T is for *Treu*, but how do you pronounce that? Using C-LARA to create phonetic texts for Kanak languages*

Pauline Welby
Aix Marseille Université, CNRS
Laboratoire Parole et Langage
and
Université de la Nouvelle Calédonie
pauline.welby@cnrs.fr

Fabrice Wacalie
Université de la Nouvelle Calédonie
fabrice.wacalie@unc.nc

Manny Rayner
University of South Australia
manny.rayner@unisa.edu.au

ChatGPT-4 C-LARA-Instance
University of South Australia
chatgptclarainstance@proton.me

Abstract

In Drehu, a language of the indigenous Kanak people of New Caledonia, the word *treu* 'moon' is pronounced [tʃe.u]; but, even if they hear the word, the spelling pulls French speakers to a spurious pronunciation [tʁø]. We implement a strategy to mitigate the influence of such orthographic conflicts, while retaining the benefits of written input on vocabulary learning. We present text in "phonetized" form, where words are broken down into components associated with mnemonically presented phonetic values, adapting features from the "Comment ça se prononce ?" multilingual phonetizer. We present an exploratory project where we used the ChatGPT-based Learning And Reading Assistant (C-LARA) to implement a version of the phonetizer strategy, outlining how the AI-engineered codebase and help from the AI made it easy to add the necessary extensions. We describe two proof-of-concept texts for learners produced using the platform, a Drehu alphabet book and a Drehu version of "The (North) Wind and the Sun"; both texts include native-speaker recorded audio, pronunciation respellings based on French orthography, and AI-generated illustrations.

1 Introduction

1.1 A challenge: Reading a word and hearing it is often not enough

. In Drehu, a language of the indigenous Kanak people of New Caledonia in the South Pacific, the word *treu* means 'moon', as we can learn in a classic, illustrated children's book (Atti et al., 1995). Books like these can be very useful to the beginning learner; seeing the spelling is a powerful aid for learning words and their meanings (Bürki et al., 2019; Pattamadilok et al., 2022; Welby et al., 2022). "[W]hen faced with speech, which is inherently [] highly variable and fleeting, the orthographic form offers L2 speaker-listeners something stable to 'grab onto"' (Welby et al., 2022).

But how do you pronounce *treu*? Teachers designing classroom activities or heritage learners trying to reclaim their language need to know. Of course, it helps to hear words spoken out loud. Multimodal texts are common in mainstream platforms like LingQ[1]: words are annotated with audio files that can be played by clicking or hovering. Sound has been integrated into online dictionaries and other resources for dozens of endangered or underresourced languages (e.g. A Speaking Atlas of Indigenous Languages of France and its Overseas (Boula de Mareüil et al., 2019), the Swarthmore Talking Dictionaries Project[2], the online dictionary of the Académie Tahitienne[3] and the 50 Words Project for the Indigenous languages of Australia[4], allowing users to read a printed word or phrase and hear its pronunciation. Having these two forms of the word may be sufficient, depending on the goals of a given project and on the user's language and literacy background.

For language learners, however, seeing (reading) the written form of a word and hearing it spoken will often not be enough. This is likely to be the case, for example, for French

* Authors in anti-alphabetical order.

[1] https://www.lingq.com/
[2] https://talkingdictionaries.swarthmore.edu/
[3] https://www.farevanaa.pf/dictionnaire.php/
[4] https://50words.online/languages/

speakers who wish to learn Drehu (or any other Kanak language) as an L2 or heritage language. This is because when there is a conflict between what we hear and what we read, the written input often wins out. *Treu* is pronounced [tʃe.u] (*chay-OO*), but knowledge of the letter-to-sound correspondences of French, the main language of schooling and of literacy-based activities in New Caledonia, will pull many Caledonians toward a spurious, French-like pronunciation, a single syllable with T' and 'R' sounds and rhyming with *feu* [fø] 'fire'. We routinely encounter this phenomenon in our experience teaching Kanak languages and linguistics (FW) and living in New Caledonia (FW and PW). The influence of the spelling system of the institutional or literacy-dominant language extends beyond the Caledonian context. Similarly, for the Australian Aboriginal language Barngarla, Bédi et al. (2022) report : "the voiced retroflex plosive [ɖ] ... is written 'rd' as for example in Barngarla *yarda*, 'country'. It is however all too easy for the anglophone reader to interpret this as representing a lengthened preceding vowel followed by [d] as for example in ... 'card' or 'herd' ". It has also been shown experimentally that L1 French orthographic conventions pull L2 English speakers to spurious French-link vowel pronunciations (Bürki et al., 2019; Welby et al., 2022).

1.2 The Caledonian context and the Kanak languages and beyond

. Issues of this kind are ubiquitous for people engaging with the Kanak languages. Like other Melanesian islands and countries, New Caledonia is characterized by its linguistic diversity, with just under 30 Kanak languages, most of which are endangered or vulnerable. Many have writing systems, developed first by 19th century missionaries working with native speakers, and revised or expanded by linguists. Twelve have suggested standard orthographies, published or under development as part of the Académie des Langues Kanak's "Propositions d'écriture" series. For some languages there is a long and continuing (e.g. in social media) tradition of writing using French orthographic conventions. All Kanak languages are under-resourced, lacking not only digital resources, but also in pedagogical and reading materials, and many have no such resources at all.

As noted by Welby et al. (2023): "[e]ach of the Kanak languages has its own phonology as well as its own orthographic system.... Learning the grapheme-phoneme correspondences (GPCs) of a second language (L2) is not always straightforward... There is also likely to be interference from the GPCs of the L1 or those of the language in which one typically reads and writes." To give just one example in a handful of languages, the grapheme <j> corresponds to /ð/ in Drehu *jol* 'difficult', /dz/ in Nengone *jo* 'to suffer', /ʒ/ in French *joli* 'pretty', and /dʒ/ in English *Joe*.

The motivation for our project and its challenges extend beyond the Caledonian context. As Welby et al. (2023) note, they will resonate with "other communities where several languages are present to one degree or another (e.g. countries with migrant communities) or in which the L1 (or the literacy dominant language) and the L2 (or L2s) have very different orthographies, for example, English and Irish".

1.3 A proposed solution: including phonetic texts

In this paper, we propose a solution to mitigate the influence of cross-language orthographic conflicts on pronunciation, while retaining the clear benefits of written input on vocabulary learning. We build on theoretical and applied results from two thus far independent projects. The first of these projects, LARA (Learning and Reading Assistant; Akhlaghi et al. 2019, 2020), now reimplemented as C-LARA (ChatGPT-based LARA; Bédi et al. 2023b,a, 2024; `https://www.c-lara.org/`), is a collaborative open source project to develop tools converting plain texts into interactive multimedia language learning resources. The second, "Comment ça se prononce ?" ('How is that pronounced?') is a project to build a multilingual phonetizer bridging between spelling and pronunciation in the many languages of New Caledonia (Welby et al. 2023). Here we implement the Caledonian phonetizer strategy inside the C-LARA platform.

When we began the joint project, we were particularly curious to explore two themes. The first was to use the idea of "pronunciation respellings" based on the orthography of the common language. For example, in

the Caledonian context where French fills this role, we present the word *treu* as [tché-ou]. Learners reading e-books report appreciating having phonetic text alongside regular written text (Bédi et al., 2023a), and our experience suggests that for many learners, pronunciation respellings are more user-friendly and thus more usable than phonetic transcription (Welby et al., 2023).

The second theme was to discover what assistance we could obtain in the context of Kanak languages from the newly reimplemented ChatGPT-based version of LARA, "C-LARA". Interestingly, we found that the AI was able to make a very useful contribution, but not in the way one might have anticipated. On the negative side, the late 2023 version appears to have essentially no knowledge of Kanak languages, and is thus unable to do any language-specific work. The AI was however able to make a large contribution in its software engineer role. Its image-generation skills also turned out to be surprisingly helpful in creating high-quality picture book texts.

The rest of the paper is structured as follows. In Section 2, we briefly outline the C-LARA architecture and describe how we were able to use it to repackage the LARA "phonetizer" functionality to make it practically useful, and in particular address the requirements by the Kanak languages. Section 3 describes initial proof-of-concept examples, a "phonetized" alphabet book for Drehu with AI-generated illustrations and a Drehu version of "The Wind and the Sun". The final section summarises and suggests further directions.

2 Adding phonetic text capabilities to C-LARA

We begin this section with a few words about the C-LARA platform, then describe how we added the phonetic text capabilities. C-LARA is an international open source project initiated in March 2023 and currently involving partners in eight countries. As previously indicated, the goal was to perform a complete reimplementation of the earlier LARA project, keeping the same basic functionality of providing a flexible online tool for creating multimodal texts, but adding ChatGPT-4 as the central component. ChatGPT-4 is used in two separate and complementary ways. In the form of GPT-4, it appears as a software *component*, giving the user the option of letting it perform the central language processing operations for languages it understands sufficiently well; it also appears as a software *engineer*, working together with human collaborators to build the platform itself. As described in the initial C-LARA report (Bédi et al., 2023a), the software engineering aspect has proven very successful, with ChatGPT not only writing about 90% of the code, but greatly improving it compared to the earlier LARA codebase. In this paper, where we are interested in small languages that GPT-4 knows little about, the AI cannot play a meaningful role as a language processor. It was however able to contribute effectively as a software engineer, both by having created a well-designed architecture that was easy to extend, and by assisting in the extension process. We start by outlining the structure.

C-LARA is a web app implemented in Python/Django.[5] The code is divided into two layers: a Python core processing layer, which implements all the language processing functions, and a Django web layer which sits on top of it. Nearly all the design decisions in the web layer come from ChatGPT-4, which has created an app with a simple, entirely mainstream structure; this has made it easy for the AI to write most of the web-level code.[6] The AI has also been responsible for the greater part of the design decisions in the core layer. This consists of a set of modules, nearly all of which are of one the four generic types: internalisation (conversion of text into a class-based Python representation); repositories (storing linguistic data into database records accessed through Python classes); rendering (converting internal representations into multimedia text); and language processing (usually, invoking the AI to perform operations like segmentation, glossing, lemma tagging etc).

We now move to describing the phonetic text functionality. An initial exploration of the idea of creating phonetic texts was carried out during the earlier LARA project (Bédi et al., 2022). The basic idea was to support a second annotation mode, where instead of

[5] https://www.djangoproject.com/

[6] The project codebase is available at https://github.com/mannyrayner/C-LARA

dividing pages into segments and words, they are instead divided into words and grapheme-groups; grapheme-groups are annotated with associated phonetic information. Two methods were developed for dividing words into meaningful annotated grapheme-groups. For languages with consistent grapheme-phoneme correspondences (e.g. Arabic, Hebrew, most Australian Indigenous languages), a conversion table and a greedy parsing algorithm gave good results. For languages with complex and/or inconsistent correspondences, a more sophisticated example-based method was implemented where words were aligned against entries taken from a phonetic lexicon.

The method gave good results for English and French (Akhlaghi et al., 2022). Unfortunately, the LARA phonetic text functionality was never integrated into the LARA web platform and could only be accessed through the command-line version of the tool; also, the language-specific resources (grapheme-phoneme conversion tables, links to phonetic lexica, etc) were hard-coded. Although these issues were from a theoretical point of view trivial, in practice they meant that the functionality was hardly used.

With the better-engineered C-LARA codebase and the AI's assistance, we found it was straightforward to solve the problems involved in repackaging the phonetic text functionality to make it easily available in the new context. We carried over the core processing modules (in particular, the dynamic programming grapheme-phoneme alignment code) from the LARA codebase, defined suitable repository modules to store the necessary information, connected them to new views in the Django layer, and generalised the rendering templates so that they worked for phonetic texts as well. Nearly all of this routine work could be carried out by the AI. From the user's point of view, the functionality exposed is the following:

1. A screen on which a language expert can define the relevant phonetic resources for a language. These can be of two types: a grapheme-group-to-phoneme-group correspondence table for languages with consistent GPCs, and a phonetic lexicon, uploaded in file form, which associates words with phonetic entries.

2. A screen on which the user can invoke phonetic processing, if phonetic resources are defined for the language in question, to convert plain text into phonetically decomposed text.

3. A screen on which the user can upload audio files corresponding to the phoneme-groups occurring in a phonetic text. These are stored in a shared language-specific repository, so it is only necessary to upload new files. For languages supported by ipa-reader[7], the files are by default produced automatically and downloaded from the site.

4. A screen on which the user can invoke a rendering process to convert phonetically decomposed text and associated audio into multimedia text.

5. The final result is a flexible multimedia document which can be viewed either as normal or as phonetic text (Fig. 1) and can be posted in the C-LARA social network, where users can rate it and leave comments. It is also possible to link together multiple C-LARA documents in the same language to form a "reading history", a virtual document representing the concatenation of the component documents and including a common concordance.

All of this is described in full detail in the second C-LARA Progress Report (Bédi et al., 2024), which in particular describes more precisely the AI's role in developing the various functionalities.

In the next section, we describe how we have started using these new functionalities for the Kanak language Drehu.

3 Two proof-of-concept examples

As an initial step, we have developed two C-LARA texts for Drehu, the language of the island of Lifou. Drehu, pronounced /dʒehu/ or *jay-hoo*, is the Kanak language with the largest community of speakers (approx. 16,000). It is taught as a subject in some schools and at university, and has a proposed written standard.

[7] http://ipa-reader.xyz/

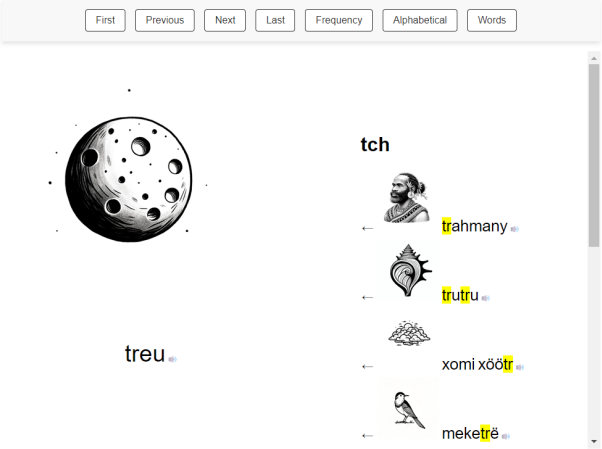

Figure 1: Top of a page from the initial Drehu alphabet book. The user has clicked on the 'tr' in *treu* ("moon"), playing audio and bringing up a concordance for the /tch/ phoneme that shows contexts where it occurs. The text has been arranged so that the associated images are part of these contexts.

The first text is a Drehu alphabet book based on the entries from an introductory booklet produced by the Académie des Langues Kanak (2023). It has almost exactly the same Drehu words and French glosses as this original resource, with 80 entries in total. The C-LARA version of the alphabet book contains one entry per page. Each page can be viewed either in "Words" mode, where words are treated as indivisible entities and hovering over a word shows a French gloss, or in "Sounds" (phonetic text) mode, where words are divided into graphemes. Clicking on the icon next to the word plays the entire word recorded by the Drehu native speaker author. Hovering over a grapheme shows the associated phoneme; clicking on it plays the audio for the associate phoneme and shows a "phonetic concordance" of words which contain the phoneme, each one displayed together with its associated image (see Fig. 1).

We used ChatGPT-4's integrated DALL-E-3 functionality to produce the images. This allowed us to test the abilities of the AI to produce images culturally appropriate to the Melanesian context. Our impression is that it has succeeded well; initial comments from the Drehu community are very positive.

The second text is a C-LARA edition of "Leu me Jö", the Drehu version of Aesop's fable "The (North) Wind and the Sun", a standard text used in studies of many languages including Kanak languages (Boula de Mareüil et al., 2019). The C-LARA functionalities add value as a teaching and learning resource.

Both texts are openly available on the C-LARA site.[8]

4 Further directions

We have three immediate goals. First, we are gathering feedback from users and community members on the proof-of-concept resources, with respect to the accuracy and the usefulness of the phonetic texts, the appropriateness of the AI-generated images, and the usability of the interface. This input will help shape subsequent work. Second, we plan to carry over several features from the "Comment ça se prononce ?" phonetiser project to the C-LARA context. These include: 1. offering two options for phonetic transcriptions: International Phonetic Alphabet (IPA) and French-based orthographic respellings (illustrated in Figure 1), enhanced with pronunciation tips, such as images or video clips of mouth shapes or tongue movement, and 2. rendering text in an interlinear form which alternates lines of plain and phonetic text. Third, we will explore the ways in which community members might participate in creating C-LARA resources, such as glossing and lemma tagging.

As in the phonetizer project, we adopt an incremental approach, developing proof-of-concept resources for one language, here Drehu, which we can then show to members of other communities. We have begun discussions with the Paicî community.

Finally, the work presented here could also serve as a first step towards giving new life to a legacy alphabet book for five Kanak languages, Drehu, Fwaî, Numèè, Paicî, and Yuanga (Atti et al., 1995). If the authors and their communities so desire, a new C-LARA, multimedia edition of the book could retain the rich, evocative illustrations of the original, while linking the written words to their pronunciations. This important cultural document would then be accessible to new generations.

[8] https://c-lara.unisa.edu.au/accounts/rendered_texts/17/phonetic/page_1.html; https://c-lara.unisa.edu.au/accounts/rendered_texts/10/phonetic/page_1.html

References

Académie des Langues Kanak. 2023. *Abécédaire illustré en drehu*. Académie des Langues Kanak, Noumea, New Caledonia.

Elham Akhlaghi, Branislav Bédi, Fatih Bektaş, Harald Berthelsen, Matthias Butterweck, Cathy Chua, Catia Cucchiarini, Gülşen Eryiğit, Johanna Gerlach, Hanieh Habibi, Neasa Ní Chiaráin, Manny Rayner, Steinþór Steingrímsson, and Helmer Strik. 2020. Constructing multimodal language learner texts using LARA: Experiences with nine languages. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 323–331.

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

Elham Akhlaghi et al. 2022. Reading assistance through LARA, the Learning And Reading Assistant. In *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI) within the 13th Language Resources and Evaluation Conference*, pages 1–8.

Solange Atti, Chantal Bouanou, Helene Diahioue, Jean-Pierre Diahioue, Michele Grynagier, Pierre Hnacipan, Yvette Lepigeon, Paulette Prevaut, and Marcko Waheo (illustrator). 1995. *Et toi, comment-dis tu?* Centre Territorial de Recherche et de Documentation Pedagogique, Noumea, New Caledonia.

Philippe Boula de Mareüil, Frédéric Vernier, Gilles Adda, Albert Rilliard, and [J]acques Vernaudon. 2019. A speaking atlas of indigenous languages of france and its overseas. In *Proceedings of the Language Technologies for All (LT4All), Paris. France*, pages 155—-159.

Audrey Bürki, Pauline Welby, Mélanie Clément, and Elsa Spinelli. 2019. Orthography and second language word learning: Moving beyond "friend or foe?". *The Journal of the Acoustical Society of America*, 145(4):EL265–EL271.

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*.

Branislav Bédi, ChatGPT-4, Belinda Chiera, Cathy Chua, Catia Cucchiarini, Christèle Maizonniaux, Claudia Mărginean, Neasa Ní Chiaráin, Chadi Raheb, Manny Rayner, Annika Simonsen, Manolache Lucreția Viorica, Pauline Welby, Zhengkang Xiang, and Rina Zviel-Girshin. 2024. ChatGPT-Based Learning And Reading Assistant: Second report. Technical report. ResearchGate preprint.

Branislav Bédi, ChatGPT-4, Belinda Chiera, Cathy Chua, Catia Cucchiarini, Neasa Ní Chiaráin, Manny Rayner, Annika Simonsen, and Rina Zviel-Girshin. 2023a. ChatGPT-Based Learning And Reading Assistant: Initial report. Technical report. ResearchGate preprint.

Branislav Bédi, ChatGPT-4, Belinda Chiera, Cathy Chua, Neasa Ní Chiaráin, Manny Rayner, Annika Simonsen, and Rina Zviel-Girshin. 2023b. ChatGPT + LARA = C-LARA. Presented at SLaTE 2023.

Chotiga Pattamadilok, Pauline Welby, and Michael D Tyler. 2022. The contribution of visual articulatory gestures and orthography to speech processing: Evidence from novel word learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 48(10):1542.

Pauline Welby, Brigitte Bigi, Antoine Corral, Fabrice Wacalie, and Guillaume Wattelez. 2023. A visit to the Cliffs of Jokin: A role for phonetizers in second language pronunciation and word learning, with an example from the languages of New Caledonia. In *Proc. 2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023)*, pages 19–23.

Pauline Welby, Elsa Spinelli, and Audrey Bürki. 2022. Spelling provides a precise (but sometimes misplaced) phonological target. orthography and acoustic variability in second language word learning. *Journal of Phonetics*, 94:101172.