Local Languages, Third Spaces, and other High-Resource Scenarios

Steven Bird
Northern Institute
Charles Darwin University
Darwin, Australia

Abstract

How can language technology address the diverse situations of the world's languages? In one view, languages exist on a resource continuum and the challenge is to scale existing solutions, bringing under-resourced languages into the high-resource world. In another view, presented here, the world's language ecology includes standardised languages, local languages, and contact languages. These are often subsumed under the label of 'under-resourced languages' even though they have distinct functions and prospects. I explore this position and propose some ecologically-aware language technology agendas.

1 Introduction

This paper is about the world's *local languages*, by which I mean small, primarily-oral languages, often Indigenous or endangered, including the original and emerging languages of Africa, Asia, Australia, the Americas, the Pacific, and the minority languages of Europe. Local languages are often called *under-resourced* because they lack what is required for creating speech and language technologies (Krauwer, 2003). Some have been called *acutely* under-resourced, because they are spoken by few people and are rarely written down (Jimerson and Prud'hommeaux, 2018). From here, it is a small step down to *zero expert resources* and the *zero resource scenario* (Dunbar et al., 2017).

I depict this situation in Figure 1. In the middle we have standardised languages, including 'high-resource' languages (e.g. English, Spanish, Mandarin, and Arabic), and 'under-resourced' languages where there are community aspirations for language technologies, and where commercial, or social, or political resources are being leveraged to create the missing language resources (e.g. Irish, Zulu). I represent these languages with hard boundaries in Figure 1 to remind us that standardisation delimits languages. With standardisation comes writing (Joseph, 1987), along with a standardised orthography, written literature, formal education, widespread literacy, and mass media.

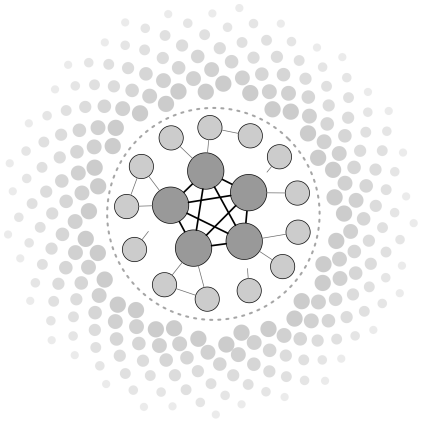


Figure 1: The Central-Peripheral Model: fully-translatable high-resource languages occupy the centre (large dark circles), surrounded by standardised but less translatable under-resourced languages (smaller circles), and outside the resource horizon of the global information society (dotted circle), we have unstandardised languages: under-resourced languages going out to acutely under-resourced languages.

Figure 1 represents what I believe to be the mindset of people who are working in 'low-resource scenarios' and seeking one-size-fits-all solutions. The vision of 'Language Technology for All' (LT4All) is to expand the resource horizon and deliver language technologies like machine translation and speech recognition to all languages. The hope is that, where political will and economic incentive have failed, technological mastery will succeed in delivering digital language equality. Regardless of what one thinks about such prospects, I believe that this agenda is misguided because it does not address the ecology of the world's languages.

In this paper I describe a *multipolar* view of language ecology. I call on researchers working on local languages to make a *local turn*, working from the ground up with speakers to identify new opportunities for language technologies.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 7817 - 7829

2 Poverty-Conscious Language Technology

In the central-peripheral model (Fig. 1), languages outside the high-resource centre are regarded as deficient. In language after language, we problematise complex socio-political situations purely in terms of missing data, and we prioritise solutions that target this shortcoming. I will refer to this as 'poverty-conscious language technology'. Poverty-conscious language technology views the high-resource language situation as normative. It sets up language technologists as the ones who will come to the rescue of deficient languages. This position is a form of Eurocentrism, a colonial worldview centred on Western civilisation. It is marked by several beliefs and values which I illustrate here.

Efficiency. The goal of the DARPA LORELEI program was to "develop methods that apply to languages of any type from any language family, eliminating the need to tailor specific technologies to a narrow set of input languages" (Tzoukermann et al., 2021). The architects of this scheme sought to capture public imagination with the scale of their vision: "Tool kit would work for every language (all 7,000 of them)" (McCaney, 2015).

Language equality. In the present context, this is the belief that languages are equally deserving of technology, that language technology is *for all* languages. It is reflected in the label "Machine Translation For All",[1] and in a manual to "help every language digitize and share equally in the benefits of a connected digital world, ensuring that 'no language is left behind'."[2]

Technologisation. The computer is presented as a neutral tool for manipulating data and implementing and testing theories (Garvin, 1963; Lawler and Aristar Dry, 1998; Bird, 1999; Hanke, 2017; Barnbrook, 2022). We provide computational tools to support language documentation, since "documentation as language salvation has become the operative metaphor used by language experts" (Perley, 2012). We might aim to help society directly, allocating our technical capabilities for maximal social good (Jin et al., 2021), using "language technology [as] the key to achieve full digital language equality in the new multilingual and interconnected world" (Steurs, 2021). Observe that it is *us* who will empower marginalised communities by introducing *our* disruptive language technologies (Joshi et al., 2019), while unwittingly reinforcing the central-peripheral model (cf. Schelenz and Pawelec, 2022).

Scriptism. There is a position that writing is "a more ideal form of linguistic representation than speech" (or 'scriptism', Harris, 1980). It appears in the belief that saving languages involves reducing them to writing (Moore, 2006; Kornai, 2013; Anderson et al., 2019). It appears in the impulse to standardise the writing of indigenous languages so that we can apply language technologies to them (e.g. Mager et al., 2018). It appears when labels such as 'Machine Translation for All' and 'European Language Equality' are used in ways that exclude oral languages.

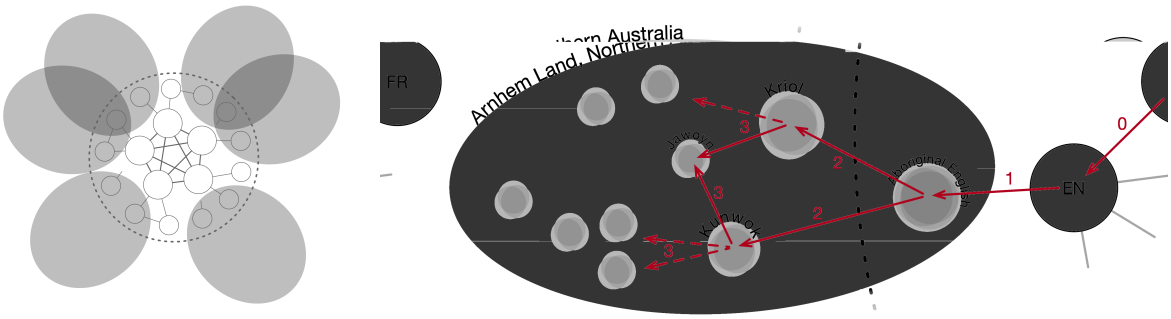
General-purpose solutions. A field linguist documented the request of a speech researcher for his data: "The scenario was that nothing was known about the language, and the data set consisted solely of audio recordings of sentences plus translations into other languages. Thus, the challenge was to automate all the following tasks: (i) establishing the phoneme inventory, (ii) generating phoneme-level alignments for the audio data, (iii) training an acoustic model, and (iv) identifying words and their pronunciation in the target language. In short, the aim was to make a language accessible for speech technology by only using audio recordings and written translations, bypassing the need for transcriptions, pronunciation dictionaries, and even phoneme set definitions. From the point of view of computer science, this ambitious objective was much more interesting than the creation of a high-quality automatic speech recognition tool" (Michaud et al., 2018, 400f). In the foreground here is the speech technologist and their skill in tackling an artificial problem, for which they need the linguist's data. They show little interest in delivering locally meaningful products. This is a widespread situation, where we apply our *savoir faire* and do more with less (e.g. Bird et al., 2014; Kempton and Moore, 2014; Vetter et al., 2016; Dunbar et al., 2017; Müller et al., 2017).

These beliefs and values underlying the central-peripheral model contain unhelpful assumptions about *language ecology* (cf. Haugen, 1972; Calvet, 2006; Lewis and Simons, 2016, 63ff).

One assumption is the 'monolingual mindset'. At least half of the world's population speaks more than one language, employing different languages

[1] https://sigul-2022.ilc.cnr.it/mt4all-shared-task/

[2] https://translationcommons.org/impact/language-digitization/



(a) The centre is ringed by 'culture areas', or 'zones of translatability', each containing local languages, and having a linguistic overlap with the centre due to historical contact and mass media.

(b) A message originating in French is translated into English (step 0), after which local linguistic expertise takes over, in expressing the message in spoken form in the contact language (Aboriginal English, step 1), and interpreting it into local languages (steps 2 and 3); many paths exist thanks to the rich language ecology; local expertise does most of the work (see Sec. 4.2).

Figure 2: A Multipolar Model of Linguistic Diversity: The centre consists of (would-be) standardised languages as in Figure 1, but the periphery contains complex, fine-grained structure

in different domains (Grosjean, 2021). "People who belong to a predominantly monolingual culture are not used to seeing the world in this [multilingual] way, because their mindset has been established through centuries of being part of a dominant culture, in which other people learn your language and you do not learn theirs. It is notable that the nations which are most monolingual in ability and attitude are those with a history of major colonial or religious expansion" (Crystal, 2000, 45).

In reality, many speech communities have a repertoire of languages, each one playing a different role in the local linguistic ecosystem. A common situation is to have 'high' and 'low' prestige varieties (Fishman, 2001), also known as vehicular and vernacular languages, one for participation in commerce and education and one for participation in the local lifeworld.

Another assumption is that written culture is normative. "Fully literate persons can only with great difficulty imagine what a primary oral culture is like... Try to imagine a culture where no-one has ever 'looked up' anything." (Ong, 1982, 31). This assumption does harm: "There is an urgent need to forefront the cultural divide between Aboriginal oral cultures and western literate cultures. The divide is disempowering Aboriginal people because literacy is argued to be a 'passport to success' in the dominant culture... Aboriginal people talk of reviving languages by returning to how the old people passed on the knowledge and the languages, on country and through the spoken word" (Kimberley Language Resource Centre, 2010).

A third assumption concerns the powerlessness of people whose languages are under threat, of speakers "relegated to the role of unwitting casualties victimised by processes greater than themselves" (Perley, 2012). Yet language shift is inevitable, and we can observe the agency of many Indigenous communities who bring epistemic resources – including grammatical distinctions and lexical items – from an ancestral language into a new language (Dickson, 2015; Ponsonnet, 2019).

The model in Figure 1 is an instance of "the central-peripheral model that dominates most technocratic thinking about technology, media, and culture" (Srinivasan, 2017). The main parameter is the quantity of language resources, and whether they are sufficient to bring a language over the line into the highly-connected, global information society.

The language we use gives us away: *for all* projects an agenda on the world; *resource* presumes machine-readability; *expert* means a specific type of western expertise; *language* in 'language resource' implies the ideology of language as data; *scaling* in "scaling up the current language technologies for the rich diversity of human languages" assumes that we have already identified the technological solutions.

3 A Multipolar Model

As an alternative to the central-peripheral model, consider the multipolar model shown in Figure 2(a). The centre contains the standardised languages, i.e., major international languages that are fully translatable. It is ringed by less well-resourced

languages, with differing strength of connection the centre. Some of these are regional spoken varieties of standardised languages, which include 'contact languages' (also known as trade languages, vehicular languages, or languages of wider communication). Contact languages connect people to other linguistic regions (cf. Fishman, 1998; Crystal, 2003). These regions are indicated using grey ovals in Figure 2(a).

What are these regions? "In linguistic ecology, one begins not with a particular language but with a particular area, not with selective attention to a few languages but with comprehensive attention to all the languages in the area" (Voegelin and Voegelin, 1964, 2). This is a notion from linguistic anthropology known as a 'culture area' (Newman, 1971). Each culture area contains many local languages, usually languages with primary orality. Translation between these languages is facilitated by a shared geography, culture, and lifeworld, plus a long history of language contact, and so we might also refer to these as 'zones of translatability'.

I avoid the term 'high-resource' when referring to the centre of the multipolar model as this valorises a particular state of a language. It reifies our technological commodities as attributes of a language. The notion of 'standardised' language is pre-existing, suggests standardised orthography, and an institutionally delimited, prestige variety. It reminds us of the existence of complexities and compromises (Ferguson, 1962; Joseph, 1987).

The terms 'under-resourced' and 'low-resource' conflate would-be standardised languages with those having purely local functions.[3] "The term 'low-resource language' is a barrier to understanding. It is applied to languages like Tamil, with 75 million speakers, most of them literate in the language, and a history of written texts that goes back thousands of years. It is ridiculous to use the same term to describe the 'biggest' Indigenous language in Canada, Cree, with 75,000 speakers and few written texts" (Kuhn, 2022, 89). Writing is key to differentiating the two.[4]

I advocate limiting the scope of 'under-resourced' and 'low-resource' labels to just the *would-be standardised* languages. I propose that the community deprecate labels like 'acutely under-resourced' because they are a myopic way to view the linguistic creation of oral cultures. I further propose that we retire the sense of 'zero resource scenario' when referring to local languages (as distinct from child language acquisition). The 'local' descriptor might also supersede others such as 'heritage', 'indigenous', 'endangered', 'threatened', or 'unwritten', which may be seen as valorising, pejorative, or Eurocentric (cf. Grinevald and Pivot, 2013). The 'local' descriptor is apt in reminding us of the local lifeworld and culture area.

In observing three primary linguistic spaces, I do not seek to confine a given language to one of the three spaces. Local languages have diasporas, such as the Nahuatl, Quechua and Hawaiian communities in New York (Kaufman and Perlin, 2018). Regional spoken varieties of a single language may have markedly different functions in different culture areas, e.g. Spanish in Mexico *vs* New Mexico (cf. Lewis and Simons, 2016, 46). Language development efforts may bring local languages into the centre without compromising their local functions.

Even the term 'local language' is problematic insofar as it seems to individuate bounded, homogeneous varieties. If the boundary of a language is unclear, it is not because Western science has not finished its job, but because human languages are not bounded codes in the first place (Dobrin et al., 2009). Diversity within a single language is sometimes problematised as deficit: "*lack* of an orthographic normalization... large dialectal variation, and *missing* standardization" (Mager et al., 2018, 57), yet this diversity within a language is the natural state and only a problem for those who would seek to scale technologies built on the assumptions of a standardised language.

Finally, the 'zero resource scenario' builds in another Eurocentrism which needs to be rooted out. It is the positivist position that we arrive at true knowledge by induction, generalising over cases. When we look at local languages and ask what we can be sure of having for our general purpose models, the answer is raw speech with translations, i.e., the zero resource scenario. This is a lowest-common-denominator approach, and it inevitably brings us back to poverty-conscious language technology. Researchers in the centre need new ways of learning technology lessons concerning local languages (cf. Sec. 5).

[3] This is not to say that there are not languages having both aspirations, e.g., contact languages and languages undergoing development.

[4] This is made explicit in the Sustainable Use Model, where 'sustainable literacy' is distinguished from 'sustainable orality' (Lewis and Simons, 2016).

4 Language Technology Agendas

The multipolar model presents an opportunity to consider the agenda of language technology in three primary linguistic spaces. The first space is the global information society, with its standardised and would-be standardised languages (Sec. 4.1). The second space consists of the culture areas, their local languages, and primary orality (Sec. 4.2). The third space is where the first and second spaces intersect. Here we have contact languages, along with local languages undergoing active development (Sec. 4.3). We consider each of these in turn.

4.1 The global information society

From the centre, we want to continue to expand the reach of language technologies to more languages, to serve the purposes of economic integration (Rivera Pastor et al., 2018). This is a version of the original agenda of under-resourced language processing (cf. Fig. 1), restricted to languages with a realistic prospect of standardisation. For example, the goal of the European Language Equality Project is "to enable all [European] languages, regardless of their specific circumstances, to realize their full potential, supporting them in achieving full digital equality in the coming decade" (Gaspari et al., 2021, 2). We can therefore chart the progress of an individual language such as Irish towards digital language equality (Lynn, 2022).

Let us consider language technology in the context of a humanitarian crisis. When it comes to messages like "tsunami warning, move to higher ground", there is global reach through standardised languages alone. We may just need translation between standardised languages (step 0 in Fig. 2(b)). From here on, we can rely on the expertise of speakers of regional varieties who – thanks to historical contact and mass media – readily understand the standardised language (step 1). Some people are highly mobile in this intercultural space, and thanks to their command of both local languages and contact languages, serve as connectors. They can interpret broadcast messages into the local lifeworld (step 2), where there is further expertise to take it to speakers of other varieties (step 3).

Conversely, when a speaker of a local language delivers information in a crisis situation, they will often use a contact language. They will not be hampered by the lack of language technology in their local language, but by the lack of support for their variety of the contact language (e.g. Lewis, 2010; Lewis et al., 2011; Anastasopoulos et al., 2020). This situation can arise even when the person is speaking a major language like English, simply because local spoken varieties of English are still not well supported (cf. Koenecke et al., 2020; Markl and Lai, 2021).

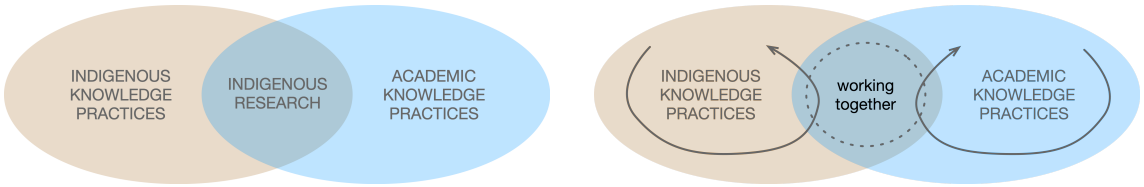
There is an opportunity here: support for contact languages, including creoles and regional spoken varieties of standardised languages, including and their rendering into non-standardised orthography, is a promising pathway for widespread language technology enabled participation in the global information society. Communications beyond these standardised languages and contact languages do not require LT4All, because there is local expertise in bridging lifeworlds and in interpreting between contact languages and local languages.

There is still the risk that broadcast messages may be misunderstood or even cause harm. The need may not be for one-shot translation of a fixed message, but for dialogue and two-way education (Sec. 4.3). Dialogue reduces the chance of messages which – while trivial to translate – are not context aware: e.g. the instruction to Australian Aboriginal people living in overcrowded housing to "stand apart from each other" instead of a more locally aware instruction to "stay in your family groups"; or the instruction to villagers in Flores to "run to higher ground" where they would only be killed by landslides.[5] This points to opportunities in the intercultural space (Sec. 4.3) and to the importance of working with local experts (Sec. 5.2).

4.2 Culture areas

Much computational work already exists for local languages and is being brought together by the ACL SIG for Endangered Languages (SIGEL) and the ISCA SIG in Under-resourced Languages (SIGUL), including workshops on Computational Methods in the Study of Endangered Languages, Spoken Language Technologies for Under-Resourced Languages, Collaboration and Computing for Under-Resourced Languages, and NLP for Indigenous Languages of the Americas. It includes tasks associated with such topics as computer-supported collaborative language documentation, and NLP for polysynthetic languages (e.g. Hanke, 2017; Lane et al., 2022). It includes support for wider participation in NLP (e.g. Nekoto et al., 2020; Mirza-

[5] https://indosasters.org/2017/08/20/a-critical-reflection-on-running-to-higher-ground-narrative-myth-and-reality-in-tsunami-warning-and-response/



(a) Indigenous research as the intersection between knowledge practices (following Christie, 2006).

(b) Working together in the intersection of indigenous and academic knowledge practices.

Figure 3: A Third Space at the Intersection of Knowledge Practices

khalov et al., 2021), and language resources with the prospect of connecting local languages in ways that are not mediated by standardised pivot languages (e.g. Madonsela et al., 2016). This work varies in the degree to which it is locally conceived.

In some places there is institutional support for developing a local language, including standardising an orthography, teaching literacy, and translating literature to and from a standardised language (e.g. Zulu, Haitian Creole). Language development may shift a language into the overlap between a culture area and the global information society.

A promising approach for work with speakers of a local language is offered by constructivism (e.g. Charmaz, 2014). A set of methods which have been successful in Arnhem Land is known as *Ground-Up*. It has grown from the observation that Indigenous knowledge is local and performed, and it employs methods that are emergent and situated. Early applications of Ground-Up methods involved content management and health communication (Cass et al., 2002; Verran et al., 2007; Lowell et al., 2021). In the language space, we can work from the ground up to explore the ecology of local speech varieties (cf. Haugen's 'ecological questions', Haugen, 1972, 65). we can explore the language ideology, the practices that support (and draw support from) local languages, and the country itself as a language resource. From this place we can seek new opportunities for language technologies.

Perhaps this will still lead to such agendas as economic participation and multilingual information access. However, where I work in Arnhem Land, people tend to see language as coupled with identity, culture, ancestors, and country. They do not tend to see language as data, or language as lexico-grammatical code. Our conversations about learning centre on human learning not machine learning. When it comes to working with technology, people prefer culturally meaningful work to passive participation in a Western process (cf. Le Ferrand et al., 2022). Many people are passiomate about intergenerational transmission of knowledge, and do not obsess about getting everything transcribed and translated. "Apart from the Inuit, no Indigenous community we've spoken with has shown much interest in machine translation (MT) between their ancestral language and English (or French, in Quebec). Communities are typically more interested in tools to encourage learning and use of their ancestral language" (Kuhn, 2022, 89).

An approach to technology engagements in culture areas is suggested by work on codesign (e.g. Verran and Christie, 2007; Verran et al., 2007; Bidwell et al., 2008; Bidwell and Browning, 2010; Winschiers-Theophilus et al., 2010; Brereton et al., 2013; Winschiers-Theophilus and Bidwell, 2013; Brereton et al., 2014; Soro et al., 2016; Taylor et al., 2018, 2019). We could apply such methods to the study of language technologies in culture areas.

4.3 Third spaces

The third space is a hybrid place, an intersection of worlds. It has been discussed under such headings as the 'contact zone', the 'recognition space', the 'intercultural space', the 'arena', and the 'research interface' (Bhabha, 2012; Pratt, 1991; Somerville and Perkins, 2003; Taylor, 2008; Hunt et al., 2008; Jasper and Duyvendak, 2015; Ryder et al., 2020). One framing is *Indigenous research* (Fig. 3(a)), defined as "that part of an Indigenous knowledge tradition which is recognisable or legible from a Western research perspective... [or conversely] as that part of the Western academic research tradition which is at the same time conceived, shaped, governed and understood within Indigenous knowledge traditions. The area in the middle of the diagram is Indigenous research because it fulfils the criteria for both Indigenous knowledge production and academic research" (Christie, 2006, 80).

Here, my frame of reference is 'working together' (Fig. 3(b)). As a participant in an Aus-

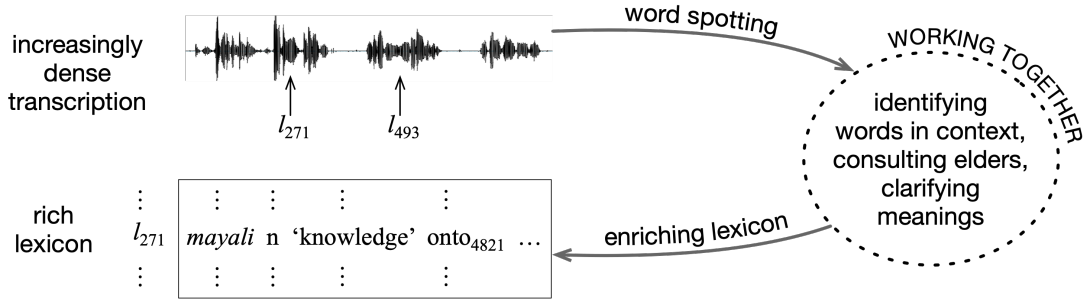


Figure 4: Expert Transcription: High value words are spotted in the audio, discussed by Indigenous and western experts working together, and added to a rich lexicon (Bird, 2020).

tralian Aboriginal community, I need pretexts for sitting with local people, and this comes from the established activities of a ranger program and a school. Here there are opportunities for computer assisted language learning and for spoken document retrieval from an archive of untranscribed media. There may be other opportunities for technology to augment traditional learning processes (Harris, 1984; Trudgen, 2012, 200ff), and for computer supported cooperative work that privileges local languages and knowledge systems (Christie and Verran, 2014; Carew et al., 2015; Hanke, 2017; Bettinson and Bird, 2021b,a).

Part of the dynamic of working together on a language resource is the diverse meanings this may have for participants (cf. Star and Griesemer, 1989; Star, 2010). The externally-driven, telic work of compiling a dictionary may sit alongside local people's atelic, day-to-day participation in exploring the meanings of words with elders and visiting the places where the associated stories can be told. The resulting bound volume or mobile app might be a learning resource to one person and an emblem of prestige to another.

5 Stories of Expertise in a Third Space

The perspective I have articulated in this paper has arisen from living and working in a Kunwinjku-speaking community situated in Arnhem Land, Aboriginal country in the far north of Australia (cf. Fig 2(b)). Here my attempts to pursue my Eurocentric practices in data collection have foundered. Over a period of several years, and with the patient guidance of many local people, I, a western-educated middle-aged white male, have learnt about the local lifeworld, glimpsed local expertise, and borne witness to systemic injustice.

In this section, I describe local responses to western academic practices in transcription and translation, practices where the agendas of language technology and language documentation fortuitously align. From my centralised perspective, the task of rendering speech into text, and the task of translating that text into another language, are disjoint. The technologies of speech recognition and machine translation are similarly distinct. However, I found that matters were different at the local level.

The task of working together on a recording and deciding what was said turns out to be a two-way practice that merges transcription and translation (Sec. 5.1). The task of working together on an emergency broadcast to interpret it into a local language turns out not to be conventional one-shot translation of a fixed message but a two-way practice of "understanding the true stories" (Sec. 5.2). I recount these experiences to reveal the contingent, situated nature of work in a third space, and to suggest that a suitable way to learn lessons from such stories is not induction to lowest common denominator scenarios leading to one-size-fits-all solutions, but abduction to deeper accounts of speech communities and language technologies.

5.1 Expert transcription

Transcription for 'acutely under-resourced' languages has depended on recruiting participants to transcribe speech recordings and to provide phrase-aligned translations into a standardised language, a practice that focusses on surface forms, quantity, and efficiency. Yet transcriptional practices on the ground are far from mechanical, and there is no simple ground-truth transcription (Hermes and Engman, 2017; Himmelmann, 2018; Bird, 2020). On numerous occasions, I have found that there is no local interest in the tedious work of rendering speech recordings into text. When I look at local people's 'transcriptions' I see a practice akin to note-taking or inscription. People write down

In Arnhemland, Yolŋu [Aboriginal] people live in extended family groups with traditional authority structures. When Balanda [Westerners] don't understand or respect our way of governance, they often come up with ways of dealing with problems that undermine the authority of our Elders and their ways of keeping people and places safe.	When trying to spread the word in Yolŋu communities, the Balanda authorities told everyone to wash their hands and *stand apart from each other*. This way of sharing the story had the effect of by-passing the Elders, and of prioritising ways to keep ourselves safe as individuals. It cared for the 'biomedical body' threatened by the virus, but not the 'Yolŋu body' which includes our family and clan groups. They picked one person to take the news to the people in the community, but this did not involve negotiating among ourselves what the right story for Yolŋu should be, and the best way for it to be shared.	There are ways we can work together, beginning with the authority of Elders, to understand the true stories of this virus. We have traditional ways of doing that sort of work and sharing out the right responsibilities to the right people. We know the right ways to keep our relationships strong, including our relations to other clan groups and to our homelands. When we are able to remain connected with each other and our places, this is how we remain healthy.

Figure 5: Caring for Yolŋu and Ways of Life during COVID 19 (exerpt from Wanambi et al., 2021)

enough so that they can reconstruct the story or perform the knowledge. In the process, we discuss the form and meaning of key words and phrases. Here is where local interests intersect with a newcomer's need to expand their vocabulary and improve their ability to recognise words in connected speech.

How can we privilege local interests and expertise, and flip this transcriptional practice from a deficit scenario to a strength scenario? My answer is 'sparse transcription', schematised in Figure 4. We give up the slavish left-to-right phoneme level transcription practice, and instead prioritise our agency in identifying words of interest and discussing their significance.

Thus, on the top left of Figure 4 we have a sparse transcription, where some tokens of lexical items have been found, manually or automatically. We are not concerned about narrow transcription of those items, only with identifying tokens of a lexeme in connected speech. On the right we have the practice of working together where local experts and western learners clarify the meaning of words, and enlarge the lexicon. Through multiple iterations, the transcription of a corpus gets denser. Our ability to automatically spot topic words and to retrieve relevant spoken documents improves.

5.2 Expert translation

Following the onset of the COVID-19 pandemic, the Australian government broadcast "simplistic directives about behaviour change" (Lowell et al., 2021, 172). The assumption seems to be that all knowledge lies in the centre, and when it comes to reaching communities who speak other languages, it is a question of translation.

To us in the language technology community, the government's approach presents a golden opportunity. We could obtain funding, collect a parallel corpus, and build a translation system. We would measure success in terms of the quantity of data collected and the performance of the system on gold translations. Over time, we would bring another language into the centre.

However, in our success we would have missed the point: this is not a translation problem. Consider the response of some local elders to the government's communication strategy (Fig. 5). The elders touch on many issues. What is the utility of an instruction to self-isolate – or what came across in Yolŋu as "stand apart from each other" – in communities with chronic overcrowding?[6] Where is the sense in transmitting messages through a person who is not locally recognised as a knowledge authority, a practice which harms the Yolŋu body?

A likely response in the language technology community would be to collect more data and build a better system. Yet how would we hope to learn, via "the mere exercise of matching words or phrases in one language with those of another" (Duranti, 1997, 154), that Yolŋu have a different metaphysics for an apparently simple term like 'body'? Our approach to translation works best when there is a shared lifeworld, where lexicalised concepts, metaphors and tropes line up across languages, i.e., within a zone of translatability (Fig.2(a)).

The government's practice of COVID communication is more Eurocentrism, and a consequence of "the West's view of itself as the centre of legitimate knowledge, [and of] science as the all-embracing method for gaining an understanding of the world" (Smith, 2012). The Yolŋu elders delivered a sophisticated response to the government's simplistic directives. They identified metaphysical issues, and asked to "work together to understand the true stories". This practice has been called *two-way learning* in Australia (Harris, 1990), cf. *two-eyed seeing* in Canada (Wright et al., 2019).

[6] https://www.creativespirits.info/aboriginalculture/land/overcrowded-houses

In a more culturally aware approach, "Balanda [Western] educators discussed with the Yolŋu participants how to explain each concept in their own language as it was introduced. This triggered active and collaborative engagement in the learning process and provided opportunities for misunderstandings to be revealed and repaired... This strategy of continual collaborative interpreting of each new concept introduced by the Balanda educators, as well as Yolŋu sharing their knowledge, facilitated a more in-depth understanding than passive listening to an explanation in English" (Lowell et al., 2021, 171). This suggests a new opportunity for language technology, not how to improve translation for 'under-resourced' languages, but how to support people to work together in a third space, and to navigate a metaphysical divide (Fig. 3(b)).

6 Conclusion

The field of language technology has placed the world's languages on a spectrum according to the available machine-readable resources, a self-serving position that I have called poverty conscious language technology. Our category of 'under-resourced' languages conflates the qualitatively different situations of local languages and would-be standardised languages. Our talk of technology for languages of *any type* and of language technology *for all* betrays our Eurocentrism. When we speak of 'acutely under-resourced' languages and 'zero expert resources' we commit an epistemic injustice.

I have described a multipolar model which respects local language ecologies with their orality and multilingualism, and I have articulated implications for the agenda of language technology. I have suggested ways that we can take a local turn and work with local speech communities from the ground up. We still need to be on guard for the colonial impulse in its many guises (cf. Dourish and Mainwaring, 2012). We still need to properly theorise language technology development outside the space of standardised languages.

The result of this program, I hope, will be language technologies that address the distinct opportunities presented in three high-resource scenarios: the global information society with its standardised languages, the culture areas with their local languages, and their intersection in third spaces with their contact languages and local language development activities.

Acknowledgements

I acknowledge the Bininj people of the *Kuwarddewardde* 'Stone Country' and thank them for welcoming me into their community. *Karrimurrngrayekmen kunwok!* I am particularly grateful to Dean Yibarbuk and Lois Nadjamerrek and the Warddeken Rangers for their support. My work in Arnhem Land has been approved by traditional owners, the board of Warddeken Land Management, a research permit from the Northern Land Council, and a human research ethics protocol approved by Charles Darwin University. An earlier version of this material was presented in a keynote address at the 2021 Conference on Empirical Methods in Natural Language Processing, and I am grateful to several participants for their feedback. I thank the following people for discussions that have helped my thinking, and for their feedback on earlier versions of the material presented here: Antonios Anastasopoulos, Laurent Besacier, Mat Bettinson, Michael Christie, Éric Le Ferrand, William Lewis, Teresa Lynn, Helen Verran, Fei Xia, and several anonymous reviewers. This work was supported by a grant from the Australian Research Council.

References

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the translation initiative for COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19*. ACL.

Deborah Anderson, Lee Collins, Craig Cornelius, and Craig Cummings. 2019. *Indigenous Languages: Zero to Digital – A Guide to Bring Your Language Online*. Translation Commons.

Geoffrey Barnbrook. 2022. *Language and Computers*. Edinburgh University Press.

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

Matthew Bettinson and Steven Bird. 2021b. Designing to support remote working relationships with indigenous communities. In *Proceedings of the 33rd Australian Conference on Human-Computer Interaction*.

Homi K. Bhabha. 2012. *The Location of Culture*. Routledge.

Nicola Bidwell and David Browning. 2010. Pursuing genius loci: interaction design and natural places. *Personal and Ubiquitous Computing*, 14:15–30.

Nicola Bidwell, Peta-Marie Standley, Tommy George, and Vicus Steffensen. 2008. The landscape's apprentice: lessons for place-centred design from grounding documentary. In *Proceedings of the 7th Conference on Designing Interactive Systems*, pages 88–98. ACM.

Steven Bird. 1999. Multidimensional exploration of online linguistic field data. In Pius Tamanji, Masako Hirotani, and Nancy Hall, editors, *Proceedings of the 29th Annual Meeting of the Northeast Linguistics Society*, pages 33–47. GLSA, University of Massachussetts at Amherst.

Steven Bird. 2020. Sparse transcription. *Computational Linguistics*, 46:713–44.

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.

Margot Brereton, Paul Roe, Thomas Amagula, Serena Bara, Judy Lalara, and Anita Lee Hong. 2013. Growing existing Aboriginal designs to guide a cross-cultural design project. In *IFIP Conference on Human-Computer Interaction*, pages 323–330. Springer.

Margot Brereton, Paul Roe, Ronald Schroeter, and Anita Hong. 2014. Beyond ethnography: engagement and reciprocity as foundations for design research out here. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1183–86. ACM.

Louis-Jean Calvet. 2006. *Towards an Ecology of World Languages*. Polity.

Margaret Carew, Jennifer Green, Inge Kral, Rachel Nordlinger, and Ruth Singer. 2015. Getting in touch: Language and digital inclusion in Australian indigenous communities. *Language Documentation and Conservation*, 9:307–323.

Alan Cass, Anne Lowell, Michael Christie, Paul Snelling, Melinda Flack, Betty Marrnganyin, and Isaac Brown. 2002. Sharing the true stories: improving communication between Aboriginal patients and healthcare workers. *Medical Journal of Australia*, 176:466–471.

Kathy Charmaz. 2014. *Constructing Grounded Theory*. Sage.

Michael Christie. 2006. Transdisciplinary research and Aboriginal knowledge. *Australian Journal of Indigenous Education*, 35:78–89.

Michael Christie and Helen Verran. 2014. The Touch Pad Body: A generative transcultural digital device interrupting received ideas and practices in Aboriginal health. *Societies*, 4:256–264.

David Crystal. 2000. *Language Death*. Cambridge University Press.

David Crystal. 2003. *English as a Global Language*. Cambridge University Press.

Greg Dickson. 2015. *Marra and Kriol: The Loss and Maintenance of Knowledge across a Language Shift Boundary*. Ph.D. thesis, Australian National University.

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

Paul Dourish and Scott Mainwaring. 2012. Ubicomp's colonial impulse. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 133–142.

Ewan Dunbar, Xuan Nga Cao, Juan Benjumea, Julien Karadayi, Mathieu Bernard, Laurent Besacier, Xavier Anguera, and Emmanuel Dupoux. 2017. The Zero Resource Speech Challenge 2017. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, pages 323–330. IEEE.

Alessandro Duranti. 1997. *Linguistic Anthropology*. Cambridge University Press.

Charles A Ferguson. 1962. The language factor in national development. *Anthropological Linguistics*, 4:23–27.

Joshua A Fishman. 1998. The new linguistic order. *Foreign policy*, pages 26–40.

Joshua A. Fishman. 2001. Why is it so hard to save a threatened language? In Joshua A. Fishman, editor, *Can Threatened Languages be Saved?: Reversing Language Shift, Revisited: a 21st Century Perspective*, pages 1–22. Multilingual Matters.

Paul Garvin, editor. 1963. *Natural Language and the Computer*. McGraw Hill.

Federico Gaspari, Andy Way, Jane Dunne, Georg Rehm, Stelios Piperidis, and Maria Giagkou. 2021. Digital language equality (preliminary definition). Technical Report D1.1, European Language Equality. https://european-language-equality.eu/deliverables/.

Colette Grinevald and Bénédicte Pivot. 2013. On the revitalization of a 'treasure language': The Rama Language Project of Nicaragua. *Keeping languages alive: Documentation, pedagogy and revitalization*, pages 181–197.

François Grosjean. 2021. *Life as a bilingual: Knowing and using two or more languages*. Cambridge University Press.

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

Roy Harris. 1980. *The Language-Makers*. Cornell University Press.

Stephen Harris. 1984. Aboriginal learning styles and formal schooling. *The Australian Journal of Indigenous Education*, 12(4):3–23.

Stephen Harris. 1990. *Two-way Aboriginal schooling: Education and cultural survival*. Canberra: Aboriginal Studies Press.

Einar Haugen. 1972. The ecology of language. In Anwar Dil, editor, *The Ecology of Language, Essays by Einar Haugen*, pages 325–339. Stanford University Press.

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

Nikolaus Himmelmann. 2018. Meeting the transcription challenge. In *Reflections on Language Documentation 20 Years after Himmelmann 1998*, number 15 in Language Documentation and Conservation Special Publication, pages 33–40. University of Hawai'i Press.

Janet Hunt, Diane Smith, Stephanie Garling, and Will Sanders. 2008. *Contested Governance: Culture, Power and Institutions in Indigenous Australia*. ANU Press.

James M. Jasper and Jan Willem Duyvendak. 2015. *Players and Arenas: The Interactive Dynamics of Protest*. Amsterdam University Press.

Robert Jimerson and Emily Prud'hommeaux. 2018. ASR for documenting acutely under-resourced indigenous languages. In *Proceedings of the 11th Language Resources and Evaluation Conference*, pages 4161–66.

Zhijing Jin, Geeticka Chauhan, Brian Tse, Mrinmaya Sachan, and Rada Mihalcea. 2021. How good is NLP? a sober look at NLP tasks through the lens of social impact. In *Findings of the Association for Computational Linguistics*, pages 3099–3113. ACL.

John Joseph. 1987. *Eloquence and Power: The Rise of Language Standards and Standard Languages*. Blackwell.

Pratik Joshi, Christain Barnes, Sebastin Santy, Simran Khanuja, Sanket Shah, Anirudh Srinivasan, Satwik Bhattamishra, Sunayana Sitaram, Monojit Choudhury, and Kalika Bali. 2019. Unsung challenges of building and deploying language technologies for low resource language communities. In *Proceedings of the 16th International Conference on Natural Language Processing*, pages 211–219. NLP Association of India.

Daniel Kaufman and Ross Perlin. 2018. Language documentation in diaspora communities. In *Oxford Handbook of Endangered Languages*, pages 399–418. Oxford University Press.

Timothy Kempton and Roger K Moore. 2014. Discovering the phoneme inventory of an unwritten language: A machine-assisted approach. *Speech Communication*, 56:152–166.

Kimberley Language Resource Centre. 2010. Whose language centre is it anyway? In John Hobson, editor, *Re-awakening Languages: Theory and Practice in the Revitalisation of Australia's Indigenous Languages*, pages 131–145. Sydney University Press.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117:7684–89.

András Kornai. 2013. Digital language death. *PloS One*, 8(10).

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.

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, Atticus Harrigan, and Antti Arppe. 2022. Interactive word completion for Plains Cree. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

John M. Lawler and Helen Aristar Dry, editors. 1998. *Using Computers in Linguistics*. London: Routledge.

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*.

Paul Lewis and Gary Simons. 2016. *Sustaining Language Use: Perspectives on Community-Based Language Development*. SIL International.

William Lewis. 2010. Haitian Creole: How to build and ship an MT engine from scratch in 4 days, 17 hours, & 30 minutes. In *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*.

William Lewis, Robert Munro, and Stephan Vogel. 2011. Crisis MT: Developing a cookbook for MT in crisis situations. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 501–511.

Anne Lowell, Elaine Läwurrpa Maypilama, and Rosemary Gundjarranbuy. 2021. Finding a pathway and making it strong: Learning from Yolŋu about meaningful health education in a remote Indigenous Australian context. *Health Promotion Journal of Australia*, 32:166–178.

Teresa Lynn. 2022. Report on the Irish language. Technical Report D1.20, European Language Equality. https://european-language-equality.eu/deliverables/.

Stanley Madonsela, Munzhedzi James Mafela, Mampaka Lydia Mojapelo, and Rose Masubelele. 2016. African WordNet: A viable tool for sense discrimination in the indigenous African languages of South Africa. In *Proceedings of the Eighth Global WordNet Conference*, pages 25–29.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018. Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69.

Nina Markl and Catherine Lai. 2021. Context-sensitive evaluation of automatic speech recognition: considering user experience and language variation. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 34–40. ACL.

Kevin McCaney. 2015. Tool kit would work for every language (all 7,000 of them). Defense Systems, https://defensesystems.com/it-infrastructure/2015/10/tool-kit-would-work-for-every-language-all-7000-of-them/190978/, Accessed Mar. 2022.

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.

Jamshidbek Mirzakhalov, Anoop Babu, Duygu Ataman, Sherzod Kariev, Francis Tyers, Otabek Abduraufov, Mammad Hajili, Sardana Ivanova, Abror Khaytbaev, Antonio Laverghetta Jr., Bekhzodbek Moydinboyev, Esra Onal, Shaxnoza Pulatova, Ahsan Wahab, Orhan Firat, and Sriram Chellappan. 2021. A large-scale study of machine translation in Turkic languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5876–90. ACL.

Robert Moore. 2006. Disappearing, Inc.: Glimpsing the sublime in the politics of access to endangered languages. *Language and Communication*, 26:296–315.

Markus Müller, Jörg Franke, Alex Waibel, and Sebastian Stüker. 2017. Towards phoneme inventory discovery for documentation of unwritten languages. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 5200–04. IEEE.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, and others. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2144–60. ACL.

James Newman. 1971. The culture area concept in anthropology. *Journal of Geography*, 70:8–15.

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

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

Maïa Ponsonnet. 2019. *Difference and Repetition in Language Shift to a Creole: The Expression of Emotions*. Routledge.

Mary Louise Pratt. 1991. Arts of the contact zone. *Profession*, pages 33–40.

Rafael Rivera Pastor, Carlota Tarín Quirós, Juan Pablo Villar García, Toni Badia Cardús, and Maite Melero Nogués. 2018. Language equality in the digital age: Towards a human language project. Technical report, European Parliament. https://data.europa.eu/doi/10.2861/834747.

Courtney Ryder, Tamara Mackean, Julieann Coombs, Hayley Williams, Kate Hunter, Andrew Holland, and Rebecca Ivers. 2020. Indigenous research methodology – weaving a research interface. *International Journal of Social Research Methodology*, 23:255–267.

Laura Schelenz and Maria Pawelec. 2022. Information and communication technologies for development (ICT4D) critique. *Information Technology for Development*, 28:165–188.

Linda Tuhiwai Smith. 2012. *Decolonizing Methodologies*, 2nd edition. Zed Books.

Margaret Somerville and Tony Perkins. 2003. Border work in the contact zone: Thinking indigenous/non-indigenous collaboration spatially. *Journal of Intercultural Studies*, 24:253–266.

Alessandro Soro, Margot Brereton, Jennyfer Lawrence Taylor, Anita Lee Hong, and Paul Roe. 2016. Cross-cultural dialogical probes. In *Proceedings of the First African Conference on Human Computer Interaction*, pages 114–125. ACM.

Ramesh Srinivasan. 2017. *Whose Global Village?: Rethinking how Technology Shapes Our World*. NYU Press.

Susan Leigh Star. 2010. This is not a boundary object: Reflections on the origin of a concept. *Science, Technology and Human Values*, 35:601–617.

Susan Leigh Star and James R. Griesemer. 1989. Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907–39. *Social Studies of Science*, 19:387–420.

Frieda Steurs. 2021. Language techonology: the key to achieve full digital language equality in the new multilingual and interconnected world. Presentation at the 50th Poznań Linguistic Meeting, https://lirias.kuleuven.be/3552049.

Jennyfer Lawrence Taylor, Wujal Wujal Aboriginal Shire Council, Alessandro Soro, Paul Roe, and Margot Brereton. 2019. A relational approach to designing social technologies that foster use of the Kuku Yalanji language. In *Proceedings of the 31st Australian Conference on Human-Computer-Interaction*, pages 161–172.

Jennyfer Lawrence Taylor, Alessandro Soro, and Margot Brereton. 2018. New literacy theories for participatory design: Lessons from three case studies with Australian Aboriginal communities. In *Proceedings of the 15th Participatory Design Conference*, pages 1–18. ACM.

John Taylor. 2008. Indigenous peoples and indicators of well-being: Australian perspectives on United Nations global frameworks. *Social Indicators Research*, 87:111–26.

Richard Trudgen. 2012. *Why Warriors Lie Down and Die*. Why Warriors Pty Ltd.

Evelyne Tzoukermann, Jason Duncan, Caitlin Christianson, and Boyan Onyshkevych. 2021. Advances in low-resource and endangered languages. In *Proceedings of the Second Workshop on Computational Methods for Endangered Languages*, pages 24–30.

Helen Verran and Michael Christie. 2007. Using/designing digital technologies of representation in Aboriginal Australian knowledge practices. *Human Technology: An Interdisciplinary Journal on Humans in ICT Environments*, 3:214–227.

Helen Verran, Michael Christie, Bryce Anbins-King, Trevor Van Weeren, and Wulumdhuna Yunupingu. 2007. Designing digital knowledge management tools with Aboriginal Australians. *Digital Creativity*, 18:129–142.

Marco Vetter, Markus Müller, Fatima Hamlaoui, Graham Neubig, Satoshi Nakamura, Sebastian Stüker, and Alex Waibel. 2016. Unsupervised phoneme segmentation of previously unseen languages. In *Proceedings of the 17th Annual Conference of the International Speech Communication Association*, pages 3544–48.

Charles F Voegelin and Florence Marie Voegelin. 1964. Languages of the world: Native america, fascicle one. *Anthropological Linguistics*, 6(6):1–149.

Gawura Wanambi, Joy Bulkanhawuy, Stephen Dhamarrandji, and Rosemary Gundjarranbuy. 2021. Caring for Yolŋu and Ways of Life During COVID-19. https://indigenousx.com.au/caring-for-yolnu-and-ways-of-life-during-covid-19, Accessed Mar. 2022.

Heike Winschiers-Theophilus and Nicola J Bidwell. 2013. Toward an Afro-Centric indigenous HCI paradigm. *International Journal of Human-Computer Interaction*, 29:243–255.

Heike Winschiers-Theophilus, Shilumbe Chivuno-Kuria, Gereon Koch Kapuire, Nicola Bidwell, and Edwin Blake. 2010. Being participated: a community approach. In *Proceedings of the 11th Biennial Participatory Design Conference*, pages 1–10. ACM.

A. L. Wright, C. Gabel, M. Ballantyne, S. M. Jack, and O. Wahoush. 2019. Using two-eyed seeing in research with indigenous people: an integrative review. *International Journal of Qualitative Methods*, 18:1–19.