Innovative approaches to speech and language technologies for Oceania, the world's most linguistically diverse region

Abstracts
(alphabetical order by family name)

Let's not be framed by dictionaries! Sharing lexical elicitations to improve lexicographic documentation, examples from the Nisvai-Bislama lexical documentation, Vanuatu

Jocelyn Aznar, contact@jocelynaznar.eu
Chargé de cours à l'Universität Bern, associé à l'ISW, Switzerland
Membre associé au CREDO, Marseille, France

Abstract:

Writing the dictionary of an oral language is a daunting endeavour, often requiring years of effort with limited academical recognition. This challenge likely contributes to lexicographic description being less popular than grammatical description; therefore any improvement in lexicographic processes can contribute to making the task more accessible and rewarding. Inspired by empirical approaches such as DoReCo (Seifart, Paschen & Stave 2022), I wondered about what were my actual data when doing lexicography with Nisvai speakers, South-East Malekula, Vanuatu (Aznar 2019; Aznar & Gala 2020). If narratives often provide occurrences of lexical items, it is only through their translation that evidence of their meaning is acquired, evidence which relies on the translation interpreter's skills. But since I am doing the translation for the Nisvai narrative corpus, the evidence value of the narrative corpus does not originate from the narrators, but from the interactions I had with Nisvai speakers to understand the stories. In order to be able to do those translations to French, the language taught at the local primary school, I engaged with the speakers in annotation interviews, a deeply ethnographic experience (Telban 1997). If the discussions I had to understand each sentence, each word, were sometimes monolingual definitions, more often they were multilingual explanations intertwined with translations in Bislama, the vehicular language of Vanuatu.

In this talk, I will discuss lexical elicitations, an essential aspect of the lexicography of oral languages, and the discourse interlocutors provide during these sessions. Often framed as "folk definition" in the language documentation literature (Casagrande & Hale 1967; Dingemanse 2015; Grimm 2022), the discourses speakers have when answering lexical inquiries are not only made up of definitions, especially of the forms expected from a dictionary, but instead are better approached as metalinguistic discourse (Culioli 1990; Canut 1998; Taylor 2014). In these discourses, the speakers often intertwine both the intermediate language and the documented language to produce explanations, a multilingualism that does not match our monolingual representations of what a dictionary discourse should be like, even in the case of multilingual resources, where the different languages are strictly distinguished and cannot occur in the same sentence. The principle of monolingualism of "proper dictionary sources" is so pervasive that it prevents lexicographers, in particular linguists documenting a language that is not their native language, to acknowledge the actual sources of their knowledge.

To better understand the metalinguistic discourse to which I am referring, here is an example:

Extract from a lexical elicitation with Androng Manmaldou	**Signal_Androng_bebete_2024-07-09_limace.acc**
Nymagmag, Nymagmag ga vi graon… sleta, sleta avyn gur klah lyn nran ag, somtaem mi luk insaed lo stumba blo banana. Inin ara kal nymagmag.	
"Slug, slug is ground, slug, slug which crawls onto the ground here, sometimes I see it inside the banana trunk. That's what they call "nymagmag"."	

The sections underlined with dots are in Bislama, the others are in Nisvai

In this example of a metalinguistic discourse, we see that Androng Manmaldou, knowing that my Bislama skills are still better than my Nisvai skills, uses both languages to describe what the word /nymagmag/ : "slug". The first Bislama word he uses graon: "ground, is actually very surprising, but my interlocutor corrects himself very quickly, and proposes sleta: "slug" afterwards. The follow-up operational explanation, to reuse Casagrande and Hale's terminology (1967, p. 168), is in Nisvai, then the speaker switches to Bislama for a personal observation. He then concludes the discourse with a sentence typical to the Nisvai explanatory discourse: he concludes using the third-person plural, which in this context refers to the previous elders, the now ancestors, who are the people who actually passed down this knowledge to him.

In the presentation, I argue that recording the metalinguistic discourses should be part of the lexicographer's workflow. I start by *discussing* the concept of folk definition, and why the term *metalinguistic discourses* enables us to accept a greater variety of sources. I then discuss two issues that can arise when producing a dictionary: first on an ethical level, the issue of writing such an authoritative object, especially as a foreigner (Cameron 1995). The second issue is the reproducibility, as the actual source of our lexical understanding is not provided to the readers. To face those issues, I discuss examples from my work with the Nisvai community and show how including the Nisvai metalinguistic discourses help facing those issues, firstly by including the voices and the perspectives of speakers in the resources, and secondly, by expliciting the source of our knowledge and our interpretations, and thus providing information that another person, speaker or researcher, could use to understand the reasons behind a translation or a definition choice.

In the conclusion, I acknowledge that, while integrating metalinguistic discourses into our lexicographic resources is not a definitive solution to solve the issues raised previously, it still contributes to improving the quality of the lexical resources produced following this methodology. And improving the reproducibility of lexicographic work will also contribute to making the task more academic, and thus potentially ease lexicographic writing and publications. Finally, it can also improve the versatility of our lexical resources as recording makes the content of the resource more accessible to speakers who cannot read, more interactive as it multiplies the modalities available and improves the reusability of the data for other purposes, academic or not.

Acknowledgement

This work has benefited from the support of the Gesellschaft für bedrohte Sprachen (https://gbs.uni-koeln.de/)

References

Aznar, Jocelyn (2019): *Narrer une nabol : La production des textes nisvais en fonction de l'âge et de la situation d'énonciation, Malekula, Vanuatu*. EHESS.

Aznar, Jocelyn & Núria Gala (2020): The Nisvai Corpus of Oral Narrative Practices from Malekula (Vanuatu) and its Associated Language Resources. *Language Resources and Evaluation for Language Technologies (LREC)*. Marseille, France.

Cameron, Deborah (1995): *Verbal Hygiene* (The Politics of Language). eBook. London and New York: Routledge.

Canut, Cécile (1998): Pour une analyse des productions épilinguistiques. *Cahiers de praxématique*. Presses universitaires de la Méditerranée. (31). 69–90. doi:10.4000/praxematique.1230.

Casagrande, Joseph B. & Kenneth L. Hale (1967): Semantic relationships in Papago folk-definitions. *Studies in Southwestern ethnolinguistics*, vol. 165, 193.

Culioli, Antoine (1990): *Pour une linguistique de l'énonciation: opérations et représentations* (Collection L'Homme dans la langue). Gap, France: Ophrys.

Dingemanse, Mark (2015): Folk definitions in linguistic fieldwork. In James Essegbey, Brent Henderson & Fiona Mc Laughlin (Hrsg.), *Culture and Language Use*, vol. 17, 215–238. Amsterdam: John Benjamins Publishing Company. doi:10.1075/clu.17.09din.

Grimm, Nadine (2022): Documentary Approaches to Lexicography. *Current Issues in Descriptive Linguistics and Digital Humanities: A Festschrift in Honor of Professor Eno-Abasi Essien Urua* (Springer Nature Singapore), 551–567. Singapore.

Seifart, Frank, Ludger Paschen & Matthew Stave (Hrsg.) (2022): *Language Documentation Reference Corpus (DoReCo) 1.2*. Berlin & Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2).

Taylor, Talbot J. (2014): Language in its own image: on epilinguistic and metalinguistic knowledge. In Sylvie Archaimbault, Jean-Marie Fournier & Valérie Raby (Hrsg.), *Penser l'histoire des savoirs linguistiques*, 135–143. ENS Éditions. doi:10.4000/books.enseditions.32065.

Telban, Borut (1997): Mutual Understanding: Participant observation and the transmission of information in Ambonwari. *Canberra Anthropology* 20(1–2). 21–39. doi:10.1080/03149099709508380.

Designing language technologies with indigenous speech communities

Steven Bird, Charles Darwin University, Australia

Abstract:

Ethical engagement with Indigenous communities begins with respectful relationships and a commitment to Indigenous self-determination and leadership. However, the dominant impulse from the language technology community has been to construct universal solutions and then recruit minoritised speech communities to participate in extractive work. In this presentation I will report on several years of experiences in a remote community in Arnhem Land and how local people shaped my work to suit their ends which were centered on cultural survival, caring for Country, and bringing children up strong in ancestral knowledge. I will describe several use cases along with a general-purpose design pattern that may be applicable for other Indigenous peoples/places where locals and newcomers come together to do language work. I hope to provoke discussion on new possibilities for community engagement in Oceania.

C-LARA: disseminating resources for Iaai, an under-documented Oceanic language

Anne Laure-Dotte and Stéphanie Geneix-Rabault, Université de la Nouvelle-Calédonie, Eralo, New Caledonia

Abstract:

Iaai is one of the Kanak languages spoken in the archipelago of New Caledonia, more specifically in the centre and north-east of Ouvéa island (Oceanic language from the Austronesian family), but also in Nouméa, Greater Nouméa and other parts of Grande Terre (Dotte et al., 2017). With around 3,700 speakers (ISEE, 2021), Iaai is a living language, which is taught as an optional subject in the local education system, but for which teaching resources and written productions are still rare. On the Internet, resources are even more sporadic, while demand from teachers, artists and the community itself is strong, linked to the fear of seeing their language disappear, and to the isolation sometimes experienced by speakers of this Kanak language, who find themselves in a living environment where their mother tongue is in the minority and dominated by other Kanak languages, but also by French.

In this context, we developed a collection of resources with the C-LARA tool which is innovative and relevant not only for young learners of the language, but also for language revitalisation efforts. It is composed of resources from oral literature, recorded with two main Iaai speakers in Ouvea, and digitised by researchers from the University of New Caledonia. Thanks to a fruitful international collaboration with a team of colleagues from Flinders University and University of South Australia, we were able to upload seven texts into C-LARA platform ('Fonds Pacifique' grant from the Agence Francaise de Developpement, Maizonniaux et al., 2023-2024).

Several challenges were faced and will be addressed in this paper, such as the particularity of using C-LARA with human-recorded audio, sung voice segmentation issues, the orthographic choices for a language in the process of being standardized, the relevance of the images proposed by DALL-E 3 (CHAT-GPT) to illustrate a rarely documented context, the prospects for pedagogical use of these resources and collaborative enrichment, the lack of resources on languages and our cultural context at the present time in the AI system.

The Little Kids' Word List: A fair vocabulary assessment tool for young Indigenous multilingual children's home languages

Carmel O'Shanessy, Australian National University
Vanessa Davis, Tangentyere Research Hub, Australian National University
Jessie Bartlett and Alice Nelson, Yuendumu Bilingual Development Resource Unit

Abstract:

In Central Australia many Indigenous children grow up hearing and learning more than one language, but these are not usually represented in assessments of their language and cognitive development. If children are not assessed fairly, there is a risk of not understanding their language skills accurately. A multilingual MacArthur Bates Communicative Development Inventory (CDI) for five of the traditional languages spoken by young children in Central Australia, plus English, has been developed. The CDI is available as an online tool with audio, in Eastern & Central Arrernte, Western Arrarnta, Warlpiri, Pitjantjatjara, Luritja and English. A team of Indigenous and non-Indigenous researchers worked with Indigenous families in data collection. In this talk we explain the process and the tool.

Empowering Language Revitalization in Oceania — Make your own language apps!

Vanessa Raffin, E-Reo, Tahiti, French Polynesia

Abstract:

Oceania mirrors a global trend where many indigenous languages lack documentation, pedagogical resources, and effective tools for passing them on to younger generations. E-Reo provides a no-code platform that allows indigenous communities and linguists to create, publish, and maintain mobile applications for language and cultural preservation. From a single, even limited database of content (words, sentences, audio), users can easily build engaging apps for learners to explore and practice languages through interactive play. This solution simplifies technical aspects, addresses geographical dispersion, and engages youth through digital means.

C-LARA 1 and 2: Hands-on sessions

Manny Rayner, University of South Australia, Adelaide

Abstract:

ChatGPT-based Learning And Reading Assistant (C-LARA; https://www.c-lara.org/) is an AI-based platform which allows users to create multimodal texts designed to improve reading skills in second languages. GPT-4/ChatGPT-4 is central to the project: as well as being the core language processing component, it has in collaboration with a human partner developed the greater part of the codebase.

In these two hands-on sessions, we will show you how to use C-LARA to create multimodal texts. The first session will focus on "Simple C-LARA", a wizard-style interface that allows the non-expert user to create a short illustrated multimodal text for any of the several dozen langages that GPT-4 supports. All the user needs to do is enter an initial prompt and approve default choices a few times, though a little post-editing often helps improve the initial result.

In the second session, we will introduce "Advanced C-LARA", the full version of the platform, and show how to make multimodal texts for low-resource langages the AI does not know. Here, glosses and other annotation must be entered by hand. The AI can however still assist by automatically creating illustrations, working off user-supplied translations in a langage it knows.

You can see many examples of C-LARA-generated texts at https://clara.unisa.edu.au/accounts/public_content_list/.

ChatGPT-Based Learning And Reading Assistant: Initial Report. https://www.researchgate.net/publication/372526096_ChatGPTBased_Learning_And_Reading_Assistant_Initial_Report

ChatGPT-Based Learning And Reading Assistant (C-LARA): Second Report. https://www.researchgate.net/publication/379119435_ChatGPTBased_Learning_And_Reading_Assistant_C-LARA_Second_Report

Making Picture Book Texts with C-LARA (interim report). https://www.researchgate.net/publication/381323238_Making_Picture_Book_Texts_with_CLARA_interim_report

The online dictionary of the Tahitian Academy and its functionalities

Jacques Vernaudon, University of French Polynesia (Eastco, MSH-P)

Abstract:

Since 2017, the online dictionary of the Tahitian Academy, which originally only allowed a simple search from Tahitian into French, has been enriched with numerous features, including allowing a reverse search from French into Tahitian, providing access to the etymology of numerous words, making their pronunciation audible, revealing word occurrences in corpora of Tahitian texts, or providing iconography and up-to-date scientific names associated with vernacular plant names. These improvements have been made possible by structuring the dictionary within the Anareo digital database hosted by the University of French Polynesia, with the help of several other scientific and technical partners. My presentation will offer a tour of the dictionary, its main functions and the underlying digital architecture.

Learning Kanak languages in the digital age

Fabrice Wacalie, Université de la Nouvelle Calédonie, Laboratoire Interdisciplinaire de Recherche en Éducation (LIRE EA 7418), New Caledonia

Abstract:

Linguistic diversity has been under threat in New Caledonia for several decades as a result of socio-economic, cultural and environmental changes within the archipelago (Wacalie, 2010). While there had previously been little development of teaching tools, the Covid crisis gave a major boost to the introduction of innovative distance learning systems for Kanak languages.

These are welcome initiatives, given that half of these languages are in danger of extinction according UNESCO criteria. In response to this loss, the school has become a place of transmission and sometimes revitalisation in New Caledonia, a French overseas collectivity, since the integration into the curriculum of Kanak languages (Vernaudon, 2013) and "Fundamental Elements of Kanak Culture" (Minvielle, 2019). These subjects are elements specific to the New Caledonian education system, which has been keen to give its schools their own identity since the successive Matignon-Oudinot (1988) and Nouméa (1998) agreements, and which took on a new dimension with the implementation of the New Caledonia Education Project (PENC) (2016).
Despite all the measures in place, the loss of languages continues. The development of self-learning tools then proves to be a path to explore in the digital age and social networks. Over the years, local institutions have made efforts to produce open-access digital resources for teachers of these new subjects. With the assistance of the Académie des Langues Kanak, we coordinated the creation of a tutorial for learning Drubea (http://www.drubea.com), one of the 28 Kanak languages, spoken in the far south of New Caledonia, with 1022 speakers according to the last census. The tutorial was modelled after online courseware for another Kanak language, Nengone, the language of the island of Maré (http://nengone.univ-nc.nc/, Bearune & Vernaudon, 2014). We are currently developing an online dictionary in the digital language.
This communication will be an opportunity to present the work carried out on Kanak languages.

References:

Bearune, S. & Vernaudon, J. (2014). Un didacticiel en ligne d'initiation au nengone, langue de Maré, Nouvelle-Calédonie. Langues et cité n°26.

Minvielle, S. (2022). École et diversité culturelle en Nouvelle-Calédonie, L'enseignement des éléments fondamentaux de la culture Kanak et sa contribution à la formation du citoyen calédonien in Sociétés inclusives et reconnaissances des diversités, 139-158. Presses Universitaires de Rennes.Rennes.

Sam, L., D. (2022). Historique de l'Enseignement des langues kanak (ELK) en Nouvelle-Calédonie. Langues et cité, 31, 16-17.

Wacalie, F. (2010). La diversité linguistique calédonienne, BNF, Chemin d'accès. Regards en Archipel. Paris.

Documenting Pacific languages, methods for longterm access to language records

Nick Thieberger, University of Melbourne and PARADISEC, Australia

Abstract:

The Pacific region is a major locus for linguistic diversity, but for most languages there are few records available. When a linguist or ethnographer works to record aspects of the culture or language there needs to be a way of ensuring the recordings are made available for the source community and for future reseachers. To do this, the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) has, since 2003, been digitising earlier analog recordings and putting them into an online service with searchable metadata, applying appropriate access licences. It is training new researchers to create records that can be described and archived for future re-use, and archives the output of current research. These citable records are also the basis for verifiable research creating a virtuous circle that takes account of both research and community-facing use of the records.

The development of MPAi: a tool to learn the pronunciation of Māori vowels

Catherine I. Watson , Justine C. T. Hui, Brooke Ross, Isabella Shields, and Peter J. Keegan, University of Auckland, Aotearoa/New Zealand

Abstract:

This paper outlines the pronunciation tool (MPAi) for the Māori language, the language of the indigenous people of Aotearoa/New Zealand. Māori is threatened and after a break in transmission the language is currently undergoing revitalization. The data for MPAi has come from a corpus of 60 bilingual speakers of Māori (men and women). MPAi allows users to model their speech against exemplars from young speakers or older speakers of Māori. This is important, because of the status of the elders in the Māori speaking community, but it also recognizes that Māori is undergoing substantial vowel change. MPAi arose from requests from the Māori community. MPAi gives feedback on vowel production via formant analysis. We discuss the evolution of MPAi , and the evaluations we have done, the resources we have developed as a consequence of those.

LaPasserelle.nc: a gateway to the pronunciation of the Kanak languages of New Caledonia

Guillaume Wattelez, Université de la Nouvelle Calédonie, LIRE, New Caledonia
Fabrice Wacalie, Université de la Nouvelle Calédonie, LIRE, New Caledonia
Pauline Welby, Aix Marseille Université, CNRS, Laboratoire Parole et Langage, France and Université de la Nouvelle Calédonie, LIRE, New Caledonia

Abstract:

Encountering written words and names in one of the almost 30 languages of the indigenous Kanak people of New Caledonia is an everyday experience, for example, on class lists, road signs, and in news articles. Pronouncing these words is often a challenge, since each of the languages has its own phonology and its own orthography. To address this challenge, we have developed a web application, hosted at LaPasserelle.nc, to serve as a gateway to the pronunciation of Kanak languages (*la passerelle* means 'the gateway' in French, the common language in New Caledonia). The main goal of LaPasserelle.nc is to provide an idea of how to pronounce a written word in Oceanic languages, especially Kanak languages. We will start with a demo of the phonetizer, which in this first version handles two languages, Drehu (the language of the island of Lifou) and Paicî (one of the languages of the north of the Grande Terre, the main island). We will show its functionalities and briefly described their implementation.

The central functionality is the phonetization. The user types or pastes a text in the input box. This input is then processed and converted to X-SAMPA phonetic transcription, using grapheme-phoneme correspondences, and a suggested pronunciation is given as output in one of two formats. The first is an International Phonetic Alphabet (IPA) transcription of the input text. However, since most people (alas!) do not understand IPA, a pronunciation respelling, based on the spelling of French (a *franétique* transcription, < *français* + *phonétique*) is also provided, enhanced by pedagogical tips. For example, for the language name *Drehu*, the phonetizer gives the IPA transcription [dʒehu] or the franétique transcription *djé-hou*. Other display options are also available, for example, to allow the user to see syllable breaks. This is useful in the parsing of particularly long words, as well as in the correct pronunciation of vowel-vowel sequences, e.g. in Drehu *kaloi* [ka.lo.i] 'good' (not [ka.lwa] as in French).

The user is "taken by the hand" thanks to a clear association between input and output, in which each sound of a word is visually connected to its pronunciation. When a phoneme is not found in the inventory of the French language, a pedagogical tip gives advice (as might a teacher or a native speaker friend) on how to pronounce the grapheme. For instance, the <x> grapheme pronounced [x] in many Kanak languages, as in the Drehu word *xen* [xen] 'to eat', does not exist in French and is considered difficult to pronounce for French speakers. However, a more intuitive pronunciation respelling (e.g. *xen* [KHén] 'to eat'), with a picture showing place of articulation (midsagittal plane showing a velar constriction), sounds recordings or references to other, familiar languages containing these phonemes should help.

We will conclude by briefly discussing ongoing and future work on and with the LaPasserelle.nc multilingual phonetizer. This includes the extension to other languages, the implementation of the franétique transcriptions in C-LARA online texts, and potential applications of "semantic web" or "web of data" technologies and concepts to phonetization.