a sand-|bank, underneath
the root of a very big fir-|tree.||
```

Figure 1: Example of text (first page of *Peter Rabbit*) with LARA markup. Segment boundaries are marked by double vertical bars (`||`). Words marked as consisting of more than morpheme are subdivided by single vertical bars (`|`). Inflected words are tagged with a morpheme enclosed in hashes (`#...#`)

easily available for purchase.

While this is more than exists for many other Australian Indigenous languages, there is still a need for Pitjantjatjara language-learning materials—particularly with audio—to be more accessible, for both non-Indigenous people working and living with Anangu, and Anangu who grew up speaking English and are looking to reconnect with their heritage language.

1.2 LARA

LARA (Akhlaghi et al. 2019; https://www.unige.ch/callector/lara/) is an open source platform, under development since 2018 by an international consortium with partners in countries including Australia, Iceland, Iran, Ireland, Israel, the Netherlands, Poland, Slovakia and Switzerland. The goal of the project is to develop tools that support the conversion of texts into a multimodal annotated form which supports learner readers. Annotations typically include audio, translations, and a concordance. Text is divided into longer units (typically, but not always, sentences), and then into smaller units (typically, but not always words). Audio and translation annotations can be attached at either level (i.e. sentence or word), and are accessed by clicking or hovering. The LARA examples page (https://www.unige.ch/callector/lara-content) contains links to LARA texts in many languages. The LARA tools can be downloaded and used from the command-line, as described in the online documentation (Rayner et al., 2020). Much of the functionality is also available through an online portal; however, the new functionality described here is currently only available in the command-line version. Though not originally designed for this purpose, LARA has turned out to be a good fit to the requirements of endangered languages. Initial work is described in papers presented at the last two ComputEL workshops (Zuckerman et al., 2021; Bédi et al., 2022).

In the present paper, we will be mostly concerned with the process of adding audio and translation annotations to LARA texts. In previous LARA projects, the workflow to do this has been as follows. First, the text is marked up to add segment and morpheme boundaries; Figure 1 illustrates. Second, a script is invoked to create files listing the audio and translation annotations that need to be created. Third, the files from the second step are used to create the required annotations. This can be done in several ways, as described in the documentation. In particular, lists of missing audio annotations can either be passed to an online recording tool or to a TTS engine; both services are integrated with LARA. Finally, another script is called to combine all the pieces and create the multimedia document.

2 Importing legacy audio into LARA

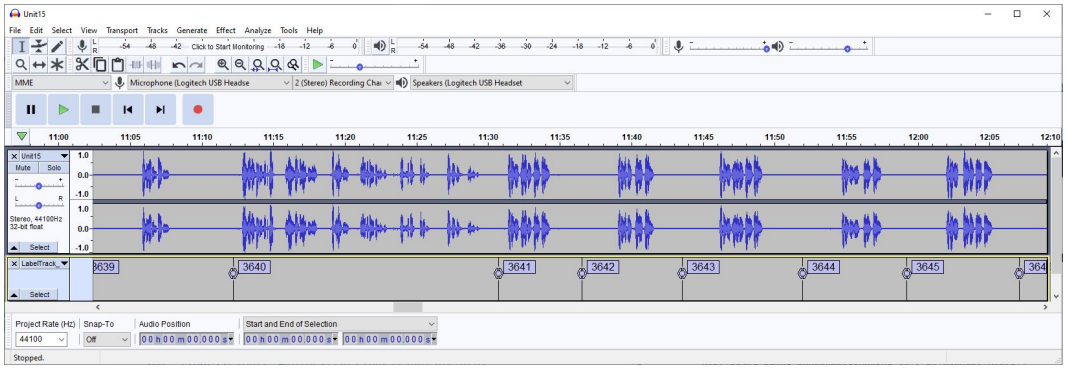
The above workflow was designed for the case where audio is to be created for an existing LARA resource, but does not fit the situation we are interested in here, where audio already exists and we are cutting it up to add it to the LARA resource. We developed a simple method to address these requirements and integrated it into LARA. As usual with such tasks, the challenge is to make the workflow straightforward and efficient; the central question is how to review the audio to find the cutting points. We quickly ascertained that it would be difficult or impossible to implement something better than the popular Audacity audio editor[2], so we organised the process around that. Starting with the audio file and a piece of LARA source text marked up as in Figure 1, the process is as follows:

1. Run a script which transforms the marked up

[2] https://www.audacityteam.org/.

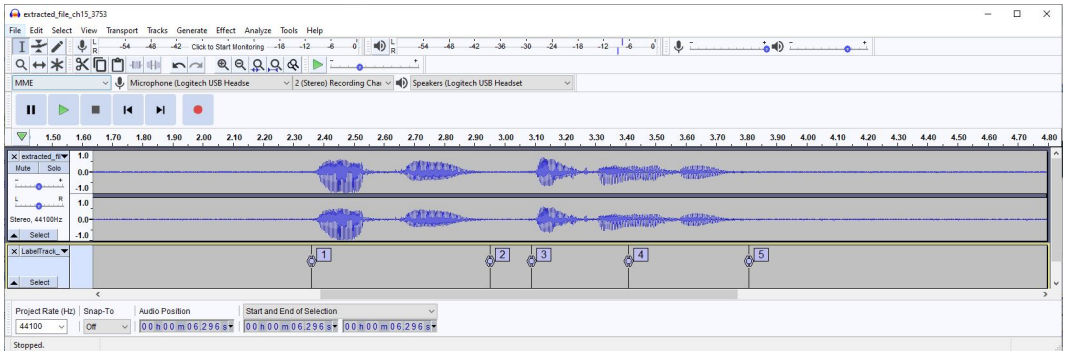
```
<i>Now you'll hear a sentence of the type "the man made a boomerang
for me", first in the long form, followed by the short form.</i>||
wati|ngku#-ngku# ngayuku#ngayulu# kali palyanu#palyani#||
wati|ngku#-ngku#|tju#ngayulu# kali palyanu#palyani#||
wati|ngku#-ngku# nyuntumpa#nyuntu# kali palyanu#palyani#||
wati|ngku#-ngku#|nku#nyuntu# kali palyanu#palyani#||
```
(a) Marked-up LARA text.

```
Now you'll hear a sentence of the type "the man made a boomerang
for me", first in the long form, followed by the short form.|3641|
watingku ngayuku kali palyanu|3642|
watingkutju kali palyanu|3643|
watingku nyuntumpa kali palyanu|3644|
watingkunku kali palyanu|3645|
```
(b) Text with numbered labels.



(c) Matching labels to audio in Audacity tool for passage in (a) and (b).

```
{ "audio_file":  "extracted_file_ch15_3753.mp3",
"label_file":  "LabelTrack_extracted_file_ch15_3753.txt",
"labelled_text":  "|1|wati|2| |3|pulka|4|ngku|5|",
"relevant_word":  "ngku" }
```
(d) Generated metadata to extract morpheme-level audio for sample morpheme *-ngku*.



(e) Matching labels to audio in Audacity tool for segment in (d).

Figure 2: Different processing stages for converting a short section of the Pitjantjatjara course.

LARA source text into a version where each segment boundary is labelled with a number.

2. Use Audacity to add a "label track" where the labels are placed at the points in the audio corresponding to marked segment boundaries.

Figure 3: Screenshot showing the short section of the Pitjantjatjara course used in Figure 2, as it appears in the final LARA document. Clicking on an audio control plays the preceding segment. Hovering over a morpheme highlights it and shows a popup gloss: thus, for example, in the word *watingkunku* it is possible to hover over any of the three component morphemes *wati* ("man"), *-ngku* (ergative case suffix) and *-nku* ("for you"). Clicking on a morpheme plays audio and shows a concordance in the other pane. Here, the user has clicked on *palyaṉu* ("made"). The document is freely available online, a link is given in §4.

3. Export the label track from Audacity.

4. Run a second script, which uses the labelled source text and the Audacity label track to extract relevant segments of audio and save them together with associated metadata. The extraction is performed using `ffmpeg`[3].

5. Optionally run a third script to create labelled data for further cutting up of segment audio into audio for individual words, and repeat steps (2)–(4) for each segment using word-level versions of the segment-level scripts.

Figure 2 illustrates. In practice, it is common to find errors in the segmentation while carrying out step (2). When this happens, the annotator corrects the errors, reruns step (1), and continues. As described below, the process works well enough that it is practically feasible to create LARA documents based on several hours of audio.

3 Rescuing the 1960s course

The Pitjantjatjara course (Amery, 2012) was developed at Adelaide University between 1966 and 1968 by Jim Downing, Ken Hale and Gordan Ingkatji. Pitjantjatjara audio was recorded by Gordon Ingkatji, a Pitjantjatjara native speaker, and two unknown female speakers, one a native Pitjantjatjara speaker, the other a non-native speaker with a strong Australian English accent.[4] The course consists of 28 units, of which the first 16 are designated as the "Elementary" course, and the remaining 12 as the "Advanced" course. There is a Xeroxed text handout of 3–6 pages and a recorded audio segment of 15–35 minutes for each unit; the text is approximately, though not exactly, a transcript of the audio. Materials for the course were digitised by Paul Eckert and Mary-Anne Gale, and have been preserved by the Australian Society for Indigenous Languages (AuSIL)[5], who distribute them on USB sticks sold through outlets including Red Kangaroo Books[6] and the Goldfields Aboriginal Language Center[7]. The USB stick is accompanied by a printed notice saying that the material is "now out of print, so legally copied".[8]

The emphasis in the "Elementary" course is on basic Pitjantjatjara grammar, introduced through various kinds of drill exercises. The main topics are demonstratives, common and proper nouns, singular, dual and plural pronouns in both the "long" and the "short" forms[9], present, past, future

[3] https://ffmpeg.org/.

[4] We have reached out to many people who might have been able to identify the female Pitjantjatjara speaker, without success.

[5] https://ausil.org.au/

[6] https://redkangaroobooks.com/

[7] https://wangka.com.au/

[8] According to Paul Eckert (personal communication), he and Mary-Anne Gale made extensive attempts to determine whether any person or institution had ever claimed copyright, concluding that no one had done so.

[9] Most pronouns in Pitjantjatjara have a "long" form (free pronoun) and "short" form (a second position clitic which can be attached to any part of speech).

and imperative forms of verbs, ergative, locative, allative, ablative, and genitive/dative case markers, YN- and WH-questions, and conjunctions. About 115 lemmas of basic vocabulary are presented.

The "Advanced" course alternates grammar units with ones organised around short stories; there are six of each. The main topics in the grammar units are the past continuous, habitual, serial, participle, purposive, nominalized, imperfective imperative and negated forms of verbs, locative and reflexive pronouns, and the avoidance case marker. The story units become successively longer as the course progresses, starting at about 75 words and ending at about 240 words. About 190 more lemmas of vocabulary are presented, giving a total of about 305 lemmas.

To convert the course to LARA form, we first ran the text content though an English OCR site[10] and then cleaned up the output manually. After adding LARA markup, we followed the workflow outlined in §2 to cut up and attach the audio (cf. Figure 2), and added word glosses. End-to-end, a unit typically required 6–12 person-hours of effort to convert, the most laborious part of the process being the manual audio annotation. The main work was carried out by the first author, a person who had extensive LARA experience but no prior exposure to Pitjantjatjara, and then corrected by the second author, a linguist with expertise in the language. Typical issues corrected in this stage included the segmentation, lemmatisation, glossing, and spelling of particular words.

An unforeseen complication resulted from the presence of audio recorded by the non-native speaker. After reviewing the initial draft of the course, we decided that the difference in quality, compared to the two native speakers, was so marked that it was necessary to minimize her contribution. We addressed this issue by manually labelling her audio segments and dispreferring them when attaching audio for individual words, adding generic infrastructure to support these tasks.

4 Accessing the course and the tools

The finished LARA version of the course, which contains a total of about 4.7 hours of Pitjantjatjara audio, is freely available online.[11] A sample passage from the "Elementary" course is shown in Figure 3.

The online LARA documentation (Rayner et al., 2020) describes how to use the tools. The details of the annotation process from §2 are presented in the section "Creating segment audio by cutting up MP3s".[12]

5 Summary and further directions

We have described a simple and general method, implemented within the open source LARA platform, which allowed us to transform a substantial legacy course for Pitjantjatjara into a modern multimedia format. The conversion process required about two person-months of effort.

The course was put online in early November 2022, and so far we only have a small amount of feedback from users. Anecdotally, the response is quite positive. In particular, AuSIL immediately asked us for permission to distribute the new version of the course in the same USB packaging as the original one, since many of their clients do not have good internet. We expect this to be available through their usual resellers by the time of the conference.

It is to be noted that the process of cutting up legacy audio and incorporating it into a LARA document can be performed much more efficiently when an adequate speech recogniser is available for the language in question, making it possible to use automatic alignment methods; we present initial results in a recent paper (Rayner et al., 2022). At the moment, no such recogniser exists for Pitjantjatjara, but it is not out of the question that one could be trained.

Acknowledgements

We would very much like to thank Paul Eckert for providing us with the background to the 1960s course and generally giving us the benefit of his enormous experience with the Pitjantjatjara language and community.

References

Elham Akhlaghi, Branislav Bédi, Matthias Butterweck, Cathy Chua, Johanna Gerlach, Hanieh Habibi, Junta Ikeda, Manny Rayner, Sabina Sestigiani, and Ghil'ad Zuckermann. 2019. Overview of LARA:

[10] https://ocr.space/

[11] https://www.issco.unige.ch/en/research/projects/callector/pitjantjatjara_courses_1_and_2vocabpages/_hyperlinked_text_.html. View in Chrome or Firefox.

[12] https://www.issco.unige.ch/en/research/projects/callector/LARADoc/build/html/cutting_up_audio.html

A learning and reading assistant. In *Proc. SLaTE 2019: 8th ISCA Workshop on Speech and Language Technology in Education*, pages 99–103.

Rob Amery. 2012. The history of Aboriginal languages and linguistics at the University of Adelaide. In Nick Harvey, Jean Fornasiero, Greg McCarthy, Clem Macintyre, and Carl Crossin, editors, *A History of the Faculty of Arts at the University of Adelaide: 1876-2012*, pages 265–298.

Rob Amery. 2020. Teaching Aboriginal Languages at University: To What End? In Jean Fornasiero, Sarah M. A. Reed, Rob Amery, Eric Bouvet, Kayoko Enomoto, and Hui Ling Xu, editors, *Intersections in Language Planning and Policy: Establishing Connections in Languages and Cultures*, Language Policy, pages 475–489. Springer International Publishing, Cham.

Branislav Bédi, Hakeem Beedar, Belinda Chiera, Nedelina Ivanova, Christèle Maizonniaux, Neasa Ní Chiaráin, Manny Rayner, John Sloan, and Ghil'ad Zuckermann. 2022. Using LARA to create image-based and phonetically annotated multimodal texts for endangered languages. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*.

Department of Infrastructure, Transport, Regional Development and Communications, Jacqueline Battin, Jason Lee, Douglas Marmion, Rhonda Smith, Tandee Wang, Yonatan Dinku, Janet Hunt, Francis Markham, Denise Angelo, Emma Browne, Inge Kral, Carmel O'Shannessy, Jane Simpson, and Hilary Smith. 2020. National Indigenous Languages Report. *https://www.arts.gov.au/what-we-do/indigenous-arts-and-languages/national-indigenous-languages-report*.

Jim Downing, Ken Hale, and Gordon Ingkatji. 1967. Pitjantjatjara language course materials for use at the University of Adelaide.

Paul Eckert and Joyce Hudson. 1988. *Wangka Wiṟu: A Handbook for the Pitjantjatjara Language Learner*. South Australian College of Advanced Education, Underdale, S. Aust.

Mary-Anne Gale, Dan Bleby, Nami Kulyuṟu, and Sam Osborne. 2020. The Pitjantjatjara Yankunytjatjara Summer School: Kulila! Nyawa! Arkala! Framing Aboriginal Language Learning Pedagogy within a University Language Intensive Model. In Jean Fornasiero, Sarah M. A. Reed, Rob Amery, Eric Bouvet, Kayoko Enomoto, and Hui Ling Xu, editors, *Intersections in Language Planning and Policy*, volume 23, pages 491–505. Springer International Publishing, Cham.

Cliff Goddard and Rebecca Defina. 2020. *Pitjantjatjara/Yankunytjatjara to English Dictionary*, 2nd rev. ed., updated edition. IAD Press, Alice Springs, N.T.

Brian Kirke. 1985. *Wangka Kulintjaku: Talk so as to Be Understood. Introductory Self-Instruction Course in Basic Conversational Pitjantjatjara*. Faculty of Education and Humanities, South Australian College of Advanced Education.

Manny Rayner, Belinda Chiera, and Cathy Chua. 2022. Using public domain resources and off-the-shelf tools to produce high-quality multimedia texts. In *Proceedings of the 20th Annual Workshop of the Australasian Language Technology Association*, Adelaide, Australia.

Manny Rayner, Hanieh Habibi, Cathy Chua, and Matt Butterweck. 2020. *Constructing LARA content*. https://www.issco.unige.ch/en/research/projects/callector/LARADoc/build/html/index.html. Online documentation.

Ghil'ad Zuckerman, Sigurður Vigfússon, Manny Rayner, Neasa Ní Chiaráin, Nedelina Ivanova, Hanieh Habibi, and Branislav Bédi. 2021. LARA in the service of revivalistics and documentary linguistics: Community engagement and endangered languages. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1, pages 13–23.