Leveraging Speech Recognition for Interactive Transcription in Australian Aboriginal Communities

Eric Le Ferrand

CHARLES DARWIN UNIVERSITY

DOCTORAL THESIS

Leveraging Speech Recognition for Interactive Transcription in Australian Aboriginal Communities

---

Author:
Eric LE FERRAND

Supervisor:
Pr. Steven BIRD
Pr. Laurent BESACIER

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

College of Indigenous Future, Education and the Arts
Charles Darwin University

April 3, 2023

CHARLES DARWIN UNIVERSITY

Abstract

Leveraging Speech Recognition for Interactive Transcription in Australian Aboriginal Communities

by Eric LE FERRAND

Speech recognition is a field of Natural Language Processing that involves automatically recognising speech data and converting it into text, such as an orthographic transcription or a sequence of phones or other kinds of labels. These technologies are predominantly used in Western societies and serve specific purposes, such as voice dictation, smart home devices, and automatic subtitling, for dominant languages such as English, French, and Mandarin. Low-resource languages refer to languages for which available data is too scarce to solve a specific task.

Indigenous communities worldwide have a history of Western scholars visiting their lands for research purposes. These visits generally have two goals: the first is the collection and analysis of Indigenous data to help Western scholars gain a better understanding of Indigenous contexts, and the second is the deployment of Western methods to address local issues. These methods are part of a long tradition of colonialist practices in which Westerners extract Indigenous resources without local benefit or disregard Indigenous bodies of knowledge and ways of knowing.

Recent calls have encouraged scholars to decolonise their research practices, which involves following a set of principles to enable more ethical research. In response to this call, research projects need to benefit the community and respect its social, cultural and political components. These principles have been translated into the adoption of Indigenous research methods, negotiations with the community

regarding the outcome of each project, and working with the community to build self-determination.

The research around speech recognition has mainly focused on generalising generic computational methods across different languages, without adequately addressing ongoing colonial practices. In this thesis, we begin with the language documentation agenda and explore how speech recognition technologies can fit into Indigenous communities, following decolonising principles. Specifically, we focus on a single Indigenous Australian community and explore how to create a speech recognition system that takes into account the constraints of this context. We also examine how to use this system in a cross-cultural activity and shape it to respect the community's desired outcomes in terms of language work.

Acknowledgements

Writing this thesis, I received support from every direction while I was in an unfamiliar country and in the middle of a pandemic. I cannot express how grateful I am for the support provided by both of my supervisors. First, Prof. Steven Bird has been my main mentor throughout this project. He has been able to keep me on the right path despite my tendency to to lose myself in old patterns in which I felt comfortable. He was also able to relieve my anxiety in moments I felt like giving up. Then, Prof. Laurent Besacier was able to remind me that my work was valuable when nothing seemed to make sense to me. He was also able to help me refocus when I got too excited about too many projects.

I would like to thank William Lane, who shared my office for most of my PhD. He was one of my greatest supports. I will truly miss having him around. Dr. Mat Bettinson provided great support in the early stages of my project, specifically in designing the different apps he made for my fieldwork. Many thanks to the rest of the Top End Language Lab, from Darwin and elsewhere: Brandon, Valerie, Tereza, Cat, and Josh, I will miss our discussions, picnics, and sunset drinks. Nikki, Annie, Julie, and Henry, it was nice chatting with you every week. I would also like to thank Dr. Cathy Bow for her good advice. She was a great co-worker and will be missed.

Throughout this project, I had the opportunity to collaborate with scholars and travel to different parts of Australia. I am grateful for the mentorship of Dr. Jennifer Lawrence Taylor during fieldwork and for her guidance in my research. I also appreciate the hospitality of Maïa Ponsonnet and Trevor Cohn who welcomed me into their academic groups during my visits to the University of Western Australia and the University of Melbourne.

This thesis involved several trips to remote Indigenous communities, and I would like to thank all the stakeholders who facilitated this work. First and foremost, I would like to thank the Bininj and Daluk people who welcomed me onto their Country and collaborated with me on this project. I extend special thanks to Kamanj

Jean from Mamadawerre, Bangardi Gavin, Kodjdjan Mary, and Balang Keith from Kabulwarnamyo, my Yabok her partner Kela Lewis, Kamarrang Ross from Manmoyi, and the artists from Gunbalanya Bulanj Sprotin, Kamarrang Roland, Bulanj Theo, and Kela Isaac. I am also grateful for the assistance of balandas in Bininj Country who facilitated this work, specifically Alex Ressel and Celina Ernst.

Finally, I would like to express my deep appreciation to my family who I missed greatly while being away: my mother Catherine, my father Philippe, my sister Emmanuelle, my brother Olivier, and my newborn nephew Leonard whom I finally had the chance to meet. I am also grateful to my Darwin friends who became my family in my new home. Josh and Polly were my biggest emotional support throughout these past years, as were my fantastic housemates Emily, Dani, Caitlin, Caz, Ellen, Ammy, Sarah, and all the others who shared my mansion life. My friend Clem, Flav, Elise, and Alex kept a part of France in my daily life, and I thank all my other Darwin friends. I want to thank my friends Nina and Alexia, who maintained their connection with me while I was away. Finally thank you to Jeremy Garnett, Aidan Mannion, Polly Henri, Marilyn Higson and Caroline Ziada for proofreading this thesis.

Remerciements

En écrivant cette thèse, j'ai reçu un grand nombre de soutiens alors que j'étais dans un pays étranger et au milieu d'une pandémie. Je n'ai pas les mots pour dire à quel point je suis reconnaissant pour le soutien que m'a offert mes directeurs de thèse. D'abord Pr. Steven Bird qui a été mon mentor principal pendant ce projet. Il a su me garder sur le droit chemin alors que j'avais tendance à me perdre dans mes vieilles habitudes dans lesquelles je me sentais confortable. Il a su calmer mon anxiété dans des moments où j'ai eu envie d'abandonner. Ensuite Pr. Laurent Besacier qui a su voir du bon dans mon travail quand rien ne semblait faire sens. Il a aussi su m'aider à me recadrer quand j'ai eu tendance à me perdre dans trop de projets.

Quand je suis arrivé à Darwin en 2019, nous n'étions que quatre. J'aimerais remercier particulièrement William Lane, mon co-bureau pendant la grande majorité de mon doctorat. Il a été l'un des plus gros soutiens que j'ai reçu quand on se perdait mutuellement dans un océan de tristesse. Sa présence au quotidien va vraiment me manquer. Ensuite Dr. Mat Bettinson qui m'a grandement aidé au début de ma thèse en créant les outils pour mes premiers travaux de terrain. Un grand merci au reste de l'équipe TELL à Darwin et ailleurs. Brandon, Valérie, Tereza, Cat et Josh. Nos discussions, pique-nique et apéros vont me manquer. Nikki, Annie, Julie et Henry, c'était sympa de vous voir pendant nos discussions hebdomadaires. J'aimerais aussi remercier Dr. Cathy Bow pour ses bons conseils. Elle était une très bonne collègue et va nous manquer.

Pendant ce projet, j'ai eu l'opportunité de rencontrer de nombreux chercheurs à travers plusieurs collaborations et voyages en Australie. D'abord Dr. Jennifer Lawrence Taylor qui m'a beaucoup appris alors que j'enchaînais les difficultés pendant mes travaux de terrain. Ensuite Maïa Ponsonnet et Trevor Cohn qui m'ont accueilli dans leurs équipes pendant mes visites à l'Université d'Australie Occidentale et l'Université de Melbourne.

Cette thèse a impliqué plusieurs voyages dans des communautés Aborigènes reculées. J'aimerais remercier tous les collaborateurs qui ont facilité ce travail. D'abord tous les Bininj et Daluk qui m'ont accueilli dans leur terre et qui ont collaboré avec moi dans ce projet. Merci à Kamanj Jean de Mamadawerre, Bangardi Gavin, Kodjdjan Mary et Balang Keith de Kabulwarnamyo. Ma Yabok, son partenaire Kela Lewis et Kamarrang Ross de Manmoyi. Merci aux artistes de Gunbalanya Bulanj Sprotin, Kamarrang Roland, Bulanj Theo et Kela Isaac. Grands mercis aux Balandas travaillant en terre Bininj qui ont facilité ce travail, en particulier, Alex Ressel et Celina Ernst.

Enfin, merci à ma famille qui m'a tellement manqué pendant que j'étais en Australie : ma mère Catherine, mon père Philippe, ma sœur Emmanuelle, mon frère Olivier et mon nouveau-né neveu Léonard que j'ai enfin eu la chance de rencontrer. Merci à tous mes amis à Darwin qui ont été ma famille dans mon nouveau chez moi : Josh et Polly qui ont probablement été mon plus grand soutien émotionnel pendant ces trois années, mes fantastiques colocataires Emily, Dani, Caitlin, Caz, Ellen, Ammy, Sarah et tous les autres qui ont partagé avec moi ma vie de palace. Merci à mes amis Clem, Flav, Elise et Alex qui ont gardé dans mon quotidien un peu de France. Merci à tous mes autres amis de Darwin. Merci pour mes amies Nina et Alexia qui ont su garder un lien avec moi alors que j'étais loin. Enfin, merci à Jeremy Garnett, Aidan Mannion, Polly Henri, Marilyn Higson et Caroline Ziada pour avoir relu cette thèse.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AE	**A**uto **E**ncoder
AP	**A**verage **P**recision
ASR	**A**utomatic **S**peech **R**ecognition
AUC	**A**rea **U**nder the **C**urve
BERT	**B**idirectional **E**ncoder **R**epresentations from **T**ransformers
BLEU	**Bil**angual **E**valuation **U**nderstudy
BLSTM	**B**i-directional **L**ong **S**hort **T**erm **M**emory
BNF	**B**ottle**n**eck **F**eatures
cAE	**c**orrespondence **A**uto **E**ncoder
DNN	**D**eep **N**eural **N**etwork
DTW	**D**ynamic **T**ime **W**arping
EGIDS	**E**xpended **G**raded **I**ntergenerational **D**isruption **S**cale
FSA	**F**inite **S**tate **A**utomaton
fMLLR	**f**eature-space **M**aximum **L**ikelihood **R**egression
GETALP	**G**roupe d' **E**tude en **T**raduction **A**utomatique traitement automatisé des **L**angues et de la **P**arole
GMM	**G**aussian **M**ixture **M**odel
GPU	**G**raphic **P**rocessing **U**nit
HCI	**H**uman **C**omputer **I**nteraction
HMM	**H**idden **M**arkov **M**odel
ICALL	**I**ntelligent **C**omputer **A**ssisted **L**anguage **L**earning
IPA	**I**nternational **P**honetic **A**lphabet
LIG	**L**abortoire **I**nformatique de **G**renoble

mBNF	**m**ultilingual **B**ottle**N**eck **F**eature
MVN	**M**ean and **V**ariance **N**ormalisation
MFCC	**M**el-**F**requency **C**epstral **C**oefficient
NHMRC	**N**ational **H**ealth and **M**edical **R**esearch **C**ouncil
NLP	**N**atural **L**anguage **P**rocessing
PER	**P**hone **E**rror **R**ate
PLP	**P**erceptual **L**inear **P**rediction
POS	**P**art **O**f **S**peech
P2W	**P**hone **2** **W**ord
QbE	**Q**uery-**b**y-**E**xample
ROC	**R**eceiver **O**perator **C**haracteristic
ROUGE	**R**ecall-**O**riented **U**nderstudy for **G**isting **E**valuation
TELL	**T**op **E**nd **L**anguage **L**ab
VTLN	**V**ocal **T**ract **L**ength **N**ormalisation
WER	**W**ord **E**rror **R**ate
w2v	**w**av **2** **v**ec
XAI	e**x**plainable **AI**

Chapter 1

Introduction

1.1 Introduction

Not everything you will experience is going to make sense to you. Does everything you do always make sense anyway? This piece of advice was given to me prior to my first trip in Bininj Country, a remote Indigenous community in West Arnhem land in Northern Australia. I flew to Darwin, the capital of the Northern Territory, Australia, from Paris, France, and started to get ready for this first trip by reading, attending cross-cultural research training and taking advice from peers with similar experience. I was mainly told to keep an open mind in a cultural setting far from mine. I was entering Bininj Country where I was obviously going to encounter a different way of living, different languages, different social rules and a different way of knowing.

In Western cultures, we share similar norms and values, such as a sense of private property and public spaces, and a shared approach to knowledge represented in national and international institutions, such as education systems and international conferences. However, when I travelled to Bininj Country, I knew I needed to question many of my social and cultural assumptions. *Not everything you will experience is going to make sense to you.* While it's important to acknowledge cultural differences, navigating through them is not an easy task. Moreover, in the context of a research project, particularly one involving technology, cross-cultural collaboration can feel like hiking in high heels: it's uncomfortable, takes more time to reach the

end, and we're unsure why we're wearing those heels in the first place. Language technologies in Indigenous contexts can seem out of place, like wearing heels in King's Canyon. Despite this, many computer science sub-fields view technology as a means of addressing local issues.

Does everything you do always make sense anyway? Probably not, hence this advice to encouraging me not to overthink everything I experienced on Indigenous land. Yet this question became even more relevant in the context of a research project where the motivations behind the project were not always clearly defined. Despite receiving advice on how to avoid falling into the common trap of becoming a white saviour who aims to save Indigenous languages, I still found myself succumbing to this cliché. Initially, I designed methods solely based on academic literature, without considering the specific needs and perspectives of the local community. It wasn't until I faced obstacles for two years that I realised the importance of exploring new paths and designing methods that were relevant and meaningful to the local context

Numerous scholarly works discuss the global presence of over 7000 languages, and the looming threat of extinction faced by a significant number of them. In this context, language technologies are regarded as a potential means of safeguarding these languages by integrating them into a language documentation framework (Adda et al., 2016; Duong, 2017; Jimerson and Prud'hommeaux, 2018; Stuker et al., 2009). Sometimes, assumptions are also made about languages' needs or what speech communities want (Hasegawa-Johnson et al., 2020).

These ideas are based on the conventional methods employed by field linguists who collect linguistic data for the purpose of documenting language phenomena (Bouquiaux and Thomas, 1992). However, this direction has limitations. Firstly, there is no evidence that the work of the linguist has any influence on the preservation of languages (Nathan and Fang, 2013). Instead, it seems to serve only an academic agenda that has no direct relationship with language preservation (Hanke, 2017). Secondly, the documentation process often fails to mention the language community and their level of involvement, which may indicate a manifestation of colonialist

approaches: the idea being that "because nobody does the job, we are going to do it ourselves". Special workshops, such as the "zero resource challenge" (*Zero resource speech challenge* n.d.), with the introduction of a "surprise language", push in this direction, bypassing the need for language experts and language communities. In Natural Language Processing (NLP) research, we create systems, train models, and evaluate them using standard metrics. A contribution then emerges based on its improvement from a previous system or model. But what do we learn? Because high-resource languages are generally grounded in a familiar cultural and socio-political setting, the purpose behind a system is often assumed to be implicit. There are examples of evaluated systems that have led to final products. Cultural differences, different ways of knowing, different language types, or vitality levels imply different agendas and requirements that should prompt reflection on the usability of the systems that are built.

Scholars from various fields including public health, anthropology, and human-computer interaction, have addressed the ongoing use of colonialist methods. Indigenous communities have a history of Western researchers coming into their lands to harvest Indigenous data in order to boost their academic careers, with no concrete benefits for the community. Equity and reciprocity are among the main principles dictated by ethical guidelines. Therefore, local populations should not be considered as passive objects to be studied, but instead as collaborators where the needs of a research project have direct benefits for the community according to their values. In other words, a research project should be community-based, conducted with the community and for the community.

More space has been provided in major workshops, conferences, and journals for research projects that are specific to individual languages. These projects are not necessarily in conflict with generic computational methods, as they allow for the identification of real-life issues. For example, a large number of Indigenous languages have complex morphology that is poorly represented in the few high-resource languages, and therefore requires special treatment. These unique features

have been explored in various contexts, and have often led to the development of morphological analysers (Lane and Bird, 2019; Schwartz et al., 2019) that work progressively toward language development in collaboration with speech communities (Lane and Bird, 2020; Schreiner et al., 2020).

Language-specific projects also allow for clear conceptualisation of the amount and quality of data available, a core element of any Natural Language Processing related project. The definition of data, its shaping or annotation, the amount of time it took to be collected, or its availability, bound the range of the possible outcomes. No single context can serve as a reference, but the description of different contexts allows for the refinement of computational methods used for low-resource languages, enabling realistic design.

The development of language technology and language development is often only considered through writing. Similar works involving speech technology are relatively rare. Yet Indigenous languages are generally only spoken, with writing often the result of colonisation and contact with Western institutions. While teaching through writing might make sense when sustainable literacy is achieved (Section 2.3.1), many Indigenous communities maintain a strong oral tradition with almost no need for writing. While the final objective might be keeping the language strong, the forced use of writing might have the opposite effect, discouraging the speakers from being involved in language technology design projects. Usually mentioned from within the language documentation perspective, speech recognition could be relevant in Indigenous contexts in a community-based setting. Speech recognition offers the potential to respond to a general concern of keeping the language strong while mantaining traditional language use practices.

1.2 Research questions

The main question we will attempt to answer in this thesis is: "In what ways can speech recognition technologies support community-based language preservation?" This

question pertains to three aspects of performance, communication, and engagement, which will be addressed through the following questions.

RQ1. Performance: how can speech recognition technologies find their place in Indigenous contexts?

The evaluation of the systems created, as currently done using standard metrics, does not provide informative insights into the applicability of such systems in real-life scenarios. Through this first question, we aim to explore **the connection between mainstream system evaluation and real-life application** and **how to define success for a given method in real-life scenarios within Indigenous contexts?** These questions are partially related to traditional NLP research based on experimental protocols applied to language data. However, the lessons learned from experimental results will be put into perspective within an Indigenous context (Section 3.2 and 3.4).

RQ2. Comprehension: how can speech recognition technologies be incorporated into an activity that makes sense to the local community?

In Section 1.1, we briefly mentioned how Indigenous communities exist in a cultural environment that can present challenges for collaboration, particularly due to their different way of organising and transmitting knowledge that may not align with Western epistemology. The introduction of language technology also adds a level of complexity. These cultural differences offer the opportunity to explore methodological paths to facilitate cross-cultural collaboration. Through this second research question, we aim to explore **how language preservation and language technology themes can be bridged in a cross-cultural context** and **what insights can be gained from existing practices to design collaborative workflows?**

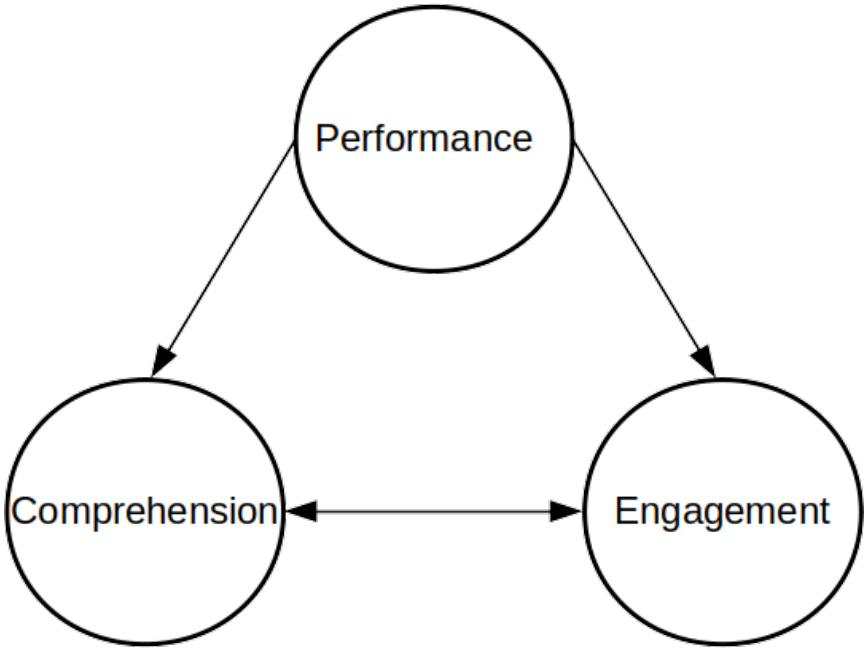


FIGURE 1.1: Three research components

The three main axes of this thesis are represented in Figure 1.1. **Comprehension** is related to cross-cultural collaboration and different ways of knowing, **Engagement** to the decolonisation of research practices, and **Performance** to the application of technological solutions. These three axes have been explored individually through anthropological, ethical, and computational perspectives. Recently, interdisciplinary research has started to apply cross-cultural and decolonising principles to their own fields, including linguistics, public health, and human-computer interaction. The contribution of this thesis is situated at the intersection of speech recognition, cross-cultural research, and decolonising research methods.

1.3 Fieldwork

1.3.1 Bininj Country and language

At the national level, the number of Indigenous people in Australia is around 3% of the total Australian population. In the Northern Territory, this population reaches 30% (Australian Bureau of Statistics, 2016). The Northern Territory is mostly made up of

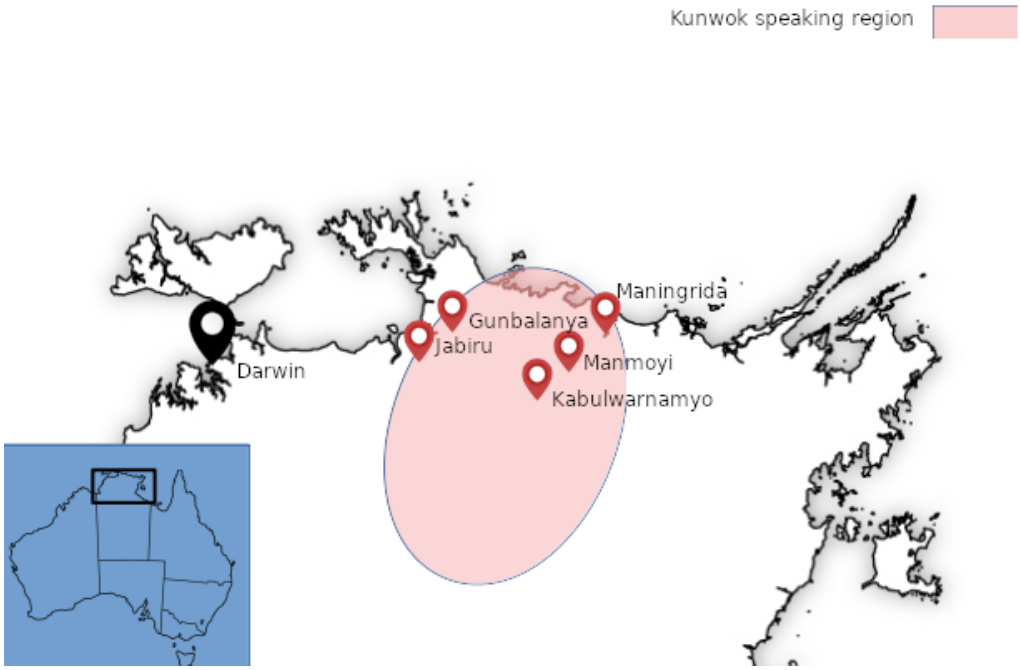


FIGURE 1.2: Map of Kunwok speaking region in the Northern Territory, Australia

regional locations and Darwin (traditionally called Garramilla), the capital and largest city, has a population of less than 150,000.[1] The traditional Country of the Bininj people, where my fieldwork took place, is situated on the Arnhem Land plateau, 300 km east of Darwin (Figure 1.2). Bininj land probably used to be smaller but expanded following colonisation with the settlement of cattle stations and the establishment of missions, which led to the population being displaced from their traditional lands to Western settlements(Remote Area Health Corps, 2010).

The Bininj people speak Bininj Kunwok (ISO-gup), or simply Kunwok, a Gunwinyguan language. The term Kunwok is primarily used by linguists, with the people referring to one of its mutually intelligible dialects: Kunwinjku (the primary dialect), Kundjeyhmi, Manyalluluk Mayali, and Kundedjnjenghmi. Kunwok is spoken on the eastern side of Kakadu National Park and around the Arnhem Land plateau. The main settlement is Gunbalanya (also written Kunbarlanja). Some Kunwok is also spoken in Maningrida, alongside other languages. The other locations indicated in Figure 1.2 are outstations (traditional Indigenous homelands), including Kabulwarnamyo

[1] Census 2019

and Manmoyi, where we were the most involved. The region where most Kunwok is spoken is an Indigenous Protected Area, and thus entrance is restricted to permit holders.

Bininj Kunwok (referred to here as Kunwinjku) is a polysynthetic language. A polysynthetic language is a highly synthetic language in which morphemes are put together to form long words which are translated into full sentences in most European languages (e.g. *Karribidyikarrmerrimen* - let's help each other). Kunwok is traditionally only spoken. Writing only exists in those rare places where there is collaboration with a Western institution (e.g. ranger program, school, tourism, academia). Kunwok speakers rarely write in their own language and are often not confident in doing so. However, they are literate in English through the Australian education system.

1.3.2 Skin name and relationship

Bininj people have social norms on how to address another person, and using the wrong term could be disrespectful. This relationship system (also known as kinship system) has an important place in the Bininj world and is one of the first things that a *balanda* (non-Indigenous person) needs to learn. This system was a common topic of discussion within the community and was part of one of the contributions. Therefore, it is important that the reader has some understanding of its mechanism.

In Bininj Country and most Aboriginal communities in Australia, each person is assigned both a European name (such as Dean, Isaac, Stuart, etc.) and a skin name (known as *Kunkurlah*). There are a total of eight skin names organised by skin group, with a male and female variant (e.g., *Nawakadj* for a man, *Ngalwakadj* for a woman). The skin group to which one is assigned affects one's relationships with others, choice of marriage partner, and moiety. The skin name is inherited from one's mother according to the $A \rightarrow B \rightarrow C \rightarrow D \rightarrow A$ order (as shown in Figure 1.3). The same eight skin names are the same in every family, and disambiguation is achieved

by clarifying one's relationship with another or the clan to which one belongs. The same skin names are not used everywhere but have variants which correspond across communities. For instance, someone named *Nakangila* in the Gunbalanja region would be called *Bulanj* in the Maningrida region.

FIGURE 1.3: Bininj Kunwok skin Names; Gunbalanya Region, Region, Northern Territory, Australia [2]

I have been adopted as a brother by an Elder in Manmoyi and have inherited the male version of her own skin name: *Nawakadj*. Following the skin name order, my adoptive mother's skin name is *Ngalkodjok*, and my sisters' sons and daughters' names are *Nabangardi* and *Ngalbangardi* respectively. Her daughters' children are *Nakangila* and *Ngalkanila*.[3]

Based on the skin name, we are part of a kinship system in which we need to address people in the community according to the relationship we share. For instance, I call my adoptive sister *yabok* (older sister), and she calls me *ngadburrung* (sibling). I call my adoptive mother and her sisters *karrang* (mummy), they call me *djedje* (son). The kinship system is a core component of Bininj culture. Misnaming someone can sound disrespectful. Simple kin terms are only the first layer of a complex relational system that I will not detail here.

Discussion about the kinship system proved to be an effective method of engagement with the Bininj people. As someone from Brittany, France, I found it challenging to find topics of discussion to build relationships for this research. Kin terms were a way to firstly show my interest and then respect for local knowledge and initiate

[3] https://kunwok.org/wiki/Skin_Names consulted 02/02/2022

conversation about ways to address people in the community and find out about our common relationships.

FIGURE 1.4: Discussion about relationships based on a family tree

1.4 Positionality statement

I come from Brittany, which is located on the western side of France, and French is my native language. I was born in the Rennes region, which is located in the eastern side of Brittany where the Gallo language is traditionally spoken. My grandparents on my father's side came from the countryside of Brittany, including Pays Vannetais and Pays de St-Malo. Breton and Gallo were their native languages, as their parents did not speak French. However, neither of these languages was transmitted to my father or myself. Like many other languages in the world, Breton and Gallo have faced the pressure of a dominant language and the political decisions of the central state, which have led to a large language shift. This project allowed me to explore language preservation themes, drawing on my background in NLP and speech processing. I moved to Darwin, Australia, in 2019 specifically to work on this thesis project. I positioned myself in this research as a white male European researcher on Indigenous

Land: Larrakia Country, where the main campus of Charles Darwin University is based and Bininj Country, where my fieldwork took place.

My research took place at the Northern Institute at Charles Darwin University. I am part of the newly created Top End Language Lab (TELL). The group was originally created in 2019 by Professor Steven Bird and consisted of one professor, one post-doctoral research fellow, and three PhD students, including myself. Originally, our research focused on topics such as computational linguistics, language revitalisation, and software design. Our research focus progressively expanded to cover community well-being and Indigenous topics, including language, tourism, and archives. I was also affiliated with the GETALP group in the Laboratoire d'Informatique de Grenoble (LIG) located at Grenoble Alpes University in France. My second supervisor, Professor Laurent Besacier, was originally affiliated with this group.

1.5 Research scope

I approached this work from an NLP perspective, specifically speech recognition. Contributions in this space (speech processing for data-restricted languages) have explored various topics where the contributions involve increasing the amount of annotated resources or indexing unlabelled data. While research contributions in this space include Indigenous data, the connection between the systems created and their benefits for the speech community have not been made clear. Starting from the language documentation agenda that is often mentioned for small languages, this thesis explores how speech recognition can find its place in an Indigenous context using a community-based approach.

The link between a system and its final usage is rarely described. While standard metrics are generally used to assess the performance of a method, its deployment in each context could pose challenges that need to be addressed. Another aim of this thesis was to identify the benefits and flaws of speech recognition methods in an Indigenous context for an activity negotiated with Indigenous stakeholders.

Working with Indigenous participants implies a cross-cultural collaboration. In this kind of collaboration, there is a need to work through different ways of knowing in order to enable better comprehension and to build a space in which Indigenous participants feel safe and comfortable. This kind of collaboration has been explored in other fields but has been relatively ignored in research around NLP topics. This research is an opportunity to explore what strategies could be built to facilitate the collaboration between western scholars and Indigenous stakeholders regarding technological topics.

This research took place on Bininj Country (Section 1.3.1). Some of the actions taken during this project are bound to this context and are not necessarily translatable to others. However, the methodological paths taken were intended to illustrate ethical principles for research regarding speech recognition, for instance, to identify what decolonising methods or cross-cultural collaboration means in research around speech recognition and language development.

1.6 Thesis outline

We presented the introduction of this project in Chapter 1, which included the research questions addressed in this thesis, the research scope for this project, and a positionality statement.

In Chapter 2, we provide a literature review and articulate the gap in knowledge at the intersection of speech recognition, language development, and technology design from a community-based perspective. This section covers research on ethical research in Indigenous spaces, language classification and documentation, technology design, and speech recognition.

Chapter 3 contains the contributions of this thesis articulated within four conference articles. This chapter includes the final versions of peer-reviewed and published articles. The contributions include the first implementation and deployment of our

sparse transcription model, the design of a new spoken term detection method, and the design of a community-based transcription activity.

Chapter 4 contains the discussion and conclusion of this thesis. This chapter explores how this thesis responds to the research questions around the three poles of **Performance**, **Comprehension**, and **Engagement**. It also includes the limitations and opportunities for future work.

Chapter 2

Literature Review

2.1 Introduction

This chapter provides a review of the literature that addresses the main topics covered by this project and describes the gap in knowledge at the intersection of NLP and sustainable community-based language development.

In Section 2.2, I present how past research practices have been undertaken in Indigenous populations, detailing how these past practices have potentially been harmful to Indigenous communities. However, an emerging community of research practice, undertaken in collaboration with Indigenous populations, has been trying to undertake research, decolonising practices.

In Section 2.3, I start by clarifying the standard labels put onto languages, thus justifying the use of specific terms that establish the scope of this project. I discuss the usage of standard terms often used in the NLP literature, such as 'endangered' or 'low-resource' languages, and how they cover a large range of contexts. I provide some key concepts related to documentary linguistics and detail how common practices have inadequately responded to the language communities' needs. I then describe the key concepts of a recently published language development model that enables the revitalisation of a language according to its level of vitality and the communities's aspirations.

In Section 2.4, I review existing projects at the intersection of Human-Computer

Interaction (HCI) and community-based research. I then present existing projects regarding technologies that support language work in remote Indigenous communities, highlighting the gap in research regarding speech recognition, language preservation and community-based design.

Finally, in Section 2.5.1, I present an overview of the existing research regarding speech recognition for Indigenous languages, which are often labelled as low-resource languages. I present traditional methods and describe teh ways in which they are inadequate for Indigenous contexts. I then present alternative speech recognition technologies and the sparse transcription model for computer-assisted speech transcription.

2.2 Community oriented research

"Researchers are like mosquitoes; they suck your blood and leave.", said a Native Alaskan in Cochran et al. (2008). Indigenous people have been the subject of research across multiple fields, often with the purpose of enhancing western knowledge about the Indigenous world. Research in Indigenous lands has a history of unethical and colonial practices, where researchers collected Indigenous resources without providing any clear benefits to local communities. In this section, I detail the history of western research methods in Indigenous spaces, and discuss how recent calls for decolonial practices have attempted to address the injustices inflicted upon these communities.

2.2.1 Research practices in Indigenous contexts

Following the British invasion of Australia in 1788, British researchers began studying Aboriginal languages. Initial reports described these languages as primitive forms of communication (Dixon, 2011). Such discourses were rooted in an idea of racial superiority that allowed scientists to determine what knowledge was legitimate (Foley, 2003). This led to the elimination of Indigenous cultural science, social systems, and traditions (Rigney, 2001).

Research projects in Indigenous spaces have often followed a top-down approach, with methods designed without consultation with the target community, and the methods and results following western epistemology (Singer et al., 2015). These practices follow colonial methods, where western theories could either be tested in a new context, or western methods could be applied to new populations for the well-being of local populations (Samarin, 1984).

Research is obligated to cause no harm, yet, even today, research in Indigenous spaces has been causing distress to Indigenous populations due to unsuitable practices (Cochran et al., 2008; Galla, 2018). While it seems evident that ethical research should have principles of non-maleficence (Schwartz, 2022), respectful behavior and taboos may not be obvious to an inexperienced researcher. For example, during the Covid-19 pandemic, the same regulations were implemented worldwide to limit the spread of the virus (i.e., washing hands, social distancing, testing). However, Wanambi et al. (2021) (cited in Bird (2022)) explained that such regulations bypassed local authorities. Regulations were imposed on individuals instead of educating the local population and letting them handle the situation their way. From a western point of view, there was no desire to harm; regulations were put in place to limit the spread of the virus and to protect vulnerable populations. From the Australian Indigenous point of view, the local governance systems were ignored in favor of colonial practices, where rules from outsiders were imposed on local populations who had ways of living that conflicted with the regulations (Murphy, 2020). Local authorities, existing bodies of knowledge, and social situations were ignored. Western thinking tend to be hegemonic and have a universal and supposedly objective view of reality (Jullien, 2021, p. 156). Indigenous populations have built an approach to science that has generally been ignored or disregarded by western scholars (Cochran et al., 2008).

the adoption of western approaches in Indigenous spaces is problematic, as this ignores Indigenous bodies of knowledge and ways of knowing. Indigenous populations have expertise in their language, culture, land (e.g. Cochran et al., 2008) or medicine (e.g. Oliver, 2013). Western practices tended to impose their approach to

science which disregarded Indigenous knowledge (Cochran et al., 2008). Research is also culturally biased, and science has often constructed a reality from Indigenous input embedded into a western discourse made by non-Indigenous participants for a non-Indigenous audience (Budby, 2001). Across the world, we can find many different epistemologies where western approaches have not always been acceptable for Indigenous populations (Descola, 2005). Western research methods categorise, and each concept must be classified under a fixed label, contrasting with Australian Indigenous approaches to knowledge, which have been described as relational (Foley, 2003; West, 1998). Becoming familiar with ways of knowing, bodies of knowledge and research practices in Indigenous communities is an ethical obligation when engaging with Indigenous communities (Schwartz, 2022).

Where western methods have failed, participatory research has started to take place, fully involving the community, and *both-ways learning* has been adopted (Haynes et al., 2019). In doing so, the local communities' participation enables a better comprehension of a given problem and the co-design of research approaches. Engagement with Indigenous populations is common across many fields, and the challenges linked to this kind of interaction have been the topic of some research contributions (e.g. Christie, 2013a; Baskin, 2006). Initially, data collection methods were bound to an acknowledgement of cultural differences that led to ad-hoc processes (Crowley, 2007), where different approaches would be explored depending on the target community, its history, and its location (Hanke, 2017). Deploying western methods in Indigenous spaces has often led to negative responses. For instance, the typical interview setting in which participants must respond to a set of questions has been described as irritating (Maar et al., 2011; Ober, 2017). Ways of transmitting knowledge in Indigenous Australia are based on storytelling (Foley, 2003). Accordingly research methods should be built that consider epistemological differences. Yarning, for instance, has been described as "a conversational process that involves listening to storytelling that creates new knowledge and understanding" (Terare and Rawsthorne, 2020). Yarning is an Australian Indigenous research method that has been applied in

many projects to enable more ethical research and respectful engagement (Fredericks et al., 2011; Rodríguez Louro and Collard, 2021; Walker et al., 2014).

2.2.2 Decolonising methods

> The most significant impact of insensitive research is the perpetuation of the myth that indigenous people represent a 'problem' to be solved and that they are passive 'objects' that require assistance from external experts.
>
> Smith (1999) as cited in Cochran et al. (2008)

The statement made by the author about public health can be applied to other fields, such as Natural Language Processing. In the zero-resource paradigm of NLP, the scientist does not require the expertise of a linguist or intervention from the language community. A fully automatic solution, applied to sanitised data, is the preferred approach for preserving languages (Bird, 2020a). It is important to note that research can be culturally biased. When conducting research on Indigenous topics without involving relevant Indigenous communities, the aim is only to create knowledge for a non-Indigenous audience and promote the scholar's career. This approach does not contribute to building social justice or bringing concrete community benefits.

The NHMRC Guidelines for Ethical Conduct in research with Aboriginal and Torres Strait Islander Peoples and Communities include equity and reciprocity as essential research values. Both these values emphasise the importance of ensuring a "reasonable distribution of benefit for Aboriginal and Torres Strait Islander people and communities". The guidelines also acknowledge that some benefits may not relate directly to the research project at hand (p. 6 *NHMRC* n.d.). Therefore, the trend has been to return the outcome of the research project to the community. However, reciprocity also means that the community has the right to define what "benefits" mean according to their values. To achieve this, the researcher needs to provide extra

work and negotiate with the community to align the research project's outcome with their expectations (Chelliah and De Reuse, 2010).

There is a need for consultation and strong community participation that acknowledges Indigenous ways of knowing, thereby enabling an ethical research approach (Cochran et al., 2008; Lignos et al., 2022). Reciprocity is often interpreted as remunerating participants. Lavallée (2009) notes that adopting Indigenous methods empowers Indigenous people and advances Indigenous ways of knowing. However, research involving Indigenous populations should also improve peoples' lives and provide justice to communities. Ford et al. (2021) assert that ethics committees emphasise the need for research to uniformly strengthen communities through the project outcome such that the benefits of the project are not limited to the researcher or specific stakeholders. For instance, reciprocity can take the form of cultural preservation (Ford et al., 2021), economic empowerment (Winschiers-Theophilus et al., 2020), or health improvement (VanderBurgh et al., 2014).

Participation alone is not enough. Recently, various fields have called for the decolonisation of current research practices (Bird, 2020a) and a move towards self-agency (Smith, 1999, p.204). Enabling self-determination begins with building relationships. There is a long history of superficial interactions between researchers and Indigenous participants, where the relationship is limited to data collection that only serves to enhance the researcher's career (Cochran et al., 2008). Relationships cannot be superficial (Bird, 2020a), and a climate of trust needs to be established. Relationship building progressively leads to the creation of a cross-cultural space where the priorities of locals and researchers can be discussed and understood (Lyons, 2011). Shared tasks can then be designed to support both.

2.3 Languages and documentation

NLP scientists have often supported linguists in justifying their work (e.g. Adams et al., 2017; Foley et al., 2018a; Godard et al., 2018b; Godard et al., 2018a). It is

believed that designing language technology can assist linguists in saving languages. However, the goals of linguists do not necessarily align with the goal of preserving languages (Hanke, 2017; Nathan and Fang, 2013). Different languages have varying official statuses, amounts of resources, and vitality levels. When discussing language technologies, the term 'low-resource' is often used, which may not provide sufficient information. In this section, I will explore methods of classifying languages that provide more clarity on various objectives of NLP. I will then examine the role of language documentation in language development.

2.3.1 Language classification

Research in NLP has used the term "low resource" for languages with insufficient data to build statistical models (in our case, speech-to-text, though it is also valid for other NLP topics). The term is used through several variants such as relatively low-resource (Xu et al., 2021), "very low-resource", (Khare et al., 2021), "extremely low-resource" (Xu et al., 2020) and "zero resource" (Kumar et al., 2022). All these terms refer to the amount of annotated language data. None of these terms have a clear definition (Liu et al., 2022a), even "zero resource", for which the notion of resource is not always clearly defined. In many cases, the simple "low-resource" label is used to describe multiple different realities. Experiments are described using datasets from 40 hours of training speech to only 12 minutes (Khare et al., 2021; Pulugundla et al., 2018; Rosenberg et al., 2017; Saeb et al., 2017; Westhuizen et al., 2022). The issue is that *low-resource* is used in opposition to *high-resource*, a term that covers only standardised dominant languages (Bird, 2022). From an NLP perspective, only the data matters and *low-resource* is used when the data is scarce, which can even lead to strategies to simulate this kind of scenario by sub-sampling English corpora (e.g. (Scharenborg et al., 2017)). Details about the type of speech, the target languages, the size of the vocabulary, and complex socio-political situations are often avoided (Bird, 2022).

Several attempts at language classification from an NLP perspective have been observed. For instance, Berment (2004) presents a method to assess the level of computerisation for a language based on digitised resources (dictionaries, ASR models, translation models, etc.), attributing a score in consultation with the speech community. A more recent classification has been proposed by Joshi et al. (2020), who organise the world's languages into six categories from "The Left-behind" to "The Winners" based on the labelled and unlabelled resources available. Besides the elitist aspect of the labels used to classify languages, this classification is symptomatic of an inclination only to consider data in its written form, where languages traditionally spoken with a few or non-written resources are the losers of the "race". Yet, it opens a door to considering the potential offered by advances in machine learning, specifically with the emergence of massively multilingual models that leverage unlabelled data, which has so far been ignored because of the prevalence of supervised learning techniques.

Outside of the simple NLP views on languages, the UNESCO (2011) classification is considered a reference for assessing the level of vitality of a language. This classification is structured around six categories, ranging from "safe" to "extinct". Languages are placed in this classification based on various criteria, such as the number of speakers, the amount of documentation, or their socio-political positions. While the diversity of criteria used for the classification is relevant, the labels provide little information beyond the level of endangerment. The most informative scale proposed so far, and the one used by the *Ethnologue* database (*Ethnologue, Languages of the world* n.d.), is the Expanded Graded Intergenerational Disruption Scale (EGIDS; Lewis and Simons (2016)). This classification extends the GIDS, which was designed by Fishman (1991), and is structured around 13 levels ranging from "International" to "Dormant". A detailed description of the classification process considers the level of institutionalisation, the evolution of the speaker (both first language and second language), and the intergenerational language transmission (whether it is disrupted or not).

There are four sustainable language use levels associated with this classification.

Firstly, "Sustainable Literacy" (EGIDS 4) is when "the language is in vigorous use, with standardisation and literature being sustained through a widespread system of institutionally supported education." "Sustainable Orality" (EGIDS 6a) is when "the language is used for face-to-face communication by all generations, and the situation is sustainable." "Sustainable Identity" (EGIDS 9) is when "the language serves as a reminder of heritage identity for an ethnic community, but no one has more than symbolic proficiency." Finally, "Sustainable History" (EGIDS 10) is when "the language has been adequately documented, and the documentary materials are safely and reliably archived for future access.

The EGIDS is used to assess the level of vitality of a language, but it also provides informative categories for the computational agenda that can be applied to different languages. For instance, category 0 ("International languages") refers to languages like English, French, and Spanish, where state-of-the-art NLP techniques can be applied due to the massive amount of data available for supervised learning (e.g. Le et al., 2020; Simoulin and Crabbé, 2021). Categories 1 and 2, respectively "National" and "Provincial" languages, include languages like Latvian, Slovenian, Marathi, Bengali, etc., which do not have as much data as international languages but still have a decent amount for supervised learning (e.g. Ishmam and Sharmin, 2019; Znotins and Barzdins, 2020). These languages also have a large number of unlabelled resources, including spoken and written materials from various sources like social media, radio broadcasts, and literature, which can be used for corpus collections (e.g. Abraham et al., 2020; Salimbajevs, 2018).

Levels 3, 4, and 5, respectively "wider communication", "educational", and "developing" languages, begin to fall into the low-resource spectrum, where there is little available data, both written and spoken. Nevertheless, sustainable literacy is achieved, and a large proportion of the speech community has expertise in the language. ASR models can still be trained, but techniques are used to cope with the lack of data (detailed in Section 2.5.1; Gauthier et al. (2016a), Gupta and Boulianne (2020a), and Shi et al. (2021)).

From EGIDS 6a to 7, sustainable literacy is not achieved; very little data is collected, but the speech community still has expertise in the language. The previous categories have some institutional support and have the potential to be progressively brought towards a higher level in the EGIDS scale. For lower levels, where the community is still strong, different agendas are set up and grounded in local logic. The documentation schedule needs to be adapted to local expectations and ways of knowing (detailed in Section 2.4; Kuhn et al. (2020) and Lane and Bird (2020)). In lower levels, the language is nearly or completely extinct. This is where language documentation done by external experts makes the most sense, as only a few resources and fluent speakers are available. In some cases, the community still exists and conserves a sense of identity with some manifestation of the language (e.g. code-switching).

While this last classification is more informative, the research associated with each category still uses the term "low-resource," which is not informative. It creates an unwanted gap in the literature where a given computational method covers languages in one EGIDS category but is not relevant elsewhere and not properly suited.

“Les taxinomies [...] relèvent de l’arbitraire d’une raison qui prétend normer la souplesse du vivant” (Classifications [...] are arbitrary and made of rationale that pretends to standardise the flexibility of the living.”) (Hüe, 2010)). No classification is perfect, and different realities can be observed within the same language. The Breton language (ISO-bre) spoken in western Brittany (France) is a great example. According to the Ethnologue database, Breton is classified at level 8a on the EGIDS scale (Moribund),[1] meaning that the only remaining speakers are from the grandparent generation. I was born in Brittany and lived there most of my life. My grandfather’s first language was Breton, and his parents did not speak French. However, the language has not been transmitted to either my father or me. We match the EGIDS classification where intergenerational language transmission has been broken. There is no sustainable orality or literacy, and only a sense of identity remains. Yet, Breton

[1] https://www.ethnologue.com/language/bre

has a written tradition, with the first known dictionary printed in 1499 (Abalain, 2000). It is recognised as a regional language by national authorities, (*Légifrance. Le service publique de la diffusion du droit* n.d.) and it is used as a language of education through the Diwan schools (which offer bilingual education in French and Breton; *Rouedad Skolioù Diwan* (n.d.)) and for the national high-school exam. It also has a media representation with Breton-language radio (*Radio Kerne* n.d.) and television (*Tébéo. Télé Bretagne Ouest* n.d.). With these elements, Breton should be at the fifth level of the EGIDS scale, where

> there are adequate materials in this language to support literacy instruction in the language, and some members of the community are successfully learning to read and write about some bodies of knowledge in the language.
>
> Lewis and Simons (2016, p.174)

Thus, this is in a situation of multi-polar language practice, where different vitality levels co-exist in the same reality.

Ultimately, different phrasings are used to address different languages depending on the perspective from which we see them. The EGIDS scale seems to be the broadest since it is informative in terms of vitality and socio-political situation, and by extension, gives an estimation of the type of processes that can be transferred from one language to another, from an NLP perspective. However, the lack of data is a central aspect in NLP-related topics and the use of one computational method instead of another strongly depends on the amount of data available. For that reason, the term low-resource will still be used throughout this thesis.

The focus will, however, be put on oral languages, - those languages for which sustainable literacy is not achieved and where sustainable orality is achieved to some extent.

2.3.2 Language documentation

Natural Language Processing researchers often justify their work on low-resource languages by mentioning the need for world languages to be documented before extinction, and explaining how their methods can support the linguist. The role of the linguist is often two-fold: on one side, documentary linguistics concerns the collection of language data (Himmelmann, 1998). On the other side, descriptive linguistics concerns the analysis of the primary data of the language (Himmelmann, 1998; Woodbury, 2003). The two disciplines communicate with each other since the documentation serves the interest of the description (Chelliah and De Reuse, 2010). However, language documentation per se does not focus on preserving language use, but rather focuses on creating data for other purposes (Boas, 2006; Nettle and Romaine, 2000).

Documentary linguistics is a time and labour-consuming process that includes going to the language community to gather raw speech data and a phase of annotation of the data collected (Crowley, 2007; Hanke, 2017). The annotation phase includes, but is not limited to, transcription, part-of-speech tagging, translation and metadata annotation (language, speakers, content etc; Thieberger et al. (2016)). However, the processes are not systematically applied (Gawne et al., 2017). In computer science, researchers are often reluctant to share their code (Moraila et al., 2014) possibly because it is unusable or not adequately commented. Revising a script, like formatting a corpus, is labour extensive and language collections are not systematically made available, probably for the same reason. There are efforts to facilitate the dissemination of language data, such as online archives (e.g. Bird and Simons, 2001). From here, new limitations arise. Firstly, as mentioned before, the creation of resources often follows a descriptive linguistics agenda, whereby the written forms become the data (Ochs, 1979) and the quality of the speech recorded is not a priority as long as it is intelligible. Furthermore, transcribers are often selective in their work (Ochs, 1979). Depending on their goals, only parts of the speech collection are annotated. Collecting

data from Indigenous participants might imply the restriction of data either between the participants and the linguist, or to the institution to which the project is linked (e.g. San et al., 2022).

As previously mentioned, the primary goal of linguists in academia is to describe languages and, by extension, collect the data necessary for this purpose (Hanke, 2017). The linguist's work focuses only on language description and analysis and does not necessarily align with the priorities of the local speech community (Dorian, 2010; Hanke, 2017). Extra work is often required to bridge the gap between the scholar's agenda and the community's needs (Chelliah and De Reuse, 2010). This usually involves a single linguist going to the language community to conduct research (Chelliah and De Reuse, 2010; Crowley, 2007), which creates a bottleneck as there are often more people in a language community than a single linguist can work with (Hanke, 2017). Moreover, First Languages Australia (2014) has pointed out the lack of collaborative methods for language documentation, with the community only involved in the recording step of the workflow.

On the one hand, a language cannot be revitalised or preserved without the contribution of the community (Besacier et al., 2014). On the other hand, regarding these notions, we see that language is too far removed from the concrete daily issues of life (Lewis and Simons, 2016). In their book on sustainable language use, Lewis and Simons (2016) discuss language development instead of language documentation. The idea is to first demonstrate to the community how language can help meet their needs in all areas of life, in line with its level of vitality and in collaboration with the community. The most common case is revitalising or preserving intergenerational language transmission, but there is no stopping point. Language development can also involve the creation of new language functions and the expansion of people's capabilities (Hinton, 2015). This could mean, for entirely oral languages, approaching sustainable literacy through the deployment of language use spaces dedicated to this purpose.

2.4 Technology design and Indigenous communities

Human-Computer Interaction (HCI) is a field concerned with the "design, evaluation, and implementation of interactive computing systems" (Hewett et al., 1992). While this project is not grounded in the HCI spirit, most of the contributions made at the intersection of computer science and cross-cultural research have been based on HCI. Gessler (2022) argues that better communication needs to occur between NLP scientists and linguists to enable better software for documentary linguistics. However, this statement is grounded in a western-centric view, where western NLP scientists and western linguists need to find an agreement to collect Indigenous resources more easily. In this section, I will explore projects involving the design of technology in cross-cultural contexts and the design of technologies around language work.

2.4.1 Cross cultural design

Research projects involving technology design and cross-cultural collaboration have taken two directions in two different fields: NLP and HCI. Research in NLP in this space is relatively small and has attempted to explore how we can design tools to address the lack of data usually mentioned in projects on data-restricted languages. HCI projects have taken cultural differences into consideration when designing tools and have worked their way through challenges related to cross-cultural design to design community-based technologies.

On the NLP side, the focus has been on the linguist's workflow and the collection of language resources. Institutional support and media coverage give some languages online representation that can be used for data collection purposes (e.g. Abraham et al., 2020; Salimbajevs, 2018). High numbers of speakers also increase the potential for creation of more resources. Smaller communities, don't have the same number of human resources, which limits the amount of data available. From there, NLP researchers initiated the creation of tools to facilitate the creation of language data with Indigenous populations.

The idea was to design software that was easy to use without formal training. A trend that has begun in recent years is to take advantage of the widespread use of mobile devices to collect linguistic data (Bettinson and Bird, 2017; Bird et al., 2014; De Vries et al., 2014; Gauthier et al., 2016c; Moeller, 2014). The purpose is to enable large-scale speech collection by speech communities to document the world's languages. This aligns with the usual linguist workflow where traditional field linguistics data needs to be gathered (recording, transcription, translation, and metadata). There is a concept of universality where the same tool can be used in any language with minimal training (e.g. Moeller, 2014). These projects also considered possible work in remote locations where internet access is not guaranteed, enabling offline usage (e.g. Bettinson and Bird, 2017).

More recent research actions have started to consider community engagement and relationships, specifically in pandemic periods where trips to remote protected locations have been compromised. For instance, the 'lingobox' project enables remote interaction between scholars and Indigenous participants, pursuing the idea of universal design (Bettinson and Bird, 2021). The design of these tools has been thought to take place in cross-cultural interaction to be used by both western scholars and Indigenous participants.

Natural Language Processing research is data-centric, and contributions with a direct influence on other NLP projects are usually valued (better models, methods or corpora). Accordingly, projects that enable data collections are often based on theoretical frameworks that address the bottleneck situation (Himmelmann, 1998) and traditional NLP expectations, such as collecting data to train better models (e.g. Gauthier et al., 2016b). Conversely, Human-Computer Interaction (HCI) projects often go through a testing process, and the contribution is translated into the adoption of a design by a particular cohort or population. In other terms, NLP focuses on the data while HCI focuses on the people.

The introduction of technologies in Indigenous spaces has been described as a double-edged sword (Galla, 2018). While technologies can support language

revitalisation strategies and language documentation (Ogie, 2010; Ovide and García-Peñalvo, 2016) or teaching and learning (Ndulue and Orji, 2021; Wigglesworth et al., 2021; Zaman et al., 2015), new technology can also appear as social pressure for speakers of minority languages (Lackaff and Moner, 2016) or even cause harm (Verran, 2007).

Working with small communities implies cross-cultural collaboration. Therefore, in many cases, participatory design (or co-design) has been preferred to enable a "culturally tailored, culturally enriched, and trustworthy environment for participation" (Peters et al., 2018). Such an approach allows, on the one hand, the limitation of the risk of harm inflicted on the community through misconceived design, and, on the other, enabling more sustainable projects through the full involvement of Indigenous users in the design process.

Co-design methods have been used in a variety of projects in Indigenous spaces, including community initiatives where locals were the users and co-designers of the products created (Hardy et al., 2016; Leong et al., 2019; Soro et al., 2017). In these projects, the relationship with Indigenous stakeholders was often assumed or already existed prior to the project, but many details about the challenges of cross-cultural collaboration and methods to facilitate the discussion were not provided. Projects were often introduced from a problem or work opportunity raised by an Indigenous population, which led to collaborative design, considering cultural requirements. However, due to cultural differences, many Indigenous communities do not have a long history of using digital devices or may have different ways of knowing and sharing information, making collaboration for the creation of technologies challenging. Only a few projects have provided methodological strategies to facilitate engagement with Indigenous participants. For instance, Zaman et al. (2015) explored the use of sketching to induce design ideas. Another example of co-design and community engagement was the "crocodile friend" project built alongside the Kukku Yalanji community in northern Queensland, Australia (Taylor et al., 2020). This project described the "Design Non-Proposal" method, an engagement method that involves presenting a

portfolio of existing technologies to show participants the field of possibilities.

While cross-cultural collaboration is challenging, building a design with machine learning or high performance computing at its core adds a level of complexity. It has only been a few years since the issues regarding the lack of accessibility between machine learning and user experience were raised (Loi et al., 2019). While some research has addressed explainable AI (XAI) for more comprehensible model visualisation (Gunning, 2017; Hohman et al., 2019; Mohseni et al., 2021), these models are usually designed for AI scientists (Miller et al., 2017), and few efforts are made to extend these designs to a non-technical public (e.g. Benjamin et al., 2020). Increased comprehension of the mechanisms behind machine learning models could help the application of these models in real-life situations (e.g. Ayobi et al., 2021).

2.4.2 Technologies and language work

Research into minority languages follows two different trends. We aim to (1) document languages to record them before they become extinct or to prevent their loss, or (2) develop learning software with a language preservation or revitalisation perspective. NLP pursues these two agendas through the creation of sub-tasks. The predominant trend is to create processes that generalise across languages, taking advantage of the available data from other languages (e.g. Eskander et al., 2022; Gupta, 2022; Lee et al., 2022; Li et al., 2020a). While less common, language-specific projects are emerging to address classic NLP tasks. These include the creation of language-specific Part-Of-Speech (POS) taggers (Finn et al., 2022; Lamb and Danso, 2014), morphological modelling (Lane and Bird, 2019; Pugh and Tyers, 2021; Schwartz et al., 2019; Tyers et al., 2019), syntactic parsing (Dyer, 2022; Zhang et al., 2022), machine translation (Góngora et al., 2022; Nicolai et al., 2021), and speech synthesis (Harrigan et al., 2019; Pine et al., 2022b). Recent universal solutions have also been used to solve language-specific problems (e.g. Siminyu et al., 2021). This raises the question of the extent to which these projects are translated into community-based projects.

A limited range of tools has dominated the field of documentary linguistics. Among the most used, are ELAN (Sloetjes et al., 2013), a multimodal annotation software, and Praat, a speech annotation and processing software. Both of these tools contain computational components that support annotation (such as the automatic detection of silences [Praat]) and allow post-processing (e.g. language-specific POS tagging [ELAN]). While the complexity of these tools has been the subject of much criticism, responses have been provided through the creation of simpleELAN, for instance, or tools that are better adapted for fild linguistics (cf. Section 2.4.1). The NLP processes contained in these tools remain very generic. The tools are designed to be robust in any language, which makes language-specific solutions limited to a few high-resource languages.

Efforts to translate speech recognition models for language documentation applications have been limited. There is a common assumption that speech recognition is probably not accurate enough to benefit oral languages (Section 2.5.1). The ELPIS software (Foley et al., 2018b) is the main example, which encapsulates the Kaldi pipeline (Povey et al., 2011) in a user-friendly interface. However, there is a paradox in that a certain amount of expertise is required to run a speech recognition pipeline, particularly in terms of data requirements, and people with such expertise are probably already familiar with Kaldi. To my knowledge, speech recognition for language documentation has been attempted using a two-stage process whereby a system is trained, and, subsequently, the output is post-edited by language experts (Shi et al., 2021). Speech recognition is language-specific, so the construction of a universal tool that works for any language is unlikely to occur. The best option would be to insert pre-trained or fine-tuned models into a language-specific transcription app. For instance, Lane and Bird (2020) developed a transcription interface following the sparse transcription model (Bird, 2020b), which incorporated voice activity detection, phone recognition and word discovery to support manual transcription in Kunwinjku. Both this and the ELPIS project have potential for real-world applications. However, as with much NLP research, a solution is provided without a real-life implementation.

The potential of a given method is described without involving or considering the potential users. Consequently, whether these tools can be adopted by real users is uncertain.

Several researchers have been exploring the intersection between speech technology and language documentation through a decolonial framework (Prud'hommeaux et al., 2021; Liu et al., 2022b). Engagement with the speech community for language-related software is, however, more often considered for language learning software from a revitalisation perspective. Regarding language technologies and community engagement, noteworthy projects include those focused on speech synthesis for language acquisition in a Canadian context (Pine et al., 2022b), as well as the development of linguistic resources (Tyers and Henderson, 2021; Schwartz et al., 2021).

Computer Assisted Language Learning (CALL) refers to any instance in which a computer is used to facilitate language learning (Stickler and Shi, 2016). In such a context, success is usually defined as a person's ability to enhance their skills in the target language, such as pragmatic, creative, and lexical acquisition (e.g. Culbertson et al., 2017; Rankin and Edwards, 2017; Zhang et al., 2018). People are at the core of the evaluation, and engagement with the target community is therefore mandatory. The design of language learning software is a small but committed area of research in which careful attention is given to cultural components to ensure the design is relevant to the target users (Bontogon et al., 2018; Xu et al., 2022), and where co-design approaches are even more relevant (e.g. Hardy et al., 2016). In response to the lack of flexibility of language learning software using a predefined set of questions, answers, and exercises, research in ICALL (Intelligent Computer Assisted Language Learning) has tried to incorporate language technology into software design. For instance, Bontogon et al. (2018) and Kazantseva et al. (2018) have used Final State Transducers to facilitate the learning of polysynthetic structures in Plains Cree and Kanyen'héha, respectively. Xu et al. (2022) have used a POS tagger in their Irish Gaelic learning software to facilitate error induction. The language processing component is not used as an end but instead to facilitate the primary goal of the design: learning the language.

While supporting the user's goal and enabling language preservation, collecting data is not part of the design. However, combining gamification and data collection has been used for high-resource languages, where user responses are used for research projects (Lafourcade, 2007; McNaught and Lam, 2010). Introducing such methods in Indigenous spaces could respond to the wishes of the speech communities and collect the data needed to enhance language technology in a virtuous cycle where more data leads to better design performance.

2.5 Speech Recognition

In this section, I explore the range of speech recognition technologies and their applicability to oral languages. I first detail the main traditional Automatic Speech Recognition (ASR) methods and the ways in which they are not adequately suited for oral languages. I then briefly explore the potential of phone recognition and finish by presenting the latest research around spoken term detection and its potential for transcription.

2.5.1 Traditional automatic speech recognition

Traditional ASR pipelines are based on Hidden Markov Models (HMMs), which rely on a large amount of speech data aligned with an orthographic transcription and a pronunciation lexicon. The required data includes a large vocabulary, clear speech, and a wide range of speakers. The outcomes of this pipeline are an acoustic model that associates speech features (discussed in Section 2.5.3) with linguistic units (commonly phones), and a language model that represents the likelihood of a given sequence of words (usually based on n-grams). Toolkits such as Sphinx (Lee et al., 1990) or Kaldi (Povey et al., 2011) have facilitated this process, allowing for the creation of models based on Gaussian Mixture Models (HMM-GMM) or hybrid models based on Deep Neural Networks (HMM-DNN). These models are progressively being replaced by

end-to-end architectures based on transformers, bypassing the need for a pronunciation lexicon (Watanabe et al., 2018).

As mentioned earlier (in Section 2.3), creating such models requires a large amount of data that is often not available for oral languages. A branch of the speech recognition literature has addressed modelling for languages with limited data, often referred to as low-resource.

Most papers that apply an ASR pipeline on low-resource languages have at least ten hours of training speech. The lack of data for such languages has led to alternative strategies. For example, some projects have taken advantage of the resources available in better-resourced languages (Anoop et al., 2021; Baevski et al., 2021; Gupta and Boulianne, 2020a). The use of self-supervised speech features extracted from models trained with high-resource languages has also enabled the reduction of training data (Baevski et al., 2021). Other projects have used voice distortion methods for data augmentation (Matsuura et al., 2020; Thai et al., 2019) or the use of data in closely related languages (Juan et al., 2015; Samson Juan et al., 2014). While probably outperformed by multilingual pretraining, strategies have been adopted to cope with speaker variability such as Vocal Tract Length Normalisation (VTLN; Tüske et al. (2014)) or the feature-space maximum likelihood regression speaker adaptation method (fMLLR; Gales (1998) and Gauthier et al. (2016b)). In terms of general system architectures, systems for low-resource languages tend to follow the example of systems for higher resource languages by deploying HMM-GMM and HMM-DNN (Gauthier et al., 2016b), Bi-directional Long Short Term Memory (BiLSTM; Pulugundla et al. (2018)), or end-to-end (Shetty and NJ, 2020; Shi et al., 2021).

However, in the case of limited data ASR, there doesn't seem to be a perfect architecture that fits all contexts (Morris et al., 2021), and methods that have proven effective with some data can result in high Word Error Rates (WER) in other contexts (e.g. Gupta and Boulianne, 2020a; Gupta and Boulianne, 2020b).

There are a variety of different performances for ASR applied to low-resource languages that are not informative. While the size of the datasets in high-resource

settings erases many biases in the final models, low-resource settings do not have that chance. The size of the lexicon, the typology of the language, the number of speakers, the mode of recording or the noise are parameters to take into consideration while evaluating a system. While experimenting on the low-resource spectrum, the information related to the data is often limited to the duration, the number of utterances and sometimes the number of speakers. While great progress is being made to address the lack of data for acoustic modelling, no robust ASR system is possible without a large vocabulary if a transcription at the word level is expected.

2.5.2 Phone recognition

Phone recognition methods have been applied to spoken languages, with system architectures providing a low phone error rate (PER) (Adams et al., 2018; Michaud et al., 2018). Increasingly better performances have been achieved, with systems relying on multilingual training where less and less data is required (Li et al., 2019; Li et al., 2020a; Li et al., 2020b; Thompson et al., 2019). The justifications behind the use of a phone recogniser include bypassing the need for a pronunciation lexicon and a language model. Automatic phone transcription can provide a canvas transcription for manual correction (Adams et al., 2018; Michaud et al., 2018). The linguist can use the phone transcript to then provide a corrected phone transcription and the associated translation (e.g. Michaud et al., 2018). Such methods have been explored to support linguists as some training is required to be able to access and produce phonemic transcripts and the final products serve primarily scholars' interests.

Most spoken languages only have a recently developped orthography, often conceived by linguists, based on language phonology. Recent frameworks have attempted to enable transliteration from phones to graphs, and vice versa, using rule-based algorithms (Pine et al., 2022b) or more complex language-specific statistical models (Mortensen et al., 2018). Combining universal phone recognition and phone-to-graph mapping offers opportunities for under-resourced languages that have barely been

explored (e.g. Leong and Whitenack, 2022).

2.5.3 Spoken term detection

Spoken term detection is a sub-category of ASR which consists of the search and detection of isolated terms in a speech collection. Spoken term detection is traditionally used for high-resource languages to enable the detection of keywords in noisy speech collection where traditional ASR would provide high error rates.

In evaluation campaigns for spoken term detection systems, the standard architectures usually have three components: an indexing component that relies on traditional ASR a detector and a decision-maker subsystem (Tejedor et al., 2019). The ASR subsystem outputs word lattices stored as indexes. The detector searches for potential hits in the index, and the decision-maker will then filters out the hits based on a confidence measure.

As we have seen in Section 2.5.1, traditional ASR is hardly applicable to oral languages both because of the lack of aligned speech and text when training an acoustic model and the lack of textual data when training a language model. For this reason, research around spoken term detection for under-resourced languages has mainly focused on ASR-free methods. Such methods have an advantage over traditional spoken term detection systems in that they are not restricted by language models and rely purely on acoustic features.

We can find two distinct research directions for spoken term detection for under-resourced languages. One focuses on search algorithms, the other on the exploration of speech features. Research in this area also mostly focuses on query-by-example (QbE) approaches, the detection of terms from a spoken example.

Dynamic time warping (DTW) (Sakoe and Chiba, 1978) is a method of alignment between two sequences with a non-linear time normalisation effect. It is usually used to find the optimal alignment between two-word exemplars expressed as vectors. In terms of spoken term detection, the DTW algorithm can be used as a similarity

measure between two acoustic sequences (Park and Glass, 2005). A basic approach to spot terms using DTW is to slide a query term over utterances both converted into acoustic features, with a given step. In doing so, we can use the DTW score to estimate the locations of new exemplars of the terms we try to retrieve. More complex methods based on DTW have been described. Park and Glass (2005), for instance, introduce segmental DTW where the search segments are sliced into smaller segments and then compared with the query term. Jansen and Van Durme (2011) use the Locality Sensitive Hashing and Point Location in Equal Balls algorithms to enable a pre-selection of similar terms before applying DTW.

The growing interest in neural architectures has led us to new methods for recognising terms based on statistical models. The main idea is to train models to convert terms expressed as speech features into fixed-dimension acoustic word embeddings. From there, it is possible to spot words by applying simple cosine distances between a query term and segments of audio to estimate the position of possible new exemplars of the query term.

The main research direction has been to use existing data in the target language to train a classifier based on a Siamese architecture (Mazumder et al., 2021; Settle and Livescu, 2016; Settle et al., 2017). Classifiers are usually trained from pairs of similar terms with a same-different loss function to create clusters of similar words. While some research has tried to reduce the amount of training data (Mazumder et al., 2021), a lot of pairs of terms are required, which constrains the application of such architectures. Other papers have explored unsupervised learning of acoustic word embeddings. Among the popular methods, we can find embedding extractors based on encoder decoder recurrent neural networks (Chung et al., 2016; Kamper, 2019) trained on similar speech segments in the input and output.

Spoken terms detection based on DTW has been considered as the state-of-the-art method for oral languages (San et al., 2021). From there, another path in the literature has been exploring speech features applied on DTW for spoken-term detection. Mel-Frequency Cepstral Coefficients (MFCC) is often used as a baseline. They are based

on the inverse Fourier transform of the log-spectrum (Davis and Mermelstein, 1980). A few authors have explored using AutoEncoders (AE) for feature extraction (Kamper, 2017; Menon et al., 2018; Menon et al., 2019), using a feed-forward encoder-decoder network. The model was trained with the same speech features (generally MFCC) at the input and the output and then use the representation created by the last hidden layer as final speech features.

The concept of correspondence Auto Encoder (cAE) was introduced by Kamper et al. (2015) using the same kind of architecture but trained with similar segments identified in an unsupervised fashion using the algorithm of Jansen and Van Durme (2011). Similarly, other projects have focused on multilingual Bottleneck Features (mBNF). Regular Bottleneck Features (BNF) are features extracted from the lower dimension layer of a neural network trained on acoustic features (MFCC or filterbanks). Large multilingual corpora are used to then obtain mBNF (Fer et al., 2017; Silnova et al., 2018). A combination of cAE and mBNF is the apparent the state-of-the-art in speech representation for spoken term detection purposes in almost zero-resource settings. Yet, over the past few years, a new variety of speech representation has emerged based on predictive coding (Baevski et al., 2020; Conneau et al., 2021; Schneider et al., 2019). The efficiency of such features has mostly been shown for ASR pretraining purposes, allowing a reduction of the amount of data necessary for acoustic modelling. While use of these features with QbE methods does not seem conclusive (San et al., 2021), they appear to allow great efficiency for some ASR-based spoken term detection methods in very-low resource settings (Macaire et al., 2022).

Spoken term detection alone, while responding to possible indexing problems of Indigenous languages, does not address the transcription bottleneck. The absence of a language model does not allow for dense transcriptions, transcriptions where every item is transcribed. As seen in Section 2.3.1, the documentation of most low-resource languages remains ongoing which means that the recognition of every term is not an accessible goal. Bird (2020b) introduces the sparse transcription model, a framework

which combines speech technologies and manual intervention to support transcription in an iterative process. The core idea is to take advantage of what can be automated (identification of breath groups or recognition of known lexical items) and leave those parts of the data that cannot be automatically annotated to be annotated manually.

2.6 Conclusion

Identified in this literature review, are the several key gaps and research opportunities that are addressed in this thesis.

2.6.1 Performance: real-life speech recognition

Research in speech recognition has been divided into low-resource and high-resource language domains. As seen in section 2.3, while the term "low-resource" is not informative, it leads scholars to provide a single computational solution for a large variety of languages, or to provide language-specific solutions without context of usage. Languages are often solely considered in terms of available resources, with no clear definition of what resources mean. Several different solutions can be imagined depending on the amount of labeled and unlabeled resources, the amount of written and spoken resources, and the variety of data collected. characterising a language using only the the term "low-resource" also ignores the language community's strength.

The performance of a method is often only considered in terms of scores. NLP researchers compute BLEU for automatic translation, ROUGE for automatic generation, and a combination of ROC, AUC, F-scores, precision, and recall for other processes. Due to the familiarity with the socio-political context of high-resource languages, it is easy to see the benefits of a given method at a larger scale, for instance, speech recognition for voice dictation software.

The tasks applied to these languages are often mapped to low-resource languages, ignoring their socio-political status, as well as the needs and agenda of the language communities.

Focusing on a specific category of languages (Bininj Kunwok on level 6a/b of the EGIDS scale), this thesis explores to what extent theoretical computational frameworks can be mapped to this specific context. The sparse transcription model (Bird, 2020b) proposes a new conception of language development where the language community is included. The co-design of an implementation of this model in a small language community presents a unique opportunity to show the potential of speech recognition in a community-based approach. Doing so goes outside traditional evaluation methods by considering a system evaluation in partnership with the users.

2.6.2 Comprehension: explainable natural language processing

Several research fields have decolonised their research practices by adopting Indigenous methodologies. These methods have been explored, from data collection through Yarning, for instance, to implementing public health programmes respecting the community's body of knowledge and self-determination. Human-Computer Interaction has been the leading computer science discipline to interact with Indigenous communities, focusing on designing culturally appropriate technologies.

Regarding co-design technologies with Indigenous stakeholders, very few details are given regarding the extent of their participation or the negotiation process. Yet cross-cultural design poses many challenges. As explained in Section 2.2, Indigenous communities have different ways of knowing and transmitting knowledge that can make western research methods hard to use. Consequently, working at the intersection of technology design and the Indigenous space can be challenging, and technological literacy should not always be assumed. Besides a few research methods to facilitate community engagement, such as design sketching or design-non-proposal, we are navigating in murky water. Extra challenges are added when computing or machine learning components are added, which have a complexity hardly explainable even in a western context.

There is a clear lack of research in technology design, specifically with speech

recognition components, in Indigenous spaces. Following the community-based practices, we can explore engagement methods with Indigenous communities around language documentation and technology design. To properly respond to the lack of documentation in this area, documentation of unsuccessful experiences would also be beneficial.

2.6.3 Engagement: community-based natural language processing

Speech recognition for low-resource languages has often been a self-contained field, producing research projects that only serve other research projects without concrete benefits for the language community. Models are trained to that yield a score that outperforms previous models or systems. However, once the models are created, they neither return to or present a clear benefit for the community.

The primary pretext given for the creation of speech recognition models, language documentation is bound to language description (Hanke, 2017). However, there is no clear evidence that either of these fields have an influence on preventing the loss of languages (Nathan and Fang, 2013). With a history of western scholars describing Indigenous knowledge, it is time to render unto Caesar that which is Caesar's. A lack of collaborative language documentation methods has been acknowledged in the past (First Languages Australia, 2014). the use of speech technologies for language documentation purposes in a community-based fashion could be the starting point from which to initiate a discussion about their use in Indigenous contexts. It could also be an opportunity for providing examples of decolonising methods for research around speech technologies.

Chapter 3

Research Contribution

3.1 Introduction

This chapter presents the primary contributions of the thesis, which are focused on four conference articles (see Figure 3.1). The PhD journey begins with the newly published sparse transcription model (Bird, 2020b), and its first simulation is discussed in Section 3.2. Section 3.3 describes the deployment of the model in Bininj Country and the insights gained from it. Building upon these insights, Section 3.4 aims to improve computational efficiency by proposing a new method for spoken term detection based on phone recognition. Section 3.5 is a direct response to the lessons learned in Section 3.3, employing the method developed in Section 3.4.

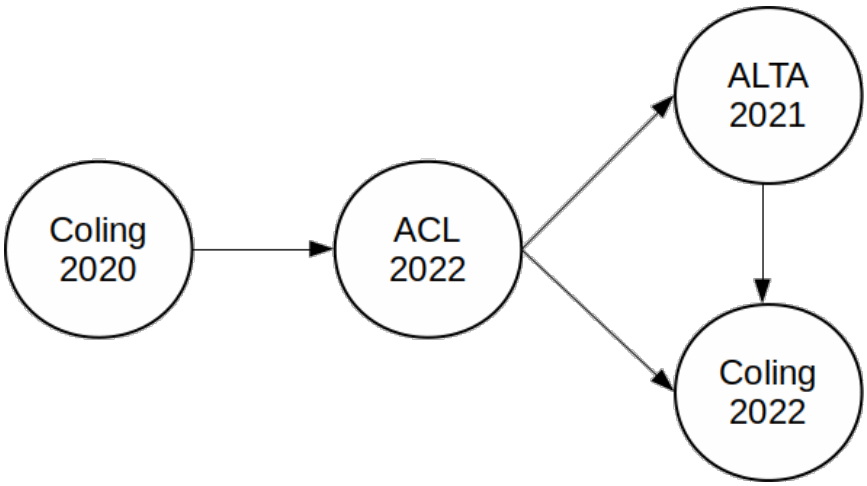


FIGURE 3.1: Concatenation of the four contributions

3.2 COLING 2020: Simulation of the sparse transcription model

Éric Le Ferrand, Steven Bird and Laurent Besacier. Enabling Interactive Transcription in an Indigenous Community. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)*. pp. 3422-3428, Barcelona, Spain

Research process

When this PhD project started in 2019, it began with two observations: although there are increasing numbers of research projects that address oral languages, none of this progress is translated into real-life applications in Indigenous northern Australia. Thus, the main motivation behind most of these projects in NLP was to support the work of linguists rather than the community agenda. As mentioned in Caselli et al., 2021, my paper was one of the only ones in the ACL anthology to mention community engagement at that time.

It has been noted that the model of a sole linguist documenting an entire language is not sustainable (Hanke, 2017). The language community is often available for such work, and people are generally willing to be involved in documentation work. However, research on community-based language documentation is significantly limited. The newly published sparse transcription model appeared to be a good starting point for enabling community participation.

In 2019, speech features based on predictive coding emerged (Schneider et al., 2019). While most of the papers discussing spoken term detection were focused on feature exploration, this new kind of feature presented an excellent opportunity to explore a community-based language transcription pipeline and its potential in our transcription setting.

Enabling Interactive Transcription in an Indigenous Community

Éric Le Ferrand,[1,2] **Steven Bird,**[1] **and Laurent Besacier**[2]
[1]Northern Institute, Charles Darwin University, Australia
[2]Laboratoire Informatique de Grenoble, Université Grenobles Alpes, France

Abstract

We propose a novel transcription workflow which combines spoken term detection and human-in-the-loop, together with a pilot experiment.

This work is grounded in an almost zero-resource scenario where only a few terms have so far been identified, involving two endangered languages. We show that in the early stages of transcription, when the available data is insufficient to train a robust ASR system, it is possible to take advantage of the transcription of a small number of isolated words in order to bootstrap the transcription of a speech collection.

1 Introduction

In remote Aboriginal communities in Australia, many efforts are made to document traditional knowledge including rock art, medicinal plants, and food practices. While it may be relatively straightforward to capture spoken content, transcription is time-consuming and has been described as a bottleneck (Brinckmann, 2009). Transcribing is often seen as an obligatory step, to facilitate access to audio. Efforts have been made to speed up this process using speech recognition systems, but the amount of data available in Indigenous language contexts is usually too limited for such methods to be effective.

Recent research has shown the efficacy of spoken term detection methods when data are scarce (Menon et al., 2018a; Menon et al., 2018b). Taking advantage of the transcription of a few words would allow us to propagate it through the speech collection and thus assist language workers in their regular transcription work. So-called "sparse transcription" would be also a way to navigate a speech collection and allow us to be selective about what needs to be transcribed (Bird, 2020). Several tools exist for manual transcription, such as Elan and Praat (Wittenburg et al., 2006; Boersma and Weenink, 1996). However such transcriptions are often made in isolation from the speech community (First Languages Australia, 2014), and so we miss out on the opportunity to take advantage of the interests and skills of local people to shape and carry out the transcription work.

We present a fieldwork pipeline which combines automatic speech processing and human expertise to support speech transcription in almost-zero resource settings. After giving the background, we detail the workflow and propose a pilot experiment on two very low-resource corpora.

2 Background

Existing approaches to automatic transcription of endangered languages involve methods that have been developed for automatic speech recognition. While a few hours of transcribed speech can be enough to train single-speaker models (Gupta and Boulianne, 2020b), speaker-independent models require a large amount of training data to produce useful transcriptions (Gupta and Boulianne, 2020a; Foley et al., 2018). Moreover, they draw language workers and speakers into the time-consuming task of exhaustive transcription, forcing them to transcribe densely, including passages that may be difficult or impossible given the early state of our knowledge of the language. A more suitable approach, we believe, involves

beginning with stretches of speech where we have the greatest confidence, and only later tackling the more difficult parts.

Spoken term detection involves retrieving a segment in a speech collection, given an example. With this method, it is possible to sparsely transcribe the corpus, i.e., take advantage of an existing transcription (or a list of spoken words) and identify tokens throughout the collection (Bird, 2020).

Dynamic Time Warping (DTW) (Sakoe and Chiba, 1978) and its more advanced versions (Park and Glass, 2005; Jansen and Van Durme, 2011) can be used to match a query in a speech collection, but they can be computationally expensive. DTW "aligns two sequences of feature vectors by warping their time axes to achieve an optimal match" (Menon et al., 2018a). A common method is to compute DTW, sliding the spoken query over each utterance in the corpus. In addition, several speech representations have been considered in the literature for the task of word-spotting. Kamper et al. (2015) and Menon et al. (2019) explore feature extraction from autoencoders (AE) and correspondence autoencoders (cAE) and show how top-down constraints can improve the quality of the hidden representation for spoken term detection tasks. Schneider et al. (2019) introduce wav2vec, a self-supervised model for speech representation learning which is based on contrastive predictive coding, and apply this to supervised ASR.

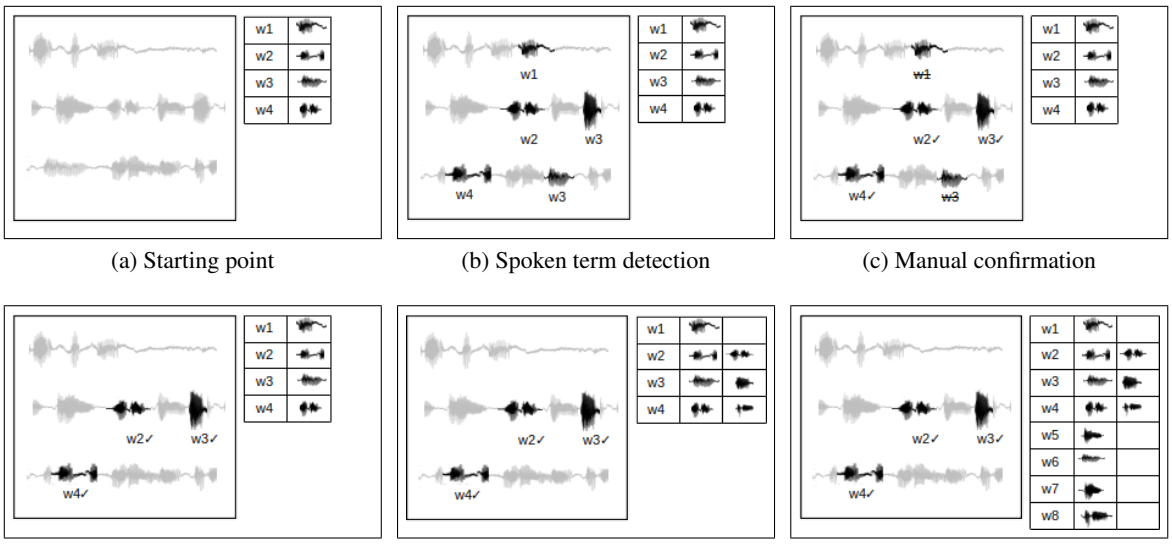
3 Proposed Workflow

3.1 Interactive and Sparse Transcription

The key idea of sparse transcription is to use spoken term detection methods to sparsely transcribe a speech collection, beginning with a small collection of spoken terms.

A spoken term is defined as a chunk of speech considered meaningful by a speaker, which may be a morph, a word, or a multiword expression. This collection of terms can be a list of keywords (or morphs) recorded in isolation by a speaker. Equally, it can be obtained by extracting audio clips from a speech collection.

We set up the workflow as shown in Figure 1. We begin with a lexicon of size *s* and a speech collection (Fig. 1a). This lexicon is composed of speech terms and their orthographic transcriptions $w_1 \dots w_n$. We use a spoken term detection method to retrieve speech terms in the collection that match those in the lexicon (Fig. 1b). The system presents the n most confidently identified terms and presents them for verification (Fig. 1c). False positives are corrected, i.e., erased from the transcription (Fig. 1d). We clip out from the utterances the speech terms correctly retrieved and add them to the lexicon as extra samples of a given entry (Fig. 1e). We allow a single entry to have maximum of *m* extra examples. We manually collect new speech units with a speaker, add them to the lexicon (Fig. 1f) and start a new



(a) Starting point (b) Spoken term detection (c) Manual confirmation

(d) Confirmed hits (e) Extra examples added to lexicon (f) Expanded lexicon

Figure 1: One iteration of our Interactive Sparse Transcription Workflow

iteration (Fig. 1b). We apply this for *i* iterations with a lexicon growing each time. The size *s* of the lexicon, *n* number of words checked, maximum *m* of extra examples for each word, and number *i* of iterations are hyperparameters which vary according to the contingencies of the fieldwork.

4 Pilot experiment: simulating this interactive scenario

Speech data. We apply the pipeline to two corpora. The first one is a 4h30m corpus in Mboshi[1] (Godard et al., 2018), a Bantu language spoken in Congo Brazaville (ISO mdw). It consists of 5,130 utterances sentence and word-aligned with an orthographic transcription. The utterances are elicited from text. This corpus contains only three speakers, of which one is responsible for 70% of the corpus. The second corpus is a very small (0h20m) corpus in Kunwinjku, an Australian Aboriginal language (ISO gup). It consists of 301 utterances aligned with an orthographic transcription. A forced alignment at the word-level has been created using the MAUS aligner (Kisler et al., 2017). The corpus contains 4 guided tours of the same town and one guided tour of another Aboriginal site. Each tour has been produced by a different speaker. To create initial and expanded lexicons, we select the 100 and 60 most frequent words bigger than 3 syllables for Mboshi and Kunwinjku respectively, in order to avoid words which are too short. A speech occurrence of each entry is extracted from the speech collection using the word-level alignments.

Acoustic features. In this interactive process, we explore several speech representations to identify those that are most suited to the pipeline, namely mel-frequency cepstral coefficients (MFCC) and perceptual linear prediction (PLP) features. We also use the hidden representation of an auto-encoder (AE) and correspondence auto-encoder (cAE). For this we use architecture of (Kamper et al., 2015) to self-train a 5-layer stacked AE (instead of 8) on 4h of Mboshi and 2h of Kunwinjku (YouTube videos). The cAE is trained using similar segments extracted from the speech collections with an unsupervised term discovery tool (Jansen and Van Durme, 2011). Finally we also use wav2vec representations (Schneider et al., 2019). The wav2vec model is either trained from scratch on Mboshi and Kunwinjku (w2v_mb and w2v_kun) or it is adapted from an original English model (w2v_en_mb and w2v_en_kun). We experiment with these features, with and without mean and variance normalisation (MVN) (Strand and Egeberg, 2004).

Sparse transcription experiments. The workflow described in section 3.1 is applied for 5 iterations for Mboshi and 3 iterations for Kunwinjku with an initial lexicon consisting of 20 words. 20 new words are added at the lexicon at each iteration. We end up with a 100 word lexicon for Mboshi and 60 words for Kunwinjku. The 10 best hits per word are checked ($n = 10$) and a maximum of 5 extra examples per word ($m = 5$) are added to the lexicon for both corpora. In addition, in order to avoid unnecessary verification in case the DTW score is too low, the worst score of the correct words checked during the first iteration is used as a threshold for the following iteration. To simulate human verification (Fig. 1c) we directly compare the hits output by the system with the gold transcriptions of the corpora.

5 Results: impact of speech feature representations in the workflow

We report the results of this new workflow in Table 1. The average precision scores (AP) correspond to the mean of the precision of the workflow computed at each iteration. The final recall is defined as the number of items retrieved from the full corpus X out of the number of all retrievable items in the corpus, i.e., the intersection of the lexicon L and the corpus C, $X/(L \cap C)$. In other words, this recall corresponds to the coverage of the lexicon related to its tokens in the speech collection. The impact of MVN is detailed only for the basic representations (PLP and MFCCs) since normalisation did not show any major influence on the neural representations (AE, cAE and w2v). We can make the following comments from the results shown in Table 1: (a) mean and variance normalisation (MVN) is important to improve results of basic (PLP and MFCC) features. Figure 2 shows that normalisation improves retrieval when the query term and the search term are pronounced by different speakers; (b) neural AE and cAE features are normalized-by-design but do not lead necessarily to better performance than PLP and MFCC;

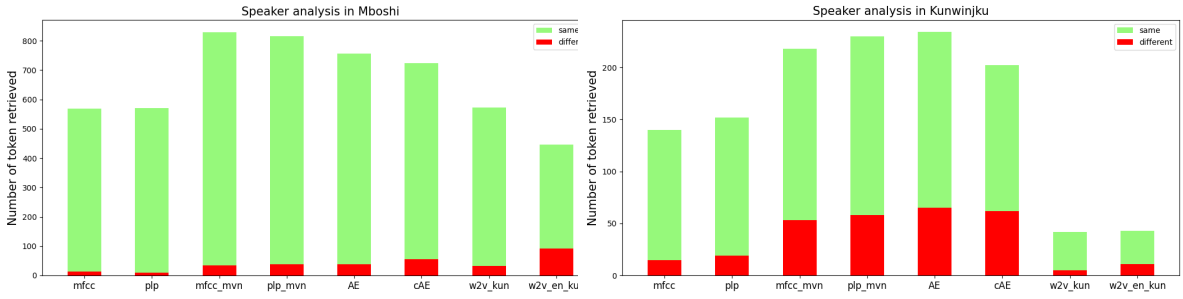
[1] https://github.com/besacier/mboshi-french-parallel-corpus

Features	MVN	AP	final recall
mfcc	no	23.87	16.78
mfcc	yes	**32.67**	**23.37**
plp	no	23.89	16.86
plp	yes	31.64	23.03
w2v_mb	no	24.57	16.88
w2v_en_mb	no	19.23	13.72
AE	no	31.93	21.51
cAE	no	27.31	20.70

Features	MVN	AP	final recall
mfcc	no	15.42	30.82
mfcc	yes	20.82	42.84
plp	no	15.21	32.67
plp	yes	**22.55**	44.89
w2v_kun	no	5.39	15.72
w2v_en_kun	no	5.45	15.87
AE	no	22.07	**45.30**
cAE	no	21.88	40.37

Table 1: Results for the Mboshi corpus (left) and for the Kunwinjku corpus (right)

and (c) representations provided by wav2vec are not efficient. The small size of the corpora might be the main obstacle when training efficient self-supervised models for learning speech representations.



(a) (b)

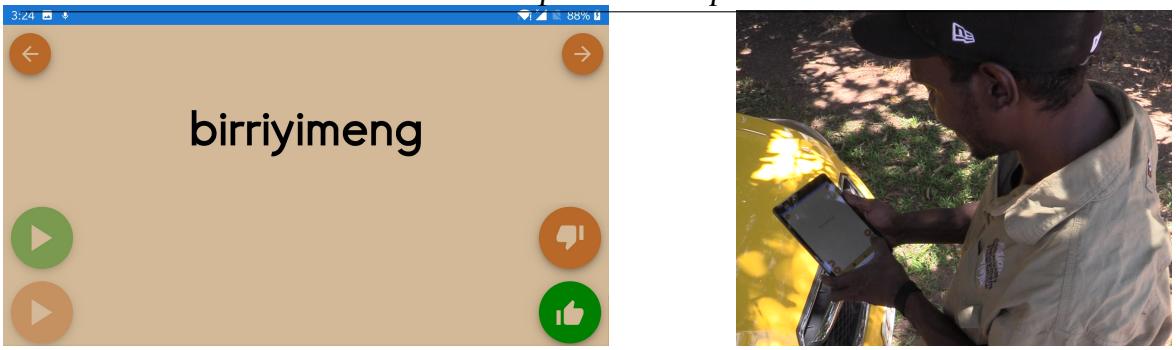
Figure 2: Proportion of same-speaker/different-speaker retrieval for each representation

Speaker diversity in the two corpora (Sec. 4) has an impact on the final results. For Mboshi 70% of the words to be retrieved in the speech collection are pronounced by the same speaker in the final lexicon and 37% for Kunwinjku. Figure 2 reports the proportion of tokens retrieved between same speakers and different speakers over the total number of tokens retrieved for Mboshi and Kunwinjku. The first observation is that the pipeline mostly retrieves terms pronounced by the speaker. It is clear that MVN improves performance for both same-speaker and different-speaker retrieval. We also note that, while wav2vec representation is overall less performant, it seems promising for extracting more speaker-independent representations, as illustrated by the results obtained with w2v_en_mb features. Future work will investigate this by leveraging more raw speech in Mboshi and Kunwinjku for training better self-supervised (wav2vec) models.

Regarding false positives (Table 2), the first errors we can observe are different spellings of words (in the gold transcription) referring to the same meaning (e.g., *namekke / nemekke* "that one"). Since

Query	False Positive	query translation	hit translation
nahne	mahne	it (pronoun)	this (demonstrative)
nemekke	namekke	that one	that one (other spelling)
balanda	balanda-ken	white man	of the white man (genitive)
bininj	bininj-beh	man	from the man
nemekke	yekke	that one	dry season
mahni	mahne	this (demonstrative)	this (other spelling)

Table 2: Top false positive generated by the spoken term detection system in Kunwinjku



(a) screenshot of the app

(b) deployment of the app

Figure 3: Lexical confirmation app

Kunwinjku is primarily a spoken language, variable spellings are common. Moreover, because of the morphological complexity of the language, many of the top false positives are actually inflectional variants of the query term (e.g., *balanda / balanda-ken* "white man / from the white man"). Allowing matching at a smaller granularity could be a way to achieve wider coverage of the speech collection.

6 Deployment

Building on prior work developing mobile tools for language documentation (Bettinson and Bird, 2017), we have begun to explore methods for deploying the pipeline in a remote community. While the first step of identification of new words is straightforward, the task of lexical confirmation might be much more complex to apply. The members of an Aboriginal community might not be familiar with technologies and they are not necessarily literate (in the narrow western sense). Taking into account these constraints, we built a lexical confirmation app and trialled it on a small lexicon (Fig. 3).

The idea would be to load the output of the spoken term detection system into the app. Then a speaker can listen to the query, listen to the utterance where we expect the query to be found, and confirm if the utterance contains the query, (Fig.1c)

7 Conclusion

We investigated the use of spoken term detection methods as an alternative to the usual methods that have been inspired by automatic speech recognition, and which require exhaustive transcription even for passages which exceed the present state of our knowledge about the language. Instead we devised a workflow based on spoken term detection, and simulated it for two small corpora: one in Mboshi, one in Kunwinjku. The simulations of this workflow show that, with well chosen speech representations, we may have a viable approach for rapidly bootstrapping transcriptions of large collections of speech in endangered languages. The next step of this work would be to design methods to involve Indigenous people in tasks such as the construction of the lexicon or the confirmation of the output of our system for a deployment of this workflow in a remote community.

Acknowledgements

We are grateful to the Bininj people of Northern Australia for the opportunity to work in their community, and particularly to artists at Injalak Arts and Craft (Gunbalanya) and to the Warddeken Rangers (Kabulwarnamyo). Our thanks to several anonymous reviewers for helpful feedback on earlier versions of this paper. The lexical confirmation app presented in this paper has been designed by Mat Bettinson, at Charles Darwin University. This research was covered by a research permit from the Northern Land Council, ethics approved from CDU and was supported by the Australian government through a PhD scholarship, and grants from the Australian Research Council and the Indigenous Language and Arts Program.

References

Mat Bettinson and Steven Bird. 2017. Developing a suite of mobile applications for collaborative language documentation. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 156–164.

Steven Bird. 2020. Sparse transcription. *Computational Linguistics*, 46(4).

Paul Boersma and David Weenink. 1996. Praat, a system for doing phonetics by computer, version 3.4. *Institute of Phonetic Sciences of the University of Amsterdam, Report*, 132:182.

Caren Brinckmann. 2009. Transcription bottleneck of speech corpus exploitation. *Proceedings of the 2nd Colloquium on Lesser Used Languages and Computer Linguistics*, pages 165 – 179.

First Languages Australia. 2014. Angkety Map: Digital resource report. Technical report. `https://www.firstlanguages.org.au/images/fla-angkety-map.pdf`; accessed Nov 2020.

Ben Foley, Joshua T Arnold, Rolando Coto-Solano, Gautier Durantin, T Mark Ellison, Daan van Esch, Scott Heath, Frantisek Kratochvil, Zara Maxwell-Smith, David Nash, et al. 2018. Building speech recognition systems for language documentation: The CoEDL Endangered Language Pipeline and Inference System (ELPIS). In *Proceedings of The 6th International Workshop on Spoken Language Technologies for Under-Resourced Languages*, pages 205–209.

Pierre Godard, Gilles Adda, Martine Adda-Decker, Juan Benjumea, Laurent Besacier, Jamison Cooper-Leavitt, Guy-Noel Kouarata, Lori Lamel, Hélène Bonneau-Maynard, Markus Mueller, et al. 2018. A very low resource language speech corpus for computational language documentation experiments. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, pages 3366–70.

Vishwa Gupta and Gilles Boulianne. 2020a. Automatic transcription challenges for Inuktitut, a low-resource polysynthetic language. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2521–27.

Vishwa Gupta and Gilles Boulianne. 2020b. Speech transcription challenges for resource constrained indigenous language Cree. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-Resourced Languages and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 362–367.

Aren Jansen and Benjamin Van Durme. 2011. Efficient spoken term discovery using randomized algorithms. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 401–406. IEEE.

Herman Kamper, Micha Elsner, Aren Jansen, and Sharon Goldwater. 2015. Unsupervised neural network based feature extraction using weak top-down constraints. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5818–22. IEEE.

Thomas Kisler, Uwe Reichel, and Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech and Language*, 45:326–347.

Raghav Menon, Herman Kamper, John Quinn, and Thomas Niesler. 2018a. Fast ASR-free and almost zero-resource keyword spotting using DTW and CNNs for humanitarian monitoring. In *Proceedings of Interspeech 2018*, pages 2608–12.

Raghav Menon, Herman Kamper, Emre Yilmaz, John Quinn, and Thomas Niesler. 2018b. ASR-Free CNN-DTW keyword spotting using multilingual bottleneck features for almost zero-resource languages. In *Proceedings of the 6th International Workshop on Spoken Language Technologies for Under-Resourced Languages*, pages 182–186.

Raghav Menon, Herman Kamper, Ewald van der Westhuizen, John Quinn, and Thomas Niesler. 2019. Feature exploration for almost zero-resource ASR-free keyword spotting using a multilingual bottleneck extractor and correspondence autoencoders. *Proceedings of Interspeech 2019*, pages 3475–3479.

Alex Park and James R Glass. 2005. Towards unsupervised pattern discovery in speech. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 53–58. IEEE.

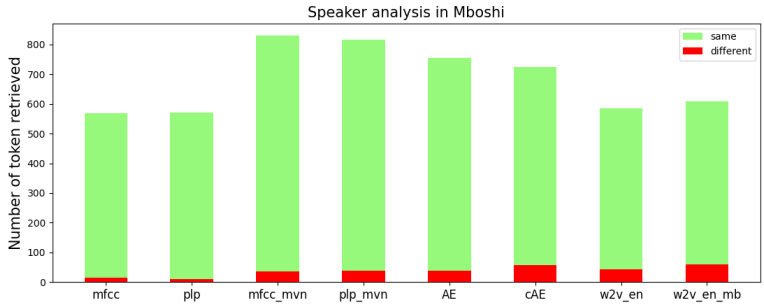
Hiroaki Sakoe and Seibi Chiba. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26:43–49.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. Wav2vec: Unsupervised pre-training for speech recognition. *Proceedings of Interspeech 2019*, pages 3465–69.

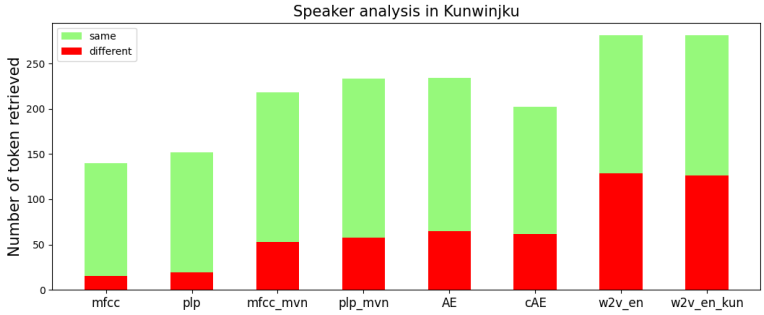
Ole Morten Strand and Andreas Egeberg. 2004. Cepstral mean and variance normalization in the model domain. In *ISCA Tutorial and Research Workshop on Robustness Issues in Conversational Interaction*.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. Elan: a professional framework for multimodality research. In *5th International Conference on Language Resources and Evaluation*, pages 1556–15.

Retrospective view



(A)



(B)

FIGURE 3.2: Proportion of same-speaker/different-speaker retrieval for each representation

Following the publication of the paper, an error in the results was discovered. The wav2vec model was not properly used. First, the benefit of such a model was its multilingual aspect, which allowed the leveraging of the small amount of data available with oral languages. Training a model from scratch with the limited amount of data available did not make much sense. Additionally, we realised we had misused the feature extraction function. While the impact of our mistake was not significant for Mboshi, it had a major influence on Kunwinjku. We report the speaker analysis in Figure 3.2 and the updated results, with the previous results labelled as *old*, in Table 3.1. The results for wav2vec trained on individual languages have been replaced by results provided with the original model trained in English, and the results with the English model fine-tuned have been replaced by the results after the corrected feature extraction function.

Features	MVN	AP	final recall
mfcc	no	23.87	16.78
mfcc	yes	**32.67**	**23.37**
plp	no	23.89	16.86
plp	yes	31.64	23.03
w2v_en	no	25.28	19.45
w2v_en_mb	no	25.67	20.16
w2v_mb (old)	no	24.57	16.88
w2v_en_mb (old)	no	19.23	13.72
AE	no	31.93	21.51
cAE	no	27.31	20.70

Features	MVN	AP	final recall
mfcc	no	15.42	30.82
mfcc	yes	20.82	42.84
plp	no	15.21	32.67
plp	yes	22.55	44.89
w2v_en	no	27.20	52.54
w2v_en_kun	no	**27.39**	**52.54**
w2v_kun (old)	no	5.39	15.72
w2v_en_kun (old)	no	5.45	15.87
AE	no	22.07	45.30
cAE	no	21.88	40.37

TABLE 3.1: Results for the Mboshi corpus (top) and for the Kunwinjku corpus (bottom)

The difference in performances between Mboshi and Kunwinjku can be explained by the speaker distribution in each corpus. In Mboshi, one speaker was responsible for 70% of the utterances of the corpus and about the same number of terms to be retrieved and query terms in the lexicon were pronounced by the same speaker. Conversely, the corpus in Kunwinjku was much more balanced and most of the terms to be retrieved in the corpus were pronounced by a different speaker in the lexicon. The use of features based on the spectrum provide better performances for same-speaker recognition, while the use of features based on multilingual pretraining slightly decreased the performances in same-speaker recognition but improved it in different speaker recognition.

The configuration of the current workflow presents a few limitations in terms of performance and efficiency. Firstly, all the speech representations used provided

low accuracy (between 20% and 30%). Future studies would benefit from exploring different optimisation strategies to achieve accuracy that could be suitable for a human-in-the-loop workflow, even if the final recall is decreased. Additionally, DTW is slow to compute, particularly with wav2vec features of 512 dimensions. In Mboshi, one iteration would take between 20 minutes (to spot 100 terms in 5 hours with normalised MFCC) to 9 hours (to spot 800 terms in 5 hours of speech with wav2vec). Finally, query-by-example approaches restrict the coverage of the transcription. Methods relying on acoustic comparison are to some extent biased by speaker information and more generally by the quality of the query terms. A careful selection of those terms is necessary to ensure decent performance, and the selection of those terms takes time. In our case, we semi-automatically extracted the spoken lexicon from the transcribed data with the forced alignment. A forced aligner is not available in all languages, and while such a tool helped with the selection of the terms, manual verification of the terms was necessary to ensure quality.

Ethical collaboration in an Indigenous context involves a phase of negotiation. If the shape or purpose of an activity is not satisfactory of all parties, it seems logical that some stakeholders refuse to be involved. A technology to support language documentation should consider this. The selection of the query terms and the computing time of DTW are two elements that can cause major obstacles to the deployment of such methods on the field.

3.3 ACL 2022: deployment of the sparse transcription simulation

Éric Le Ferrand, Steven Bird and Laurent Besacier. Learning From Failure: Data Capture in an Australian Aboriginal Community. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*. pp. 4988–4998.

Research process

After designing the first sparse transcription simulation and its associated prototype, the next logical step was to replace simulated manual verification with actual verification by Indigenous participants. In July 2019, we visited Bininj Country for a week, starting in Gunbalanya and then moving to Kabulwarnamyo, to present and discuss our work with the Traditional Owners of the land. We returned to Gunbalanya a few weeks later to test our initial app design. We recorded a few videos at the time but we did not analyse them immediately.

A year passed before we could return to Bininj Country due to Covid-19 restrictions. In September 2020, we travelled to Kabulwarnamyo and Manmoyi for a two-week trip with the aim of further testing the app with a wider range of participants. Josh Yang, a final year undergraduate student, joined us for this trip and recorded videos of the participants' interactions with the app. As the paper explains, the work was unsuccessful, and we ceased this deployment.

Initially, I was hesitant to write about this experience as it involved interdisciplinary work and described a negative outcome. It did not have a place in HCI conferences that were already familiar with cross-cultural design, and its anthropological component was not a good fit for NLP conferences. However, the special theme track organised by ACL 2022 provided a unique opportunity to reflect on this failure and offer the NLP

research community a real-life perspective on the application of speech processing technologies in Indigenous contexts.

A former colleague in the HCI field encouraged me to write about this experience. Cross-cultural interactions involving technology are rare and can provide valuable insights for the community to pursue more meaningful work. I transcribed all the videos we recorded, including verbal and nonverbal interactions, and began a thorough analysis of my failures.

Learning From Failure: Data Capture in an Australian Aboriginal Community

Éric Le Ferrand,[1,2] **Steven Bird,**[1] **and Laurent Besacier**[2]
[1]Northern Institute, Charles Darwin University, Australia
[2]Laboratoire Informatique de Grenoble, Université Grenoble Alpes, France

Abstract

Most low resource language technology development is premised on the need to collect data for training statistical models. When we follow the typical process of recording and transcribing text for small Indigenous languages, we hit up against the so-called "transcription bottleneck." Therefore it is worth exploring new ways of engaging with speakers which generate data while avoiding the transcription bottleneck. We have deployed a prototype app for speakers to use for confirming system guesses in an approach to transcription based on word spotting. However, in the process of testing the app we encountered many new problems for engagement with speakers. This paper presents a close-up study of the process of deploying data capture technology on the ground in an Australian Aboriginal community. We reflect on our interactions with participants and draw lessons that apply to anyone seeking to develop methods for language data collection in an Indigenous community.

1 Introduction

For decades, the work of collecting data for Indigenous languages has been the province of documentary and descriptive linguistics (Bouquiaux and Thomas, 1992; Vaux and Cooper, 1999; Meakins et al., 2018). This work has involved various kinds of elicitation, e.g. of word lists, phrases, etc, to support description of the phonology, morphosyntax, and grammar of the language. It has also involved the collection of unrestricted text, through recording and transcription. In most cases, the result is audio with aligned text. Many software tools have been developed for supporting these activities (Boersma, 2001; Clark et al., 2008; Hatton, 2013; Sloetjes et al., 2013).

Within the field of natural language processing, established practice is to support the linguist's work (Michaud et al., 2018; Seifart et al., 2018; Foley et al., 2018; Cox et al., 2019). In some cases, this includes the participation of speakers in activities using apps controlled by linguists (Bird et al., 2014; Hanke, 2017; Bettinson and Bird, 2017). However, the premise is basically the same: obtain a substantial quantity of audio and transcribe it, or post-edit the output of an automatic transcription system.

We believe that these approaches do not adequately address a fundamental reality of small languages: they are *oral*. There may be an official orthography, but it has no place in the local language ecology where any written business takes place in a language of wider communication. As a result, local people are usually not confident in the orthography of the language. Furthermore, there may be low confidence in using computers and text editors, and inadequate support for the language in terms of keyboarding and spelling correction. Add to all this the fact that the whole space of rendering an oral language into standardised orthography can be alienating (Dobrin et al., 2009; Hermes and Engman, 2017).

There is no particular reason for NLP approaches to Indigenous languages to follow the long-established practices of linguists. After all, there is an equally long history of algorithmic approaches being profoundly different to the human tasks they replicate. For instance, a human sorting a hand of cards may use insertion sort, but a machine might use Quicksort, with better average-case complexity (Levitin, 1999). Computational approaches may be inspired by analogy, e.g. simulated annealing, genetic algorithms, neural networks, but they are not required to adhere to the human defined process. Accordingly, we can ask, what is an idiomatic computational approach to collecting data for Indigenous languages that is a better fit to the capabilities of human participants? In the case of associating text and speech, we believe that the answer might be keyword spotting. This is because, in our experience, speakers and learners are attuned to identifying whole words, rather than obsessing

about the idiosyncratic phonetic makeup of individual tokens as required for phone transcription (cf. Bird, 2020b, 718f).

Accordingly, we investigate an approach to transcription based on word spotting known as "sparse transcription" (Bird, 2020b). This would seem to be an easier, less specialised task than direct, contiguous transcription. If more people can participate, we can hope to establish a virtuous circle with more data, better models, less correction, even more data, and so on. The idea is that transcription can be accelerated by identifying the tokens of high-frequency terms all at once, then playing them back in quick succession for confirmation by participants.

This paper reports on the deployment of a lexical confirmation app which supports human confirmation of system hypotheses. We begin by describing the background to this work (Sec. 2), including related work on designing technology for use in Indigenous places. We also describe the site where we work and the design of the lexical verification app. Next, we report what happened when we deployed the app in two field tests, including detailed accounts of interactions with participants (Sec. 3). In the discussion section, we reflect on the field experience from a variety of perspectives, trying to draw out lessons that may be applicable to other places where NLP researchers seek to design technologies for language data collection (Sec. 4). The paper concludes with a summary and prospects for further research.

2 Background

2.1 Designing in an Indigenous context

Designing in the Indigenous space is a small but growing area within the field of Human-Computer Interaction (HCI). Projects in this space often begin with ethnographic research to identify local priorities. Co-design is advocated as a way to establish a "culturally-tailored, culturally-enriched and trustworthy environment for participation" (Peters et al., 2018). The focus of this work includes traditional knowledge (Verran, 2007), language revitalisation (Hardy et al., 2016) or media sharing (Soro et al., 2017). Recent research mentioned the need to involve stakeholders in a system design (Lynch and Gregor, 2004) highlighting the challenges related to the transparency of the mechanism of a given system, specifically when machine learning is involved (Loi et al., 2019) and the difficulty to explain to the users such mechanism (Abdul et al., 2018). The lack of published accounts of experiences collecting language data in Indigenous contexts, specifically in the intersection of NLP and documentary linguistics, makes it difficult for newcomers like us to devise approaches that are likely to work. We address this shortcoming by reporting and reflecting on our field experience.

Deploying speech technologies in remote Aboriginal communities is challenging, not primarily because of low technological literacy on the part of local people, but because of low interactional literacy on the part of NLP researchers who enter indigenous places to gather data.

2.2 Working in an Indigenous place

Our work is grounded in Bininj country in Arnhem land in the north of Australia. The biggest town is Gunbalanya with 1,100 inhabitants where we can find primary and secondary schools in which teaching is done in English. A few remote satellite communities, or "outstations," can be found throughout this country in which education of young people takes place in a bi-cultural environment both in Kunwok and English.

Kunwok (ISO gup) is the main language of communication here, and Kunwinjku is the prevalent dialect. It is spoken by some 2,500 people and is one of the few Australian languages which is gaining speakers (Evans et al., 2003). While a standard orthography exists, most community members do not write at all. When pressed, some of them are able to leverage their knowledge of English literacy in order to decode Kunwok texts (cf. Feinauer et al., 2013; August et al., 2009).

In prior work in Bininj country, we discussed our work with traditional owners (heirs of a given tract of Aboriginal land and leaders of the community). We described and demonstrated prior work involving transcription, and how it can be used to transcribe Kunwok. They raised their concerns about intergenerational knowledge preservation and transmission and access to the resources created by westerners. While it is not clear to us that the nature of our work had been thoroughly understood, we could identify through this interaction topics which are addressed by current speech processing and HCI research projects (San et al., 2021; Taylor et al., 2020). Our work took place in Gunbalanya and Manmoyi, a remote community situated 5 hours drive from Gunbalanya.

Australian Aboriginal communities are far from uniform. The experiences and challenges we describe here may be relevant for the Australian Top End, but they cannot be directly applied to Indigenous communities in other places.

Figure 1: Screenshot of the Spoken Term Detection based Lexical Verification App

2.3 Lexical verification app

We designed a lexical verification app that bridges the output of a spoken term detection system to people. It was built following the design of Bettinson and Bird (2017). We focused on a simple design without any textual component besides the transcription of the query term. The idea is to first load in a web app the query/utterance pairs generated by our spoken term detection system. We then ask speakers of the target language to confirm for each pair if the query word (i.e. the term we are trying to retrieve) is pronounced in the search utterance (i.e. the sentence in the speech collection in which the query term was detected).

The participants have six buttons available to perform the task. They have two play buttons at the bottom left: One to play the query term, the other to play the search utterance. Once the two audio files have been listened to, two feedback buttons appear at the bottom right to allow the user to confirm if the query term is included or not in the utterance. We also added two arrows on each side of the top of the screen to allow the user to jump to the previous or the next example. When a new example is displayed on the screen, the query term is played automatically When the utterance is played, the transcription of the query term is highlighted around the timestamps in which the query term was detected. The terms are spotted in the utterances beforehand following the parameters of the sparse transcription simulation proposed by Le Ferrand and Bird (2020). Because of the challenges posed by the remote Aboriginal context such as the lack of reception or proper working facilities (e.g. a table), we needed to find solution in terms of data storage and activity design. Based on the work of (Bettinson and Bird, 2021), we stored the query/utterance pairs output by our spoken term detection system in a JSON file and loaded them in a Raspberry Pi with the app. The Pi acts as a WiFi hotspot to which any device can connect. We can then then connect a tablet to the Pi and, doing so, the feedback provided by the participant can directly be stored in the associated database.

3 Fieldwork

We tested our approach with two trials in two Aboriginal towns, with three people in each place. While the number of participants seems small, larger trials are difficult to arrange in Aboriginal contexts due to the small number of speakers. At the beginning of each elicitation session, the first author explained our intention to teach a machine to transcribe the language automatically, and that we wanted help to correct system guesses. There is actually no direct translation of transcription in Kunwok and the concept is usually given by the formulation *karribimbun kure djurra*, “we’re drawing on paper”.

In both places, we recruited the participants with the support of two local institutions, the art centre in Gunbalanya and the ranger organisation in Manmoyi. At the start of our trips, the first author introduced himself to the communities and explained that he was looking for people to support him for language work. Then the people interested came to find him throughout the day. Each session lasted approximately 15 minutes and was part of other language work including recordings or language learning. Each participant was paid at the regular rate for language work.

3.1 Trial 1: Gunbalanya

For our first trial, we recorded source audio from a three hour guided tour of a local site. We transcribed a few minutes of this recording and used this transcription to build a lexicon. We used voice activity detection to segment the recording into breath groups. Finally, we automatically spotted terms from the lexicon in these breath groups. Since the speaker of the lexicon and the speech collection overlap, most of the terms spotted by the system were correctly retrieved. In the data presented to participants, the query term was present in the supplied phrase in 57% of the instances.

This configuration was tested with three Gunbalanya residents: SB (20s), TM (30s), and RB (40s). This last participant was also the speaker of the recordings.

SB appeared nervous and said little in response to our explanations and questions. When an audio clip was played, he translated, even though this was not the instruction. It was as if he projected his assumption about the purpose of the task, namely for the researchers to understand the content. At one point he respoke the query term and the target phrase in a single utterance, before explaining his knowledge about the associated place. The interface itself was not legible to him: faced with a choice of two play buttons – one for the query term and one for the phrase – he was never clear which one to press. He never used the thumbs up/down feedback buttons.

Here is an example of the confusing situation set up by our approach (we use "App" to indicate audio produced by the app, along with speaker initials, and ELf for the first author. "Play1" refers to the button that plays the query term and "play2" the utterance).

ELF	<press play1>
App	*manyilk*
ELF	<press play2>
App	*menekke mandjewk karuy*
ELF	*manyilk? larrh*. Because he says *mandjewk*
SB	*manyilk*, first <press play1>
App	*manyilk*

Notice that the query term *manyilk* "grass" is not contained in the utterance *menekke mandjewk karuy* "this wet season he dug it". When we demonstrate the use of the app by giving the expected response of *larrh* "no", SB asserts that manyilk is present, contradicting us. He presses on the query term play button to show us.

The following day, when we discussed with another participant, we heard that SB thought that our task was an attempt to test his memory.

RB was more confident than SB. He seemed intrigued at hearing his own voice on the device. For each audio segment we played, RB gave an interpretation of the content. We offered the device to him to control, but he declined. After we pressed the two play buttons, he waited, and we had to follow up with overt questions: "does he say <query term>?", or "can you hear <query term> in this sentence?" He answered as expected, with: "yes, <query term>" or "no, he doesn't say <query term>." Consider the following example:

ELF	<press play1>
App	*marnbom* ("he made")
ELF	<press play2>
App	*kumekke* artist *marnbom kadi*
ELF	do you hear *marnbom*?
RB	*marnbom* that's painting making the painting
ELF	but do you hear *marnbom* in the sentence?
RB	yeah

Unlike SB and RB, TM readily took the device and used the controls. Sometimes, when the query term was not contained in the utterance, he not only translated the audio, but he also offered an example sentence containing the query term. In the following example, "confirm" refers to one of the feedback button which automatically display the next example and play the query term:

TM	<press confirm>
App	*karrikadjung* ("we follow it")
TM	*karrikadjung*, ("we are following") <press play2>
App	*karrikadjuy* road ("we followed the road")
TM	he says *karrikadjuy*, it means we went this way road, he should have say we are following this one, *karrikadjung*

In this case, the difference between the query term *karri-kadju-ng* "we-follow-PRES" and the utterance *karri-kadju-y* "we-follow-PAST" is only in verb tense. The whole query term appears in the sentence, except for the tense marker. Should the speaker say yes or no? This points to a shortcoming of the task definition.

When the term was correctly retrieved, TM would respeak the audio and press the thumbs-up button. When the term was not correctly retrieved, TM offered extensive explanations.

3.2 Trial 2: Manmoyi

For the second trial, we visited the Manmoyi outstation. We used five short audio recordings from previous fieldwork, including guided tours and traditional stories. One of the recordings was transcribed and we extracted the words to use as our lexicon. As before, we segmented the source audio into breath groups and ran word spotting against

this set.

Since the speaker of the lexicon and those of the rest of the collection did not overlap, there was much lower precision; often a query term matched noise or mumbling. In the data presented to participants, only in about 10% of cases was the query term present in the supplied phrase.

This configuration was tested with three residents of Manmoyi: LY (60s), LB (50s), RG (50s).

LB and LY participated together, with LB taking an active role and LY only participating by talking to LB during the task. With each round, LB listened to the query term and the utterance then appeared to associate them as a single linguistic event, and he would recount a story that included both the term and the utterance. After this, he would give feedback (thumbs up or down) depending on how easy he found it to link the two semantically:

LB	<press play1>
App	*wirrihmi* (“dislike/wrong”)
LB	that’s “wrong one”
LB	<press play2>
App	*wanjh manjbekkan manmanjmak*
LY	it tasted sweet
LB	it tasted like you know this, it might have been a little bit funny or something like that
LB	yeah like for us they say: “no I can’t eat” because he tasted it and they say “try it” and they gave it, and he says “aah yeah it tasted nice”
LB	*yoh*, that’s the one, that’s good, *kamak*

LB often interpreted the audio segment. At one point, he recognised the speaker for the queries, and he told us about her and began to recount the same story:

ELF	<press play1>
App	*nawernwarre* (“big brother”)
ELF	<press play2>
App	*birribonguni birri...* (“they were drinking, they...”)
LY	*nawernwarre*
LB	*yoh, nawernwarre*
LY	*nawernwarre*, or *manekke* might be... lonely boy (story)
LB	lonely boy *yoh* that’s the lonely boy (story)

Towards the end of the session, we asked about LB’s understanding of the task:

ELF	Can you tell me in English what do you think I am trying to do?
LB	You are trying to... you are making like Kunwok and English translating, but if you are making straight like Kunwok you’re making straight and English making straight, that’s the all same.
ELF	well, not really
LB	no it’s real, we are talking, we know everything. Not all these, we’ve seen these people, they don’t know anything about it, myself and LY we know everything about it.

LB understood this to be a translation activity. When we disagreed, he re-asserted his standing as a knowledge authority. Later, we explained our ultimate purpose of automatically transcribing the language. LB rephrased transcription as “make it together.” We realised afterwards that LB may have been referring to his semantic linking process.

RG was our final participant, and this session revealed many issues. Given the low number of correct query-utterance pairs, we found ourselves needing to manually skip over utterances that were too hard to understand out of context. Each time we abandoned a round and moved on to the following round, the next query term played automatically (this feature was added before any testing with the assumption that it would speed up the verification process).

Such automation turned out to be confusing for RG. For a few instances, RG responded “yes” when the query term was not literally present in the utterance, maybe because the query term was morphologically related to a term that was present, e.g. *birri-m-h-ni* “they-towards-immediate-were” (query) and *birri-ni* “they-were”. Another interpretation of this behaviour is that RG was focussing on meanings not forms. In this and other cases, it seems that RG was not clear about what we were asking for.

RG	The old woman is talking about country and the young fellow is talking about what creation was.
RG	It’s all a bit confusing. They are not even saying *kunred* it means home, the young other fellow is talking about dreamtime story, so it is not, well it’s connect but it is not pronouncing.

Sometimes, RG asked about the speakers and the

overall context of the out-of-context audio segment, asking, e.g. "Is this <name> speaking? I don't know what they're talking about here."

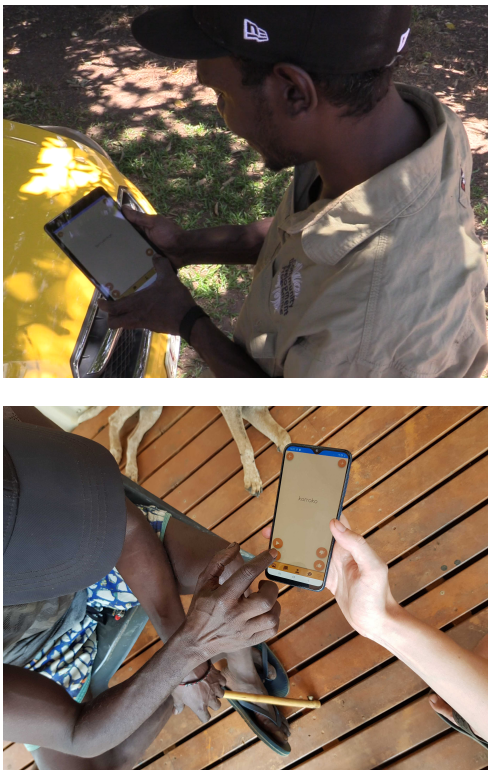


Figure 2: Use of the Word verification app

4 Discussion

There were many issues in the design and conduct of this elicitation activity, and it is clear that our approach need to be completely rethought. In this section, we analyse the above interactions and try to identify some principles to inform NLP elicitation methodology, hoping to avoid such problems occurring in future.

Task motivation. SB, RB and LB understood us to be interested in interpreting the content. SB thought we were testing his memory. TM offered detailed explanations. LB said things that we interpret as asserting authority. It appears that our attempt to explain our purpose in automatic transcription, and the activity of confirming or refuting system guesses, was unsuccessful.

Task definition. Participants were not clear about what we were asking of them. The notion of "word" was not clearly defined, and there were a variety of responses when the query term was not identical yet morphologically or semantically similar to a word in the corresponding utterance.

Naturalness of the task. When it comes to collaboration with western language workers, Aboriginal people in these communities are accustomed to participating in interviews, recordings, transcription, and translation activities. This may explain people's readiness to respeak or interpret the content or supply additional cultural information. We entered with a different task, one where the overt activity of human confirmation/rejection of system guesses was not transparently related to a recognisable transcription task. We explained and demonstrated the activity, but TM was the only participant to instantly grasp this task. Even so, he provided extensive explanations when the system guess was wrong in an effort to teach us.

Utterance context. From our perspective, the components of the device were clear. We have a query term that needs to be detected, and an utterance that should contain the query term. From this, we just need two feedback buttons to confirm whether the query term is included in the utterance. However, to the participant listening to the audio produced by the app and not following our use of the controls, the query term and utterance may be perceived as a single utterance. Everything put into the aural space appears to be concatenated by listeners, and our non-conventional metalinguistic context is not interpretable. When endeavouring to explain the task in Kunwok, we were hampered by the lack of words for "word" and "sentence".

Teaching. The participants generally provided much more information than the simple yes/no response we requested. Each instance was another opportunity to teach us about the language or the country. The design of the task only limited the space for this style of participation. The activity itself was not particularly engaging, taking utterances out of context and asking for a mechanical response to a seemingly pointless question. It seems to be a kind of resilience that participants made the most of the opportunity to pursue their own ends of educating newcomers. Further discussion with community members highlighted their concerns about knowledge preservation, access to archival recordings, and learning literacy.

Knowledge transmission. George et al. (2010) explains that the way in which westerners and Australian Aboriginal people transmit their knowledge varies in that one extracts, identifies and, categorizes while the other needs the information to be

embedded in a system of kinship relationships. For example, in Bininj country, every individual has a kinship relationship to every other individual, and they address each other accordingly (Glowczewski, 1989). Stories do not exist in isolation but are connected to an individual who tells them, and the country it comes from. We ran up against this when participants needed to connect isolated utterances back to their rightful cultural context, not just consider them as arbitrary linguistic material for which they can answer an unmotivated question: "does this utterance contain this word?" We can see this in Trial 2 where LB ignores the utterance and uses his knowledge of the speaker of the query term to link the content back to the story.

Yarning. Recent fieldwork methods research has shown that adopting Aboriginal-led approaches leads to more culturally appropriate practices and better feedback from Aboriginal consultants (Louro and Collard, 2021). Yarning has been described as a research method and the traditional way for Aboriginal people in Australia to pass knowledge. It can be defined as "a conversational process that involves listening to storytelling that creates new knowledge and understanding" (Terare and Rawsthorne, 2020). Adopting this to engage with participants could lead to better participation and a more appropriate way to collaborate. Here, the Aboriginal consultant would occupy a teaching role and the function of the technology would be to capture, support, and organise natural ways of transmitting knowledge.

Spoken term detection performance. The spoken term detection method delivered markedly different results in the two trials. Presenting data with 50% accuracy (first trial) makes the user's task seem most worthwhile, otherwise, the user is mostly confirming or refuting system guesses (refuting in 90% of cases in the second trial). If this reasoning is correct, then we predict that a trial involving 90% accuracy would also be challenging to motivate and teach. The low accuracy of the system probably contributed to the challenges encountered during the second trial. However similar behaviour in both trials was observed (e.g. the systematic translation after an audio was played or the semantic linkage process) which makes us think that the sole performance of a system is not the main source of the misinterpretation of the task.

App design. The design of the app was based on preliminary thinking about how collection could proceed fluidly. We did not consider the confusion that might be caused by having two play buttons on the screen (one for the query term, and one for the corresponding utterance). In the interests of efficiency, with each new round, the query term was played automatically. It was as if the thumbs up/down button from the previous round caused playback, and this turned out to be confusing. When we wanted to skip forward by a few examples using the right or left arrow keys at the top of the display (Fig. 1), the app would play a series of seemingly random words. Such automation should have been avoided, specifically in the early stage of our work when there was a lot of uncertainty regarding people reaction towards our activity.

Design improvements. Besides the elements we already mentioned, a few paths can be explored to address the challenges we have faced. Removing the query play button could have the effect of reducing the number of contexts and avoid the linkage process we have observed with LB and SB. Limiting the activity to a single story and playing the utterances in chronological order can make the context clear, and the participant would not need to clarify it. Using bottleneck features instead of MFCCs to spot words could improve the precision of the system (Menon et al., 2019).

Such modifications, however, cannot address the biggest flaw of our proposed task: it does not respond directly to people's agenda in terms of language work, but simply tries to leverage people's skills to respond to westerners' expectations. Pushing the proposed pipeline for several iterations would risk alienating our participants and compromising further collaboration. We believe that a complete reshaping of our method is necessary to enable a sustainable and community-based model for language and knowledge documentation.

5 Further Reflection

Our first attempt in this space was unsuccessful on many levels. Most superficially were issues with the task definition and the app interface. The task focused on the notion of "word" and on deciding whether a given word occurred in a given utterance. Yet the notion of word was not established; as an oral language, there was no *a priori* shared understanding between the participant's notion of spoken

word and our notion of orthographic word.

Throughout our interactions with participants, our attempts to explain the method and the purpose were unsuccessful. Local perception was fixed on the idea that we had entered the community to learn the language and culture, and that the purpose of participating in the study was to teach us and to interpret the texts for us.

Consequently, the narrow focus of our activity on eliciting a binary, thumbs up/down response was unsuccessful. This is hardly surprising as many people have noted that engaging Aboriginal people with direct questions requiring a yes or no response is seen as testing people's knowledge or memory, and potentially irritating (Maar et al., 2011; Ober, 2017). We observed this ourselves, when SB reported that he felt like he was being tested, or when LB responded as if his authority was being questioned.

Clearly, our style of engagement was not the expected kind of collaboration on a linguistic task. Aside from one participant (TM), no one would participate in the abstract and apparently pointless task of confirming whether a word was present in a sentence. Instead, all participants sought to create meaning from any language fragments they were presented with. On the basis of an isolated word, and person, place or story would be detected, and people would seek to teach us about these aspects of their lifeworld. This took various forms: repeating, paraphrasing, translating, interpreting, or offering extensive cultural commentaries.

In retrospect, this response to our approach comes across as resilient and generous. In comparison, our narrow focus on data collection, and on getting across the specialised task of lexical confirmation may have come across as disconnected from local interests, and potentially disrespectful.

Of course, we can hope to recruit more people like TM. However, the story about scalable creation of language resources involves working with whoever is available. The tasks need to be locally comprehensible and motivating. In moving forward, we believe it is necessary to rethink the collaborative transcription task. The starting point is to understand local participants as teachers and cultural guides, occupied with their own knowledge practices and with passing these on. Special focus need to be given on the creation of a third space between the several stakeholders of a project with benefits that serve both Indigenous participants and external actors (Bird, 2020a). Could we view the task of putting an audio recording into textual form as a way to help a newcomer make progress with the language and culture, and with getting the pronunciations and meanings correct? The answer to this question depends on further research.

6 Conclusion

Outside the major languages, the development of language technologies is considered to be held up by the general lack of data (Krauwer, 2003). In the case of the world's small, oral languages, the usual approach has been to follow the long-established practice of linguists and record and transcribe audio and elicit wordlists and paradigms. Many computational tools were developed to support this approach. However, algorithmic approaches to working with small languages do not need to be limited by these past practices, and so we believe it is worth considering other approaches to data collection that might simultaneously support computational methods while engaging effectively with members of the speech community.

Accordingly, we took a recently proposed approach to transcription based on keyword spotting, and developed an app for confirming system guesses. We anticipated that this app would be more accessible to local participants than the conventional linguist-driven tasks. We ran trials in two Aboriginal towns, with speakers of the Kunwok language.

In this paper, we report the description of the several interactions we had with locals around a lexical verification activity. We present the many challenges we encountered, including a reflection around the technical and cultural issues of the task design, and the flaws around our general approach in terms of collaborative language work.

For the present, we offer our findings as a candid report on the experience of deploying data capture technology in an Indigenous community, in the hope that others will succeed where we have failed. We hope others will also follow our lead and share their own experiences of data collection, and make visible more of the real work of NLP (cf. Star, 2007). Perhaps it is possible for an externally-defined task such as transcription to be aligned to local agendas. Just as often, we expect that it will be necessary to let go of such tasks and do something different. Something that makes sense locally.

Acknowledgements

This research was covered by a research permit from the Northern Land Council, and ethics approved from Charles Darwin University. We are grateful to the Australian government for a PhD scholarship to the first author, and for grants from the Australian Research Council and the Indigenous Language and Arts Program to the second author.

The recruitment of participants was done with the support of the local organisation: Injalak Arts and Craft in Gunbalanya, and Warddeken Land Management in Manmoyi. The shape and purpose of the work was explained in English and oral consent has been obtained by all the participants. Additional approval has been given by Manmoyi traditional owners, concerning the collection and use of the data. All the participants have been paid at the regular rate for Aboriginal people consultancy. We would like to thank Mat Bettinson for his involvement in the design of the lexical verification App and Joshua Yang for the video recording of the trials.

References

Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Diane August, Timothy Shanahan, and Kathy Escamilla. 2009. English language learners: Developing literacy in second-language learners—report of the national literacy panel on language-minority children and youth. *Journal of Literacy Research*, 41:432–452.

Mat Bettinson and Steven Bird. 2017. Developing a suite of mobile applications for collaborative language documentation. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 156–164.

Mat Bettinson and Steven Bird. 2021. Collaborative fieldwork with custom mobile apps. *Language Documentation & Conservation*, 15:411–432.

Steven Bird. 2020a. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519.

Steven Bird. 2020b. Sparse transcription. *Computational Linguistics*, 46:713–744.

Steven Bird, Florian Hanke, Oliver Adams, and Haejoong Lee. 2014. Aikuma: A mobile app for collaborative language documentation. In *Proceedings of the Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 1–5. ACL.

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

Luc Bouquiaux and Jacqueline M. C. Thomas. 1992. *Studying and describing unwritten languages*. Dallas: Summer Institute of Linguistics.

Jonathan Clark, Robert Frederking, and Lori Levin. 2008. Toward active learning in data selection: Automatic discovery of language features during elicitation. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*.

Christopher Cox, Gilles Boulianne, and Jahangir Alam. 2019. Taking aim at the 'transcription bottleneck': Integrating speech technology into language documentation and conservation. Paper presented at the 6th International Conference on Language Documentation and Conservation, https://instagram.com/p/Buho4Z0B7xT/.

Lise M Dobrin, Peter K Austin, and David Nathan. 2009. Dying to be counted: The commodification of endangered languages in documentary linguistics. *Language Documentation and Description*, 6:37–52.

Nicholas Evans et al. 2003. *Bininj Gun-wok: a pandialectal grammar of Mayali, Kunwinjku and Kune*. Pacific Linguistics, Research School of Pacific and Asian Studies, The Australian National University.

Erika Feinauer, Kendra M Hall-Kenyon, and Kimberlee C Davison. 2013. Cross-language transfer of early literacy skills: An examination of young learners in a two-way bilingual immersion elementary school. *Reading Psychology*, 34:436–460.

Ben Foley, Josh Arnold, Rolando Coto-Solano, Gautier Durantin, T. Mark Ellison, Daan van Esch, Scott Heath, František Kratochví, Zara Maxwell-Smith, David Nash, Ola Olsson, Mark Richards, Nay San, Hywel Stoakes, Nick Thieberger, and Janet Wiles. 2018. Building speech recognition systems for language documentation: The CoEDL Endangered Language Pipeline and Inference System. In *Proceedings of the 6th International Workshop on Spoken Language Technologies for Under-Resourced Languages*, pages 205–209. ISCA.

Reece George, Keith Nesbitt, Patricia Gillard, and Michael Donovan. 2010. Identifying cultural design requirements for an australian indigenous website. In *Proceedings of the Eleventh Australasian Conference on User Interface-Volume 106*, pages 89–97.

Barbara Glowczewski. 1989. *A topological approach to Australian cosmology and social organisation*, volume 19. Proquest Social Sciences Journals.

Florian Hanke. 2017. *Computer-Supported Cooperative Language Documentation*. Ph.D. thesis, University of Melbourne.

Dianna Hardy, Elizabeth Forest, Zoe McIntosh, Trina Myers, and Janine Gertz. 2016. Moving beyond "just tell me what to code" inducting tertiary ict students into research methods with aboriginal participants via games design. In *Proceedings of the 28th Australian Conference on Computer-Human Interaction*, pages 557–561.

John Hatton. 2013. SayMore: Language documentation productivity. Paper presented at the Third International Conference on Language Documentation and Conservation, http://hdl.handle.net/10125/26153.

Mary Hermes and Mel Engman. 2017. Resounding the clarion call: Indigenous language learners and documentation. *Language Documentation and Description*, 14:59–87.

Steven Krauwer. 2003. The Basic Language Resource Kit (BLARK) as the first milestone for the Language Resources Roadmap. *Proceedings of the International Workshop on Speech and Computer*, pages 8–15.

Eric Le Ferrand and Steven Bird. 2020. Enabling interactive transcription in an indigenous community. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3422–3428.

Anany Levitin. 1999. Do we teach the right algorithm design techniques? In *The Proceedings of the Thirtieth SIGCSE Technical Symposium on Computer Science Education*, pages 179–183.

Daria Loi, Christine T Wolf, Jeanette L Blomberg, Raphael Arar, and Margot Brereton. 2019. Co-designing AI futures: Integrating AI ethics, social computing, and design. In *Companion Publication of the 2019 on Designing Interactive Systems Conference 2019 Companion*, pages 381–384.

Celeste Louro and Glenys Collard. 2021. Working together: Sociolinguistic research in urban aboriginal australia. *Journal of Sociolinguistics*.

Teresa Lynch and Shirley Gregor. 2004. User participation in decision support systems development: influencing system outcomes. *European Journal of Information Systems*, 13(4):286–301.

MA Maar, NE Lightfoot, ME Sutherland, RP Strasser, KJ Wilson, CM Lidstone-Jones, DG Graham, R Beaudin, GA Daybutch, BR Dokis, et al. 2011. Thinking outside the box: Aboriginal people's suggestions for conducting health studies with aboriginal communities. *Public Health*, 125(11):747–753.

Felicity Meakins, Jenny Green, and Myfany Turpin. 2018. *Understanding Linguistic Fieldwork*. Routledge.

Raghav Menon, Herman Kamper, Ewald van der Westhuizen, John Quinn, and Thomas Niesler. 2019. Feature exploration for almost zero-resource ASR-free keyword spotting using a multilingual bottleneck extractor and correspondence autoencoders. *Proceedings of Interspeech 2019*, pages 3475–3479.

Alexis Michaud, Oliver Adams, Trevor Cohn, Graham Neubig, and Séverine Guillaume. 2018. Integrating automatic transcription into the language documentation workflow: experiments with Na data and the Persephone Toolkit. *Language Documentation and Conservation*, 12:481–513.

Robyn Ober. 2017. Kapati time: storytelling as a data collection method in indigenous research. *Mystery Train*, 2007.

Dorian Peters, Susan Hansen, Jenny McMullan, Theresa Ardler, Janet Mooney, and Rafael A Calvo. 2018. "participation is not enough" towards indigenous-led co-design. In *Proceedings of the 30th Australian Conference on Computer-Human Interaction*, pages 97–101.

Nay San, Martijn Bartelds, Mitchell Browne, Lily Clifford, Fiona Gibson, John Mansfield, David Nash, Jane Simpson, Myfany Turpin, Maria Vollmer, et al. 2021. Leveraging neural representations for facilitating access to untranscribed speech from endangered languages. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU)*.

Frank Seifart, Harald Hammarstroöm, Nicholas Evans, and Stephen C. Levinson. 2018. Language documentation twenty-five years on. *Language*, 94:e324–45.

Han Sloetjes, Herman Stehouwer, and Sebastian Drude. 2013. Novel developments in Elan. Paper presented at the Third International Conference on Language Documentation and Conservation, http://hdl.handle.net/10125/26154.

Alessandro Soro, Margot Brereton, Jennyfer Lawrence Taylor, Anita Lee Hong, and Paul Roe. 2017. A cross-cultural noticeboard for a remote community: design, deployment, and evaluation. In *IFIP Conference on Human-Computer Interaction*, pages 399–419. Springer.

Susan Leigh Star. 2007. Living grounded theory: Cognitive and emotional forms of pragmatism. *The Sage Handbook of Grounded Theory*, pages 75–94.

Jennyfer Lawrence Taylor, Wujal Wujal Aboriginal Shire Council, Alessandro Soro, Michael Esteban, Andrew Vallino, Paul Roe, and Margot Brereton. 2020. Crocodile language friend: Tangibles to foster children's language use. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Mareese Terare and Margot Rawsthorne. 2020. Country is yarning to me: Worldview, health and well-being amongst australian first nations people. *The British Journal of Social Work*, 50(3):944–960.

Bert Vaux and Justin Cooper. 1999. *Introduction to Linguistic Field Methods*. Lincom Europa.

Helen Verran. 2007. The educational value of explicit non-coherence. *Education and Technology: Critical Perspectives, Possible Futures*, pages 101–124.

Retrospective view

We faced much criticism during the process of publication of this paper, including, during the first round of reviews, comments about its alleged lack of ethics. The encountered failures have been attributed to shortcomings in the design of the methods, such as inadequate consideration of contextual factors, as well as suboptimal communication with the community, which, according to the reviewers, could have been mitigated by engaging the services of an interpreter. The discomfort reported by community members has been regarded as an ethical lapse, stemming from a failure to convey the full scope of the project to stakeholders in a clear and transparent manner. While writing this paper, I was sceptical about the relevance of such work as community engagement and community-based practices have been explored in other fields, such as HCI, linguistics, or even public health. This failure could have, indeed, simply been the result of a shallow scientific approach and a poor design of the app and activities. Yet, as seen in Section 2.4.1, the extent of the participation of the community in projects involving co-design is often unclear, and the experience we went through was probably not avoidable. What I have been calling failure was an excellent opportunity to learn about the community. The audio used during the activity allowed us to initiate conversations and learn more about the local culture.

The footage made of the interactions also turned out to be a great help to our understanding of the community's ways of knowing. Descriptions of the Aboriginal epistemology can be very technical for non-experts (cf. Foley, 2003; West, 1998). Moreover, related work is very theoretical, making real-life applications of this knowledge difficult to envisage.

3.4 ALTA 2021: towards more flexible spoken term detection

Éric Le Ferrand, Steven Bird and Laurent Besacier. Phone Based Keyword Spotting for Transcribing Very Low Resource Languages. In *Proceedings of the 19th Annual Workshop of the Australasian Language Technology Association (ALTA 2021)*. pp. 79–86.

Research process

Several lessons have been learnt from the deployments of the lexical verification app. We first started to reflect on the technical side and on the obstacles linked to DTW. We needed to find a method that was more precise, more robust to speaker variability, but also quicker to process and more flexible in terms of the needed data. Several phone recognition systems based on multilingual pre-training were released during the course of this project including Allosaurus (Li et al., 2019). Allosaurus allowed us first to take advantage of the data available to fine-tune the model and, secondly, to take advantage of the direct mapping between phones and graphs. We initially compared the results of our approach with the results obtained in Section 3.2, arguing that the proposed methods outperformed MFCCs. The release of the Shennong Library (Bernard et al., 2021), which made possible the extraction of multilingual bottleneck features, allowed us to compare our method with another multilingual approach. While the difference of performance was not outstanding, it was interesting to see the difference of benefits between the two approaches and their complementarity.

Phone Based Keyword Spotting for Transcribing Very Low Resource Languages

Éric Le Ferrand,[1,2] **Steven Bird,**[1] **and Laurent Besacier**[2]
[1]Northern Institute, Charles Darwin University, Australia
[2]Laboratoire Informatique de Grenoble, Université Grenobles Alpes, France

Abstract

We investigate the efficiency of two very different spoken term detection approaches for transcription when the available data is insufficient to train a robust speech recognition system. This work is grounded in a very low-resource language documentation scenario where only a few minutes of recording have been transcribed for a given language so far. Experiments on two oral languages show that a pre-trained universal phone recognizer, fine-tuned with only a few minutes of target language speech, can be used for spoken term detection through searches in phone confusion networks with a lexicon expressed as a finite state automaton. Experimental results show that a phone recognition based approach provides better overall performances than Dynamic Time Warping when working with clean data, and highlight the benefits of each methods for two types of speech corpus.

1 Introduction

Efforts are made across Australia to preserve, document and revitalize Aboriginal languages. These languages exist primarily in spoken form, and even if there often is an official orthography available, it is not widely used by local people. Making recordings of speakers has been a widespread practice for documenting traditional knowledge. However, such recordings are often not transcribed, making them hard to access.

Manual transcription is time consuming and is often described as a bottleneck (Brinckmann, 2009). While automatic speech recognition (ASR) has seen great improvements in recent years (Povey et al., 2011; Watanabe et al., 2018), it relies on a large amount of annotated data. Attempts to build ASR systems for low-resource languages end up with high word error rate or single-speaker models making them of limited use in Indigenous contexts (Gupta and Boulianne, 2020a,b).

Such methods assume that everything should be transcribed. Bird (2020) describes a sparse transcription model where we only transcribe the words we can confidently recognize, using word-spotting, while leaving the transcription of more difficult sections for later, perhaps when a speaker is available (Bird, 2020). Based on this model, Le Ferrand et al. (2020) proposed a workflow which combines spoken term detection and a human-in-the-loop to support transcription in under-resourced settings. Such a workflow avoids the use of a language model which requires too much textual data, data that we cannot find in most Aboriginal contexts, and which only needs a few spoken terms to be annotated. While they show through their simulation the capability of iterative transcription in remote communities, the precision of their method depends on the quality of the spoken queries, and the density of the resulting transcription is limited by the size of the lexicon.

Automatic phone recognition has seen progress with minimal data (Gupta and Boulianne, 2020b; Li et al., 2020). While Bird (2020) argues that phonetic transcriptions do not stand in for the speech data and cannot be segmented to generate the required higher-level word units, we can nevertheless view phone transcriptions as a speech encoding, retaining our commitment to the sparse transcription model. Such an approach has an advantage over traditional query-by-example methods in that a simple word list can be used instead of a spoken lexicon which can be challenging to collect. In this paper we show how this can be done, and compare it with dynamic time warping (DTW) (Sakoe and Chiba, 1978) commonly used for keyword spotting for Indigenous languages. We consider both methods as applied to two very low-resource languages, Kunwinjku (gup) spoken in the far north of Australia and Mboshi (mdw) spoken in Congo Brazzaville.

2 Background

Traditional ASR systems are not well suited to Aboriginal languages. The lack of data for such languages does not allow us to train an acoustic model or a language model. Additionally, the type of data usually recorded is often spontaneous and noisy which makes it difficult to transcribe, regardless of the amount of annotated data available.

Bird (2020) describes the sparse transcription model, which combines spoken term detection with a human-in-the-loop, in an iterative process. Using spoken term detection as a transcription method allows us to avoid traditional components of an ASR system, specifically the language model, to focus only on the recognition of isolated words.

Traditional Spoken Term Detection systems rely on text-based search in lattices extracted from ASR systems (Lleida et al., 2019; Saraclar and Sproat, 2004). Attempts to train ASR systems in low-resource contexts have so far provided poor results for single speaker systems (Gupta and Boulianne, 2020a,b). This makes traditional spoken term detection approaches questionable in very low-resource settings. A few papers linked to the Babel Project have explored lattice search using ASR systems trained in low-resource settings (Gales et al., 2014; Rosenberg et al., 2017). However, they work with much larger data collections than what is available in Indigenous contexts.

Query-by-Example methods have been preferred in very low-resource contexts since they only rely on acoustic comparison between spoken queries and utterances. Le Ferrand et al. (2020) explore feature representation using DTW in an iterative pipeline following the sparse transcription model (Bird, 2020), and have been able to transcribe up to 42% of a lexicon in their speech collections. This method, however, has shown limitations in terms of robustness in the face of speaker variability. Research around speech features for spoken term detection has explored the use of bottleneck features, or the hidden representation of an auto-encoder (Menon et al., 2019; Kamper et al., 2015, 2020). Such research highlights the benefits of multilingual approaches for spoken term detection when transcribed data are limited in the target language. Others have exploited neural approaches to train word classifiers from word pairs using a Siamese loss (Settle and Livescu, 2016; Settle et al., 2017), however pairs of words are required, limiting the selection to words that can be searched.

Query-by-example relies on a spoken lexicon and, by extension, a comparison between two acoustic vectors. A difference of speakers or recording channel between the query term and the speech collection has an influence on the likelihood of a given term to be retrieved. Moreover, a spoken lexicon is not simple to gather and this therefore limits the amount of terms we can retrieve. Using a lexicon made of terms recorded in isolation for spoken term detection purposes will lead to poor precision. Another solution would be to manually extract the terms of the lexicon from a speech collection which is time-consuming. Phone recognizers, like ASR systems, also need a few hours of annotated speech to provide acceptable performance (Gupta and Boulianne, 2020b; Adams et al., 2018). However, recent work has shown how multilingual phone recognizers can be fine-tuned with minimal data to work on a new language (Li et al., 2020). Raw phone transcriptions are hard to obtain as they require the skills of a trained linguist, and they cannot help directly for retrieving higher level-units (Bird, 2020). However, the orthography of most Indigenous languages is based on their phonology and there is usually a simple mapping from graphemes to phonemes can be obtained to train a phone recognizer, even with a shallow knowledge of the phonology. A spoken term detection method based on a phone recognizer could allow us to rely only on written queries following a traditional lattice-search method.

3 Methods

We begin with a lexicon of size s consisting of audio clips of spoken words, along with orthographic transcriptions, plus a speech collection in which more instances of those words may be found.

Two spoken term detection approaches, involving a multilingual component, are investigated here: (a) a baseline method based on DTW applied on multilingual BottleNeck Features (mBNF); and (b) a method based on a textual search in phone confusion networks extracted from a universal phone recognizer (P2W).

3.1 Baseline: Sparse Transcription using DTW

We first extract acoustic features from both the corpora and lexicons. Based on general performance scores reported in the literature, and in order to compare our method with another multilingual ap-

proach, we have chosen multilingual bottleneck features. These are extracted from a model trained on the Babel corpus and consist of 80 dimension acoustic vectors. They have been extracted with the Shennong library.[1] We slide each term of the lexicon along the utterances of the corpus with a step size of 30 milliseconds. We then select the best matches for each utterance-word pair based on DTW distance and retain all matches above a threshold *m* for evaluation.

3.2 Sparse Transcription using Phone Recognition (P2W)

Li et al. (2020) introduced *Allosaurus*, a universal phone recognition system which combines a language independent encoder and phone predictor, and a language dependent allophone layer with a loss function, associated with each language (Fig. 1). Allosaurus models are trained using standard phonetic transcriptions and the *allovera* database (Mortensen et al., 2020), a multilingual allophone database that can be used to map allophones to phonemes. The model first encodes speech with a standard ASR encoder which computes the universal phone distribution. Then an allophone layer is initialized with the allophone matrix and maps the universal phone distribution into the phoneme distribution for the given target language. The resulting model can be fine-tuned and applied to unseen languages.

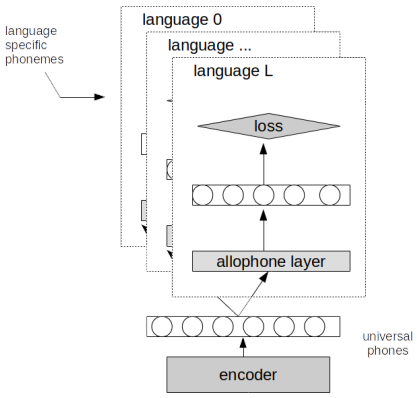


Figure 1: Allosaurus model (Li et al., 2020)

In the current context, since we only have an orthographic transcription for Kunwinjku, we transliterate it into IPA with the mapping shown in Table 1. The transcription contains some English words which will be mapped as if they were Kunwinjku words (e.g., school is written /s?kool/ instead of /skʊl/). For Mboshi, the orthographic transcription already mostly matches the corresponding phonetic transcription.[2]

graphs	a	b	d	h	e	i	ch	y	o	k	dj	s	r	rr
phones	ɑ	b	d	ʔ	ɛ	i	ʃ	j	ɔ	k	ɟ	s	ɻ	r
graphs	ng	rd	rl	nj	rn	u	f	l	m	n	w	p	t	
phones	ŋ	ɖ	ɭ	ɲ	ɳ	u	f	l	m	n	w	p	t	

Table 1: Grapheme to phoneme mapping for Kunwinjku

We fine-tuned the original pretrained model with the training and validation subsets described in Section 4 following the mapping described above, resulting in one new phone recognition model per language. We used the resulting models to automatically extract phones in confusion networks from the validation and test sets of the two languages (Mboshi and Kunwinjku) (Fig. 3).

The graph extracted is a confusion network (confnet) and consists of a size *s* sequence of phones and the top *k* likely alternatives for each phone (see Fig. 3). For each phone in the graph a probability score between 0 and 1 is assigned. We also map the lexicons into phones and convert them into a finite state automaton (FSA) in which each final state corresponds to the end of a given word (Fig.2). We explore, in the confusion networks related to our collection, every path which corresponds to a valid transition in the FSA and has a probability strictly greater than zero. If a path reaches a terminal state in the FSA, we extract the word and a score corresponding to the mean of the accumulated likelihood scores. Like the baseline with DTW, we then select the best match for each pair utterance/word pairs based on the likelihood score and keep for evaluation the matches above a threshold *n*. For both systems, we do not keep for evaluation the pairs which correspond to the query instances used to build the lexicons.

4 Data

We are using a corpus of spontaneous speech in Kunwinjku built from several sources. The training, validation and test set are described in Table 2. The training and validation sets are built from transcribed recordings made for language descrip-

[1] https://docs.cognitive-ml.fr/shennong/

[2] The tones are marked in the orthographic transcription but this feature is not taken into account in the Allosaurus model. We thus decided to treat the orthographic transcription as a phonetic transcription so the accentuated vowels are considered as new phones.

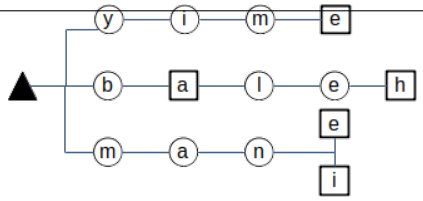


Figure 2: Example of lexicon converted into a FSA

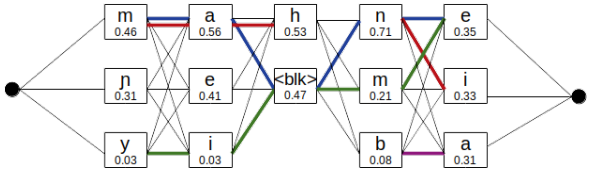


Figure 3: Example of search in a graph confusion network

tion purposes around language and emotion. They also contain some recording of guided tours of an Aboriginal town. The test set contains exclusively guided tour recordings. The orthographic transcription has been force-aligned using the MAUS forced aligner (Kisler et al., 2017). The train and validation sets contain the same 5 speakers and the test set has a non-overlapping set of 5 speakers.

We are also using a corpus of Mboshi speech which consists of 4.5 hours of speech elicited from text with orthographic transcription and a forced alignment at the word level (Godard et al., 2017). Training, validation and test sets have been extracted from the corpus and are described in Table 2. The same three speakers are represented among the three partitions.

The lexicon queries (for spoken term detection) are made of 100 words for Mboshi and 60 words for Kunwinjku. We randomly selected in the test set words which occur at least 3 times in the corresponding corpus. For each word, we manually selected examples clearly pronounced, respecting the speaker distribution of the test set (Table 3 and 4), and clipped them out.

Partitions	train	valid	test
Kunwinjku	35min45	7min39	19min43
Mboshi	21min10	10min03	3h56min

Table 2: Partition duration

Speaker	RB	TG	GN	SG	MM
Distribution	10%	25%	15%	38%	12%

Table 3: Speaker distribution across Kunwinjku lexicon

Speaker	AB	KO	MA
Distribution	63%	33%	4%

Table 4: Speaker distribution across Mboshi lexicon

5 Results

5.1 Phone Error Rate (PER)

We first evaluate the PER for both languages on the validation set. For Kunwinjku the PER started at 55.45%, and we obtained 38.82% after the system early stopped at the 24th epoch. For Mboshi the PER started at 59% and reached 38.72% at the 29th epoch. Although the PER is low considering the small amount of data used for fine-tuning Allosaurus, we would expect a bigger difference between Kunwinjku and Mboshi considering that Mboshi is read speech without foreign words and Kunwinjku is spontaneous speech containing English words. To estimate the performances for each language, we computed the PER on the test set between the top 1 phones generated by Allosaurus and the gold standard. For Kunwinjku the PER is at 39% and for Mboshi at 44%.

5.2 System performances

We evaluate the proposed methods using precision, recall and F-score.

We provide for each language the scores based on a threshold that is optimized on the respective validation sets. For the P2W method, the optimized threshold is set at 0.77 for Kunwinjku and 0.631 for Mboshi. For the DTW baseline, it is set at 0.217 for Kunwinjku and 0.174 for Mboshi. The results are detailed in Table 5. In Mboshi, the method outperforms the baseline with DTW with recall and precision. In Kunwinjku, the method does not outperform the baseline in terms of F-scores. We can see that while the baseline brings more candidates than P2W, our method is more precise. While it is clear that a phone recognition based method provides better overall performance on clean speech, the gap between the F-scores of each method in Kunwinjku is small which can make them both beneficial.

The Kunwinjku corpus contains spontaneous speech. We can observe elision phenomenon and fast speech which are not well supported by an approach based on recognition of canonical, lexical phone sequences. Figures 4 and 5 show that, while our approach seems to be more consistent across

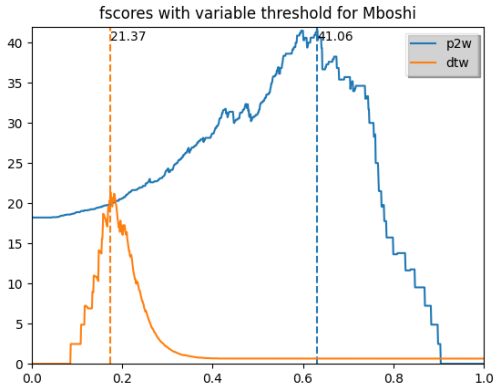


Figure 4: F-scores for Mboshi with variable thresholds on validation set

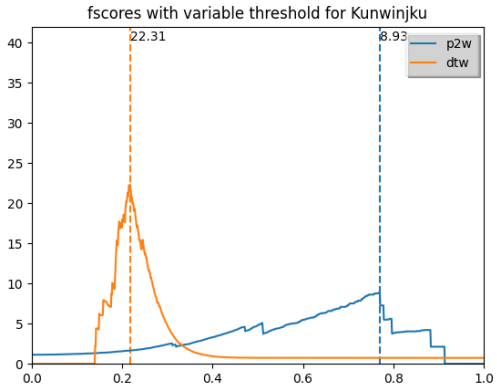


Figure 5: F-scores for Kunwinjku with variable thresholds on the validation set

	recall	precision	F-score
DTW_mb	14.55%	20.46%	17.01%
P2W_mb	22.61%	45.97%	30.31%
DTW_kun	42.09%	22.81%	29.59%
P2W_kun	17.41%	62.50%	27.23%

Table 5: Performance of spoken term detection on the test set with the optimized threshold

thresholds, it is less efficient than DTW for noisy and spontaneous speech corpora.

We present in Table 6 the top 5 false positives across methods and languages. We could only report the top 4 for P2W in Kunwinjku since most of the errors were isolated cases. We can see for P2W that the errors are made between very similar words. For Mboshi, the top 5 only includes tonal differences between the query and the hit. For Kunwinjku, the errors are made between similar words, some of which are morphologically related (balanda (man), balandaken (of the man); karrire (we-INCL go), ngarrire (we-EXCL go)). For DTW, the errors are not as consistent and the hits seem to only match subparts of the query terms (wa, **wá**re; marn**bolh**, bonj).

5.3 Speaker analysis

Le Ferrand et al. (2020) pointed out the limitation of their method in terms of cross speaker spoken term detection. To compare the two approaches on this aspect, we analyze each true positive that is output by each system: we check if the word matched is pronounced by a same or different speaker that the query term. Even if we only use the written forms of the queries for P2W, we also make the same analysis.

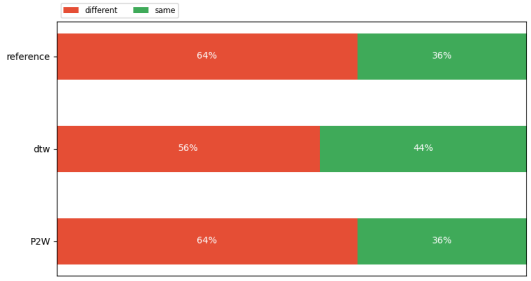


Figure 6: Proportion of same-speaker/different-speaker retrieval in Kunwinjku

Figure 7: Proportion of same-speaker/different-speaker retrieval in Mboshi

Figures 6 and 7 present the proportion of spoken terms retrieved from same-speaker or different-speaker For a fair comparison, we also compute the distribution of same/different speaker between the lexicons and all the words to be retrieved in the corpora (reference). We can see that P2W method follows the general distribution in the corpora while the baseline DTW retrieves mostly terms pronounced by the same speaker.

Mboshi P2W		Kunwinjku P2W		Mboshi DTW		Kunwinjku DTW	
Query	Hit	Query	Hit	Query	Hit	Query	Hit
ádzá	ádza	balanda	balandaken	abvúa	wa	munguyh	bonj
ádzá	adzá	birrimarnbom	birrimanbun	mwána	wa	kahdi	konhda
ngala	ngalá	mani	yiman	mvúá	wa	kunak	konhda
ngaa	ngáa	karrire	ngarrire	wáre	wa	kunred	konhda
okándá	ωkándá			ngaa	ngá	marnbolh	bonj

Table 6: Top 5 false positives

6 Combining the methods

We mentioned in Section 2 that DTW and P2W each have their own strengths. As we know, DTW will cope more easily with spontaneous speech and co-articulation effects such as assimilation and elision. Phone recognition allows us to avoid gathering spoken queries and retrieving terms with exact matching between written forms. To highlight the complementarity of the methods, we analyse the intersection of their true positives in Figure 8. We show that across both corpora the intersection of the true positives is small, and so combining the two methods can help us increase the coverage of the transcription to reach up to 49.99% for Kunwinjku and 32.16% for Mboshi.

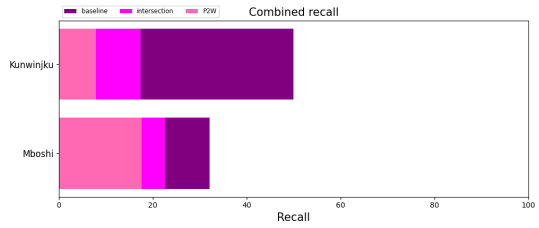


Figure 8: Relative coverage of the combined methods

We analysed the most common terms retrieved by DTW which have been ignored by P2W. For Kunwinjku, the glottal stop and doubled consonant are the phones the least properly recognized (*wanjh* written *wanj kunwardde* written *kunwarde* for example). More generally, since the data used in Kunwinjku is spontaneous speech, most of the missed hits by P2W are due to highly mistaken phone transcriptions by allosaurus. For Mboshi, beyond the main easily-confusable phones (*o* / *ω*, *e* / *ϵ* for instance) the main missed hits are due to tones or long vowels not being correctly recognized.

The baseline provides a match for every utterance/query pair if no threshold is applied. However, since P2W is restricted by the phones output by the phone recognizer, we have a limited amount of candidates regardless of the threshold. As mentioned before, this has the advantage of being more precise, but can easily miss a match if the phone lattices contain many mistakes. In view of this, we combine the two methods as follows. For each utterance/query pair brought by P2W, we first keep for evaluation the candidates which have a score greater than the P2W threshold. Then we send to evaluation every pair having a distance less than the DTW threshold. We provide in Table 7 the results for the same optimized thresholds mentioned before.

	recall	precision	F-score
comb_mb	24.89%	45.54%	32.19%
P2W_mb	22.61%	45.97%	30.31%
comb_kun	35.76%	31.48%	33.48%
P2W_kun	17.41%	62.50%	27.23%

Table 7: Performance of the combined methods

The described way of combining the methods outperforms both P2W and DTW approaches in terms of F-score. For Mboshi, we can observe a small increase of the recall with a precision barely affected. For Kunwinjku, the results are less clear. While the F-score outperforms both the baseline and P2W, combining the methods double the recall but decreases by half the precision.

7 Conclusion

This paper compares two methods of spoken term detection, one based on DTW with bottleneck features, and one based on on phone recognition. Both methods have been applied on two very low-resource languages, namely, a corpus in Mboshi recorded in a controlled environment, and a corpus of spontaneous speech in Kunwinjku recorded in remote communities. Experimental results shown that a few minutes of transcribed speech can be

used to fine-tune a universal phone recognizer. Then searching terms in a confusion network with a lexicon expressed as a FSA outperforms the baseline for Mboshi but not for Kunwinjku.

A text-based approach has the advantage over traditional Query-by-example that a set of written queries is easier to gather than spoken queries. Further analysis has shown that the proposed phone recognition approach is more robust to speaker variability and tends to be more accurate than DTW overall. However, the baseline seems to have a better coverage over the corpora and to be more suitable with noisy data.

One method relies on canonical orthography while the other relies on acoustic comparison. Both methods have their own benefits depending on the type of data they are applied to. Experimental results have shown that it is possible to take advantage of both methods to increase the overall recall while maintaining precision at an acceptable rate.

References

Oliver Adams, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird, and Alexis Michaud. 2018. Evaluating phonemic transcription of low-resource tonal languages for language documentation. In *LREC 2018 (Language Resources and Evaluation Conference)*, pages 3356–3365.

Steven Bird. 2020. Sparse transcription. *Computational Linguistics*, 46:713–744.

Caren Brinckmann. 2009. Transcription bottleneck of speech corpus exploitation. *Proceedings of the 2nd Colloquium on Lesser Used Languages and Computer Linguistics*, pages 165 – 179.

Mark JF Gales, Kate M Knill, Anton Ragni, and Shakti P Rath. 2014. Speech recognition and keyword spotting for low-resource languages: Babel project research at cued. In *Fourth International workshop on spoken language technologies for under-resourced languages (SLTU-2014)*, pages 16–23. International Speech Communication Association (ISCA).

Pierre Godard, Gilles Adda, Martine Adda-Decker, Juan Benjumea, Laurent Besacier, Jamison Cooper-Leavitt, Guy-Noël Kouarata, Lori Lamel, Hélène Maynard, Markus Müller, et al. 2017. A very low resource language speech corpus for computational language documentation experiments. *arXiv preprint arXiv:1710.03501*.

Vishwa Gupta and Gilles Boulianne. 2020a. Automatic transcription challenges for Inuktitut, a low-resource polysynthetic language. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2521–27.

Vishwa Gupta and Gilles Boulianne. 2020b. Speech transcription challenges for resource constrained indigenous language Cree. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-Resourced Languages and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 362–367.

Herman Kamper, Micha Elsner, Aren Jansen, and Sharon Goldwater. 2015. Unsupervised neural network based feature extraction using weak top-down constraints. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5818–22. IEEE.

Herman Kamper, Yevgen Matusevych, and Sharon Goldwater. 2020. Multilingual acoustic word embedding models for processing zero-resource languages. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6414–6418. IEEE.

Thomas Kisler, Uwe Reichel, and Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech and Language*, 45:326–347.

Eric Le Ferrand, Steven Bird, and Laurent Besacier. 2020. Enabling interactive transcription in an indigenous communitỳ. In *COLING 2020*.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R Mortensen, Graham Neubig, Alan W Black, et al. 2020. Universal phone recognition with a multilingual allophone system. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8249–8253. IEEE.

Eduardo Lleida, Alfonso Ortega, Antonio Miguel, Virginia Bazán-Gil, Carmen Pérez, Manuel Gómez, and Alberto De Prada. 2019. Albayzin 2018 evaluation: the iberspeech-rtve challenge on speech technologies for spanish broadcast media. *Applied Sciences*, 9(24):5412.

Raghav Menon, Herman Kamper, Ewald van der Westhuizen, John Quinn, and Thomas Niesler. 2019. Feature exploration for almost zero-resource ASR-free keyword spotting using a multilingual bottleneck extractor and correspondence autoencoders. *Proceedings of Interspeech 2019*, pages 3475–3479.

David Mortensen, Xinjian Li, Patrick Littell, Alexis Michaud, Shruti Rijhwani, Antonios Anastasopoulos, Alan Black, Florian Metze, and Graham Neubig. 2020. Allovera: a multilingual allophone database. In *LREC 2020: 12th Language Resources and Evaluation Conference*.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The kaldi speech recognition toolkit.

In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

Andrew Rosenberg, Kartik Audhkhasi, Abhinav Sethy, Bhuvana Ramabhadran, and Michael Picheny. 2017. End-to-end speech recognition and keyword search on low-resource languages. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5280–5284. IEEE.

Hiroaki Sakoe and Seibi Chiba. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26:43–49.

Murat Saraclar and Richard Sproat. 2004. Lattice-based search for spoken utterance retrieval. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 129–136.

Shane Settle, Keith Levin, Herman Kamper, and Karen Livescu. 2017. Query-by-example search with discriminative neural acoustic word embeddings. *Proc. Interspeech 2017*, pages 2874–2878.

Shane Settle and Karen Livescu. 2016. Discriminative acoustic word embeddings: Tecurrent neural network-based approaches. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 503–510. IEEE.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson-Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, et al. 2018. Espnet: End-to-end speech processing toolkit. *Proc. Interspeech 2018*, pages 2207–2211.

Retrospective view

While the proposed method did not provide outstanding performance in terms of f-score, it did respond to our prerequisites. Our method is more precise than DTW, and more robust to speaker variability. It is more flexible as it does not require the prior collection of spoken queries (only written) and is quicker to process. The publication of related work, such as the use of a phone recogniser for the identification of named entity (Leong and Whitenack, 2022), or the creation of a toolkit of the conversion between phones and graphs (Pine et al., 2022a) also gave credibility to our approach.

The evaluation of the paper primarily focused on statistical evaluation, which occluded the main benefit of the proposed method: its flexibility. First, it allowed us to properly use the available data to train our model (i.e. a very small amount of spontaneous narrative transcribed at the sentence level). The size of our spoken lexicon no longer limited us, and we no longer needed to extract the queries for pre-recorded utterances or record them manually. A simple written request was enough. Then, once the model was trained (which did not take longer than a day), spotting terms in the test sets was much faster than DTW. This difference allowed us to directly incorporate the spoken term detection step inside the activity and obtain instant results (Section 3.5).

In terms of precision and recall, a recent publication appears to highly outperform our method (Macaire et al., 2022). As with our method, their amount of training data was small, and their retrieval mechanism relied on written queries. While their experimental setup is bounded to creole languages, it would be interesting to see if such results are reproducible with the data we used and what processing time is to be expected.

3.5 COLING 2022: second attempt participatory language development

Éric Le Ferrand, Steven Bird and Laurent Besacier. Fashioning Local Designs from Generic Speech Technologies in an Australian Aboriginal Community. *Proceedings of the 29th International Conference on Computational Linguistics (COLING 2022)*. pp. 4274–4285, Gyeongju, South Korea

Research process

This final paper presents the conclusions drawn from the lessons learned in Section 3.3 and the potential offered by the spoken term detection method proposed in Section 3.4. Our goal was to enable community-based language documentation using spoken term detection to overcome possible issues with literacy. As a spoken lexicon was no longer necessary, I opted for syllable spotting to avoid over-interpreting the isolated words observed in Section 3.3. Constructing a spoken lexicon would have been a significant investment of time, and there was no guarantee of success. Since we were uncertain whether the syllable-spotting approach would be well-received, we designed a basic interface (with Tkinter[1]) for the pilot study.

[1] https://docs.python.org/fr/3/library/tkinter.html

Fashioning Local Designs from Generic Speech Technologies in an Australian Aboriginal Community

Éric Le Ferrand,[1,2] **Steven Bird,**[1] **and Laurent Besacier**[2]
[1]Northern Institute, Charles Darwin University, Australia
[2]Laboratoire Informatique de Grenoble, Université Grenoble Alpes, France

Abstract

An increasing number of papers have been addressing issues related to low-resource languages and the transcription bottleneck paradigm. After several years spent in Northern Australia, where some of the strongest Aboriginal languages are spoken, we could observe a gap between the motivations depicted in research contributions in this space and the Northern Australian context. In this paper, we address this gap in research by exploring the potential of speech recognition in an Aboriginal community. We describe our work from training a spoken term detection system to its implementation in an activity with Aboriginal participants. We report here on one side how speech recognition technologies can find their place in an Aboriginal context and, on the other, methodological paths that allowed us to reach better comprehension and engagement from Aboriginal participants.

1 Introduction

A consistent theme in recent NLP research has been *doing more with less* (Wiesner et al., 2022; Gao et al., 2021; Baevski et al., 2021; Schneider et al., 2019; Menon et al., 2019). It is popular to describe new pipelines to solve a wide range of tasks for under-resourced languages (Godard et al., 2018; Anastasopoulos et al., 2018; Settle et al., 2017; Mitra et al., 2016; Lane and Bird, 2019). However, the motivations behind the design of a computational method are not systematically well justified according to the needs of the target speech communities.

The category of under-resourced languages encompasses a wide range of contexts, not simply in terms of the quantity of data available but also in terms of local speech communities' sociolinguistic and political situation (Bird, 2022). Often, the focus has been to generalise a given method across languages, where the proposed system is at the core of the argument instead of the benefits that it could have for the speakers. We could ask whether the same language technology would be equally applicable to Marathi, spoken by millions in a major metropolis, and Miriwoong, with only a few elderly speakers in a remote Australian Aboriginal community (cf. Kuhn, 2022).

Universal solutions dominate NLP: research and results are often provided without taking into account the global situation of the languages involved or the views of the speech communities about the preservation of their language. Instead, it is common to assert that an improvement in Word Error Rate yielded by a given speech recognition system is the answer to the transcription bottleneck and, therefore, the problem of scaling up language documentation (van Esch et al., 2019; Foley et al., 2018).

Most of the world's languages are primarily oral (Ong, 1982; Walsh and Yallop, 1993). Writing is often not a priority, and very few people are skilled in transcribing their language. Written resources often only exist in limited spaces where there is a collaboration between westerners and local communities, such as schools, ranger programs, tourism, and academia. In such cases, writing would seem to primarily serve institutional agendas (cf. Dobrin et al., 2009; Perley, 2012; Nevins, 2013). Accordingly, we must ask ourselves to what extent automatic transcription technologies have a place in research that respects local self-determination. Bird (2022) calls for a *local turn*, for the need to work with local speech communities from the ground up. In other words, outsiders who enter communities with their expertise need to begin with local concerns and local knowledge practices, and only later begin to explore ways in which language technologies can be added into the mix. For example, a local person might want non-indigenous colleagues to learn and use the local language, rather than assuming that all work is conducted in English. We have found that such an approach enlarges the opportunities

for collaboration, while simultaneously generating language resources.

This paper extends our previous work on collaborative transcription (Le Ferrand et al., 2022), where the language documentation pipeline we designed failed. We were confronted with different ways of knowing and different expectations in terms of language work. In this work, learning from our past failure, we describe our approach, from the training of a transcription system to the design of collaborative transcription activities with Aboriginal participants. We first describe our speech recognition method based on syllable spotting. We then present the design of the app used that bridges the output of the syllable spotting system to the people, taking into account existing practices. We also explain our method to engage with participants to address their interests in terms of language work. Finally, we detail the application of the proposed transcription activities and discuss the success and flaws of this work.

2 Background

2.1 Decolonising practices

Research contributions around speech processing for low-resource languages have often followed the work of documentary linguistics, where some automation is added to support manual annotations (Adams et al., 2018; Godard et al., 2018; Foley et al., 2018). The 7000+ world languages are often mentioned and language technologies appear as a way to prevent their loss (Adda et al., 2016; Duong, 2017; Jimerson and Prud'hommeaux, 2018). Special workshops like the zero resource challenge[1] and the introduction of a surprise language have pushed in this direction allowing the creation of computational solutions that bypass the need of the speech communities of language experts. Recent studies have also shown that the languages (Schwartz, 2022) or the speech communities (Caselli et al., 2021) are rarely ate the core of the argument in the ACL anthology's publications.

Documentary linguistics is often the preliminary step of language description and analysis (Hanke, 2017). Documentation and description communicate with each other to allow western scholars to have a better comprehension of Indigenous languages. There are no clear benefits for the speech community, and extra work needs to be provided to share the benefits of a research project (Chelliah and De Reuse, 2010). The NHMRC Guidelines[2] for Ethical Conduct in research with Aboriginal and Torres Strait Islander Peoples and Communities set out principles of equity and reciprocity, where the outcome of the research should benefit both parties. Recent research practices, including documentary linguistics, started to fully commit to these standards by adopting a community-based approach (e.g. Rodríguez Louro and Collard, 2021; Ryder et al., 2021; Taylor et al., 2020). Community-based research has the community at its core and is meant to be conducted for and with the participation of community members (Rice, 2011).

2.2 Community-based projects

Community-based research around software design is a small but growing area. Projects have been based on research Human-Computer Interaction (HCI) or NLP from a language learning perspective. On the HCI side, research has contributed to responding to local issues by designing tools in collaboration with the community (Soro et al., 2017; Hardy et al., 2016; Leong et al., 2019). Cross-cultural collaboration is challenging. From this kind of project have also emerged engagement methods to facilitate the conversation with Indigenous communities about technology design (Zaman et al., 2016; Taylor et al., 2020). On the NLP side, the research contributions have been language-specific or bounded to a specific context. For instance, Pine et al. (2022) have described speech synthesis systems in several Indigenous Canadian languages responding to a call from the language learners. Projects that did not initially have a community-based component sometimes ended up serving community-based projects. Uí Dhonnchadha and Van Genabith (2006) for instance, created a POS tagger for gaellig Irish. The system has been then incorporated into an Irish learning game (Xu et al., 2022). In either case, the majority of the work done in this area is based on writing (e.g. Lane and Bird, 2019; Schwartz et al., 2019; Finn et al., 2022). The only speech-based projects are around speech synthesis (Harrigan et al., 2019; Pine et al., 2022). Speech recognition seems to be rarely involved in community-based projects.

[1] https://www.zerospeech.com/

[2] https://www.nhmrc.gov.au/about-us/resources/ethical-conduct-research-aboriginal-and-torres-strait-islander-peoples-and-communities#block-views-block-file-attachments-content-block-1

2.3 Context

Our work is grounded in Bininj country in West Arnhem, Northern Territory in the Australian Top End. Bininj country is part of the Indigenous Protected Area of Arnhem Land where the land and sea are managed by Aboriginal groups.[3] The main language of communication is Kunwinjku (ISO gup) which is spoken by approximately 2500 people (Marley, 2021). There is a standard orthography that has been introduced by linguists but it is not widely used by the members of the community.

The first and second authors have several years of experience with the Bininj community, have some expertise in Kunwinjku, the local language, and have both been adopted by Traditional owners of the land. In this case, adoption means the attribution of a *skin name* that connects an individual to the rest of the community (cf. Christie, 2008, p.35).

2.4 Learning from failure

This work is the continuation of Le Ferrand et al. (2022). We previously designed a spoken term detection prototype to detect whole words in untranscribed speech collections in Kunwinjku. We then used an app to bridge the output of our prototype to the people to allow local communities to verify the guesses of our system and therefore be part of transcription works. We faced many challenges that we tried to build on in this work.

This previous work focused on the collection of data to enhance the performance of the system. The design ended up being irrelevant and redundant for the participants. From here we realised the need for further discussion with the community to set up activities that are relevant to their agenda, interests and practices.

The app presented displayed only four buttons: one to play the query, one to play the utterance and two to give a feedback on whether the query has been spotted in the utterance or not. While testing the app, we realised how the audio files extracted from their contexts were confusing for the participants. Besides, the fact of validating system guesses in random utterances was disconnected from the idea of transcription which led most participants to overthink the task.

In projects around cross-cultural technology design, shallow information is provided about the extent of the collaboration and the challenges encountered. Yet, studies have described ways of knowing in Indigenous communities that differs from the western approach to knowledge (Descola, 2005; Foley, 2003). Such differences appear as the main reason behind the failed attempt of app design where the proposed task lose all meaning in Bininj context.

From our first failed attempt, the challenges were two-fold. We first needed to figure out a way to solve the comprehension issues we have faced. Then, we needed to improve the relevance of this work for Bininj participants. The key was to find out how to design transcription technologies based on existing practices. From the language learning sessions we had with some of our Aboriginal collaborators, we noticed, for instance, how they teach us breaking down words into syllables to decompose the pronunciation of a given item. This led us to think about replacing word spotting with syllable spotting, allowing participants to reproduce their word decomposition strategy to build up the transcription from the syllables spotted. From here, the focus needed to be given on incorporating this transcription strategy into an activity that matters to the people.

3 Transcription by syllables

3.1 Data

To build the system, we are using a corpus in Kunwinjku built from several sources. The training and validation sets consist of 35.45 min and 7.39 min respectively of spontaneous speech made of guided tours of Aboriginal towns and utterances for language description purposes. Two different sets are used for testing: one set of 19.43 min of spontaneous utterances and one set of 4.43 min of elicited words recorded in isolation.

To build our list of valid syllables, we used a word list built from the Bible in Kunwinjku. We then applied on each word syllable segmentation rules resulting in a set of 584 unique syllables with relative frequency values associated.

3.2 Experimental setup

Le Ferrand et al. (2021) introduced a method of spoken term detection for very low-resource languages based on phone recognition. Their method is based on Allosaurus (Li et al., 2020), a universal phone recognizer. We preferred this method in this work due to its flexibility in terms of query selection and its speed compared to Dynamic Time Warp-

[3] https://www.awe.gov.au/agriculture-land/land/indigenous-protected-areas

ing, which is usually used for very low-resource languages.

We first trained the phone recognizer using our train set and generated confusion matrix from the validation and test sets. A confusion matrix consists of a phone transcription and the top k (we use k=5) most likely alternatives per phone with a likelihood score associated. To spot syllables, we expressed the syllables extracted from the bible as a finite state automaton after conversion from graphs to phones and explored every possible path in the phone matrix that corresponded to a valid transition in the lexicon. Ultimately we extracted the resulting syllables with the mean of the phones' scores that are used as a likelihood measure to filter the syllables spotted based on a threshold *T*.

To increase the accuracy of the method of Le Ferrand et al. (2021) which only relies on the likelihood scores output by allosaurus, we used the frequency information in our syllable list to more precisely select our candidates. To do so, we average the likelihood score L_s of a detected syllable with its unigram probability P_s weighted with a constant α as:

$$L_s + \alpha P_s \quad (1)$$

We then optimised, on the validation set, α varying a range of values between 0 and 10 with a 0.1 step and a syllable detection threshold T between 0 and 10 with a step at 0.01. We then spotted syllables on the test set with the parameters which provided the best F-score on validation. We also report results without the frequency where only the threshold T is optimized on the validation set.

3.3 Experimental results

Our best results on the validation set have been obtained with $T = 0.39$ when unigram probability is added. For our baseline without unigram probability, the best threshold has been obtained with $T = 0.35$ We report the results in Table 1.

We can see here that the frequency information has an impact on the overall performances in both scenarios with an F-score nearly 4 points higher in the results with frequency. Better performances are obtained on the test set made of utterances. Two elements can explain it. First, the phone recognition model has been trained on similar data to the utterance test set which leads to better phone recognition performances. Then, the chance of a given syllable being pronounced several times is higher in longer utterances which means that it has higher chance to be spotted.

Results with likelihood score alone			
Sets	Recall	Precision	F-score
Words	41.71%	24.40%	30.79%
Utterances	47.21%	36.26%	41.02%
+ unigram probability			
Sets	Recall	Precision	F-score
Words	43.08%	28.09%	34.23%
Utterances	46.56%	41.50%	43.88%

Table 1: Experimental results (syllable spotting) on the two test sets

4 App design

4.1 Prototype

We designed a simple interface to display the syllables from our spoken term detection systems to our participants (see Figure 1). Our goal here was not to design a final product but to present a simple interface that works well enough to see if the proposed syllable concatenation mechanism makes sense from a Bininj perspective. We bridged the output of the system to a transcription interface by creating one button per syllable spotted for a given audio recording. The buttons display the orthography of the syllables spotted. They play the corresponding pronunciation when clicked. There is one *play* button to play the audio to transcribe and one text area with an associated *play* button to look for syllables that have not been spotted. The user needs to use the keyboard to make guesses on missing syllables and needs to click on the *play* button to check the pronunciation of their guesses.

Figure 1: Preliminary version of the app

We organised a testing session with one participant in Gunbalanya: IG, a 25 year old local artist and tour guide. We spotted syllables in a 3.35min recording made of elicited speech of Bible stories. Because of the quality of the audio, most of the syllables were correctly spotted. We explained to IG that we wanted to write down Kunwinjku and we needed his help to spell the words.

IG rapidly understood the task and started point-

ing syllables on the screen while we were writing with pen and paper IG's feedback. He clicked several times on the different syllables displayed and progressively gave feedback. When a syllable was not spotted, he could with some hesitancy, write with the keyboard syllables guesses in the text area.

The main observations made during the pilot study were IG's quick comprehension of the task, his hesitancy while using the keyboard and his confidence while reporting the orthography. At the end of the activity, he told us that he was expecting the text area to produce a new syllable button he could use.

4.2 Design and features

The quick comprehension of IG showed the potential of the proposed transcription mechanism which made us pursue this direction. Based on the first trial, we designed a proper transcription interface based on syllable spotting (see Figure 2). The core of the interface was the same that our first trial: we have a play button on the top of the screen playing the target audio to transcribe. We have one button per detected syllable associated with a wav file containing their pronunciation. The syllables can be dragged and dropped to the black box at the bottom of the screen. The user can listen to the final concatenation of the syllables with the associated play button and validate the transcription created with a thumb up button.

We needed to find a way to allow the user to add undetected syllables manually. To do so, we initially added a side menu accessible through a plus button on the side of the screen. The menu consisted of a scrolling list that contained the 584 syllables. We added a text area at the top of the list that allowed the user to retrieve a syllable from its first letters (see Figure 3). Following the principle of the regular syllable button, the user could click on the syllable to hear the pronunciation and click on the associated plus button once their choice was made. The syllable was then added as a regular syllable button. To avoid the use of the keyboard, we changed this syllable search mechanism by removing the text area and by replacing the list of syllables with expandable sub-lists labelled with the first graph[4] of the syllables it contains (see Figure 4). The user can then search for a syllable by expanding the lists and select a syllable by listening to it and clicking on the associated plus button. The app and databases were stored in a laptop accessed remotely by a tablet with wifi.

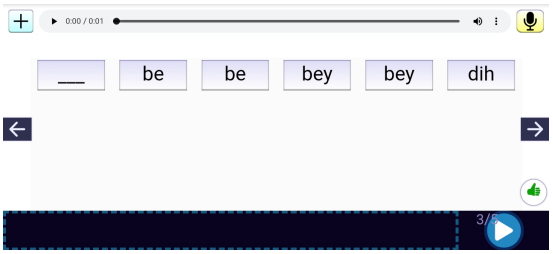


Figure 2: Final version of the app

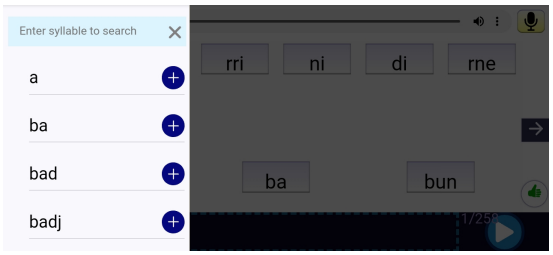


Figure 3: Initial syllable search mechanism

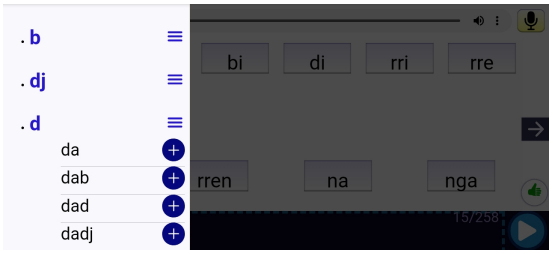


Figure 4: Updated syllable search mechanism

5 User testing

Due to Covid-19 restrictions, no trips to remote communities were possible. However, we have been able to work individually with Kunwinjku speakers in transit in Darwin at the university. We incorporated our syllable spotting based transcription task in a more global resource creation workflow. We could test it with two participants from Bininj country. In order to engage with the participants, we organised the testing phase in two sections. In the first one, we discussed and elicited knowledge about topics of interest based on previous conversations, in the second, we used the interface to transcribe the knowledge recorded. Therefore, besides the focus given to the design of the app and spoken term detection system, time of this

[4] We are not talking in terms of individual letter but graph or group of graphs that correspond to a single phone in Kunwinjku

project has been dedicated to the study of cultural elements to enable more efficient collaboration.

5.1 Activity description

Elicitation of knowledge

> *Ngabenbekken nahni wurdwurd nawu kabirrihre minj Kundebi kabirrikarrme. Burrkyak. Kabirridjalngeybun. Minj kabirridebikarren, burrkyak.*
>
> "I hear these children going about – they don't have Kundebi. No. They just use people's names. They don't use Kundebi with each other, no." (Etherington, 2006)

Language shift is not a new phenomenon. Language variation in Kunwinjku has been the subject of recent research (Marley, 2021) and has been one of the concerns raised by Bininj Elders. *Kundebi* specifically has been described as a language feature that the community is proud of and that is being progressively lost by the young generation (Garde, 2013; Etherington, 2006). It has also been mentioned in the same terms by some Elders during some of our fieldtrips. *Kundebi* refers to the way a speaker A refers to an individual C while talking to an addressee B. For example, a speaker A is talking to their elder sister's child B about their elder sister C. A is usually referring to B using the term *djedje* "nephew" and to C using the term *yabok* "sister". Listener B however usually refers to C using the term *morlah* "mother's elder sister". The kundebi term *berlunghkowarre* is then used to summarize these three relationships and could be translated as "my sister, your mother's elder sister, you are my sister's child" (Garde, 2013).

In order to respond to people's priority in terms of language work, we have decided to first focus the activity on the creation of written resources around *Kundebi*. To do so, while working with a Kunwinjku speaker, we would talk about common acquaintances, identify the way we both would refer to them and then identify and record the corresponding *Kundebi* terms. We used an activity sheet (see Figure 5) to draw the relationships we wanted to elicit (for instance, E for first author, G for the participant and J for the person we are talking about). The recording is directly stored on our laptop. The speed of the pipeline, described in Section 3.2, also allowed us to directly spot the syllables in the audio. Some of our participants expressed the fact that they were not confident with

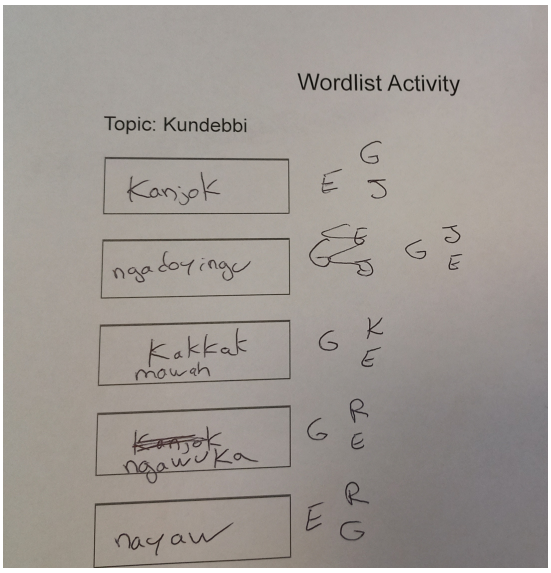


Figure 5: activity sheet filled

Kundebi and would feel more comfortable talking about *Kunbalak*. *Kunbalak* is a sub-language used for forbidden relationships to show respect. It is identical to regular Kunwinjku syntactically but would use different lexical items. For instance *Birriwam* "they went" becomes *birridokang* in *Kunbalak* (Manakgu, 1996). To elicit *Kunbalak* we would just ask for the conversion of regular Kunwinjku terms.

Use of the app

After recording a few terms with a speaker, we presented the transcription interface to them. The terms previously recorded and the syllables spotted have been automatically loaded into the app database. After showing the interface's different features to the participants, we asked them to drag and drop the syllables to build the transcription of the previously recorded terms. After actively working around Kunwinjku and building expertise about the proper way to write the language through the years, we let the participants use their own expertise on what they think is the orthography without questioning their authority.

5.2 Fieldwork

We tested our pipeline with two participants. JB (30s) and GB (30s).

We could present our activity to JB on three different occasions. We could identify and record some *kundebi* terms during the first trial. The activity has then been interrupted by upset child. During this first trial, she briefly started to point syllables on the screen without properly using the app.

She told us afterwards that the *kundebi* terms we recorded should be double-checked by an Elder, and she would feel more confident talking about *Kunbalak* instead. During the second and third trials, we could easily identify and record some *Kunbalak* terms. While using the app, we faced minor technical difficulties with the manual syllable addition feature. However, JB could take control of the tablet to transcribe some of the terms recorded. One of the issues we faced was the playback of syllables that include a glottal stop which was hard to identify in syllables in isolation (the difference between *ma* and *mah*, for instance). The activity was trialled with the first version of the syllable search (see Figure 3). The keyboard generated by the text area would take most of the space on the screen. JB needed to ask for our support to know how to proceed. At the end of the second trial, while no instruction had been explicitly given, she started to drag and drop the syllables available on the screen to explore the different words that are possible with them. We asked about her thoughts about the activity, and she responded that she liked it and would like to get more confident in writing in Kunwok and download the app later.

We could test the activity with GB, the second participant. We first recorded a few *Kundebi* terms. We wrote on paper the relationship to elicit, which made him understand the activity was about constructing a word list. After recording a few terms, we gave GB the tablet and asked him to transcribe the words. For each term, he listened to the audio first and pressed the syllable displayed on the screen. He was able to add new syllables manually without too much difficulty. For one particular term: *nayaw*, we discussed rather the term should be written *nayaw* or *nayawu*. While listening for a given syllable, he sometimes asked for confirmation about what he heard (for instance, "Is this *ka*?"). We discussed his thought about the task at the end of the activity. He showed enthusiasm about the incremental construction of the transcription. During the activity, he rephrased the syllable concatenation process by "putting pieces of language together".

No more participants were available for the time for this project. However, to sustain this work in the future, we deployed it in a laptop to be brought to the community by future scholars or language workers, as soon as COVID-19 restrictions are eased.

Figure 6: Picture of a participant using the app

6 Discussion and Limitations

The design and testing of the activity have shown promising results among a few participants, which gave us a glance at the potential of syllable spotting for the design of language related activity for Aboriginal people.

Syllable Spotting: It has been shown in the literature that traditional ASR is hardly applicable to Aboriginal languages due to the lack of resources available to train robust systems. Sub-word detection has been seen as a way to avoid out-of-vocabulary (Szoke et al., 2008; Parlak and Saraclar, 2008; Van Heerden et al., 2017) and, in our case, to allow a denser transcription than word spotting specifically for a polysynthetic language like Kunwinjku. Adding information on frequency, not surprisingly, allowed us to boost our performance (F-score) from 40% to nearly 44% for the syllables displayed on the screen for a given utterance.

Enabling mutual comprehension: Our main objective, starting from our previous work, was to enable a better comprehension in our cross-cultural setting. Part of this process consisted of getting familiar with cultural components that have been raised by the community (namely, *Kundebi* and *Kunbalak*). This also consisted of finding methods to trigger a conversation about these topics. For the rest, strategies have been found to help the participants to understand our contribution is this work. For instance, the support of the activity sheet made clear that the ultimate goal of the activity was to build a word list. Then the syllable concatenation mechanism allowed the participants to leverage existing language patterns from the aural space into writing.

Aligning agendas: Asking the participants about traditional knowledge allowed them to di-

rectly use their expertise and navigate in familiar territory. Talking about *Kundebi* and *Kunbalak* gave a sense of clarity regarding our function in Aboriginal land because of the continuation between previous conversations and the current activity. Yet the extent of our contribution being seen as beneficial for Bininj people from a language preservation perspective is still unclear. Writing in language is not a traditional practice in this community. People are often literate in English but not in their language. We then needed to find a space where the orthography made sense (Lewis and Simons, 2016). Documents written in Kunwinjku exist in Bininj country through the ranger program, the schools or in facilities where exists an interaction between Bininj and westerners (art centres, clinics, etc.). While we thought that the proposed activity could enable the continuation of the creation of these resources by Aboriginal participants, the proposed app has probably mainly been seen as a way to enhance writing skills.

App design: There were two main challenges related to the design of the app. The first one was enabling syllable concatenation, prioritising information from the oral space. Then we needed to efficiently retrieve syllables that had not been spotted. The first challenge was easily solved by the syllable playback features possible with the progressive collection of syllables throughout this project. Then we designed a basic search mechanism. The first search mechanism to add new syllables relied on the keyboard, which we knew was problematic (cf. Section 4.1). We believe that the new design would lead to better efficiency, but it could not be properly tested.

Activity flaws: The lack of good quality data available in Kunwinjku did not allow us to build a robust speech synthesis system that would have been relevant to the interface. Instead, we recorded in isolation syllables which sometimes lacked clarity. While ultimately, some of the most common syllables have been recorded by a native speaker, many were still pronounced by the first author, whose pronunciation might not be accurate. For instance, in the pilot study, while writing the word *djurra* (IPA djura) "paper", first author's pronunciation of the syllable *rra* has not been accepted by IG and selected instead "*da*" which was closer to the pronunciation of the word according to him. The case of the glottal stop has also been mentioned as a challenge in the literature (Wigglesworth et al., 2021). The glottal stops included in some syllables were not clearly audible out of context, which made them hard to differentiate from similar syllables without glottal stops (*ma* and *mah*, for instance).

Limitations: There is a limited number of Kunwinjku speakers, and recruiting a large number of participants for such work was not easy. The current pandemic did not facilitate our work, and we know that it is hard to draw final conclusions with activities conducted with only three participants. Further research needs to be done, including proper testing in Bininj country to consolidate our observations. The activity setup was also grounded for JB and GB in an academic environment with access to facilities that we do not necessarily have access to in remote locations (access to the internet, workplaces etc...). Besides, we can ask ourselves about the sustainability of such a work grounded in an interaction between Aboriginal participants and scholars in a very controlled environment. To be sure that our methods can be used in the long term, we imagine setting up a remote server to enable remote access on tablets so that people can keep interacting with the app without outside intervention.

7 Conclusion

Generic speech recognition methods for under-resourced languages offer the potential to support small speech communities. Yet the translation of such methods into community-based projects is rare. We have presented a study on the creation and testing of a syllable spotting-based transcription interface to enable the creation of written resources by the members of an Aboriginal community in the Australian Top End. Based on the challenges encountered in previous work, we went from word spotting to syllable spotting to reach a denser transcription and enabled a transcription method closer to existing practices. With the help of collaborators, we designed a transcription interface that allowed the users to build the transcription of given audio using the syllable spotted by our system. We reported the testing of the app with three participants at different stages of development, including lessons learnt from their interaction with the transcription activity and the app design.

Research guidelines push scholars to decolonise their practices and to go towards self-determination. Yet the translation of guidelines to real-life applications is unclear, specifically in cross-cultural collaborations with different ways of knowing. This

work allowed us to highlight methodological paths that improved the engagement and comprehension of the participants. The activity sheet, for instance, made clear that the activity was about creating a wordlist which was not necessarily clear based on our explanation. Dividing the activity between an elicitation part and a transcription part allowed us to hook the interest of the participants with a task they were familiar with and allowed us to clarify the context of our work in contrast to the sparse transcription of random sentences explored previously (Le Ferrand et al., 2022). All participants frequently used the playback of the syllables in isolation and their concatenation, confirming its engaging aspect.

Documentary linguistics has often been undertaken by non-indigenous linguists where the collaboration with the community did not go further than the collection of spoken data (First Languages Australia, 2014). In this work, we initially wanted to counterbalance these practices by enabling community-based language documentation. Yet keeping a language strong does not need to be about language documentation, and Bininj people who took part in this work did not seem to buy into documentary linguistics practices. Instead, they seemed to see the interface as a literacy learning tool. Keeping language strong is seen as building capabilities instead of creating and storing language material. Community-based implies an active role of the community in the work we conducted, and following their view in terms of language work is then crucial. The cross-cultural challenges we encountered required extra work to enable a common ground we could build on. Now that comprehension issues are solved, that we have a better comprehension of people agenda and COVID-19 restrictions start to be eased, more iteration can happen to allow the community to take control of the design of the proposed tool to better fit their agenda and practices.

8 Acknowledgements

We are grateful to the Bininj people of Northern Australia for the opportunity to work in their community, and particularly to artists at Injalak Arts and Craft (Gunbalanya), so as the Bininj and Daluk from Mamadawerre and Kabulwarnamyo who spent time helping us in the university facilities. Our thanks to several anonymous reviewers for helpful feedback on earlier versions of this paper. The final version of the transcription building app has been co-designed by Dr. Cat Kutay, Lecturer at Charles Darwin University and Melko Nguyen as part of her final master's project.

This research was covered by a research permit from the Northern Land Council, ethics approved from CDU and was supported by the Australian government through a PhD scholarship, and grants from the Australian Research Council and the Indigenous Language and Arts Program. All the participants have been paid at the regular rate for Aboriginal people consultancy.

References

Oliver Adams, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird, and Alexis Michaud. 2018. Evaluating phonemic transcription of low-resource tonal languages for language documentation. In *LREC 2018 (Language Resources and Evaluation Conference)*, pages 3356–3365.

Gilles Adda, Sebastian Stüker, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Hélene Bonneau-Maynard, Pierre Godard, Fatima Hamlaoui, Dmitry Idiatov, et al. 2016. Breaking the unwritten language barrier: The bulb project. *Procedia Computer Science*, 81:8–14.

Antonios Anastasopoulos, Marika Lekakou, Josep Quer, Eleni Zimianiti, Justin DeBenedetto, and David Chiang. 2018. Part-of-speech tagging on an endangered language: a parallel Griko-Italian resource. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2529–2539.

Alexei Baevski, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2021. Unsupervised speech recognition. *Advances in Neural Information Processing Systems*, 34:27826–27839.

Steven Bird. 2022. Local languages, contact languages, and other high-resource scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 7817–7829.

Tommaso Caselli, Roberto Cibin, Costanza Conforti, Enrique Encinas, and Maurizio Teli. 2021. Guiding principles for participatory design-inspired natural language processing. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 27–35.

Shobhana Chelliah and Willem De Reuse. 2010. *Handbook of Descriptive Linguistic Fieldwork*. Springer Science & Business Media.

Michael Christie. 2008. Yolngu studies: A case study of aboriginal community engagement. *Gateways: International Journal of Community Research and Engagement*, 1:31–47.

Philippe Descola. 2005. *Par-delà nature et culture*, volume 1. Gallimard Paris.

Lise M Dobrin, Peter K Austin, and David Nathan. 2009. Dying to be counted: The commodification of endangered languages in documentary linguistics. *Language Documentation and Description*, 6:37–52.

Long Duong. 2017. *Natural Language Processing for Resource-poor Languages*. Ph.D. thesis, University of Melbourne.

Steve Etherington. 2006. *Learning to be Kunwinjku: Kunwinjku People Discuss their Pedagogy*. Ph.D. thesis, Charles Darwin University.

Aoife Finn, Peter-Lucas Jones, Keoni Mahelona, Suzanne Duncan, and Gianna Leoni. 2022. Developing a part-of-speech tagger for te reo māori. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 93–98.

First Languages Australia. 2014. Angkety Map: Digital resource report. Technical report. https://www.firstlanguages.org.au/images/fla-angkety-map.pdf; accessed Oct 2021.

Ben Foley, Josh Arnold, Rolando Coto-Solano, Gautier Durantin, T. Mark Ellison, Daan van Esch, Scott Heath, František Kratochví, Zara Maxwell-Smith, David Nash, Ola Olsson, Mark Richards, Nay San, Hywel Stoakes, Nick Thieberger, and Janet Wiles. 2018. Building speech recognition systems for language documentation: The CoEDL Endangered Language Pipeline and Inference System. In *Proceedings of the 6th International Workshop on Spoken Language Technologies for Under-Resourced Languages*, pages 205–209. ISCA.

Dennis Foley. 2003. Indigenous epistemology and indigenous standpoint theory. *Social Alternatives*, 22(1):44–52.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830.

Murray Garde. 2013. *Culture, Interaction and Person Reference in an Australian Language: An Ethnography of Bininj Gunwok Communication*, volume 11. John Benjamins Publishing.

Pierre Godard, Marcely Zanon Boito, Lucas Ondel, Alexandre Berard, François Yvon, Aline Villavicencio, and Laurent Besacier. 2018. Unsupervised word segmentation from speech with attention. In *Proceedings of Interspeech 2018*, pages 2678–2682.

Florian Hanke. 2017. *Computer-Supported Cooperative Language Documentation*. Ph.D. thesis, University of Melbourne.

Dianna Hardy, Elizabeth Forest, Zoe McIntosh, Trina Myers, and Janine Gertz. 2016. Moving beyond "just tell me what to code" inducting tertiary ICT students into research methods with Aboriginal participants via games design. In *Proceedings of the 28th Australian Conference on Computer-Human Interaction*, pages 557–561.

Atticus Harrigan, Timothy Mills, and Antti Arppe. 2019. A preliminary plains cree speech synthesizer. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1, pages 64–73.

Robbie Jimerson and Emily Prud'hommeaux. 2018. ASR for documenting acutely under-resourced indigenous languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 4161–4166.

Roland Kuhn. 2022. The Indigenous Languages Technology Project at the National Research Council of Canada, and its context. In *Language Technologies and Language Diversity*, pages 85–104. Linguapax International.

William Lane and Steven Bird. 2019. Towards a robust morphological analyzer for Kunwinjku. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, pages 1–9.

Éric Le Ferrand, Steven Bird, and Laurent Besacier. 2021. Phone based keyword spotting for transcribing very low resource languages. In *Proceedings of the The 19th Annual Workshop of the Australasian Language Technology Association*, pages 79–86.

Éric Le Ferrand, Steven Bird, and Laurent Besacier. 2022. Learning from failure: Data capture in an australian aboriginal community. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 4988–4998.

Tuck Wah Leong, Christopher Lawrence, and Greg Wadley. 2019. Designing for diversity in aboriginal australia: Insights from a national technology project. In *Proceedings of the 31st Australian Conference on Human-Computer-Interaction*, pages 418–422.

M Paul Lewis and Gary F Simons. 2016. *Sustaining Language Use*. SIL International Publications.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R Mortensen, Graham Neubig, Alan W Black, et al. 2020. Universal phone recognition with a multilingual allophone system. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8249–8253. IEEE.

Andrew Manakgu. 1996. Kunbalak. *Stories for Kunwinjku young people in mother-inlaw language, ordinary Kunwinjku and English. Kunbarllanjnja (Oenpelli): Kunwinjku Language Centre*.

Alexandra Marley. 2021. "I speak my language my way!"—young people's kunwok. *Languages*, 6(2):88.

Raghav Menon, Herman Kamper, Ewald van der Westhuizen, John Quinn, and Thomas Niesler. 2019. Feature exploration for almost zero-resource ASR-free keyword spotting using a multilingual bottleneck extractor and correspondence autoencoders. *Proceedings of Interspeech 2019*, pages 3475–3479.

Vikramjit Mitra, Andreas Kathol, Jonathan D Amith, and Rey Castillo García. 2016. Automatic speech transcription for low-resource languages-the case of Yoloxóchitl Mixtec (Mexico). In *Proceedings of Interspeech 2016*, pages 3076–3080.

M. Eleanor Nevins. 2013. *Lessons from Fort Apache: Beyond Language Endangerment and Maintenance*. Wiley.

Walter Ong. 1982. *Orality and Literacy: The Technologizing of the Word*. Routledge.

Siddika Parlak and Murat Saraclar. 2008. Spoken term detection for Turkish broadcast news. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5244–5247. IEEE.

Bernard Perley. 2012. Zombie linguistics: Experts, endangered languages and the curse of undead voices. *Anthropological Forum*, 22:133–149.

Aidan Pine, Dan Wells, Nathan Brinklow, Patrick Littell, and Korin Richmond. 2022. Requirements and motivations of low-resource speech synthesis for language revitalization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7346–7359.

Keren Rice. 2011. Documentary linguistics and community relations. *Language Documentation and Conservation*, 5:187–207.

Celeste Rodríguez Louro and Glenys Collard. 2021. Working together: Sociolinguistic research in urban Aboriginal Australia. *Journal of Sociolinguistics*, 25:785–807.

Courtney Ryder, Tamara Mackean, Kate Hunter, Julieann Coombes, Andrew JA Holland, and Rebecca Ivers. 2021. Yarning up about out-of-pocket healthcare expenditure in burns with aboriginal families. *Australian and New Zealand journal of public health*, 45(2):138–142.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. In *Proceedings of Interspeech 2019*, pages 3465–3469.

Lane Schwartz. 2022. *Primum Non Nocere*: Before working with Indigenous data, the ACL must confront ongoing colonialism. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 724–731.

Lane Schwartz, Emily Chen, Benjamin Hunt, and Sylvia LR Schreiner. 2019. Bootstrapping a neural morphological analyzer for st. lawrence island yupik from a finite-state transducer. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1, pages 87–96.

Shane Settle, Keith Levin, Herman Kamper, and Karen Livescu. 2017. Query-by-example search with discriminative neural acoustic word embeddings. *Proceedings of Interspeech 2017*, pages 2874–2878.

Alessandro Soro, Margot Brereton, Jennyfer Lawrence Taylor, Anita Lee Hong, and Paul Roe. 2017. A cross-cultural noticeboard for a remote community: design, deployment, and evaluation. In *IFIP Conference on Human-Computer Interaction*, pages 399–419. Springer.

Igor Szoke, Lukás Burget, Jan Cernocky, and Michal Fapso. 2008. Sub-word modeling of out of vocabulary words in spoken term detection. In *2008 IEEE Spoken Language Technology Workshop*, pages 273–276. IEEE.

Jennyfer Lawrence Taylor, Wujal Wujal Aboriginal Shire Council, Alessandro Soro, Michael Esteban, Andrew Vallino, Paul Roe, and Margot Brereton. 2020. Crocodile language friend: Tangibles to foster children's language use. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

E. Uí Dhonnchadha and J. Van Genabith. 2006. A part-of-speech tagger for Irish using finite-state morphology and constraint grammar disambiguation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 2241–2244, Genoa, Italy.

Daan van Esch, Ben Foley, and Nay San. 2019. Future directions in technological support for language documentation. In *Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 14–22.

Charl Van Heerden, Damianos Karakos, Karthik Narasimhan, Marelie Davel, and Richard Schwartz. 2017. Constructing sub-word units for spoken term detection. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5780–5784. IEEE.

Michael Walsh and Colin Yallop, editors. 1993. *Language and Culture in Aboriginal Australia*. Aboriginal Studies Press.

Matthew Wiesner, Desh Raj, and Sanjeev Khudanpur. 2022. Injecting text and cross-lingual supervision in few-shot learning from self-supervised models. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8597–8601. IEEE.

Gillian Wigglesworth, Melanie Wilkinson, Yalmay Yunupingu, Robyn Beecham, and Jake Stockley. 2021. Interdisciplinary and intercultural development of an early literacy app in Dhuwaya. *Languages*, 6:106.

Liang Xu, Elaine Uí Dhonnchadha, and Monica Ward. 2022. Faoi gheasa an adaptive game for irish language learning. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 133–138.

Tariq Zaman, Heike Winschiers-Theophilus, Franklin George, Alvin Yeo Wee, Hasnain Falak, and Naska Goagoses. 2016. Using sketches to communicate interaction protocols of an indigenous community. In *Proceedings of the 14th Participatory Design Conference: Short Papers, Interactive Exhibitions, Workshops-Volume 2*, pages 13–16.

Retrospective view

This final research contribution was completed simultaneously with the writing and submission of this thesis. Although it is difficult to reflect on this work without repeating what has already been discussed, it can be contextualised with the other three contributions. Initially, our aim was to facilitate community-based work that was separate from the linguist's workflow, but this work shifted towards traditional language documentation. This final contribution helped us to realise the gap that exists between our conception of language preservation and what the Bininj people considered beneficial. Ultimately, the feedback we received from the participants highlighted the potential of this work, which can now be advanced to develop local capacity.

3.6 Conclusion

Despite the collaborative effort in co-designing and implementing the sparse transcription model, we encountered various challenges during our interactions with Bininj participants. I began this project by designing and deploying the sparse transcription model, initially following the linguist's workflow and employing a spoken term detection method based on DTW. However, it became clear that the initial study did not align with the Bininj community's goals. By redesigning the spoken term detection method and learning from my initial shortcomings, I conducted a new study in the Bininj community that aligned more closely with their needs and desires.

Chapter 4

Discussion and Conclusion

4.1 Introduction

This chapter provides a summary and discussion of the contributions to knowledge made by this thesis, as well as the conclusion for this project. This project was based in northern Australia, in Indigenous communities with strong identities and cultures, which have bodies of knowledge and ways of knowing that do not always align with western approaches to knowledge. Following a long history of unethical research practices in Indigenous contexts, research undertaken in these communities has started to adopt community-based methodologies and co-design approaches that aim to promote community self-determination.

Natural Language Processing research addresses some Indigenous topics, usually under the "low-resource" label. Many of these research contributions promote the use of data to provide solutions designed by external experts to solve local issues. In this thesis, I aimed to counterbalance established NLP research practices by exploring ways to enable computer-assisted and community-based language development.

I began this project with the sparse transcription model, designing the first simulation of a speech transcription workflow combining spoken term detection and human expertise (Section 3.2). This model enables the participation of the community in language-related work with the support of some automation. Following this first simulation, we conducted several trials in remote Indigenous communities using a

web app to bridge the output of our spoken term detection system to Indigenous participants (section 3.3). Throughout these trials, the challenges associated with cross-cultural research were evident, particularly with respect to speech transcription and technologies. Building on the obstacles we faced, we designed a phone recognition-based spoken term detection method that allowed for better efficiency and flexibility than our previous method based on Dynamic Time Warping (Section 3.4). Following on from this new method and the lessons learned in Section 3.3, we designed a syllable spotting-based transcription interface. We introduced this interface into a larger activity combining discussion about cultural topics with transcription work (Section 3.5).

With the emergence of research contributions that advocate decolonising methods, principles of respect or reciprocity seem obvious in projects involving Indigenous participants. Yet, beyond general recommendations dictated by ethical research guidelines, the application of ethical obligations to real-life actions is not always clear. Cross-cultural research has expanded to provide solutions for more ethical practices. For instance, in Australia, traditional Indigenous research methods have been adopted by scholars for data collection purposes, while health practitioners aim to co-design health program construction (Section 2.2). Following decolonising research principles, this research aims to contribute in the area of community-based NLP by exploring ways to use speech recognition for language development purposes in partnership with an Australian Indigenous community.

Following up on the research questions detailed in Section 1.2, the contributions of this thesis are articulated around three main themes. First, **Performance**, where we try to rethink the notion of performance and usability of speech technologies in Indigenous contexts. Second, **Comprehension**, where we detail strategies to organise information to build a common ground for efficient collaboration. Last, **Engagement**, where we summarise our approach to engage the community towards sustainable language development practices. In Section 4.5, we report the limitations of this work and discuss future research opportunities.

4.2 Performance: Technological strategies to facilitate transcription

Data availability is rarely a concern for dominant languages, but this is not always the case for others. Research on speech recognition has the potential to provide solutions to real-life issues. For example, automatic sub-titling (Palaskar et al., 2018), ASR systems for smart homes (Desot et al., 2019), and chatbots (Belenko et al., 2019) all have unique requirements in terms of training data or system architectures.

A branch of NLP has explored ways to apply processes for low-resource languages for which we do not have the data necessary to perform a given task in the standard way. The main context provided for these languages is the language documentation workflow and the transcription bottleneck. The documentary linguist records speech data, in many cases in an Indigenous language, then transcribes it and analyses it. Manual transcription is described as an onerous task and speech recognition methods are therefore applied to support this work. Experiments are then conducted, generally on sanitised corpora made for NLP experiments, to reduce a Word Error Rate, or to increase recall or precision. The translation of experimental outputs into concrete applications is barely explored.

Research contributions around speech recognition for Indigenous languages generally only mention the lack of data by which to justify their methods. However, Indigenous contexts might have other kinds of constraints. As seen in Section 2.3.1, many languages are not traditionally written, which makes solutions based on written resources questionable. In Section 1.3.1, it is explained that many Indigenous communities can be found in remote locations with few facilities and limited access to the Internet, which complicates the deployment of large computational models. Fieldwork often happens outdoors, with the language data recorded often being of poor quality as it includes background noise. Finally the willingness of the Indigenous communities to record speech data will limit the amount of non-transcribed speech available.

In this thesis, I attempted to follow decolonising methods in reviewing how speech recognition can be applied in response to the needs of an Indigenous community. To do so, we used the sparse transcription model (Bird, 2020b) by combining spoken term detection and human expertise to iteratively transcribe a speech collection. The goal was to allow the speech community to be involved in the documentation of their language by verifying the guesses of a speech recognition system. The first contribution was the design of a transcription pipeline where we tried to automatically detect isolated words from a spoken lexicon. To do so, we used DTW and explored several speech features. We rapidly questioned the proposed pipeline while trying to apply it in Indigenous contexts, facing obstacles we describe in detail in Section 4.3. These obstacles pushed us to reconsider our initial approach: instead of spotting words, we decided to use phone recognition to spot syllables instead. Among our justifications for this switch in direction, two main issues were identified, lack of consistency and real-life application.

4.2.1 Lack of consistency

Throughout this project, we conducted spoken term detection experiments on two different corpora, using two different algorithms to explore a set of different types of speech features. Throughout the experiments described in Section 3.2 and 3.4, we were able to determine which method was better based on quantitative metrics alone. While for dominant languages, the amount of language resources available allows a given system to generalise correctly; this is not the case for low-resource languages.

Two corpora were explored: one in Bininj Kunwok (also called Kunwinjku), recorded across several trips in Bininj Country was made up of spontaneous speech from various guided tours with six speakers, and one in Mboshi made up of speech elicited by only three speakers, including one responsible for 70% of the corpus. This second corpus allowed us to compare the results obtained for our target language in

another corpus collected for NLP purposes. The difference in the speaker distribution between the two corpora and the mode of recording (spontaneous/read) led to inconsistency in the evaluation.

On the computational side, two methods have been explored. The first one is based on DTW, which is a dynamic acoustic comparison of speech features. The second method is based on a universal phone recognition system and Finite State Automata, and is detailed in Section 3.4.

A method based on DTW, applied to features extracted from wav2vec models, provides the best results for Kunwinjku, due to the robustness of these features to speaker variation. However, in Mboshi, these features were less efficient than normalised MFCCs, which were more robust for same-speaker recognition.

In a second set of experiments, we compared the performance of DTW applied to multilingual bottleneck features[1] and our proposed method based on phone recognition. Although we cannot definitively state that one method was better than the other, the difference in performance across our two corpora was more pronounced with DTW than with the phone recognition-based approach. Overall, the phone recognition-based approach provided better precision, making it more reliable.

When we went to the field, we realised how the lack of consistency was problematic. A simulation on a given data set was not enough to draw conclusions and a change of corpus or lexicon could drastically affect the final results. The phone recognition-based approach was not better overall, however, it was more consistent than DTW.

4.2.2 Real life application

Speech recognition covers a range of different architectures or algorithms that can be applied to different kinds of tasks with varying requirements. In our case, we

[1]We chose not to use wav2vec due to the unsuccessful results we initially obtained.

were targeting transcription. After facing the challenges described in Section 3.3, we encountered several obstacles imposed by DTW.

A comparison between acoustic sequences takes time to process. This time increases depending on the size of the speech collection, the spoken lexicon, or the size of the speech representation. For instance, spotting words in the Mboshi corpus with a lexicon of 20 items using the wav2vec representation would take nearly 2 hours on a single GPU. Dynamic Time Warping alone does not involve any learning, which means that new computing is necessary each time a new term is added to the lexicon or the speech collection grows. This computing time was not initially an issue since the scripts were running in the background before the activities were conducted with the Bininj participants.

Changing direction for an elicitation and transcription workflow, DTW was no longer a valid method. We could not anticipate what people would be keen on doing beforehand, and the method used to spot terms needed to take into account this unpredictability. While switching to the syllable spotting-based interface, spotting syllables on a single word with the entire syllable lexicon would take a few minutes, which can be constraining when it occurs in the middle of an activity.

The second constraint imposed by DTW was the data requirements: it needed a spoken lexicon. For transcription purposes, the density of the final transcription was restricted by the size of the lexicon. Clipping out terms from the speech collection requires manual work when a forced aligner is not available in the target language. Recording terms with a speaker also takes time and is restricted by people's availability and willingness to be recorded. In the specific case of syllable spotting, because most of the data recorded was spontaneous, clipping out clearly pronounced syllables was not easy because of co-articulation effects. While we were changing direction for syllable spotting, gathering a clear lexicon of spoken syllables was difficult. We eventually needed such a lexicon for the design of the app, and we gathered it during several trips. An hour of work with one speaker was not enough to record the whole syllable set, and we ended up filling the gaps with recordings made by ourselves. Use

of the phone recognition-based method described in Section 3.4 allowed us to bypass the need for spoken queries and use a syllable list instead.

Ultimately, none of the methods used were universally better for oral languages. Implicitly, DTW appeared to be the more suitable method since it is based on acoustic features comparison and does not rely on a standard orthography. However, such a method does not allow for much flexibility regarding computing time and data requirements. Changing to a phone recognition-based approach allowed us to bypass the need for a spoken lexicon and then to incorporate the syllable spotting algorithm inside the transcription workflow, which enabled the freedom to create a wider variety of resources than word spotting.

4.3 Comprehension: Computer-assisted transcription for an oral language

Most Indigenous languages are only spoken, and their speakers are not necessarily skilled in writing in their language. Organising transcription with Indigenous participants might sound counter-intuitive. Yet, written content in these Indigenous languages still exists in limited contexts. In Bininj Country, there is written Kunwinjku for educational purposes at school, for tourism purposes at the art centre, and for traditional knowledge documentation purposes in the ranger program facilities. Bringing transcription activities to people could have multiple benefits, such as participating in the creation of written material or building writing skills.

As with many other Indigenous peoples in Northern Australia, the Bininj People have a history of western researchers coming to their land for fieldwork. In this context, they have been involved in recording sessions, interviews, and storytelling activities to support scholars in Indigenous contexts. Some contributions , for more ethical and ecological research methods, have started to use Indigenous approaches for data collection, in archaeology, anthropology, and more recently, public health

and linguistics. Yarning, for instance, is an Indigenous Australian method for data collection, based on storytelling. Such a method has been used more frequently to enable culturally appropriate interaction with local communities (Carlin et al., 2019; Geia et al., 2013; Ryder et al., 2021). Navigating in familiar territory allows Indigenous collaborators to thrive, whereas western methods impose unnecessary barriers that compromise collaboration. The different fields of computer science are unlikely to have a place in the different Indigenous bodies of knowledge. If a given project is likely to have components that are associated with Indigenous epistemology, computational aspects can present challenges that need to be addressed. Human-Computer Interaction has been a precursor with engagement methods such as 'design non-proposal', where a portfolio of existing prototypes is presented to a community to illustrate possible achievements (Taylor et al., 2020).

When entering Indigenous space as a white European researcher, I had many assumptions about what research in this context should look like. Initially, I adopted a *the simpler, the better* approach when designing the activities for the first trips, building on people's feedback. However, my first mistake was assuming that *simple* means the same thing everywhere. In western perspectives, simplicity is often conceptualised as fewer features, fewer buttons, and a simple design. However, in the Indigenous community, this approach increased the complexity on their side, resulting in an abstract activity that was disconnected from their assumptions about language work and sprinkled with meaningless cultural items. After realising this, we incorporated the community's feedback to design an activity with a more complex app that was closer to their practices.

Our first approach was to present a lexical verification app to the community, which takes the output of a spoken term detection system to enable manual verification. Initially, we opted for a simple design to initiate a discussion with the community around NLP themes. We believed that the activity would trigger some curiosity and reflection, allowing us to work progressively towards co-design while the app design was still preliminary. However, the activity was unsuccessful, frustrating, and

irritating for both us and the community. In Section 3.3, we detail the possible reasons for the failed trials of this first work, and in Section 3.5, the strategies explored to address these issues. Among the several reasons we provided, we identified two main aspects: a failure to design from the ground-up and a lack of attention to the need for approaches that work in the real world.

4.3.1 Designing from the ground-up

As previously explained, the first proposed activity was too far removed from the usual work in which Bininj people have participated while collaborating with scholars. The comprehension issues encountered while presenting the app pushed the participants to return to their usual practices of storytelling and translation, among other things. Consequently, there was a clear need to rethink our methods to address comprehension issues. Other research has argued the need to start from the ground-up when designing projects with Aboriginal communities, building on what has already been done (Christie, 2013b; Waller, 2016). This raises the question of how to translate ground-up principles into a speech recognition-related topic. Transcription tasks are generally not widely spread in Indigenous areas, although the Bininj People are familiar with such work through helping with transcription for linguists and transcription workshops organised through the ranger program. Reflection on language themes typically takes place during teaching sessions between westerners and Bininj. People from the community would occasionally yell hyperarticulated words at me, broken down into syllables, in order to learn the proper way of pronouncing a given term. We have incorporated these two elements into a syllable spotting-based transcription interface. In contrast to the first design, this new interface allows the participant to maintain their teacher-student position whilst teaching westerners about pronunciation, this time using the app as a support. Additionally, this interface allows the participant to take control of language transcription using their expertise in the aural space.

4.3.2 Working in the real world

The final product that we wanted to make was transcribed speech. Yet, in the first activity design, this product was not visible anywhere. The transcription was built up in the background through manual verification. For the same reason, this manual verification was almost never understood, as the verification did not have any visible effect. We considered that the proposed activity lacked context. This issue has also been seen in different contexts, such as the need to reconnect an utterance to the overall story. In a retrospective view of this work, it is ironic to consider an activity about transcription and, at the same time, try to avoid written content. We did learn from this first mistake in the revised app described in Section 3.5. We realised the need for the activity to make sense in the real world, making obvious the participants' contribution. Methodological strategies such as the use of activity sheets and the report of people's feedback with pen and paper allowed us to make clear the purpose of the collaboration. The syllable concatenation strategy has also been understood as 'putting pieces of language together'.

In section 3.3, we reported the obstacles we faced in the first design and then in section 3.5, we detailed a success story built on what we learned from our failure. This last contribution is the result of a year and a half of interaction and relationship building with the Bininj People. While the failure of the first proposed design was hardly avoidable due to the lack of documentation around community-based NLP, the collaboration was too preliminary, and our knowledge of the Bininj context was too shallow. Ultimately, we did not try to make people understand our purpose. Speech recognition is just a tool to solve a given task. It was not our place to set the goal but only to see how speech recognition could be incorporated into existing agendas. We tried to build, in collaboration with the community, a mutual understanding in order to have a common ground from which to enable deeper collaboration. On our side, we needed to understand what the community wanted in terms of language preservation. On their side, we needed to make clear what our contribution could be.

4.4 Engagement: towards community-based and sustainable practices

Descriptive linguistics is often the motivation behind language documentation and a linguist doing fieldwork (Hanke, 2017). The data collected serves the description of some linguistic phenomena that hardly aligns with the community's agenda. Research around speech recognition has often followed this language documentation agenda in designing methods to facilitate the transcription of speech resources. The data required by speech processing does not focus on the content of the audio, but rather on its quality, for example, clear speech without noise and meta-information such as the number of speakers and the size of the vocabulary. While the output of a research project around descriptive linguistics needs to be negotiated, the desire of the speech community in terms of language work and the goals of speech recognition can be complementary. The main obstacle was engaging people in a transcription task when transcription was not part of their usual practices. Yet, written language exists in places where there is a collaboration between Indigenous people and westerners. Speech recognition could therefore find its place in the Bininj context for language development purposes.

The NHMRC Guidelines for Ethical Conduct in research with Aboriginal and Torres Strait Islander Peoples and Communities (*NHMRC* n.d.) state that there should be a notion of reciprocity while doing research in Indigenous space where "research outcomes are equitable and of value for Aboriginal and Torres Strait Islander people and communities". The community has also the right to define what benefits mean according to them (cf. Section 2.2).

Coming from an NLP culture, I jumped into this project, focusing too much on the data I could gather to bring Bininj Kunwok a little closer to well-resourced languages. This spirit has been translated into this first design proposal, where the community would contribute on the creation of data that could be used to create a better speech recognition system. We followed this intention during the first trips

to Bininj Country, the Elders of the communities discussed their agenda, and we could demonstrate our prototypes. I assumed that this principle of reciprocity was respected since a better speech recognition system would indirectly respond to their expectation for the production of resources. The mistake we made was that the first contributions only addressed the data collection from an NLP perspective without considering the community's desires and their view regarding what benefits mean to them. One of the main flaws of the first activity was that it was not engaging. Instead it was grounded in an NLP-based framework where the manual verification part was abstract, confusing, and redundant. Most of the participants found strategies to provide, through the activity, knowledge based on their assumptions about our work, a knowledge that could not be channeled by the first design. We ultimately designed a second activity where the participation of the community was visible and, while following their expectations in terms of language work, were providing data valuable to the improvement of our system. The engagement strategies we detailed in Section 3.3 and 3.5 can be summarised in two main aspects, people's authority, and people's agenda.

4.4.1 People's authority

Whilst entering Bininj Country for the trials of our first design, the relationship with the community was biased. We were introduced as scholars looking for Indigenous consultancy to work around language topics. This general introduction built false expectations from the participants across the project. In the worst-case scenario, people would come to us asking if we were paying people to tell stories. More commonly, the interactions were grounded in a teacher-student context where the Indigenous participants wanted to teach us language and cultural knowledge. This kind of interaction was probably due in part to a desire on the part of the Indigenous community that westerners learn the language and the culture in order to behave appropriately. Whilst it was not originally designed like this, our first activity design

ended up perpetuating the usual NLP story where the scholar provides the solution to local issues. We wanted to use speech recognition to create resources in languages but ended up trying to teach the participants the western way of doing so. This was due to a lack of mutual understanding, as described in Section 4.3. We were trying too hard to push a technology and an activity to solve a very specific task that had not been negotiated and did not allow us much flexibility. The second activity design allowed us to solve comprehension issues and properly collaborate with the community regarding relevant topics. People's knowledge authority has been accidentally questioned through the first design. In the second, we focused on establishing more fluid communication by introducing a design that could capture what the participants were teaching us. As explained in Section 4.3, the primary strategy was to build on existing practices so that the participants could confidently use their skills to support transcription activities. A more concrete example was the use of playback buttons in the syllable spotting-based transcription interface, where the participants could use their knowledge of the language's pronunciation to guide us through speech transcriptions.

4.4.2 People's agenda

Language is above all, a vessel for sharing knowledge among individuals. Whilst talking to people in the communities, we quickly realised that they were not particularly interested in abstract linguistic concepts such as morphemes, noun incorporation, or polysynthesis. It was particularly evident that while presenting a task about word verification, participants were automatically trying to explain or translate the content of the recordings or link them to stories. In Bininj Kunwok, there is no metalinguistic vocabulary; everything is language. A sentence like "Can you hear the word in the sentence?" is translated into *yibekkan namekke kunwok kure kunwok* (Can you hear the language in the language?). The discussion needed to be about knowledge. While the second app design enabled better comprehension and was more aligned with

people's practices, we needed to make a connection between language and traditional knowledge in order to pique people's interest.

In conversations we had with both Elders and participants through the several activities we conducted, the recurrent topic was the inter-generational language gap, where the younger generation was progressively losing the traditional way of speaking. Just as English children need to learn how to write and speak English properly to be more easily accepted in modern society, *wurdurd* (children) need to learn *kundebbi*, *kunbalak* or *kunkurlah* to address members of the community properly and fit within Bininj social order. We addressed this topic by incorporating transcription into a larger workshop where these concepts were discussed, and lexical items were identified, recorded, and transcribed. However, as seen in Section 3.5, participants did not buy into the usual language documentation work and preferred using the final app design as a literacy learning tool. Language documentation tends to preserve languages by creating material and storing it. Keeping language strong in the Bininj context appears to be done by building capabilities such as writing skills.

Western research has a long history of unethical practices, where scholars have been compared to mosquitoes – who suck blood and leave – to illustrate the lack of reciprocity in their methods (Cochran et al., 2008). Throughout this project, while facing occasional misunderstandings, we maintained ongoing discussions with the community and sought their feedback to ensure our research remains relevant. Once a mutual understanding has been established, we incorporate people's comments into the app design or activity, progressively following recent HCI practices and working towards co-design.

4.5 Limitations and opportunities for future work

This project is multidisciplinary, with a focus on computational, linguistic, and anthropological aspects. Regular visits were organised to Indigenous land to build and maintain relationships with the Bininj community and ensure efficient, ethical,

and sustainable achievements. This was the plan when I flew to Australia in 2019. However, about eight months after the start of the project, Covid-19 began to spread worldwide, and the Australian border were shut down. Access to Indigenous land was restricted to Indigenous people and essential workers, and Indigenous people were advised to stay on their land to limit the risks of community spread. Between 2020 and 2022, restrictions fluctuated, allowing us to organise a few trips and for people to visit the university. However, during this period, visits were still restricted by local institutions to limit the risks. The core of this project depended on collaboration with Indigenous consultants, and the Covid-19 pandemic significantly affected the conduct of this project.

Alongside the difficulties linked to the sanitary situations, working within an Indigenous context requires building relationships and spending time in the community, which may not necessarily yield research outcomes that are immediately apparent. In fact, entire trips were devoted solely to immersing ourselves in the local way of life, which included activities such as bushwalking, fishing and having informal conversations. This was crucial for establishing trust and creating a conducive environment for the project stakeholders, as well as gaining an understanding of the local customs and practices. While this time was essential, it may not be readily apparent in the final contributions of the thesis.

The community was also small, with around 2,500 individuals speaking a dialect of Bininj Kunwok across Bininj Country. Throughout the three-year duration of this project, only a dozen Bininj people were involved. We intended to explore more directions than were feasible within the three-year time frame, such as **expanding** the proposed work to include more discussion on community engagement, **sharing** this work with other communities with varying needs, and **enhancing** this work from a technical standpoint.

4.5.1 Extension

A recurring criticism we received about this work was its unfinished or preliminary nature. Naive thinking might expect a testing pipeline like the one seen in some HCI publications or in large companies during the design of new technologies. However, in the context of a PhD thesis, large-scale testing was simply not possible for the reasons mentioned above. Nevertheless, the trials of the first design exceeded our expectations in terms of the number of participants, considering the size of the community. Deployment on a similar scale was planned for the second design, which was ultimately compromised. Although my involvement in this project as a PhD student is now over, algorithms and interfaces have been left in Darwin to ensure the continuity of this project. The addition of more stakeholders will ensure its validity with greater participation. During the few interactions we had with people during the trials of the second design, feedback was given on the design and the topic addressed for elicitation. Co-design has been described as the way forward for the deployment of technologies in Indigenous spaces. However, the limited visits to the community did not encourage full commitment to this project. In future work, it would be beneficial to continue the conversation with the community and enable future stakeholders to reshape activity and app design to fit their vision. Most of our attention was focused on real understanding of the task to enable proper collaboration. While we may have reached that point in this project, a major component that has not been addressed is the question of access to the data created. An extension of this project could address this question through discussion with the community regarding how they want the resources to be organised and accessed.

4.5.2 Generality

Bininj Country stands out for its strong cultural and language identity, connection to traditional practices, position in protected Indigenous land, late exposure to western culture, limited number of local institutions, and the lack of a physical language

centre. Other contexts can be found in the Australian Top End, such as the Yolŋu communities in East Arnhem land or the communities in Tiwi islands, or Groot Eylandt. Their languages would be situated at Level 6 of the EGIDS scale, where orality is still prevalent, and an appropriate direction would be sustainable literacy (Lewis and Simons, 2016). Many languages across Australia have experienced significant language loss, but revitalisation programmes are underway in language centres to revitalise these languages alongside the communities (Olawsky, 2010; Roman and Harvey, 2003). While currently only relevant in the Bininj context, it would be interesting to explore to what extent our syllable spotting-based approach is applicable in contexts where language vitality is lower.

4.5.3 Improvement

Research in NLP, and specifically speech processing, progresses quickly, and a given method is rapidly outperformed. When the project started, while we could find a variety of different methods, the main one was still DTW. However, in 2020, new methods such as allosaurus (Li et al., 2019) and wav2vec (Baevski et al., 2020) were released, offering opportunities for improvement over DTW for spoken term detection. Projects like Leong and Whitenack (2022) or Macaire et al. (2022) have explored these new methods, which could have found their space in the scope of this project. A parallel project was also taking place in the Bininj Kunwok area, specifically the creation of a morphological analyser and morph completion algorithm (Lane and Bird, 2019; Lane and Bird, 2020). The incorporation of such technologies into our workflow could have improved the performance of the spoken term detection system, especially considering the possible alternate orthography.

4.6 Concluding thoughts

Coming from an NLP background, my mind was fixated on the idea of performance where a precision, a recall, and an F-score need to be improved for low resource

languages. Coming into the Bininj Country, building on existing literature, designing a human-in-the-loop framework appeared to be a great opportunity by which to address the famous transcription bottleneck. Yet the time of this project has been underlined by a succession of unfortunate events. The three themes around which I tried to summarise this research are an attempt to address a tendency of over-generalisation of the world's languages. We tried to go outside the usual paradigm of the linguist using speech technology to support manual transcription. Instead we went towards the speech community, and realised that performances were not to be understood simply in terms of quantitative metrics, that mutual comprehension is not straightforward and needs to be built, and that language development goes beyond the translation of speech into text and discussion, and that negotiation with the speech community needs to take place to ensure sustainable engagement.

The research directions taken during this project are bound to Bininj Country and community. While using a syllable spotting method, worksheets, or eliciting kinship terms made sense in Bininj Country, other challenges may occur in different contexts that may lead to other solutions. However, throughout this specific context, this work has been an opportunity to show the obstacles and challenges one can face while applying speech recognition in a cross-cultural setting. The translation of ethics guidelines or research theories into real-life action is not straightforward. In this thesis, we could adapt computational methods into real-life scenarios with specific restrictions. We could examine how to accommodate NLP concepts to a different system of knowledge. We could explore how speech technologies could fit into local agendas. Bininj Kunwok, Gumatj, Plains Cree, Mboshi, Bislama, and many other languages have been called "left-behind" as they do not possess the data that mainstream NLP processes need to function (Joshi et al., 2020). Yet, a man who does not have a car is still able to walk. Research usually follows a linear trend where projects are built from other projects. The BERT model, for instance (Devlin et al., 2019), was built following opportunities given by the transformers architecture (Wolf et al., 2020). From BERT emerged FlauBERT in French (Le et al., 2020), PhoBERT

in Vietnamese (Nguyen and Tuan Nguyen, 2020), and even gaBERT for Irish Gaelic (Barry et al., 2022). Yet, we can wonder if a KunBERT would really make sense here or whether we could redesign the way we conduct the research. Instead of building from the technologies to the languages, we started from the ground and saw what made sense.

There is a risk that when working in NLP on languages in the low-resource spectrum, we will put on a white saviour suits and design fancy computational models in attempts to save languages before their extinction. The NLP community tends to believe implicitly in the universality of its systems and approaches – i.e., that these systems and approaches can cope with all the world's languages. A fictional physicist once said that the universe is super asymmetrical (Lorre et al., 2018), and I believe this is the same thing for languages. Certainly, the NLP community has made great progress at a general, universal level, without always taking into account the particularities of the different communities and the context in which their languages exist in order to be able to design real-life solutions to real-life problems.

Bibliography

Abalain, Hervé (2000). *Histoire de la Langue Bretonne*. Vol. 10. Editions Jean-Paul Gisserot.

Abraham, Basil, Goel, Danish, Siddarth, Divya, Bali, Kalika, Chopra, Manu, Choudhury, Monojit, Joshi, Pratik, Jyoti, Preethi, Sitaram, Sunayana, and Seshadri, Vivek (2020). Crowdsourcing speech data for low-resource languages from low-income workers. In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pp. 2819–2826.

Adams, Oliver, Cohn, Trevor, Neubig, Graham, Cruz, Hilaria, Bird, Steven, and Michaud, Alexis (2018). Evaluating phonemic transcription of low-resource tonal languages for language documentation. In: *Proceedings of the 11th Language Resources and Evaluation Conference (LREC)*, pp. 3356–3365.

Adams, Oliver, Cohn, Trevor, Neubig, Graham, and Michaud, Alexis (2017). Phonemic transcription of low-resource tonal languages. In: *Proceedings of the Australasian Language Technology Association Workshop (ALTA)*, pp. 53–60.

Adda, Gilles, Stüker, Sebastian, Adda-Decker, Martine, Ambouroue, Odette, Besacier, Laurent, Blachon, David, Bonneau-Maynard, Hélene, Godard, Pierre, Hamlaoui, Fatima, Idiatov, Dmitry, et al. (2016). Breaking the unwritten language barrier: The BULB project. In: *Procedia Computer Science* 81, pp. 8–14.

Anoop, CS, Prathosh, AP, and Ramakrishnan, AG (2021). Unsupervised domain adaptation schemes for building ASR in low-resource languages. In: *Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, pp. 342–349.

Australian Bureau of Statistics (2016). *Estimates of Aboriginal and Torres Strait Islander Australians*. Tech. rep.

Ayobi, Amid, Stawarz, Katarzyna, Katz, Dmitri, Marshall, Paul, Yamagata, Taku, Santos-Rodríguez, Raúl, Flach, Peter, and O'Kane, Aisling Ann (2021). Machine Learning Explanations as Boundary Objects: How AI Researchers Explain and Non-Experts Perceive Machine Learning. In: *Workshop on Transparency and Explanations in Smart Systems (TEXSS)*. CEUR Workshop Proceedings.

Baevski, Alexei, Hsu, Wei-Ning, Conneau, Alexis, and Auli, Michael (2021). Unsupervised speech recognition. In: *Advances in Neural Information Processing Systems* 34, pp. 27826–27839.

Baevski, Alexei, Zhou, Yuhao, Mohamed, Abdelrahman, and Auli, Michael (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In: *Advances in Neural Information Processing Systems* 33, pp. 12449–12460.

Barry, James, Wagner, Joachim, Cassidy, Lauren, Cowap, Alan, Lynn, Teresa, Walsh, Abigail, Meachair, Mícheál J. Ó, and Foster, Jennifer (2022). gaBERT - an Irish Language Model. In: *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*. Marseille, France: European Language Resources Association, pp. 4774–4788.

Baskin, Cyndy (2006). Aboriginal world views as challenges and possibilities in social work education. In: *Critical Social Work* 7.2, pp. 1–16.

Belenko, Mikhail, Burym, Nikita, Muratova, Uliana, and Balakshin, Pavel (2019). Training aspects of automatic speech recognition systems during chat bot creation. In: *International Multidisciplinary Scientific GeoConference: SGEM* 19.2.1, pp. 681–688.

Benjamin, Jesse Josua, Kinkeldey, Christoph, and Müller-Birn, Claudia (2020). Participatory Design of a Machine Learning Driven Visualization System for Non-Technical Stakeholders. In: *Mensch und Computer 2020-Workshopband*.

Berment, Vincent (2004). Méthodes pour informatiser les langues et les groupes de langues «peu dotées». PhD thesis. Université Joseph Fourier Grenoble 1.

Bernard, Mathieu, Poli, Maxime, Karadayi, Julien, and Dupoux, Emmanuel (2021). Shennong: a Python toolbox for audio speech features extraction. In: *arXiv preprint arXiv:2112.05555*.

Besacier, Laurent, Barnard, Etienne, Karpov, Alexey, and Schultz, Tanja (2014). Automatic speech recognition for Under-Resourced Languages: A survey. In: *Speech Communication* 56, pp. 85–100.

Bettinson, Mat and Bird, Steven (2017). Developing a suite of mobile applications for collaborative language documentation. In: *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pp. 156–164.

— (2021). Designing to Support Remote Working Relationships with Indigenous Communities. In: *Proceedings of the Australian Workshop on Computer Human Interaction*, pp. 165–169.

Bird, Steven (2020a). Decolonising Speech and Language Technology. In: *28th International Conference on Computational Linguistics (COLING)*. International Committee on Computational Linguistics, pp. 3504–3519.

— (2020b). Sparse transcription. In: *Computational Linguistics* 46, pp. 713–744.

— (2022). Local Languages, Contact Languages, and other High-Resource Scenarios. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 7817–7829.

Bird, Steven, Hanke, Florian R, Adams, Oliver, and Lee, Haejoong (2014). Aikuma: A mobile app for collaborative language documentation. In: *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pp. 1–5.

Bird, Steven and Simons, Gary (2001). The OLAC metadata set and controlled vocabularies. In: *Proceedings of Workshop on Sharing Tools and Resources*. Vol. 15, pp. 7–18.

Boas, Hans Christian (2006). From the field to the web: implementing best-practice recommendations in documentary linguistics. In: *Language Resources and Evaluation* 40.2, pp. 153–174.

Bontogon, Megan, Arppe, Antti, Antonsen, Lene, Thunder, Dorothy, and Lachler, Jordan (2018). Intelligent computer assisted language learning (ICALL) for nêhiyawêwin: an in-depth user-experience evaluation. In: *Canadian Modern Language Review* 74.3, pp. 337–362.

Bouquiaux, Luc and Thomas, Jacqueline M. C. (1992). *Studying and describing unwritten languages*. Dallas: Summer Institute of Linguistics.

Budby, John (2001). The academic quandary—an Aboriginal experience. In: *Postgraduate Research Supervision: Transforming Relations*, pp. 247–253.

Carlin, Emma, Atkinson, David, and Marley, Julia V (2019). "Having a quiet word": yarning with Aboriginal women in the Pilbara region of Western Australia about mental health and mental health screening during the perinatal period. In: *International Journal of Environmental Research and Public Health* 16.21, p. 4253.

Caselli, Tommaso, Cibin, Roberto, Conforti, Costanza, Encinas, Enrique, and Teli, Maurizio (Aug. 2021). Guiding Principles for Participatory Design-inspired Natural Language Processing. In: *Proceedings of the 1st Workshop on NLP for Positive Impact*. Association for Computational Linguistics, pp. 27–35.

Chelliah, Shobhana and De Reuse, Willem (2010). *Handbook of Descriptive Linguistic Fieldwork*. Springer Science & Business Media.

Christie, Michael (2013a). The box of veggies: Method and metaphysics in Yolŋu research. In: *Discourse, Power, and Resistance Down Under: Volume 2*. Brill Sense, pp. 45–56.

Christie, Michael J (2013b). Generative and 'ground-up' research in Aboriginal Australia. In: *Learning Communities: International Journal of Learning in Social Contexts* 13, pp. 3–12.

Chung, Yu-An, Wu, Chao-Chung, Shen, Chia-Hao, Lee, Hung-Yi, and Lee, Lin-Shan (2016). Unsupervised learning of audio segment representations using sequence-to-sequence recurrent neural networks. In: *Proceedings of Interspeech 2016*, pp. 765–769.

Cochran, Patricia AL, Marshall, Catherine A, Garcia-Downing, Carmen, Kendall, Elizabeth, Cook, Doris, McCubbin, Laurie, and Gover, Reva Mariah S (2008). Indigenous ways of knowing: Implications for participatory research and community. In: *American Journal of Public Health* 98.1, pp. 22–27.

Conneau, Alexis, Baevski, Alexei, Collobert, Ronan, Mohamed, Abdelrahman, and Auli, Michael (2021). Unsupervised cross-lingual representation learning for speech recognition. In: *Proceedings of Interspeech 2021*.

Crowley, Terry (2007). *Field linguistics: A Beginner's Guide*. OUP Oxford.

Culbertson, Gabriel, Shen, Solace, Jung, Malte, and Andersen, Erik (2017). Facilitating development of pragmatic competence through a voice-driven video learning interface. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 1431–1440.

Davis, Steven and Mermelstein, Paul (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In: *Transactions on Acoustics, Speech, and Signal Processing* 28.4, pp. 357–366.

De Vries, Nic J, Davel, Marelie H, Badenhorst, Jaco, Basson, Willem D, De Wet, Febe, Barnard, Etienne, and De Waal, Alta (2014). A smartphone-based ASR data collection tool for under-resourced languages. In: *Speech Communication* 56, pp. 119–131.

Descola, Philippe (2005). *Par-delà Nature et Culture*. Vol. 1. Gallimard Paris.

Desot, Thierry, Portet, François, and Vacher, Michel (2019). Towards end-to-end spoken intent recognition in smart home. In: *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. IEEE, pp. 1–8.

Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina (June 2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: pp. 4171–4186.

Dixon, Robert Malcolm Ward (2011). *The Languages of Australia*. Cambridge University Press.

Dorian, Nancy C (2010). Documentation and responsibility. In: *Language & Communication* 30.3, pp. 179–185.

Duong, Long (2017). Natural language processing for resource-poor languages. PhD thesis. University of Melbourne.

Dyer, Bill (2022). New syntactic insights for automated Wolof Universal Dependency parsing. In: *Proceedings of the 5th Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pp. 5–12.

Eskander, Ramy, Lowry, Cass, Khandagale, Sujay, Klavans, Judith L, Polinsky, Maria, and Muresan, Smaranda (2022). Unsupervised Stem-based Cross-lingual Part-of-Speech Tagging for Morphologically Rich Low-Resource Languages. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4061–4072.

Ethnologue, Languages of the world (n.d.). URL: https://www.ethnologue.com/ (visited on 10/28/2022).

Fer, Radek, Matějka, Pavel, Grézl, František, Plchot, Oldřich, Veselỳ, Karel, and Černockỳ, Jan Honza (2017). Multilingually trained bottleneck features in spoken language recognition. In: *Computer Speech & Language* 46, pp. 252–267.

Finn, Aoife, Jones, Peter-Lucas, Mahelona, Keoni, Duncan, Suzanne, and Leoni, Gianna (2022). Developing a Part-Of-Speech tagger for te reo Māori. In: *Proceedings of the 5th Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pp. 93–98.

First Languages Australia (2014). *Angkety Map: Digital Resource Report*. Tech. rep. https://www.firstlanguages.org.au/images/fla-angkety-map.pdf; accessed Nov 2020.

Fishman, Joshua A. (1991). Reversing Language Shift. In: *Clevedon, UK: Multilingual Matters*.

Foley, Ben, Arnold, Josh, Coto-Solano, Rolando, Durantin, Gautier, Ellison, T. Mark, Esch, Daan van, Heath, Scott, Kratochví, František, Maxwell-Smith, Zara, Nash, David, Olsson, Ola, Richards, Mark, San, Nay, Stoakes, Hywel, Thieberger, Nick, and Wiles, Janet (2018a). Building Speech Recognition Systems for Language Documentation: The CoEDL Endangered Language Pipeline and Inference System. In: *Proceedings of the 6th International Workshop on Spoken Language Technologies for Under-Resourced Languages*. ISCA, pp. 205–209.

Foley, Ben, Arnold, Joshua T, Coto-Solano, Rolando, Durantin, Gautier, Ellison, T Mark, Esch, Daan van, Heath, Scott, Kratochvil, Frantisek, Maxwell-Smith, Zara, Nash, David, et al. (2018b). Building Speech Recognition Systems for Language Documentation: The CoEDL Endangered Language Pipeline and Inference System (ELPIS). In: *Proceedings of The 6th International Workshop on Spoken Language Technologies for Under-Resourced Languages*, pp. 205–209.

Foley, Dennis (2003). Indigenous epistemology and Indigenous standpoint theory. In: *Social Alternatives* 22.1, pp. 44–52.

Ford, Linda, Christie, Michael, Bow, Catherine, Nasir, Tanyah, Spencer, Michaela, Campbell, Matt, Verran, Helen, and Prior, John (2021). Collaborative Research into Contemporary Indigenous Governance. In: *Leading from the North: Rethinking Northern Australia Development*. ANU Press, pp. 479–492.

Fredericks, Bronwyn, Adams, Karen, Finlay, Summer, Fletcher, Gillian, Andy, Simone, Briggs, Lyn, Briggs, Lisa, and Hall, Robert (2011). Engaging the practice of Indigenous yarning in action research. In: *ALAR: Action Learning and Action Research Journal* 17.2, pp. 12–24.

Gales, Mark JF (1998). Maximum likelihood linear transformations for HMM-based speech recognition. In: *Computer speech & language* 12.2, pp. 75–98.

Galla, Candace Kaleimamoowahinekapu (2018). Digital realities of Indigenous language revitalization: A look at Hawaiian language technology in the modern world. In: *Language and Literacy* 20.3, pp. 100–120.

Gauthier, Elodie, Besacier, Laurent, and Voisin, Sylvie (2016a). Speed perturbation and vowel duration modeling for ASR in Hausa and Wolof languages. In: *Proceedings of Interspeech 2016*, pp. 3529–3533.

Gauthier, Elodie, Besacier, Laurent, Voisin, Sylvie, Melese, Michael, and Elingui, Uriel Pascal (2016b). Collecting Resources in Sub-Saharan African Languages for Automatic Speech Recognition: a Case Study of Wolof. In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, pp. 3863–3867.

Gauthier, Elodie, Blachon, David, Besacier, Laurent, Kouarata, Guy-Noël, Adda-Decker, Martine, Rialland, Annie, Adda, Gilles, and Bachman, Grégoire (2016c). Lig-Aikuma: A Mobile App to Collect Parallel Speech for Under-Resourced Language Studies. In: *Proceedings of Interspeech 2016*, pp. 381–382.

Gawne, Lauren, Kelly, Barbara F, Berez-Kroeker, Andrea L, and Heston, Tyler (2017). Putting Practice into Words: The State of Data and Methods Transparency in Grammatical Descriptions. In: *Language Documentation & Conservation* 11, pp. 157–189.

Geia, Lynore K, Hayes, Barbara, and Usher, Kim (2013). Yarning/Aboriginal storytelling: Towards an understanding of an Indigenous perspective and its implications for research practice. In: *Contemporary Nurse* 46.1, pp. 13–17.

Gessler, Luke (2022). Closing the NLP Gap Documentary Linguistics and NLP Need a Shared Software Infrastructure. In: *Proceedings of the 5th Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pp. 119–126.

Godard, Pierre, Adda, Gilles, Adda-Decker, Martine, Benjumea, Juan, Besacier, Laurent, Cooper-Leavitt, Jamison, Kouarata, Guy-Noel, Lamel, Lori, Bonneau-Maynard, Hélène, Mueller, Markus, et al. (2018a). A Very Low Resource Language Speech Corpus for Computational Language Documentation Experiments.

In: *Proceedings of the 11th International Conference on Language Resources and Evaluation*, pp. 3366–70.

Godard, Pierre, Besacier, Laurent, Yvon, François, Adda-Decker, Martine, Adda, Gilles, Maynard, Hélène, and Rialland, Annie (2018b). Adaptor grammars for the linguist: Word segmentation experiments for very low-resource languages. In: *Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pp. 32–42.

Góngora, Santiago, Giossa, Nicolás, and Chiruzzo, Luis (2022). Can We Use Word Embeddings for Enhancing Guarani-Spanish Machine Translation? In: *Proceedings of the 5th Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pp. 127–132.

Gunning, David (2017). Explainable artificial intelligence (XAI). In: *Defense Advanced Research Projects agency (DARPA), nd Web* 2.2, p. 1.

Gupta, Akshat (July 2022). On Building Spoken Language Understanding Systems for Low Resourced Languages. In: *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Seattle, Washington: Association for Computational Linguistics.

Gupta, Vishwa and Boulianne, Gilles (2020a). Automatic Transcription Challenges for Inuktitut, a Low-Resource Polysynthetic Language. In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pp. 2521–27.

— (2020b). Speech Transcription Challenges for Resource Constrained Indigenous Language Cree. In: *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-Resourced Languages and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pp. 362–367.

Hanke, Florian (2017). Computer-Supported Cooperative Language Documentation. PhD thesis. University of Melbourne.

Hardy, Dianna, Forest, Elizabeth, McIntosh, Zoe, Myers, Trina, and Gertz, Janine (2016). Moving beyond "just tell me what to code" inducting tertiary ICT students into research methods with aboriginal participants via games design. In:

Proceedings of the 28th Australian Conference on Computer-Human Interaction, pp. 557–561.

Harrigan, Atticus, Mills, Timothy, and Arppe, Antti (2019). A Preliminary Plains Cree Speech Synthesizer. In: *Workshop on Computational Methods for Endangered Languages*. Vol. 1, p. 9.

Hasegawa-Johnson, Mark, Rolston, Leanne, Goudeseune, Camille, Levow, Gina-Anne, and Kirchhoff, Katrin (2020). Grapheme-to-phoneme transduction for cross-language ASR. In: *International Conference on Statistical Language and Speech Processing*. Springer, pp. 3–19.

Haynes, Emma, Marawili, Minitja, Marika, Brendan Makungun, Mitchell, Alice G, Phillips, Jodi, Bessarab, Dawn, Walker, Roz, Cook, Jeff, and Ralph, Anna P (2019). Community-based participatory action research on rheumatic heart disease in an Australian Aboriginal homeland: Evaluation of the 'On track watch'project. In: *Evaluation and program planning* 74, pp. 38–53.

Hewett, Thomas T, Baecker, Ronald, Card, Stuart, Carey, Tom, Gasen, Jean, Mantei, Marilyn, Perlman, Gary, Strong, Gary, and Verplank, William (1992). *ACM SIGCHI curricula for human-computer interaction*. ACM.

Himmelmann, Nikolaus P. (1998). *Documentary and descriptive linguistics*. Vol. 36. 1. de Gruyter, pp. 161–195.

Hinton, Leanne (2015). What counts as "success" in language revitalization. In: *University of British Colombia*.

Hohman, Fred, Head, Andrew, Caruana, Rich, DeLine, Robert, and Drucker, Steven M (2019). Gamut: A design probe to understand how data scientists understand machine learning models. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–13.

Hüe, Denis (2010). *Rémanences, Mémoire de la forme dans la littérature médiévale*. Honoré Champion, pp. 17–18.

Ishmam, Alvi Md and Sharmin, Sadia (2019). Hateful speech detection in public Facebook pages for the Bengali language. In: *International Conference on Machine Learning and Applications (ICMLA)*. IEEE, pp. 555–560.

Jansen, Aren and Van Durme, Benjamin (2011). Efficient spoken term discovery using randomized algorithms. In: *Workshop on Automatic Speech Recognition & Understanding*. IEEE, pp. 401–406.

Jimerson, Robbie and Prud'hommeaux, Emily (2018). ASR for documenting acutely under-resourced Indigenous languages. In: *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*, pp. 4161–4166.

Joshi, Pratik, Santy, Sebastin, Budhiraja, Amar, Bali, Kalika, and Choudhury, Monojit (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6282–6293.

Juan, Sarah Samson, Besacier, Laurent, Lecouteux, Benjamin, and Dyab, Mohamed (2015). Using resources from a closely-related language to develop ASR for a very under-resourced language: A case study for Iban. In: *Proceedings of Interspeech 2015*, pp. 1270–1274.

Jullien, François (2021). *Altérités: de l'Altérité Personnelle à l'Altérité Culturelle*. Gallimard.

Kamper, Herman (2017). Unsupervised neural and Bayesian models for zero-resource speech processing. PhD thesis. The University of Edinburgh.

— (2019). Truly unsupervised acoustic word embeddings using weak top-down constraints in encoder-decoder models. In: *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 6535–6539.

Kamper, Herman, Elsner, Micha, Jansen, Aren, and Goldwater, Sharon (2015). Unsupervised neural network based feature extraction using weak top-down constraints. In: *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 5818–5822.

Kazantseva, Anna, Maracle, Owennatekha Brian, Pine, Aidan, et al. (2018). Kawennón: nis: the Wordmaker for Kanyen'kéha. In: *Proceedings of the workshop on computational modeling of polysynthetic languages*, pp. 53–64.

Khare, Shreya, Mittal, Ashish, Diwan, Anuj, Sarawagi, Sunita, Jyothi, Preethi, and Bharadwaj, Samarth (2021). Low Resource ASR: The surprising effectiveness of High Resource Transliteration. In: *Proceedings of Interspeech 2021*, pp. 1529–1533.

Kuhn, Roland, Davis, Fineen, Désilets, Alain, Joanis, Eric, Kazantseva, Anna, Knowles, Rebecca, Littell, Patrick, Lothian, Delaney, Pine, Aidan, Running Wolf, Caroline, Santos, Eddie, Stewart, Darlene, Boulianne, Gilles, Gupta, Vishwa, Maracle Owennatékha, Brian, Martin, Akwiratékha', Cox, Christopher, Junker, Marie-Odile, Sammons, Olivia, Torkornoo, Delasie, Thanyehténhas Brinklow, Nathan, Child, Sara, Farley, Benoît, Huggins-Daines, David, Rosenblum, Daisy, and Souter, Heather (2020). The Indigenous Languages Technology project at NRC Canada: An empowerment-oriented approach to developing language software. In: *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*. Barcelona, Spain: International Committee on Computational Linguistics, pp. 5866–5878.

Kumar, Vinit, Kumar, Avinash, and Shahnawazuddin, S (2022). Creating Robust Children's ASR System in Zero-Resource Condition Through Out-of-Domain Data Augmentation. In: *Circuits, Systems, and Signal Processing* 41.4, pp. 2205–2220.

Lackaff, Derek and Moner, William J (2016). Local languages, global networks: Mobile design for minority language users. In: *Proceedings of the 34th ACM International Conference on the Design of Communication*, pp. 1–9.

Lafourcade, Mathieu (2007). Making people play for Lexical Acquisition with the JeuxDeMots prototype. In: *SNLP'07: 7th International Symposium on Natural Language Processing*, p. 7.

Lamb, William and Danso, Samuel (2014). Developing an automatic part-of-speech tagger for Scottish Gaelic. In: *Proceedings of the 1st Celtic Language Technology Workshop*, pp. 1–5.

Lane, William and Bird, Steven (2019). Towards A Robust Morphological Analyzer for Kunwinjku. In: *Proceedings of the Australasian Language Technology Association Workshop (ALTA)*, p. 1.

— (2020). Interactive Word Completion for Morphologically Complex Languages. In: *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pp. 4600–4611.

Lavallée, Lynn F (2009). Practical application of an Indigenous research framework and two qualitative Indigenous research methods: Sharing circles and Anishnaabe symbol-based reflection. In: *International Journal of Qualitative Methods* 8.1, pp. 21–40.

Le, Hang, Vial, Loïc, Frej, Jibril, Segonne, Vincent, Coavoux, Maximin, Lecouteux, Benjamin, Allauzen, Alexandre, Crabbé, Benoit, Besacier, Laurent, and Schwab, Didier (2020). FlauBERT: Unsupervised Language Model Pre-training for French. In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pp. 2479–2490.

Lee, En-Shiun, Thillainathan, Sarubi, Nayak, Shravan, Ranathunga, Surangika, Adelani, David, Su, Ruisi, and McCarthy, Arya D (2022). Pre-Trained Multilingual Sequence-to-Sequence Models: A Hope for Low-Resource Language Translation? In: *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 58–67.

Lee, K-F, Hon, H-W, and Reddy, Raj (1990). An overview of the SPHINX speech recognition system. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 38.1, pp. 35–45.

Leong, Colin and Whitenack, Daniel (2022). Phone-ing it in: Towards Flexible Multi-Modal Language Model Training by Phonetic Representations of Data.

In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5306–5315.

Leong, Tuck Wah, Lawrence, Christopher, and Wadley, Greg (2019). Designing for diversity in Aboriginal Australia: Insights from a national technology project. In: *Proceedings of the 31st Australian Conference on Human-Computer-Interaction*, pp. 418–422.

Lewis, M Paul and Simons, Gary F (2016). *Sustaining Language Use*. SIL International Publications.

Li, Xinjian, Dalmia, Siddharth, Black, Alan W, and Metze, Florian (2019). Multilingual Speech Recognition with Corpus Relatedness Sampling. In: *Proceedings of Interspeech 2019*, pp. 2120–2124.

Li, Xinjian, Dalmia, Siddharth, Li, Juncheng, Lee, Matthew, Littell, Patrick, Yao, Jiali, Anastasopoulos, Antonios, Mortensen, David R, Neubig, Graham, Black, Alan W, et al. (2020a). Universal phone recognition with a multilingual allophone system. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 8249–8253.

Li, Xinjian, Dalmia, Siddharth, Mortensen, David, Li, Juncheng, Black, Alan, and Metze, Florian (2020b). Towards zero-shot learning for automatic phonemic transcription. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05, pp. 8261–8268.

Lignos, Constantine, Holley, Nolan, Palen-Michel, Chester, and Sälevä, Jonne (2022). Toward More Meaningful Resources for Lower-resourced Languages. In: *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 523–532.

Liu, Zoey, Richardson, Crystal, Hatcher, Richard, and Prud'hommeaux, Emily (May 2022a). Not always about you: Prioritizing community needs when developing endangered language technology. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 3933–3944.

Liu, Zoey, Richardson, Crystal, Hatcher, Richard, and Prud'hommeaux, Emily (2022b). Not always about you: Prioritizing community needs when developing endangered language technology. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 3933–3944.

Loi, Daria, Wolf, Christine T, Blomberg, Jeanette L, Arar, Raphael, and Brereton, Margot (2019). Co-designing AI futures: Integrating AI ethics, social computing, and design. In: *Companion Publication of the 2019 on Designing Interactive Systems Conference 2019 Companion*, pp. 381–384.

Lorre, Chuck, Prady, Bill, Holland (Writers), Steve, and Cecil (Director), Kristy (2018). *The Citation Negation - The Big Bang Theory*. Warner Bros. Television.

Lyons, Natasha (2011). Creating space for negotiating the nature and outcomes of collaborative research projects with Aboriginal communities. In: *Études/Inuit/Studies* 35.1-2, pp. 83–105.

Légifrance. Le service publique de la diffusion du droit (n.d.). URL: https://www.legifrance.gouv.fr/loda/id/LEGITEXT000006071194/\#75-1 (visited on 10/28/2022).

Maar, MA, Lightfoot, NE, Sutherland, ME, Strasser, RP, Wilson, KJ, Lidstone-Jones, CM, Graham, DG, Beaudin, R, Daybutch, GA, Dokis, BR, et al. (2011). Thinking outside the box: Aboriginal people's suggestions for conducting health studies with Aboriginal communities. In: *Public Health* 125.11, pp. 747–753.

Macaire, Cécile, Schwab, Didier, Lecouteux, Benjamin, and Schang, Emmanuel (2022). Automatic Speech Recognition and Query By Example for Creole Languages Documentation. In: *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2512–2520.

Matsuura, Kohei, Mimura, Masato, Sakai, Shinsuke, and Kawahara, Tatsuya (2020). Generative Adversarial Training Data Adaptation for Very Low-Resource Automatic Speech Recognition. In: *Proceedings of Interspeech 2020*, pp. 2737–2741.

Mazumder, Mark, Banbury, Colby, Meyer, Josh, Warden, Pete, and Reddi, Vijay Janapa (2021). Few-Shot Keyword Spotting in Any Language. In: *Proceedings of Interspeech 2021*, pp. 4214–4218.

McNaught, Carmel and Lam, Paul (2010). Using Wordle as a supplementary research tool. In: *Qualitative Report* 15.3, pp. 630–643.

Menon, Raghav, Kamper, Herman, Quinn, John, and Niesler, Thomas (2018). Fast ASR-free and Almost Zero-resource Keyword Spotting Using DTW and CNNs for Humanitarian Monitoring. In: *Proceedings of Interspeech 2018*, pp. 2608–2612.

Menon, Raghav, Kamper, Herman, Westhuizen, Ewald van der, Quinn, John, and Niesler, Thomas (2019). Feature exploration for almost zero-resource ASR-free keyword spotting using a multilingual bottleneck extractor and correspondence autoencoders. In: *Proceedings of Interspeech 2019*, pp. 3475–3479.

Michaud, Alexis, Adams, Oliver, Cohn, Trevor, Neubig, Graham, and Guillaume, Séverine (2018). Integrating automatic transcription into the language documentation workflow: Experiments with Na data and the Persephone toolkit. In: *Language Documentation & Conservation* 12, pp. 393–429.

Miller, Tim, Howe, Piers, and Sonenberg, Liz (2017). Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. In: *arXiv preprint arXiv:1712.00547*.

Moeller, Sarah Ruth (2014). SayMore, a tool for language documentation productivity. In: *Language Documentation and Conservation* 8, pp. 66–74.

Mohseni, Sina, Zarei, Niloofar, and Ragan, Eric D (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 11.3-4, pp. 1–45.

Moraila, Gina, Shankaran, Akash, Shi, Zuoming, and Warren, Alex M (2014). Measuring reproducibility in computer systems research. In: *PLoS Comput Biol* 9.

Morris, Ethan, Jimerson, Robbie, and Prud'hommeaux, Emily (2021). One Size Does Not Fit All in Resource-Constrained ASR. In: *Proceedings of Interspeech 2021*, pp. 4354–4358.

Mortensen, David R, Dalmia, Siddharth, and Littell, Patrick (2018). Epitran: Precision G2P for many languages. In: *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*, pp. 2710–2714.

Murphy, Emma (2020). Racism multiplies COVID-19 risk for Indigenous communities. In: *Green Left Weekly* 1260, pp. 1–5.

Nathan, David and Fang, Meili (2013). Re-imagining documentary linguistics as a revitalization-driven practice. In: *Keeping Languages Alive: Documentation, Pedagogy, and Revitalization*, pp. 42–55.

Ndulue, Chinenye and Orji, Rita (2021). Heuristic evaluation of an African-centric mobile persuasive game for promoting safety measures against COVID-19. In: *3rd African Human-Computer Interaction Conference: Inclusiveness and Empowerment*, pp. 43–51.

Nettle, Daniel and Romaine, Suzanne (2000). *Vanishing voices: The Extinction of the World's Languages*. Oxford University Press on Demand.

Nguyen, Dat Quoc and Tuan Nguyen, Anh (Nov. 2020). PhoBERT: Pre-trained language models for Vietnamese. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics.

NHMRC (n.d.). URL: https://www.nhmrc.gov.au/about-us/resources/ethical-conduct-research-aboriginal-and-torres-strait-islander-peoples-and-communities\#block-views-block-file-attachments-content-block-1 (visited on 10/28/2022).

Nicolai, Garrett, Coates, Edith, Zhang, Ming, and Silfverberg, Miikka (2021). Expanding the JHU Bible Corpus for Machine Translation of the Indigenous Languages of North America. In: *Proceedings of the Workshop on Computational Methods for Endangered Languages*. Vol. 1, pp. 1–5.

Ober, Robyn (2017). Kapati time: storytelling as a data collection method in Indigenous research. In: *Mystery Train* 2007.

Ochs, Elinor (1979). Transcription as theory. In: *Developmental pragmatics* 10.1, pp. 43–72.

Ogie, Ota (2010). Using an online tool for the documentation of Edo language. In: *Proceedings of the 4th Linguistic Annotation Workshop*, pp. 109–112.

Olawsky, Knut J (2010). Going public with language: involving the wider community in language revitalisation. In: *Re-Awakening Languages*, p. 75.

Oliver, Stefanie J (2013). The role of traditional medicine practice in primary health care within Aboriginal Australia: a review of the literature. In: *Journal of Ethnobiology and Ethnomedicine* 9.1, pp. 1–8.

Ovide, Evaristo and García-Peñalvo, Francisco José (2016). A technology-based approach to revitalise indigenous languages and cultures in online environments. In: *Proceedings of the 4th International Conference on Technological Ecosystems for Enhancing Multiculturality*, pp. 1155–1160.

Palaskar, Shruti, Sanabria, Ramon, and Metze, Florian (2018). End-to-end Multimodal Speech Recognition. In: *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5774–5778.

Park, Alex and Glass, James R (2005). Towards unsupervised pattern discovery in speech. In: *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, pp. 53–58.

Peters, Dorian, Hansen, Susan, McMullan, Jenny, Ardler, Theresa, Mooney, Janet, and Calvo, Rafael A (2018). "Participation is not enough" - towards indigenous-led co-design. In: *Proceedings of the 30th Australian Conference on Computer-Human Interaction*, pp. 97–101.

Pine, Aidan, Littell, Patrick William, Joanis, Eric, Huggins-Daines, David, Cox, Christopher, Davis, Fineen, Santos, Eddie Antonio, Srikanth, Shankhalika, Torkornoo, Delasie, and Yu, Sabrina (2022a). Gi2Pi Rule-based, index-preserving grapheme-to-phoneme transformations. In: *Proceedings of the 5th Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pp. 52–60.

Pine, Aidan, Wells, Dan, Brinklow, Nathan, Littell, Patrick, and Richmond, Korin (2022b). Requirements and Motivations of Low-Resource Speech Synthesis for

Language Revitalization. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7346–7359.

Povey, Daniel, Ghoshal, Arnab, Boulianne, Gilles, Burget, Lukas, Glembek, Ondrej, Goel, Nagendra, Hannemann, Mirko, Motlicek, Petr, Qian, Yanmin, Schwarz, Petr, Silovsky, Jan, Stemmer, Georg, and Vesely, Karel (2011). The Kaldi Speech Recognition Toolkit. In: *Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE Signal Processing Society.

Prud'hommeaux, Emily, Jimerson, Robbie, Hatcher, Richard, and Michelson, Karin (2021). Automatic Speech Recognition for Supporting Endangered Language Documentation. In: *Language Documentation & Conservation* 15.

Pugh, Robert and Tyers, Francis (2021). Towards an open source finite-state morphological analyzer for Zacatlán-Ahuacatlán-Tepetzintla Nahuatl. In: *Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pp. 80–85.

Pulugundla, Bhargav, Baskar, Murali Karthick, Kesiraju, Santosh, Egorova, Ekaterina, Karafiát, Martin, Burget, Lukás, and Cernockỳ, Jan (2018). BUT system for low resource Indian Language ASR. In: *Proceedings of Interspeech 2018*, pp. 3182–3186.

Radio Kerne (n.d.). URL: https://www.radiokerne.bzh/fr/ (visited on 10/28/2022).

Rankin, Yolanda A and Edwards, Mya S (2017). The Choices we make: game design to promote second language acquisition. In: *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 907–916.

Remote Area Health Corps (2010). *Community Profile Gunbalanya Oenpelli*. Tech. rep.

Rigney, Lester-Irabinna (2001). A first perspective of Indigenous Australian participation in science: Framing Indigenous research towards Indigenous Australian intellectual sovereignty. In: pp. 1–13.

Rodríguez Louro, Celeste and Collard, Glenys (2021). Working together: Sociolinguistic research in urban Aboriginal Australia. In: *Journal of Sociolinguistics* 25.5, pp. 785–807.

Roman, Sue and Harvey, Mark (2003). *Larrakia language project: a living culture in a changing world–a tribute to Yirra*. The Australian National University.

Rosenberg, Andrew, Audhkhasi, Kartik, Sethy, Abhinav, Ramabhadran, Bhuvana, and Picheny, Michael (2017). End-to-end speech recognition and keyword search on low-resource languages. In: *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 5280–5284.

Rouedad Skolioù Diwan (n.d.). URL: https://www.diwan.bzh/ (visited on 10/28/2022).

Ryder, Courtney, Mackean, Tamara, Hunter, Kate, Coombes, Julieann, Holland, Andrew JA, and Ivers, Rebecca (2021). Yarning up about out-of-pocket healthcare expenditure in burns with Aboriginal families. In: *Australian and New Zealand journal of public health* 45.2, pp. 138–142.

Saeb, Armin, Menon, Raghav, Cameron, Hugh, Kibira, William, Quinn, John, and Niesler, Thomas (2017). Very Low Resource Radio Browsing for Agile Developmental and Humanitarian Monitoring. In: *Proceedings of Interspeech 2017*, pp. 2118–2122.

Sakoe, Hiroaki and Chiba, Seibi (1978). Dynamic programming algorithm optimization for spoken word recognition. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 26, pp. 43–49.

Salimbajevs, Askars (2018). Creating Lithuanian and Latvian speech corpora from inaccurately annotated web data. In: *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*.

Samarin, Willlam J (1984). The linguistic world of field colonialism. In: *Language in Society* 13.4, pp. 435–453.

Samson Juan, Sarah, Besacier, Laurent, and Rossato, Solange (2014). Semi-Supervised G2P Bootstrapping and its Application to ASR for a very Under-Resourced Language: Iban. In: *Spoken Language Technologies for Under-Resourced Languages (SLTU)*, pp. 66–72.

San, Nay, Bartelds, Martijn, Browne, Mitchell, Clifford, Lily, Gibson, Fiona, Mansfield, John, Nash, David, Simpson, Jane, Turpin, Myfany, Vollmer, Maria, Wilmoth, Sasha, and Jurafsky, Dan (2021). Leveraging neural representations for facilitating access to untranscribed speech from endangered languages. In: *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU)*.

San, Nay, Bartelds, Martijn, Ogunremi, Tolulope, Mount, Alison, Thompson, Ruben, Higgins, Michael, Barker, Roy, Simpson, Jane Helen, and Jurafsky, Dan (2022). Automated speech tools for helping communities process restricted-access corpora for language revival efforts. In: *Proceedings of the 5th Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pp. 41–51.

Scharenborg, Odette, Ciannella, Francesco, Palaskar, Shruti, Black, Alan, Metze, Florian, Ondel, Lucas, and Hasegawa-Johnson, Mark (2017). Building an ASR system for a low-research language through the adaptation of a high-resource language ASR system: preliminary results. In: *Proc. Internat. Conference on Natural Language, Signal and Speech Processing (ICNLSSP)*, pp. 26–30.

Schneider, Steffen, Baevski, Alexei, Collobert, Ronan, and Auli, Michael (2019). Wav2vec: Unsupervised Pre-Training for Speech Recognition. In: *Proceedings of Interspeech 2019*, pp. 3465–3469.

Schreiner, Sylvia LR, Schwartz, Lane, Hunt, Benjamin, and Chen, Emily (2020). Multidirectional leveraging for computational morphology and language documentation and revitalization. In: *Language Documentation and Conservation* 14.

Schwartz, Lane (2022). Primum Non Nocere: Before working with Indigenous data, the ACL must confront ongoing colonialism. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 724–731.

Schwartz, Lane, Chen, Emily, Hunt, Benjamin, and Schreiner, Sylvia LR (2019). Bootstrapping a neural morphological analyzer for St. Lawrence Island Yupik from a finite-state transducer. In: *Proceedings of the Workshop on Computational Methods for Endangered Languages*. Vol. 1, pp. 87–96.

Schwartz, Lane, Chen, Emily, Park, Hyunji Hayley, Jahn, Edward, and Schreiner, Sylvia LR (2021). A Digital Corpus of St. Lawrence Island Yupik. In: *Proceedings of the 4th Workshop on Computational Methods for Endangered Languages*, pp. 69–86.

Settle, Shane, Levin, Keith, Kamper, Herman, and Livescu, Karen (2017). Query-by-Example Search with Discriminative Neural Acoustic Word Embeddings. In: *Proceedings of Interspeech 2017*, pp. 2874–2878.

Settle, Shane and Livescu, Karen (2016). Discriminative acoustic word embeddings: Tecurrent neural network-based approaches. In: *Spoken Language Technology Workshop (SLT)*. IEEE, pp. 503–510.

Shetty, Vishwas M and NJ, Metilda Sagaya Mary (2020). Improving the performance of transformer based low resource speech recognition for Indian languages. In: *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 8279–8283.

Shi, Jiatong, Amith, Jonathan D, García, Rey Castillo, Sierra, Esteban Guadalupe, Duh, Kevin, and Watanabe, Shinji (2021). Leveraging End-to-End ASR for Endangered Language Documentation: An Empirical Study on Yolóxochitl Mixtec. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 1134–1145.

Silnova, Anna, Matejka, Pavel, Glembek, Ondrej, Plchot, Oldrich, Novotnỳ, Ondrej, Grezl, Frantisek, Schwarz, Petr, Burget, Lukas, and Cernockỳ, Jan (2018). BUT/Phonexia Bottleneck Feature Extractor. In: *Odyssey*, pp. 283–287.

Siminyu, Kathleen, Li, Xinjian, Anastasopoulos, Antonios, Mortensen, David R, Marlo, Michael, and Neubig, Graham (2021). Phoneme Recognition through Fine Tuning of Phonetic Representations: a Case Study on Luhya Language Varieties. In: *Proceedings of Interspeech 2021*.

Simoulin, Antoine and Crabbé, Benoit (2021). Un modèle Transformer Génératif Pré-entrainé pour le _ français. In: *Traitement Automatique des Langues Naturelles*. ATALA, pp. 245–254.

Singer, Judy, Bennett-Levy, James, and Rotumah, Darlene (2015). "You didn't just consult community, you involved us": transformation of a 'top-down' Aboriginal mental health project into a 'bottom-up' community-driven process. In: *Australasian Psychiatry* 23.6, pp. 614–619.

Sloetjes, Han, Stehouwer, Herman, and Drude, Sebastian (2013). *Novel developments in Elan*. Paper presented at the Third International Conference on Language Documentation and Conservation, http://hdl.handle.net/10125/26154.

Smith, Linda Tuhiwai (1999). *Decolonizing Methodologies: Research and Indigenous Peoples*. London, Engalnd: Zed Book Ltd.

Soro, Alessandro, Brereton, Margot, Taylor, Jennyfer Lawrence, Hong, Anita Lee, and Roe, Paul (2017). A cross-cultural noticeboard for a remote community: design, deployment, and evaluation. In: *IFIP Conference on Human-Computer Interaction*. Springer, pp. 399–419.

Stickler, Ursula and Shi, Lijing (2016). TELL us about CALL: An introduction to the Virtual Special Issue (VSI) on the development of technology enhanced and computer assisted language learning published in the System Journal. In: *System* 56, pp. 119–126.

Stuker, Sebastian, Besacier, Laurent, and Waibel, Alex (2009). Human Translations Guided Language Discovery for ASR Systems. In: *10th International Conference on Speech Science and Speech Technology*. Eurasip, pp. 1–4.

Taylor, Jennyfer Lawrence, Aboriginal Shire Council, Wujal Wujal, Soro, Alessandro, Esteban, Michael, Vallino, Andrew, Roe, Paul, and Brereton, Margot (2020). Crocodile Language Friend: Tangibles to Foster Children's Language Use. In: *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–14.

Tejedor, Javier, Toledano, Doroteo T, Lopez-Otero, Paula, Docio-Fernandez, Laura, Montalvo, Ana R, Ramirez, Jose M, Peñagarikano, Mikel, and Rodriguez-Fuentes, Luis Javier (2019). ALBAYZIN 2018 spoken term detection evaluation: a multi-domain international evaluation in Spanish. In: *EURASIP Journal on Audio, Speech, and Music Processing* 2019.1, pp. 1–37.

Terare, Mareese and Rawsthorne, Margot (2020). Country is yarning to me: Worldview, health and well-being amongst Australian First Nations people. In: *The British Journal of Social Work* 50.3, pp. 944–960.

Thai, Bao, Jimerson, Robert, Arcoraci, Dominic, Prud'hommeaux, Emily, and Ptucha, Raymond (2019). Synthetic data augmentation for improving low-resource ASR. In: *2019 IEEE Western New York Image and Signal Processing Workshop (WNYISPW)*. IEEE, pp. 1–9.

Thieberger, Nick, Margetts, Anna, Morey, Stephen, and Musgrave, Simon (2016). Assessing annotated corpora as research output. In: *Australian Journal of Linguistics* 36.1, pp. 1–21.

Thompson, Jessica AF, Schönwiesner, Marc, Bengio, Yoshua, and Willett, Daniel (2019). How transferable are features in convolutional neural network acoustic models across languages? In: *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 2827–2831.

Tüske, Zoltán, Golik, Pavel, Nolden, David, Schlüter, Ralf, and Ney, Hermann (2014). Data augmentation, feature combination, and multilingual neural networks to

improve ASR and KWS performance for low-resource languages. In: *15th Annual Conference of the International Speech Communication Association*. Citeseer, pp. 1420–1424.

Tyers, Francis and Henderson, Robert (2021). A corpus of K'iche'annotated for morphosyntactic structure. In: *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pp. 10–20.

Tyers, Francis M, Washington, Jonathan N, Kavitskaya, Darya, Gokırmak, Memduh, Howell, Nick, and Berberova, Remziye (2019). A biscriptual morphological transducer for Crimean Tatar. In: *Proceedings of the Workshop on Computational Methods for Endangered Languages*. Vol. 1, pp. 74–80.

Tébéo. Télé Bretagne Ouest (n.d.). URL: https://www.tebeo.bzh/ (visited on 10/28/2022).

UNESCO (2011). *Atlas of the world's languages in danger*. Tech. rep. UNESCO.

VanderBurgh, David, Jamieson, Rachel, Beardy, Jackson, Ritchie, SD, and Orkin, Aaron (2014). Community-based first aid: a program report on the intersection of community-based participatory research and first aid education in a remote Canadian Aboriginal community. In: *Rural and Remote Health* 14.2, pp. 215–222.

Verran, Helen (2007). The educational value of explicit non-coherence. In: *Education and Technology: Critical Perspectives, Possible Futures*, pp. 101–124.

Walker, Melissa, Fredericks, Bronwyn, Mills, Kyly, and Anderson, Debra (2014). "Yarning" as a method for community-based health research with indigenous women: The Indigenous women's wellness research program. In: *Health Care for Women International* 35.10, pp. 1216–1226.

Waller, Lisa (2016). Australian indigenous public spheres: from the ground up. In: *Critical Arts* 30.6, pp. 788–803.

Wanambi, Gawura, Bulkanhawuy, Joy, Dhamarrandji, Stephen, and Gundjarranbuy, Rosemary (2021). Caring for Yolŋu and Ways of Life During COVID-19. In: https://indigenousx.com.au/caring-for-yolnu-and-ways-of-life-during-covid-19.

Watanabe, Shinji, Hori, Takaaki, Karita, Shigeki, Hayashi, Tomoki, Nishitoba, Jiro, Unno, Yuya, Soplin, Nelson-Enrique Yalta, Heymann, Jahn, Wiesner, Matthew, Chen, Nanxin, Adithya, Renduchintala, and Tsubasa, Ochiai (2018). ESPnet: End-to-End Speech Processing Toolkit. In: *Proceedings of Interspeech 2018*, pp. 2207–2211.

West, Japanangka Errol (1998). Speaking towards an Aboriginal philosophy. In: *1st Conference on Indigenous Philosophy, 'Linga Longa' Philosophy Farm, Rollands Plains, Australia*.

Westhuizen, Ewald van der, Kamper, Herman, Menon, Raghav, Quinn, John, and Niesler, Thomas (2022). Feature learning for efficient ASR-free keyword spotting in low-resource languages. In: *Computer Speech & Language* 71, p. 101275.

Wigglesworth, Gillian, Wilkinson, Melanie, Yunupingu, Yalmay, Beecham, Robyn, and Stockley, Jake (2021). Interdisciplinary and Intercultural Development of an Early Literacy App in Dhuwaya. In: *Languages* 6.2, p. 106.

Winschiers-Theophilus, Heike, Virmasalo, Veera, Samuel, Marly M, Stichel, Brit, and Afrikaner, Helena (2020). Facilitating design for the unknown: An inclusive innovation design journey with a San community in the Kalahari Desert. In: *Proceedings of the 6th International Conference on Design Creativity (ICDC 2020)*, pp. 263–270.

Wolf, Thomas, Debut, Lysandre, Sanh, Victor, Chaumond, Julien, Delangue, Clement, Moi, Anthony, Cistac, Pierric, Rault, Tim, Louf, Rémi, Funtowicz, Morgan, Davison, Joe, Shleifer, Sam, Platen, Patrick von, Ma, Clara, Jernite, Yacine, Plu, Julien, Xu, Canwen, Scao, Teven Le, Gugger, Sylvain, Drame, Mariama, Lhoest, Quentin, and Rush, Alexander M. (Oct. 2020). Transformers: State-of-the-Art Natural Language Processing. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, pp. 38–45.

Woodbury, Anthony C (2003). Defining documentary linguistics. In: *Language Documentation and Description* 1.1, pp. 35–51.

Xu, Fan, Dan, Yangjie, Yan, Keyu, Ma, Yong, and Wang, Mingwen (2021). Low-Resource Language Discrimination toward Chinese Dialects with Transfer Learning and Data Augmentation. In: *Transactions on Asian and Low-Resource Language Information Processing* 21.2, pp. 1–21.

Xu, Jin, Tan, Xu, Ren, Yi, Qin, Tao, Li, Jian, Zhao, Sheng, and Liu, Tie-Yan (2020). Lrspeech: Extremely low-resource speech synthesis and recognition. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2802–2812.

Xu, Liang, Dhonnchadha, Elaine Uí, and Ward, Monica (2022). Faoi Gheasa an adaptive game for Irish language learning. In: *Proceedings of the 5th Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pp. 133–138.

Zaman, Tariq, Winschiers-Theophilus, Heike, Yeo, Alvin W, Ting, Lai Chiu, and Jengan, Garen (2015). Reviving an indigenous rainforest sign language: digital Oroo'adventure game. In: *Proceedings of the 7th International Conference on Information and Communication Technologies and Development*, pp. 1–4.

Zero resource speech challenge (n.d.). URL: https://www.zerospeech.com/ (visited on 10/28/2022).

Zhang, Borui, Kazemzadeh, Abe, and Reese, Brian (2022). Shallow Parsing for Nepal Bhasa Complement Clauses. In: *Proceedings of the 5th Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pp. 61–67.

Zhang, Eda, Culbertson, Gabriel, Shen, Solace, and Jung, Malte (2018). Utilizing narrative grounding to design storytelling gamesfor creative foreign language production. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–11.

Znotins, Arturs and Barzdins, Guntis (2020). LVBERT: Transformer-Based Model for Latvian Language Understanding. In: *Baltic HLT*, pp. 111–115.