Designing Speech Technologies for Australian Aboriginal English: Opportunities, Risks and Participation

Ben Hutchinson
Google
Sydney, NSW, Australia
benhutch@google.com

Celeste Rodríguez Louro
University of Western Australia
Perth, WA, Australia
celeste.rodriguezlouro@uwa.edu.au

Glenys Collard
University of Western Australia
Perth, WA, Australia
glenysdcollard@gmail.com

Ned Cooper
Australian National University
Sydney, NSW, Australia
Edward.Cooper@anu.edu.au

Abstract

In Australia, post-contact language varieties, including creoles and local varieties of international languages, emerged as a result of forced contact between Indigenous communities and English speakers. These contact varieties are widely used, yet are poorly supported by language technologies. This gap presents barriers to participation in civil and economic society for Indigenous communities using these varieties, and reproduces minoritisation of contemporary Indigenous sociolinguistic identities. This paper concerns three questions regarding this context. First, can speech technologies support speakers of Australian Aboriginal English, a local indigenised variety of English? Second, what risks are inherent in such a project? Third, what technology development practices are appropriate for this context, and how can researchers integrate meaningful community participation in order to mitigate risks? We argue that opportunities do exist—as well as risks—and demonstrate this through a case study exploring design practices in a project aiming to improve speech technologies for Australian Aboriginal English. We discuss how we integrated culturally appropriate and participatory processes throughout the project. We call for increased support for languages used by Indigenous communities, including contact varieties, which provide practical economic and socio-cultural benefits, provided that participatory and culturally safe practices are enacted.

ACM Reference Format:
Ben Hutchinson, Celeste Rodríguez Louro, Glenys Collard, and Ned Cooper. 2025. Designing Speech Technologies for Australian Aboriginal English: Opportunities, Risks and Participation. In *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25), June 23–26, 2025, Athens, Greece.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3715275.3732010

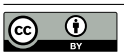
1 Introduction

Most work on speech and language technologies, a.k.a. Natural Language Processing (NLP), focuses on languages which are spoken by majority populations and heavily standardised within one or more countries [16, 77, 126]. The field of NLP is thus reproducing the minoritisation of language varieties, and in the process reinforcing barriers of exclusion for diverse language communities. These include most local languages in highly linguistically diverse countries such as Indonesia [1], as well as various creoles which combine elements of settler-colonial languages elsewhere in the world [85]. Additionally affected are varieties of national languages spoken by minoritised groups [79], such as Māori English in New Zealand, Multicultural London English in the UK, and African-American English in the US. Speakers of minoritised varieties must typically switch to a majority variety when using language technologies, and when they do they can experience negative psychological impacts, such as feeling that technology wasn't designed for them [44, 66, 102, 141].

Machine learning-driven NLP is not typically oriented towards participatory processes, driven instead by the valorisation of efficiency [27] and fundamentally constrained by tensions between participation and scale [146]. NLP evaluations emphasise quantitative and decontextualised metrics [72], while language dataset projects often reduce participation to crowdsourced platform work. These projects fail to create relationships which emphasise reciprocity and social justice [42, 132], and risk perpetuating colonial histories, especially in the Global South, of exploitation by academic institutions, states and corporations [95]. However participatory AI projects, on the other hand, are often challenged by a lack of shared understanding of good practices [64], and may risk "participation-washing" [132]. More generally, projects which aspire to participation often fail to afford decision-making power to participating communities [6], and risk co-optation when delivered through corporate structures [8, 26].

In Australia, where this research takes place, of the 400+ ancestral Indigenous languages that existed before European colonisation [29, p. 56], most are either extinct or critically endangered [9]. This is due to linguicidal practices of colonialism, which at various times included governments forcibly separating Aboriginal children from their families and coercing Aboriginal pupils to speak English while forbidding them from speaking their languages [119]. As a result, most Indigenous people in Australia speak Australian Aboriginal English, making it a strong encoder of ethno-cultural identity [122].[1] This indigenised variety of English integrates phonological,

[1] Other contact languages in Australia include mixed languages such as Gurindji Kriol, Light Warlpiri and New Tiwi [97].



FAccT '25, Athens, Greece

ACM ISBN 979-8-4007-1482-5/25/06
https://doi.org/10.1145/3715275.3732010

lexico-grammatical and discourse-pragmatic features from English and Aboriginal ancestral languages (see §2.1).

Given this context, we take three positions in this paper. First, Indigenous people should be supported to use *whichever of their linguistic varieties that they prefer*, including contact varieties, when using language technologies. Second, NLP requires not only culturally appropriate outputs, such as models, tasks and evaluations [67, 70, 91], but also culturally appropriate processes, outcomes and modes of participation [25, 110]. All are currently undervalued in NLP. Third, language technologists should be equally concerned with contemporary Indigenous needs, including supporting recent and evolving Indigenous language varieties, as they are with needs related to ancestral languages. What, then, are these opportunities, as well as the risks? And how are they distinct from the technology needs of mainstream language communities, or of ancestral Indigenous language communities?

To address these questions, this paper contributes the first detailed consideration of language technologies to support Australian Aboriginal English speakers. To our knowledge, this is the first paper to focus on technologies for this language community. We do so through a case study of an ongoing project intended to improve Automatic Speech Recognition (ASR) for Indigenous communities in Australia. In doing so, this paper follows previous NLP literature in considering a specific Indigenous language community and its context [*e.g.*, 138, 147], with a focus on design considerations in order to avoid assuming that mainstream methods and goals are locally applicable [110, 114]. As such, this paper contributes to our understanding of ethical and responsible development of technologies for minoritised communities, as well as to our understanding of how to design and build culturally sensitive technologies while incorporating participatory methods.

The paper begins with an introduction to Australian Aboriginal English, including a description of relevant linguistic and sociolinguistic factors (§2). We also present a brief survey of related work on Indigenous language technologies. Next, we discuss the opportunities and risks inherent in deploying ASR systems for this community, as well as opportunities and risks in projects building such technologies (§3). Building on this analysis, we explore strategies to mitigate risk, and facilitate opportunities, through a detailed discussion of the design considerations of a real-world project to design culturally appropriate ASR while integrating participatory methods (§4). In doing so, we provide a detailed discussion of some of the considerations that are critical for a project of this sort, including system requirements, data elicitation and transcription, system evaluation, and processes and frameworks for participation, and employment. Finally, we discuss further technological opportunities for supporting speakers of Australian Aboriginal English (§5).

2 Background and Related Work

When considering how technologies can support speakers of a specific Indigenous language variety, it is critical to have some basic knowledge of the speech community, the language variety, and their contexts [147]. In this section we introduce Australian Aboriginal English and its context, followed by a brief summary of prior work relevant to language technologies for Indigenous communities. This includes consideration of both ancestral and contact language varieties, and also of participatory methods in Indigenous language projects.

2.1 Australian Aboriginal English

Australian Aboriginal English is a contact variety of English used by Indigenous people in Australia [123]. It is particularly prevalent across southern Australia where ancestral languages are less widely spoken. In these contexts, Australian Aboriginal English is used to strengthen Indigenous ways of being and communicating [51]. Australian Aboriginal English has its origins in the product of forced contact between Indigenous people and the English-speaking settler-colonialists. The English spoken by the British was a powerful instrument used to strip Indigenous communities of their ancestral rights [94, p. 125]. Australian Aboriginal English's precursor was a mixed jargon viewed as 'broken English' by the British. By the time this jargon had transitioned to a novel pidgin (known as 'New South Wales pidgin' [139]), it began to travel outside present-day Sydney and across New South Wales. With the advent of the pastoral industry, the pidgin also spread north into Queensland and the Northern Territory, giving rise to Queensland Pidgin English and the Northern Australian creoles, including Roper River Kriol [98, p. 371].

Australian Aboriginal English has linguistic characteristics, including pronunciation, lexis, grammar, and discourse-pragmatics, that distinguish it from mainstream English [51]. Pronunciation features include: frequent /h/-dropping and /h/-addition; metathesis of 'k' and 's' in the word 'ask', which is pronounced 'aks'; and default reduction of the 'ing' verbal suffix to 'in'. Grammatical features include double negatives as in 'never done nothing'; and double subjects as in 'My niece, she did it.' Lexical features include lexical transfer such as *blackfella*, meaning 'black person/fellow', and borrowings from ancestral languages, such as *dardy* 'good/good looking' from Nyungar.

Experts remain ambivalent as to whether Australian Aboriginal English represents a single ethnolect, or whether different varieties of Aboriginal English, or 'Aboriginal Englishes', exist across Australia [46]. Australian Aboriginal English is frequently placed on a continuum that ranges from 'lighter' acrolectal varieties (closer to standardised Australian English) to 'heavier' basilectal varieties which are closer to Kriol, spoken across northern Australia [128]. The basilectal varieties are mostly unintelligible to standardised English speakers, but the acrolectal varieties are superficially similar to Australian English.

Because Australian Aboriginal English is English based, it is sometimes socially stigmatised, erroneously classified by laypeople as an incorrect version of English. As a result, L1 users of Australian Aboriginal English are often discriminated against in schooling and legal settings [52, 65], perpetuating invisibility of the variety and leading to the false perception that its speakers are uneducated and rude [53]. Stigmatisation of Australian Aboriginal English is evident in the comment below, which was offered in response to an online article on the role of Australian Aboriginal English in Indigenous health communication [121].

> With Aboriginal children all going to school, and nearly all of them being taught in English, does it still serve them to support the continued use of Pidgin English?

The commenter's deliberate use of 'Pidgin English' reflects a deficit perspective that undermines the value of Australian Aboriginal English. Rather than embracing Australian Aboriginal English as a linguistic variety many Indigenous people use daily in Australia, this perspective portrays racialised peoples' language varieties as 'lacking' vis-à-vis mainstream English [71, p. e214].

These negative attitudes are exacerbated because Australian Aboriginal English is primarily oral, with no nationally adopted guidelines as to how it should be written. Additionally, the orthographic representation of Australian Aboriginal English varies across locales, reflecting the local linguistic biographies and practices of distinct Indigenous groups. For example, the name of the Traditional Owners of the region comprising the southwest of Western Australia, as well as their language, is variably spelled 'Nyungar', 'Noongar', or 'Nyoongar' depending on community members' preferences.[2] Rather than a weakness, the lack of standardising orthographies for Australian Aboriginal English reflects Indigenous people's prerogative to differentiate—rather than homogenise—how Australian Aboriginal English is used across Indigenous Country.

Indigenous people tend to find the use of direct questions awkward or offensive [125],[3] and these pragmatic considerations continue in Australian Aboriginal English. Direct questions are also likely to elicit misleading information; Indigenous people have been known to answer 'yes' or agree with non-Indigenous people even if the question is not fully understood—a phenomenon known as 'gratuitous concurrence' [89, p. 137]. To avoid these limitations, researchers have used 'yarning', an Indigenous cultural form of storytelling and conversation, to lead unstructured data collection with Indigenous communities who speak Australian Aboriginal English [18, 124].

2.2 Technologies for ancestral Indigenous languages

The NLP literature for Indigenous languages tends to focus on ancestral languages, *i.e.*, Indigenous languages which predate colonisation.[4] Given the considerable overlap between endangered languages and Indigenous languages [32]—a consequence of colonisation—NLP literature tends to focus on application domains which support language revitalisation and preservation efforts, including assisting language learners or teachers [62, 100, 101]. This literature discusses technologies such as quiz generation, automated language assessment, language learning platforms, computer assisted translation, predictive text, and readability classification [*e.g.* 35, 92, 138, 147]. Bird and Yibarbuk [25] pose open questions regarding distinct needs of learners of oral languages without the technology working through a written form.

Language preservation or documentation is frequently cited as a goal [83, 117], and is discussed variously as a means towards revitalisation, or as an end in itself. Recognising that ancestral language communities can be minoritised in multiple ways, Carew et al. [35] discuss the role of language technologies in promoting digital inclusion. Other motivations encompass the importance of language technologies in cultural revitalisation projects, *e.g.*, helping to learn about traditional place names, stories, songs and kinship relations.

2.3 Technologies for contact language varieties

The NLP literature has to-date had marginal concern with supporting Indigenous contact language communities, such as Hawaiian Pidgin, Australian Kriol, Tok Pisin, Chavacano, *etc.* For example, Nigerian Pidgin has over 100 million speakers [90], yet is poorly supported by technology. This seems to be due to a combination of reasons, including the stigmatisation of contact languages [87, 130] and a belief that ancestral languages are more important in terms of global heritage. However Indigenous peoples are not museum artifacts or relics of an 'ancient' era [61, 68]; their cultures and societies are dynamic and evolving.

To understand the needs of Indigenous contact language communities, we first draw on the NLP literature for Creole languages. Lent et al.'s survey of six Creole communities found different levels of desire for language technologies [86]. They note that these languages often lack recognised status, are primarily oral with variable orthography and writing conventions, and are often socially stigmatised. This points to the importance of contact varieties in shaping contemporary social identities. If technology does not support contemporary forms of Indigenous expression, then it reproduces and reinforces minoritisation of Indigenous culture and identity. Contact varieties are often vehicular, used to communicate with other communities, including participation in national and regional economies, and with civic services [23]. If language technologies are to support economic and civic participation, then they need to support vehicular varieties.

Since contact varieties are primarily oral, prior work has identified speech technologies as the most wanted by both experts and communities [83, 86]. Robust Automated Speech Recognition (ASR) that can recognise Indigenous pronunciations and loan words is an important foundational technology for a broad range of applications, from voice search to automated transcription and interaction with digital assistants [2]. Design of ASR for contact varieties should be mindful of the social dimensions of orthography, in particular of variation from standardised written forms.

For some contact varieties, orthographic norms may have colonial associations, and can be alienating for speakers [96]. For this reason, and due to their primary orality, consultation with experts and communities is required to ascertain the need for text-based technologies [23, 84, 86]. When needed, writing tools supporting local scripts, fonts, and keyboards would generally be the most useful text-based technologies [55]. The lack of standardised writing conventions for many contact varieties may present challenges for the utility of technologies such as spell checkers and next word predictors.

Language technologies are increasingly multilingual, and a critical component of many systems is Language Identification (LangId). Since most Indigenous communities are multilingual, robust LangId

[2] Collard is Nyungar so we prefer this spelling.
[3] With some exceptions, including teaching contexts where instructors asking questions is heavily routinised [105].
[4] Others call these languages 'traditional' Indigenous languages, *e.g.*, [53]. The term 'ancestral' is sometimes used by others with a different meaning: a language from which modern languages descend.

and code-switching/mixing detection are needed to support local multilingual technologies [108].

2.4 Culture, and participatory methods, in language research

Recently, the NLP community has demonstrated increasing interest in the topic of culture, with, *e.g.*, three recent workshops on Cross-cultural Considerations in NLP [45, 115, 116]. This interest has been spurred by the capabilities of Large Language Models (LLMs) with regards to linguistic form, while also demonstrating lack of cultural nuance [67, 88, 106]. Threads of research have focused on measuring cultural biases [106, 135] and expression of cultural values by LLMs [7, 50, 120].

It is critical that the NLP community goes further than developing technological fixes in data and models, to consider design questions and project processes. NLP research using Indigenous language data often fails to engage with the cultural differences between technologists and the communities they aspire to support [25]. Recent community-led projects such as Masakhane [107] and the Papa Reo project [76] provide alternative models for the NLP community, by drawing on principles of participatory design to embed the cultural knowledge of language communities throughout NLP research and development processes. Rather than developing technologies *for* Indigenous contact varieties in isolation of communities speaking those varieties, the primary focus of such projects must be on developing language technologies *with* Indigenous language communities. This requires research teams, and the broader NLP community, to consider the position of NLP researchers relative to community partners, reflecting not only on performance disparities of technologies across languages or cultural biases, but also on the composition of research and development teams, as the cultural assumptions and norms of such teams influence the research we conduct [59].

2.5 Projects with Indigenous communities

The literature emphasises that working with data of Indigenous language varieties raises ethical considerations. These also apply to contact varieties such as Australian Aboriginal English, and are critical to keep in mind when considering opportunities and risks. Four questions that recur in the literature include:

Q1. What is the context? Before working with Indigenous varieties, researchers must reflect on the harmful impact of past colonial research projects on Indigenous communities and seek to understand the historical and sociolinguistic context of partner communities [20, 48, 84, 110, 129, 134, 147].

Q2. What relationships will be maintained? Relationships are fundamental to Indigenous identities, communities, and research [134, 143], and this holds for research on contact varieties. Technologists should be cognisant that relationships with partners and participants may, and in some cases should, persist beyond the scope of the project [40].

Q3. Who will control the project, and its outputs? In both ancestral and contact variety-speaking communities, Indigenous people are less likely to instigate language technology projects, enumerate technological goals, or design/control project budgets, and it is important to both acknowledge and address these power differentials. Using language data ethically is the lowest threshold for language projects [34], and involves considerations not only of IP (Intellectual Property) but also ICIP (Indigenous Cultural and Intellectual Property) [75].

Q4. Is the technology appropriate for the community? Indigenous language varieties are typically differentiated functionally within a multilingual community, and technologies may be more useful for some than others [23, 48, 86]. Technologists should understand the technology needs of Indigenous communities, which may be distinct from those of non-Indigenous users [40, 91].

3 Opportunities and Risks for ASR for Australian Aboriginal English

We now consider Automated Speech Recognition (ASR) systems for Australian Aboriginal English, acknowledging that this focus on converting spoken language to written language itself risks reproducing beliefs about the primary importance of written language vis-à-vis spoken [37]. With the contexts and questions of the previous section in mind, we consider potential opportunities and risks, informed by the literatures on ASR, Human-computer Interaction, and Indigenous language technologies. We consider both deployed ASR technologies as well as the projects which build ASR technologies; our ultimate goal is to build inclusive technologies using inclusive methods.

3.1 Deployment Opportunities

ASR supports many application domains [2], and improving ASR for Australian Aboriginal English could provide multiple specific benefits to First Nations people in Australia, their communities, and non-Indigenous people who are encountering Australian Aboriginal English. We attempt to distinguish those possible benefits with as much specificity as possible, and in contrast to prior work we adopt a human-centered framing rather than a technology-centered one.

Opportunity 1: Improved dictation for Australian Aboriginal English speakers. Dictation typically involves a single speaker consciously speaking directly to a device, and is useful in a variety of contexts in which speaking is preferred to typing. These include messaging while driving, speakers with physical or vision disabilities, speakers avoiding small keyboards on mobile devices, and speakers with dyslexia. The speaker often reviews the text output for accuracy. Since they are conscious of the interaction, some speakers may modify their speech in an attempt to improve the fidelity of the transcription, such as minimising use of informal registers, obscure or complex words, and minoritised varieties [69, 102].

We note that not all dictation applications are likely to benefit Australian Aboriginal English speakers. For example, ASR is sometimes used in *dictation practice* by second-language learners, however Australian Aboriginal English is not formally taught or studied as a second-language.

Opportunity 2: Improved transcriptions for readers of personal messages. ASR can be initiated by those within the speaker's community or social network, for example, users of a voicemail service

providing automated transcriptions. This can be useful in a variety of contexts in which reading is preferred to listening, including in quiet environments in which silence is preferred, in noisy environments in which hearing is difficult, when privacy is desirable in public environments, and for people with hearing disabilities. An Australian Aboriginal English speaker is often unaware when their voice is being transcribed, and so is unlikely to modify their voice. The reader might or might not listen to the voice recording to verify fidelity of the transcription. The reader may be able to accommodate and even correct some ASR errors, based on their familiarity with the speaker and/or context, even without listening to the recording.

Opportunity 3: Improved transcriptions for consumers of platform content. When speech and video platforms host content, they often provide closed captions. Such content can be unscripted (typical for Australian Aboriginal English on YouTube, for example, see Appendix B) or scripted (common in TV shows, movies, podcasts and vodcasts, etc—for an analysis of scripted Australian Aboriginal English see [14]). These are generated offline and stored as metadata alongside the Australian Aboriginal English voice recording. The platforms may have some access controls, but in the general case both the recordings and the transcriptions may be accessible to potentially anyone online, including listeners and readers with no prior experience with Australian Aboriginal English. Listening to the recordings might not help readers without familiarity with Australian Aboriginal English to correct some ASR errors.

Due to systemic societal biases, some applications of closed captioning may encounter relatively little Australian Aboriginal English compared to Australian English. The automated transcription of Australian parliamentary proceedings [63], for example, is one use case for Australian Aboriginal English ASR that is impacted both by the under-representation of Indigenous people in Australian parliaments and by the social stigmatisation of speaking Australian Aboriginal English in formal contexts.

Opportunity 4: Improved human-machine interactions for Australian Aboriginal English speakers. ASR is increasingly incorporated into a wide range of computational systems used by Australian Aboriginal English speakers. These include automated voice assistants, automated or semi-automated call centres including for critical banking and government services, web search using voice queries, and audio/video retrieval that indexes speech under transcribed terms. (Voice Query-By-Example technologies [113] do not require ASR but constitute a minority of audio retrieval systems.) In many cases, the transcription might not be seen by any human, including the Australian Aboriginal English speaker. Instead, the transcription is typically the input to a Natural Language Understanding component, which outputs some form of semantic representation in forms such as a task frame, a search index entry, or an expanded Information Retrieval query [54, 78]. More accurate transcriptions can lead to more relevant search results, more efficient interactions, and fewer abandoned tasks.

3.2 Deployment Risks

Schwartz has argued that the primary ethical responsibility when building language technologies for Indigenous communities is to not do any harm [129]. Although many taxonomies of risks and negative impacts of technologies exist, our goal here is be a specific and situated as possible. Hence we do not attempt to map the identified risks to prior taxonomies which often aim at universalism and generality [17].

Risk 1: Inappropriate for the contexts of Australian Aboriginal English. Researchers should not assume that minoritised language communities have the same wants and needs as mainstream language communities [23], nor that multilingual speech communities require the same technologies for all languages [25, 86]. For example, would some Australian Aboriginal English speakers prefer to switch into Standardised Australian English when interacting with machines, if it results in greater task efficiency? What about when dictating messages to family and friends? Conversely, would they prefer more support for speaking Australian Aboriginal English in contexts which are currently not well supported? There is a shortage of research on the real-world needs of Australian Aboriginal English speakers. Building systems that are neither needed or wanted is not only inefficient—it also undermines the rights of minoritised communities to determine which technologies are built for them.

Risk 2: Lack of sociolinguistic ecological validity. Like all language varieties, Australian Aboriginal English is not homogeneous, but exhibits regional variation [46]. Furthermore, and notably, Australian Aboriginal English includes both acrolectal and basilectal varieties (§2.1). If the training data for an ASR model does not represent sufficient variation, then the model will fail to recognise the breadth of forms of Australian Aboriginal English. If the training data collection protocols are not designed to be culturally safe, then speakers may switch into Standardised Australian English, reducing the validity of the data.

Risk 3: Psychological harms for Australian Aboriginal English users when dictating or interacting with systems. Prior research has found that when ASR systems make errors, some members of minoritised language communities feel that the computer 'doesn't like' how they are speaking [31], or that the technology isn't built for people like them [102]. Whereas speakers of mainstream language varieties may blame the ASR system for poor performance, speakers of minoritised communities may blame themselves, and experience a drop in self esteem or an increase in self consciousness about their identity [141]. They may feel that their language variety is being judged as illegitimate [44]. If ASR fails to meet expectations for speakers of Australian Aboriginal English, all of these impacts may be reproduced. Feelings of shame or embarrassment might also arise if speakers are self-conscious about their way of speaking being transcribed and re-contextualised into contexts where mainstream written English is the norm.

Risk 4: Torquing of Australian Aboriginal English sociolinguistic identities. ASR systems function as classifiers not just of speech, but also, by extension, of language varieties and sociolinguistic identities. As such, ASR systems can be examined through Bowker and Star's sociological lens of torque and residuals [30]. 'Residuals' refers to the 'other' categories that are relevant to the individual but missing from the classificatory system, while 'torque' refers to the painful twisting and bending that occurs when an individual's

data is forced into an inappropriate classificatory system, such as forcing individuals into racial categories in Apartheid South Africa. Previous work has explored how gender classification technologies produce torque and othering for trans people [127], and how medical and biometric IT projects can torque minoritised groups [13, 131]. We posit that a similar form of torque exists for users of non-standardised language varieties, when ASR systems systematically fail to recognise spoken forms salient to social identities, bending and twisting them into the written forms of mainstream dictionaries of colonising cultures.

3.3 Project Opportunities

We take the position that language technology projects provide the potential for a range of opportunities, especially when the project is conducted not just *for* the community but also *with* the community [34].

Opportunity 5: Project leads to greater non-Indigenous understanding of Indigenous contexts and histories. Projects involving both Indigenous and non-Indigenous participants create opportunities for 'two way learning', a phrase used in Australia to refer to Indigenous and non-Indigenous people coming together to learn from each other in a space that is safe for those with diverse perspectives and ways of knowing [118, p. xx].

Opportunity 6: Project work produces opportunities for Indigenous community. Data collection projects can provide opportunities for what Sloane et al. [132] call 'participation as work', with equitable compensation. There are large economic disparities in Australia, with median incomes for Indigenous people around 23% lower than for non-Indigenous [10], and Indigenous people have higher rates of unemployment for all education levels [12]. If equitable compensation and non-extractive processes are successfully adopted, data work can create economic opportunities for community, as well as opportunities for meaningful cultural engagement [23].

Opportunity 7: Project builds capabilities in community. Building Indigenous capabilities on technology projects has the opportunity of addressing the current digital divide in skills [35]. When projects effectively build community capacity, participants can draw on their developing technical skills—such as data collection and management capabilities—to advocate for their communities in future projects and on technology and policy issues [41]. To achieve these benefits, meaningful opportunities for skill development need to be incorporated into the project, focusing on capabilities that will remain valuable to the community beyond the project's completion.

Opportunity 8: Project increases Indigenous self-determination. Providing opportunities for Indigenous control over project outputs, for example through data governance. Design decisions can have not only consequential benefits (*e.g.*, avoiding data set misuse by third parties in ways that do not accord with Indigenous aspirations), but also constitute a form of procedural justice through greater self-determination.

3.4 Project Risks

Developing inclusive ASR *inclusively* requires more than addressing model biases [56]— it also requires researchers to integrate participatory practices throughout the project. As such, we take the position that, when developing language technologies, risks need to be considered using a wide-angle lens that takes into account not just the deployment scenarios but also the project processes used when developing the technology. For projects with Indigenous communities, 'it's how do you things that matters' [40], and the ends should not be used to justify the means.

Risk 5: Project harms relationships with community. The literature on building language technologies for Indigenous communities emphasises the foundational importance of relationships with the community [19, 20, 25, 40, 147]. If a project damages community trust and harms relationships, not only will that project become endangered, but future projects with the community may also fail due to lack of community support.

Risk 6: Project fails to partner with community where appropriate. A lack of community participation, generally, can not only lead to poorer data and models, but can also miss opportunities to address social injustices [132]. Failure to partner when appropriate with Indigenous communities, specifically, can reproduce histories of ignoring or disrespecting Indigenous sovereignty. Calls for collaborative and participatory research recognise community rights to be meaningfully included in language research that involves them [34, 35]. This is emphasised by UNESCO in the Los Pinos Declaration calling for a decade of action on Indigenous languages: 'Nothing about us without us' [140].

Risk 7: Project uses culturally unsafe methods to collect data. Recording one's own voice can be an uncomfortable experience for many people, and especially so for members of minoritised communities encountering researchers from universities or technology companies using mainstream English. Unless care is taken, prompts might not be culturally relevant, stray into territories of secret traditional knowledge, elicit personal narratives of trauma, or trigger feelings that their culture is not important and valid. Data work may also risk being meaningless to workers' social or cultural contexts, for example, transcribing long recordings of a non-standardised variety into a mainstream script can be tedious and unengaging work [22, 23].

Risk 8: Project does not respect Indigenous data sovereignty. Failure to adopt appropriate Indigenous data governance protocols can violate the principles of Indigenous people controlling Indigenous data [36, 93, 137]. This is an important aspect of generally recognising Indigenous sovereignty, which was never ceded in what we now know as Australia.

Risk 9: Project does not acknowledge Indigenous knowledge and expertise. There are many ways in which Indigenous contributions may fail to be acknowledged. Within technology research, this includes not citing or including as co-authors Indigenous experts whose ways of working fall outside Western academic norms [73].

4 Designing ASR for Australian Aboriginal English

We now discuss a real-world partnership between Google and The University of Western Australia which aims to improve ASR for speakers of Australian Aboriginal English. We focus on the design

considerations, and specifically on our efforts to relate the project goals and processes to both the sociolinguistic context (§2.1) and the technological opportunities (§3.1). We discuss not just technical questions, but the design of the project itself, including how we attempt to mitigate the system and project risks (§3.2, §3.4) while providing societal and economic opportunities (§3.3), *e.g.*, by integrating participatory processes throughout the project lifecycle.

When building technologies for Indigenous speech communities—of either ancestral or contact varieties—it is important to consider the ways in which projects face social and technical questions distinct from those that involve mainstream languages [see, *e.g.*, 21, 35, 84, 86, 110, 114, 136]. For example, ASR development pipelines typically make assumptions about speech communities and language use, including the existence of 'gold standard' spoken and written forms [96]. These assumptions break down in the case of Australian Aboriginal English, for which the written form is not standardised, and so community participation in the design process is a prerequisite for building technologies that serve the community.

4.1 System desiderata

ASR should support the use of Australian Aboriginal English in all application horizontals which incorporate ASR systems. These include: 1) dictation of messages and documents; 2) voice queries to search engines; 3) voice commands to virtual assistants; and 4) automated transcriptions of voicemail messages, meetings, podcasts, and video content [2]. It is important to distinguish applications in which the transcripts will be processed by computers (Opportunity 4 in §3, *e.g.*, commands to virtual assistants) from those where the transcripts will be read by humans (Opportunities 1–3, *e.g.*, voicemail transcriptions). In the first case, post-processing converts the transcripts to a structured or semi-structured representation, such as an expanded Information Retrieval query or a task frame [78]. This post-processing involves increasing abstraction and standardisation, mapping diverse user inputs into more computationally tractable forms. In the case of transcripts for humans, however, standardisation of linguistic forms is no longer a computational requirement.[5] Instead, standardisation of transcripts minimises the diverse ways in which social meanings are conveyed through linguistic choices. In the process, standardisation reproduces the minoritisation of non-mainstream sociolinguistic identities [96]. This is especially relevant for informal registers, such as those used in personal messages or social media, in which expressions of social identities are more prevalent and varied.

4.2 Elicitation of speech data

The collection protocols to secure the speech data needed for this project were designed by the authors with advice from the Indigenous Advisory Committee (§4.5). The project also applies participatory research methods discussed further in §7 and previously used by Rodríguez Louro and Collard [122]. However, the data to be collected for this project will be used by Google and it is expected that the corpus will be open access, in accordance to FAIR and CARE principles [36, 142], subject to further input from the Indigenous Advisory Committee. For this reason, the project explicitly excludes sharing of taboo cultural materials and sensitive information that is potentially harmful for Indigenous people and communities. To ensure cultural safety (cf. Risk 7), we designed a program of careful elicitation by using culturally relevant and sensitive prompts in consultation with the Indigenous Advisory Committee for this project, and ultimately designed by an Indigenous-owned company which specialises in producing educational resources using an anti-racism lens (Appendix A, cf. Risk 4).

The design of the recording process enables voice data providers to record their voices in their own homes, which is a more comfortable and culturally-sensitive experience than visiting a university or a corporate office. Participant payment processes were designed with both safety and convenience in mind (Opportunity 6).

Once collected, all data will be screened by a team of Indigenous RAs, led by Collard, who will vet the audio files and transcriptions for cultural sensitivity. Any culturally unsafe material (*e.g.*, stories of specific people, places, Indigenous knowledges, trauma, and loss) will be erased from the corpus before sharing the data with others.

4.3 Transcription of speech data

The introduction of writing by settler-colonialists has significantly impacted Indigenous populations [60, 111]. In Australian Aboriginal English, there are two major questions regarding orthographic practices. First, how are loan words from ancestral languages, including place names and Country names (*e.g.* 'Nyungar'), written in Latin-based script? Second, how do Indigenous people in Australia deliberately adopt their own spelling conventions, *e.g.*, 'Blak'. These are not purely technical questions, but also sociopolitical [33, 60, 96] (cf. Risk 4).

To co-develop transcription guidelines with community members, we required examples of public Australian Aboriginal English to refer to. For this, we used a corpus of over 100 hours of YouTube videos that are primarily in Australian Aboriginal English (see Appendix B). An Indigenous team, led by Rodríguez Louro, developed the corpus and used their expertise with Australian Aboriginal English to validate the variety spoken, ensuring that speakers were only coded as Australian Aboriginal English if they used at least two linguistic features that distinguish this variety from Standardised Australian English. We developed an initial version of variety-specific transcription guidelines using examples from this corpus. Example guidelines include:

(1) borrowings from Pidgin and ancestral languages should be transcribed verbatim, with an English translation included;
(2) variety-specific utterance tags should be included, *e.g.*, 'unna' and 'inni'.

The transcription guidelines are a living document, which will grow—with input from the community—as novel classes of examples are observed.

4.4 Evaluation considerations

Doğruöz and Sitaram [48] have observed that NLP metrics are usually designed with high-resource scenarios in mind. In the case of ASR, the most common evaluation metric is Word Error Rate (WER), which does not actually compare words but rather spelled

[5]The requirement for readability by non-linguists also precludes transcription into phonetic or phonemic representations, as has been done for some other work such as [103].

forms [80]. It does so in binary fashion: small differences between predictions and reference transcriptions (*e.g.*, 'naïve' vs 'naive') are penalised just as much as large ones. As such, WER is less useful as a measure of system utility in various contexts, including with language varieties without standardised written forms [3, 109], and varieties with multiple standard written forms [80, 82]. In the case of Australian Aboriginal English, WER is not robust to culturally acceptable variation in borrowings from ancestral languages (*e.g.*, 'Gadigal' vs 'Cadigal'[6]) or indigenised spelling variants of English words (*e.g.*, 'Blak' vs 'Black').[7]

With a lack of standardised orthography for Australian Aboriginal English, it is important to engage with communities, to centre their perspectives, rather than relying solely on automated metrics such as WER [21, 44, 96]. As such, the design of the evaluation stage includes interviewing Aboriginal community members, to understand their perspectives on how important fidelity to references is compared to other considerations. An Australian Aboriginal English-specific mapping of commonly accepted spelling variations will be incorporated into WER calculations and enable quantitative measures which provide increased validity with community practices [4, 109] (cf. 'lexeme error rate' of [22]).

4.5 Designing the project

The project is a partnership between Google and The University of Western Australia (UWA). Google is a technology company, many of whose products incorporate ASR technologies. The University of Western Australia is an Australian University whose Linguistics department has both expertise in Australian Aboriginal English and existing relationships with Aboriginal communities, including the Nyungar community. We recognise that the legacy of unethical and settler-colonial research has contributed to distrust of research in many Indigenous communities [134]. This legacy, along with a distrust of technology companies, can jeopardise even well intentioned projects. Google found that seeking out partnership with (socio)linguists at The University of Western Australia with existing relationships, cultural competencies, and sociolinguistic expertise, was critical, and preferable to attempting to build them from scratch [40, 48]. With regards to control of the project and its outputs (Q3 of §2.5), the project is funded by Google, who set the technological goal of the project: improving ASR for First Nations people in Australia. Google will own any ASR models that they develop using the voice recordings, however UWA will retain ownership of the recordings and serve as stewards of the collection for its appropriate use, by setting up culturally appropriate data governance. We strengthened accountability through explicit incorporation of ICIP considerations in our research partnership agreement (§7) (cf. Risk 8).

[6]The Country and people of the Sydney region.

[7]Character Error Rate (CER), another common metric, does not adequately resolve the issue. CER can avoid the overpenalisation of variants accepted by different groups in the community, *e.g.*, only penalising lightly a prediction of 'Cadigal' when the reference transcription is 'Gadigal' (both refer to the same First Nations Country). However CER is too lenient a metric in cases where the ASR system confuses the spoken word for another word which has a very different meaning or referent, *e.g.*, predicting 'Kukatj' when then reference transcription is 'Kukatja' (the names of two unrelated Aboriginal language varieties with ancestral lands thousands of kilometres apart [11]). CER would judge such an error as equally bad as the 'Cadigal' vs 'Gadigal' case.

Before funding was secured for the project, Aboriginal scholars (with expertise in education, applied linguistics, and language revitalisation) were approached and asked to participate as members of a committee advising researchers on the project. Because the research involves Indigenous participants, letters of support from advisors Professor Clint Bracknell and Mx Sharon Davis, and from AIATSIS (Australian Institute of Aboriginal and Torres Strait Islander Studies), were included in a submission to The University of Western Australia's IRB. Both The University of Western Australia's IRB and Google's privacy and legal teams reviewed the consent forms that would be used with Aboriginal people creating voice recordings. In addition, we sought and obtained the support and expertise of the Language Data Commons of Australia (LDaCA),[8] as the research involves creating a new dataset of Indigenous language materials. LDaCA advised on culturally appropriate data governance and access frameworks [58].

The collection of Australian Aboriginal English voice recordings for the purposes of this project raises Indigenous Cultural and Intellectual Property (ICIP) issues. As a team, the academic authors at The University of Western Australia negotiated with Google the inclusion of an ICIP clause in the research agreement signed in November 2023. This clause reads as follows:

> The parties acknowledge that Australian protocols may be relevant to ICIP in relation to the Research and will work collaboratively during the Research Term to agree the extent to which Australian protocols, including but not limited to the Australian Council for the Arts Protocols for using First Nations Cultural and Intellectual Property in the Arts (2019), should apply to the Research.

Most ancestral languages are extinct or endangered, and Australian Aboriginal English is spoken by around 80% of First Nations people in Australia. This project focuses on both Australian Aboriginal English varieties spoken in southwest Western Australia, where Rodríguez Louro and Collard live and work, and varieties of Australian Aboriginal English spoken in the Pilbara (northern Western Australia), Queensland and Alice Springs (Northern Territory) (cf. Risk 2).

Project members took time to reflect on power dynamics, both within the project and in its context [28, 110] (Opportunity 5). We shared and discussed critiques of extractive data practices. We discussed social injustices, including barriers to opportunity for Indigenous people in Australia, and stigmatisation of Australian Aboriginal English.

It was important from the outset to be clear about which Indigenous people and organisations would participate in the project, and how their participation would be facilitated [20, 40, 57, 86, 129, 147] (cf. Risk 6). In doing so, it was important to be cognisant of the barriers to technology and research careers facing Indigenous participants, to recognise qualifications and authority beyond that of Western institutions, and to recognise contributors as co-authors of research artefacts (145, p. 219; 99, pp. 28–29) (cf. Risk 9). While Indigenous participation is critical to the project, we acknowledge that this does not guarantee that the project will be equitable [110, 132]. We also recognise that the value given to participation and relationships is asymmetric between various project participants. We were

[8]https://www.ldaca.edu.au/

also mindful not to force control onto community partners, who are often under-resourced and overburdened [34, 43]. Establishing the Indigenous Advisory Committee (see below) was an important step in helping to negotiate these tensions. We also incorporated aspects of community knowledge and practices into the design process [144] (cf. Risk 6), including hiring Goorlil Consulting, a Canberra-based consulting firm owned and run by Indigenous experts, to curate and design culturally appropriate visual prompts for the data collection (cf. Risk 7).

An Indigenous Advisory Committee was established to provide guidance throughout the project. The advisors include Professor Clint Bracknell and Mx Sharon Davis. Bracknell is a respected Nyungar academic with a wealth of expertise in projects examining how Indigenous data should be collected and archived for maximum cultural safety. Davis is a renowned education leader from both Bardi and Kija peoples of the Kimberley in Western Australia. Davis has expertise in, and is a champion of, Australian Aboriginal English. Collaborating with Indigenous partners who have expertise in working with community using cultural safety frameworks helps us mitigate risks involving potential harm to Indigenous people and communities (cf. Risk 5). Consultation with the Advisory Committee takes place quarterly. The Advisory Committee meetings, conducted online due to members living in disparate locales across Australia, involve project updates delivered by Rodríguez Louro, followed by Q&A on project progress and outcomes. The Advisory Committee's role is to provide critical feedback on project conduct and progress. For example, our first meeting involved discussing culturally-safe ways to collect the project data, while additional meetings discussed how to share the collected data, and how to ensure that the bulk of the research assistants for this project are Indigenous.

Indigenous associates participate in the project, not just as mere providers of voice data, but as partners in the design of elicitation materials and transcription guidelines, management of the voice data collection, and validation and annotation of the collected materials (Opportunity 7). All of the research assistants (RAs) conducting fieldwork for this project are First Nations people. While Collard, our senior research assistant, is also a project leader, three additional RAs were also hired. This team includes three Nyungar fieldworkers: Hope Narrier, Katrina Cox and Lily Hayward. While Narrier was introduced to the leadership team through one of Rodríguez Louro's PhD students, Cox and Hayward were recommended to the project team by UWA's School of Indigenous Studies. This recommendation was made after Rodríguez Louro contacted the School's administration with an offer to accept Expressions of Interest for the positions of research assistants, which Cox and Hayward were successful in obtaining.

The maintenance of relationships with Indigenous stakeholders is critical (§2.5). Rodríguez Louro and Collard are in daily contact via phone messaging, a practice that dates back to 2018 when they began their research together. Rodríguez Louro also maintains regular contact with the Indigenous RAs and quarterly direct contact with the Advisory Committee, although contact on social media is more frequent.

Additionally, outreach on the project to Indigenous communities facilitates greater accountability through public visibility. The project has been reported on by Indigenous media outlets [49, 74], and Collard has been interviewed about the project on national radio stations [38]. Collard, Rodríguez Louro and Hutchinson will also jointly present on aspects of the project at the PULiiMA 2025 Indigenous Languages and Technology Conference in Darwin, Northern Territory, a forum which expects over 700 Indigenous attendees.

For convenience, in Appendix C we re-iterate and summarise aspects of the projects which mitigate the risks and facilitate the opportunities that we enumerated in §3.

4.6 Limitations

The technological goals of this project were set by Google, and this precluded Indigenous people exercising full control over the project goals. In other words, this project was not driven by or from the community, limiting the opportunities for self-determination. Although innovating on method, and embracing alternate methodologies, the project is nevertheless situated within the dominant paradigm of NLP that seeks to address technical shortcomings through richer datasets, rather than other paradigms which focus on local human agency [24]. We also acknowledge that First Nations people in Australia are less likely to be in full time research positions at companies or universities in Australia, limiting opportunities to be partner investigators on this research.

We acknowledge the plurality of Indigenous cultures and Countries in Australia, and that Indigenous co-investigators, designers, managers and transcribers cannot represent the full diversity of those cultures and Countries. Our decision to include multiple regions is important not only because of lexico-grammatical differences, but also because Australian Aboriginal English has a key emblematic function as an encoder of different ways of being and doing. We believe that the various ancestries represented wherever Australian Aboriginal English is used are as important as understanding Australian Aboriginal English itself.

Our corpus for developing the initial version of the transcription guidelines was limited to publicly shared YouTube videos. We do not claim that it is representative of spoken Australian Aboriginal English in other domains, such as personal messages, nor of written forms of Australian Aboriginal English.

As researchers based in Australia, our lived experiences with colonial contexts are also based primarily in Australia. Although we hope that this project contains aspects that generalise, we do not claim to know or represent other contexts.

5 Discussion: Further Technology Opportunities for Australian Aboriginal English

We now consider technologies supporting Australian Aboriginal English speakers beyond ASR. Given that languages go by multiple names, the NLP community typically uses language codes as identifiers in multilingual systems, datasets, and data catalogues [81]. This discrete mapping of languages to codes is central to NLP's conceptual model of what languages are [24, 133]. Australian Aboriginal English currently lacks a standardised identifier, and this would be useful for future work on this variety. However, given the power of classifying [30]—and the colonial associations of the origins of the ISO 639-3 identifiers [47], which also underpin codes

such as the BCP 47—we echo calls for involving community members and linguists, and avoiding offensive exonyms, when designing codes for Australian Aboriginal English [104].

Given the primary orality of Australian Aboriginal English, and its distinct grammatical and discourse-pragmatic features relevant to information seeking (§2.1), there are open questions around whether spoken language assistants could better support speakers. Robust support for tag-questions—rather than just questions with subject-verb inversion—might provide more natural support for information requests. While speakers of Australian Aboriginal English benefit from hearing others talk in the same way [5, p. 42], the question of whether digital assistants should try to 'sound Indigenous,' must be approached with care, to avoid potentially reproducing interactions of racialised subservience to non-Indigenous consumers [112].

Before developing writing tools tailored for Australian Aboriginal English, more understanding is required of how writing practices diverge from standardised Australian English. Social media and other domains with informal registers are more likely to have written Australian Aboriginal English, and it may be that modeling orthographic variation is a useful approach for text technologies [cf. 90].

Evidence suggests that designing health campaigns in Australian Aboriginal English has community benefits [39]. This suggests potential use cases for machine translation, particularly automated translation and rewriting of health and other social services materials in Australian Aboriginal English.

Finally, given the prevalence of switching between Australian Aboriginal English and Standardised Australian English, technologies which support the former would benefit from also supporting the latter. Technologies which aim to support English in Australia should explicitly document whether they support Australian Aboriginal English [15].

6 Conclusions

This paper has explored three questions. First, can speech technology projects support speakers of Australian Aboriginal English by affording specific opportunities? Second, what are the system deployment and procedural risks of such endeavours? Third, how can we mitigate risks, and facilitate opportunities, by integrating participatory practices which appropriately respect the context and are culturally appropriate for the community? We explored these questions through a detailed study of the design considerations of a real-world project that aims to improve ASR for Australian Aboriginal English. The case study of ASR also brings greater attention to a more general trend: the lack of technological support for Indigenous contact language varieties, and the ways that this perpetuates neo-colonialism through a continued failure to recognise and support Indigenous socio-cultural identities and ways of being. Contact language varieties are critical to contemporary Indigenous sociolinguistic identities, and to civil and economic participation, including in places such as Australia where settler-colonialism has led to the extinction of most ancestral languages. We discussed a number of ways in which speech technologies can support speakers of Australian Aboriginal English, as well as the risks of harms caused by system errors. This paper contributes to our understanding of ethical and responsible development of technologies for minoritised language communities, as well as to our understanding of practices for building culturally sensitive technologies while incorporating participatory methods. Although the project focuses on a specific technology, community and context—ASR for speakers of an indigenised variety of English in what we now know as Australia—we believe that some of the considerations and methodologies may be applicable to other technology projects.

7 Ethical considerations

The collection of Australian Aboriginal English for use by Google must be approached with extreme care. From the start, Rodríguez Louro and Collard clarified with their partners that they would not use unstructured yarning sessions like they had done for RA/4/20/4977 'Aboriginal English in the global city: Minorities and language change' (see, *e.g.*, [124]). With no pre-determined questions and clear Indigenous leadership through the research process, that project collected language as used in everyday settings, including sensitive stories about dispossession, trauma and suicide. These materials are firmly closed access with no risk of the audio files and associated transcriptions entering the public domain. However, unstructured yarning sessions tend to give rise to highly culturally sensitive material, which should be protected from use by third parties, especially large technology companies like Google. For this reason, in consultation with Hutchinson and the Advisory Committee, Rodríguez Louro and Collard recommended for this project the design of conversational prompts and story boards (see Appendix A) which seek to elicit interaction on a variety of general topics while steering clear of culturally sensitive material.

Only publicly shared YouTube videos were included in the corpus for developing transcription guidelines, and the data was not copied, modified or re-published, however the video creators were not contacted in order to understand their sentiments towards this research.

8 Positionality

Collard, Hutchinson, and Cooper are born in Australia, and Rodríguez Louro is an Australian born and raised in Argentina. Together, we have ancestral connections to Australian Indigenous Countries, Europe and South America, and we have trained and/or worked in linguistics, education, law, natural language processing, software engineering, and cybernetics. Collard is an L1 speaker of Australian Aboriginal English, and Collard and Rodríguez Louro have been active in advocating for greater appreciation of Australian Aboriginal English in civil society.

Acknowledgments

The University of Western Australia, where the second and third authors are based, is situated on unceded Nyungar Country; Google's Sydney office, where the first author is based, is situated on unceded Gadigal Country; and The Australian National University's Acton campus, where the last author is based, is situated on unceded Ngunnawal and Ngambri Country. All authors acknowledge that the Nyungar, Gadigal, Ngunnawal and Ngambri people respectively remain the spiritual and cultural custodians of their Country, on

which they continue to practise their values, languages, beliefs, and knowledge.

We would like to thank the Advisory Committee members Prof. Clint Bracknell and Mx Sharon Davis for their guidance; research assistants Hope Narrier, Katrina Cox and Lily Hayward for their data collection with members of their community; Emily Gilchrist for her help with project administration and data processing; Nefeli Perdikouli, Mitch Browne, Grace Shepherd, Madeleine Clews and Lucia Fraiese for their help identifying Aboriginal English in YouTube videos; and Vera Axelrod, Daan van Esch, Prof. Michael Haugh, Robert McLellan and Scott Riddle for their project support and advice. We would also like to thank Steven Bird and the anonymous reviewers of FAccT and other venues for their questions and suggestions.

References

[1] Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 7226–7249. doi:10.18653/v1/2022.acl-long.500

[2] Alëna Aksënova, Daan van Esch, James Flynn, and Pavel Golik. 2021. How Might We Create Better Benchmarks for Speech Recognition?. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, Kenneth Church, Mark Liberman, and Valia Kordoni (Eds.). Association for Computational Linguistics, Online, 22–34. doi:10.18653/v1/2021.bppf-1.4

[3] Ahmed Ali, Walid Magdy, Peter Bell, and Steve Renais. 2015. Multi-reference WER for evaluating ASR for languages with no orthographic rules. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 576–580.

[4] Ahmed Ali, Preslav Nakov, Peter Bell, and Steve Renals. 2017. WERd: Using social text spelling variants for evaluating dialectal speech recognition. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 141–148.

[5] Denise Angelo. 2021. *Countering misrecognition of Indigenous contact languages and their ecologies in Australia.* Ph. D. Dissertation. The Australian National University.

[6] Sherry R Arnstein. 1969. A ladder of citizen participation. *Journal of the American Institute of planners* 35, 4 (1969), 216–224.

[7] Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023. Probing Pre-Trained Language Models for Cross-Cultural Differences in Values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, Sunipa Dev, Vinodkumar Prabhakaran, David Adelani, Dirk Hovy, and Luciana Benotti (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 114–130. doi:10.18653/v1/2023.c3nlp-1.12

[8] Peter M Asaro. 2000. Transforming society by transforming technology: the science and politics of participatory design. *Accounting, Management and Information Technologies* 10, 4 (2000), 257–290.

[9] Australian Government, Australian Institute of Aboriginal and Torres Strait Islander Studies, and Australian National University. 2020. *National Indigenous Languages Report.* Technical Report. Commonwealth of Australia. https://www.arts.gov.au/sites/default/files/documents/national-indigenous-languages-report-lowres.pdf

[10] Australian Institute for Health and welfare. 2023. Income and finance of First Nations people. https://www.aihw.gov.au/reports/australias-welfare/indigenous-income-and-finance. https://www.aihw.gov.au/reports/australias-welfare/indigenous-income-and-finance Accessed: 2025-1-10.

[11] Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS). [n. d.]. AustLang Australian Indigenous Languages Database. https://aiatsis.gov.au/austlang. https://aiatsis.gov.au/austlang Accessed: 2025-4-30.

[12] Australian Institute of Health and Welfare. 2023. Employment of First Nations people. https://www.aihw.gov.au/reports/australias-welfare/indigenous-employment. Accessed: 2024-9-26.

[13] Natalia-Rozalia Avlona and Irina Shklovski. 2024. Torquing patients into data: enactments of care about, for and through medical data in algorithmic systems. *Information, Communication & Society* 27, 4 (2024), 735–757.

[14] Monika Bednarek. 2021. Australian Aboriginal English in Indigenous-Authored Television Series: A Corpus Linguistic Study of Lexis in Redfern Now, Cleverman and Mystery Road. *Journal of the European Association for Studies on Australia* 12 (2021).

[15] Emily Bender. 2019. The #benderrule: On naming the languages we study and why it matters. *The Gradient* 14 (2019), 34.

[16] Emily M Bender. 2009. Linguistically naïve!= language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction Between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?* 26–32.

[17] Glen Berman, Ned Cooper, Wesley Hanwen Deng, and Ben Hutchinson. 2024. Troubling Taxonomies in GenAI Evaluation. *arXiv preprint arXiv:2410.22985* (2024).

[18] Dawn Bessarab and Bridget Ng'andu. 2010. Yarning about yarning as a legitimate method in Indigenous research. *International Journal of Critical Indigenous Studies* 3 (2010), 37–50.

[19] Mat Bettinson and Steven Bird. 2021. Designing to support remote working relationships with Indigenous communities. In *Proceedings of the 33rd Australian Conference on Human-Computer Interaction*. 165–169.

[20] Steven Bird. 2020. Decolonising speech and language technology. In *28th International Conference on Computational Linguistics, COLING 2020*. Association for Computational Linguistics (ACL), 3504–3519.

[21] Steven Bird. 2020. Sparse Transcription. *Computational Linguistics* 46, 4 (Dec. 2020), 713–744. doi:10.1162/coli_a_00387

[22] Steven Bird. 2021. Sparse transcription. *Computational Linguistics* 46, 4 (2021), 713–744.

[23] Steven Bird. 2022. Local languages, third spaces, and other high-resource scenarios. In *60th Annual Meeting of the Association for Computational Linguistics, ACL 2022*. Association for Computational Linguistics (ACL), 7817–7829.

[24] Steven Bird. 2024. Must NLP be Extractive?. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 14915–14929.

[25] Steven Bird and Dean Yibarbuk. 2024. Centering the speech community. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. 826–839.

[26] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the people? Opportunities and challenges for participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–8.

[27] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 173–184.

[28] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 5454–5476. doi:10.18653/v1/2020.acl-main.485

[29] Claire Bowern. 2023. *The Oxford guide to Australian languages*. Oxford University Press.

[30] Geoffrey C Bowker and Susan Leigh Star. 2000. *Sorting things out: Classification and its consequences*. MIT press.

[31] Robin N Brewer, Christina Harrington, and Courtney Heldreth. 2023. Envisioning Equitable Speech Technologies for Black Older Adults. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 379–388.

[32] Lindell Bromham, Russell Dinnage, Hedvig Skirgård, Andrew Ritchie, Marcel Cardillo, Felicity Meakins, Simon Greenhill, and Xia Hua. 2022. Global predictors of language endangerment and the future of linguistic diversity. *Nature Ecology & Evolution* 6, 2 (Feb. 2022), 163–173. doi:10.1038/s41559-021-01604-y

[33] Mary Bucholtz. 2000. The politics of transcription. *Journal of pragmatics* 32, 10 (2000), 1439–1465.

[34] Deborah Cameron, Elizabeth Frazer, Penelope Harvey, Ben Rampton, and Kay Richardson. 1993. Ethics, advocacy and empowerment: Issues of method in researching language. *Language & Communication* 13, 2 (1993), 81–94.

[35] 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 (2015), 307–323.

[36] Stephanie Carroll, Ibrahim Garba, Oscar Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, et al. 2020. The CARE principles for Indigenous data governance. *Data science journal* 19 (2020).

[37] Grzegorz Chrupała. 2023. Putting Natural in Natural Language Processing. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 7820–7827. doi:10.18653/v1/2023.findings-acl.495

[38] Glenys Collard. 2025. Interview on "Australia Wide". https://www.abc.net.au/listen/programs/australia-wide/australia-wide/104974550. https://www.abc.net.au/listen/programs/australia-wide/australia-wide/104974550 Interviewed by Vanessa Mills. ABC Radio. Accessed: 2025-4-29. Start time: 25m53s.

[39] Glenys Collard and Celeste Rodríguez Louro. 2025. From Spark to Flame: Decolonising Linguistics and the Creation of First Nations Medical Media. In *Language and Decolonisation*. Routledge, 207–221.

[40] Ned Cooper, Courtney Heldreth, and Ben Hutchinson. 2024. "It's how you do things that matters": Attending to Process to Better Serve Indigenous Communities with Language Technologies. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, Yvette Graham and Matthew Purver (Eds.). Association for Computational Linguistics, St. Julian's, Malta, 204–211. https://aclanthology.org/2024.eacl-short.19

[41] Ned Cooper, Tiffanie Horne, Gillian R Hayes, Courtney Heldreth, Michal Lahav, Jess Holbrook, and Lauren Wilcox. 2022. A Systematic Review and Thematic Analysis of Community-Collaborative Approaches to Computing Research. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans, LA, USA, 1–18. doi:10.1145/3491102.3517716

[42] Ned Cooper and Alexandra Zafiroglu. 2024. From Fitting Participation to Forging Relationships: The Art of Participatory ML. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, Vol. 19. ACM, New York, NY, USA, 1–9. doi:10.1145/3613904.3642775

[43] Eric Corbett, Emily Denton, and Sheena Erete. 2023. Power and public participation in AI. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–13.

[44] Jay Cunningham, Su Lin Blodgett, Michael Madaio, Hal Daumé III, Christina Harrington, and Hanna Wallach. 2024. Understanding the Impacts of Language Technologies' Performance Disparities on African American Language Speakers. In *Findings of the Association for Computational Linguistics ACL 2024*. 12826–12833.

[45] Sunipa Dev, Vinodkumar Prabhakaran, David Adelani, Dirk Hovy, and Luciana Benotti (Eds.). 2023. *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*. Association for Computational Linguistics, Dubrovnik, Croatia. https://aclanthology.org/2023.c3nlp-1.0

[46] Greg Dickson. 2019. Aboriginal English(es). In *Australian English reimagined*. Routledge, 134–154.

[47] Lise M Dobrin and Jeff Good. 2009. Practical language development: Whose mission? *Language* 85, 3 (2009), 619–629.

[48] A Seza Doğruöz and Sunayana Sitaram. 2022. Language technologies for low resource languages: Sociolinguistic and multilingual insights. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*. 92–97.

[49] Jennifer Dudley-Nicholson. 2025. Google, researchers team to teach AI Aboriginal English. https://nit.com.au/19-02-2025/16375/google-researchers-team-to-teach-ai-aboriginal-english. https://nit.com.au/19-02-2025/16375/google-researchers-team-to-teach-ai-aboriginal-english Accessed: 2025-4-29.

[50] Esin Durmus, Karina Nguyen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv [cs.CL]* (June 2023). arXiv:2306.16388 [cs.CL] http://arxiv.org/abs/2306.16388

[51] Diana Eades. 2012. Communication with Aboriginal speakers of English in the legal process. *Australian Journal of Linguistics* 32, 4 (2012), 473–489.

[52] Diana Eades. 2012. The social consequences of language ideologies in courtroom cross-examination. *Language in society* 41, 4 (2012), 471–497.

[53] Diana Eades. 2013. *Aboriginal ways of using English*. Aboriginal Studies Press.

[54] Efthimis N Efthimiadis. 1996. Query Expansion. *Annual review of information science and technology (ARIST)* 31 (1996), 121–87.

[55] Daan van Esch, Elnaz Sarbar, Tamar Lucassen, Jeremy O'Brien, Theresa Breiner, Manasa Prasad, Evan Crew, Chieu Nguyen, and Françoise Beaufays. 2019. *Writing across the world's languages: Deep internationalization for Gboard, the Google keyboard*. Technical Report. Google. https://arxiv.org/pdf/1912.01218

[56] Siyuan Feng, Bence Mark Halpern, Olya Kudina, and Odette Scharenborg. 2024. Towards inclusive automatic speech recognition. *Computer Speech & Language* 84 (2024), 101567.

[57] Darren Flavelle and Jordan Lachler. 2023. Strengthening Relationships Between Indigenous Communities, Documentary Linguists, and Computational Linguists in the Era of NLP-Assisted Language Revitalization. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*. 25–34.

[58] Ben Foley, Peter Sefton, Simon Musgrave, and Moises Sacal Bonequi. 2024. Access Control Framework for Language Collections. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 113–121.

[59] Diana Forsythe. 2001. *Studying those who Study us: An Anthropologist in the World of Artificial Intelligence*. Stanford University Press, Palo Alto, CA. https://books.google.com/books/about/Studying_Those_Who_Study_Us.html?id=orNUzuFQeLgC

[60] Bruna Franchetto. 2008. The War of the Alphabets: Indigenous Peoples between the Oral and the Written. *Mana* 4, SE (2008), 0–0.

[61] Alice Gaby and Lesley Woods. 2020. Toward linguistic justice for Indigenous people: A response to Charity Hudley, Mallinson, and Bucholtz. *Language* 96, 4 (2020), e268–e280.

[62] Candace K Galla. 2009. Indigenous language revitalization and technology from traditional to contemporary domains. *Indigenous language revitalization: Encouragement, guidance & lessons learned* (2009), 167–182.

[63] Christian Gollan, Maximilian Bisani, Stephan Kanthak, Ralf Schluter, and Hermann Ney. 2005. Cross domain automatic transcription on the tc-star epps corpus. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, Vol. 1. IEEE, I–825.

[64] Lara Groves, Aidan Peppin, Andrew Strait, and Jenny Brennan. 2023. Going public: the role of public participation approaches in commercial AI labs. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1162–1173.

[65] Jean Harkins. 1994. *Bridging two worlds: Aboriginal English and crosscultural understanding*. University of Queensland Press.

[66] Christina N Harrington, Radhika Garg, Amanda Woodward, and Dimitri Williams. 2022. "It's kind of like code-switching": Black older adults' experiences with a voice assistant for health information seeking. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15.

[67] Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and Strategies in Cross-Cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 6997–7013. doi:10.18653/v1/2022.acl-long.482

[68] Jane Hill. 2002. "Expert Rhetorics" in advocacy for endangered languages: Who is listening, and what do they hear? *Journal of Linguistic Anthropology* 12 (2002), 119–133.

[69] Faye Holt, William Held, and Diyi Yang. 2024. Perceptions of Language Technology Failures from South Asian English Speakers. In *Findings of the Association for Computational Linguistics ACL 2024*. 4067–4081.

[70] Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Stroudsburg, PA, USA, 588–602. doi:10.18653/v1/2021.naacl-main.49

[71] Anne H Charity Hudley, Christine Mallinson, and Mary Bucholtz. 2020. Toward racial justice in linguistics: Interdisciplinary insights into theorizing race in the discipline and diversifying the profession. *Language* 96, 4 (2020), e200–e235.

[72] Ben Hutchinson, Negar Rostamzadeh, Christina Greer, Katherine Heller, and Vinodkumar Prabhakaran. 2022. Evaluation gaps in machine learning practice. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*. 1859–1876.

[73] Shannon Indigenous Archives Collective, Faulkhead, Kirsten Thorpe, Nathan Mudyi Sentance, Lauren Booker, and Rose Barrowcliffe. 2023. Indigenous Referencing Guidance for Indigenous Knowledges. (2023).

[74] Indigenous Business News. 2025. First Nations to benefit from UWA tech partnership. https://ibnews.com.au/first-nations-to-benefit-from-uwa-tech-partnership/. https://ibnews.com.au/first-nations-to-benefit-from-uwa-tech-partnership/ Accessed: 2025-4-29.

[75] Terri Janke. 2019. *True tracks: Indigenous cultural and intellectual property principles for putting self-determination into practice*. The Australian National University.

[76] Peter-Lucas Jones, Keoni Mahelona, Suzanne Duncan, and Gianna Leoni. 2023. Kia tangata whenua: Artificial intelligence that grows from the land and people. *Ethical Space: International Journal of Communication Ethics* 2023, 2/3 (Aug. 2023). doi:10.21428/0af3f4c0.9092b177

[77] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 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*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 6282–6293. doi:10.18653/v1/2020.acl-main.560

[78] Daniel Jurafsky and James H Martin. 2008. Speech and Language Processing: An introduction to speech recognition, computational linguistics and natural language processing. *Upper Saddle River, NJ: Prentice Hall* (2008).

[79] Anjali Kantharuban, Ivan Vulić, and Anna Korhonen. 2023. Quantifying the Dialect Gap and its Correlates Across Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 7226–7245. doi:10.18653/v1/2023.findings-emnlp.481

[80] Shigeki Karita, Richard Sproat, and Haruko Ishikawa. 2023. Lenient Evaluation of Japanese Speech Recognition: Modeling Naturally Occurring Spelling Inconsistency. In *Proceedings of the Workshop on Computation and Written Language (CAWL 2023)*, Kyle Gorman, Richard Sproat, and Brian Roark (Eds.). Association

for Computational Linguistics, Toronto, Canada, 61–70. doi:10.18653/v1/2023.cawl-1.8

[81] Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics* 10 (2022), 50–72. doi:10.1162/tacl_a_00447

[82] Per E Kummervold, Javier de la Rosa, Freddy Wetjen, Rolv-Arild Braaten, and Per Erik Solberg. 2024. Whispering in Norwegian: Navigating Orthographic and Dialectic Challenges. In *Proceedings of Interspeech 2024*.

[83] Eric Le Ferrand. 2023. *Leveraging Speech Recognition for Interactive Transcription in Australian Aboriginal Communities*. Ph. D. Dissertation. Charles Darwin University.

[84] 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 (Volume 1: Long Papers)*. 4988–4998.

[85] Heather Lent, Emanuele Bugliarello, Miryam de Lhoneux, Chen Qiu, and Anders Søgaard. 2021. On Language Models for Creoles. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, Arianna Bisazza and Omri Abend (Eds.). Association for Computational Linguistics, Online, 58–71. doi:10.18653/v1/2021.conll-1.5

[86] Heather Lent, Kelechi Ogueji, Miryam de Lhoneux, Orevaoghene Ahia, and Anders Søgaard. 2022. What a Creole Wants, What a Creole Needs. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 6439–6449. https://aclanthology.org/2022.lrec-1.691

[87] Heather Lent, Kushal Tatariya, Raj Dabre, Yiyi Chen, Marcell Fekete, Esther Ploeger, Li Zhou, Ruth-Ann Armstrong, Abee Eijansantos, Catriona Malau, et al. 2024. CreoleVal: Multilingual multitask benchmarks for creoles. *Transactions of the Association for Computational Linguistics* 12 (2024), 950–978.

[88] Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. CultureLLM: Incorporating Cultural Differences into Large Language Models. In *Proceedings of the Thirty-Eighth Annual Conference on Neural Information Processing Systems NeurIPS 2024*.

[89] Kenneth Liberman. 1985. *Understanding interaction in Central Australia*. Routledge.

[90] Pin-Jie Lin, Merel Scholman, Muhammed Saeed, and Vera Demberg. 2024. Modeling Orthographic Variation Improves NLP Performance for Nigerian Pidgin. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 11510–11522. https://aclanthology.org/2024.lrec-main.1006

[91] Zoey Liu, Crystal Richardson, Richard Hatcher, and Emily Prud'hommeaux. 2022. 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)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 3933–3944. doi:10.18653/v1/2022.acl-long.272

[92] Manuel Mager, Abteen Ebrahimi, Shruti Rijhwani, Arturo Oncevay, Luis Chiruzzo, Robert Pugh, and Katharina Von Der Wense. 2024. Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024). In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*.

[93] Maiam nayri Wingara. 2018. Indigenous Data Sovereignty Communique. https://www.maiamnayriwingara.org/mnw-principles. https://www.maiamnayriwingara.org/mnw-principles Accessed: 2025-1-10.

[94] Ian G Malcolm. 2000. Aboriginal English: From contact variety to social dialect. *Processes of language contact: Studies from Australia and the South Pacific* (2000), 123–144.

[95] Nina Markl. 2022. Mind the data gap(s): Investigating power in speech and language datasets. In *2nd Workshop on Language Technology for Equality, Diversity, Inclusion 2022*. Association for Computational Linguistics, 1–12.

[96] Nina Markl, Electra Wallington, Ondrej Klejch, Thomas Reitmaier, Gavin Bailey, Jennifer Pearson, Matt Jones, Simon Robinson, and Peter Bell. 2023. Automatic transcription and (de)standardisation. In *SIGUL 2023, 2nd Annual Meeting of the Special Interest Group on Under-resourced Languages: a Satellite Workshop of Interspeech 2023*.

[97] Felicity Meakins. 2013. Mixed languages. In *Contact languages: A comprehensive guide*, Peter Bakker and Yaron Matras (Eds.). Mouton De Gruyter, 159–228.

[98] Felicity Meakins. 2014. Language contact varieties. *The languages and linguistics of Australia: A comprehensive guide* 3 (2014), 365–416.

[99] Felicity Meakins, Jennifer Green, and Myfany Turpin. 2018. *Understanding linguistic fieldwork*. Routledge.

[100] Paul J Meighan. 2021. Decolonizing the digital landscape: The role of technology in Indigenous language revitalization. *AlterNative: An International Journal of Indigenous Peoples* 17, 3 (2021), 397–405.

[101] Paul J Meighan. 2022. Indigenous language revitalization using TEK-nology: How can traditional ecological knowledge (TEK) and technology support intergenerational language transmission? *Journal of Multilingual and Multicultural Development* (2022), 1–19.

[102] Zion Mengesha, Courtney Heldreth, Michal Lahav, Juliana Sublewski, and Elyse Tuennerman. 2021. "I don't think these devices are very culturally sensitive."—Impact of automated speech recognition errors on African Americans. *Frontiers in Artificial Intelligence* 4 (2021), 725911.

[103] Alexis Michaud, Oliver Adams, Trevor Anthony Cohn, Graham Neubig, and Séverine Guillaume. 2018. Integrating automatic transcription into the language documentation workflow: Experiments with Na data and the Persephone toolkit. (2018).

[104] Stephen Morey, Mark W Post, and Victor A Friedman. 2013. The language codes of ISO 639: A premature, ultimately unobtainable, and possibly damaging standardization. (2013). unpublished.

[105] Karin Moses, Colin Yallop, et al. 2008. Questions about questions. *Children's language and multilingualism: Indigenous language use at home and school* (2008), 30–55.

[106] Tarek Naous, Michael Ryan, Alan Ritter, and Wei Xu. 2024. Having Beer after Prayer? Measuring Cultural Bias in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 16366–16393. doi:10.18653/v1/2024.acl-long.862

[107] Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F P Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 2144–2160. doi:10.18653/v1/2020.findings-emnlp.195

[108] Dong Nguyen and A Seza Doğruöz. 2013. Word level language identification in online multilingual communication. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*. Association for Computational Linguistics, 857–862.

[109] Iuliia Nigmatulina, Tannon Kew, and Tanja Samardzic. 2020. ASR for Non-standardised Languages with Dialectal Variation: the case of Swiss German. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, Marcos Zampieri, Preslav Nakov, Nikola Ljubešić, Jörg Tiedemann, and Yves Scherrer (Eds.). International Committee on Computational Linguistics (ICCL), Barcelona, Spain (Online), 15–24. https://aclanthology.org/2020.vardial-1.2

[110] Kari Noe and Nurit Kirshenbaum. 2024. Where Generalized Equitable Design Practice Meet Specific Indigenous Communities. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–8.

[111] Walter J Ong. 2013. *Orality and literacy*. Routledge.

[112] Golden Marie Owens. 2023. "Hey Google, Talk Like Issa": Black Voiced Digital Assistants and the Reshaping of Racial Labor. https://soundstudiesblog.com/2023/06/05/google-talk-like-issa-black-voiced-digital-assistants-and-the-reshaping-of-racial-labor/. Accessed: 2024-10-04.

[113] Carolina Parada, Abhinav Sethy, and Bhuvana Ramabhadran. 2009. Query-by-example spoken term detection for OOV terms. In *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE, 404–409.

[114] Aidan Pine, Erica Cooper, David Guzmán, Eric Joanis, Anna Kazantseva, Ross Krekoski, Roland Kuhn, Samuel Larkin, Patrick Littell, Delaney Lothian, et al. 2024. Speech Generation for Indigenous Language Education. *Computer Speech & Language* (2024), 101723.

[115] Vinodkumar Prabhakaran, Sunipa Dev, Luciana Benotti, Daniel Hershcovich, Laura Cabello, Yong Cao, Ife Adebara, and Li Zhou (Eds.). 2024. *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*. Association for Computational Linguistics, Bangkok, Thailand. https://aclanthology.org/2024.c3nlp-1.0

[116] Vinodkumar Prabhakaran, Sunipa Dev, Luciana Benotti, Daniel Hershcovich, Yong Cao, Li Zhou, Laura Cabello, and Ife Adebara (Eds.). 2025. *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*. Association for Computational Linguistics, Albuquerque, New Mexico. https://aclanthology.org/2025.c3nlp-1.0/

[117] Emily Prud'hommeaux, Robbie Jimerson, Richard Hatcher, and Karin Michelson. 2021. Automatic speech recognition for supporting endangered language documentation. *Language Documentation & Conservation* 15 (2021), 491–513.

[118] Nola Purdie, Gina Milgate, and Hannah Rachel Bell. 2012. *Two way teaching and learning: Toward culturally reflective and relevant education.* Acer Press.

[119] Laura Rademaker. 2018. *Found in Translation: Many Meanings on a North Australian Mission.* University of Hawaii Press, Honolulu, HI, USA.

[120] Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 428–446. doi:10.18653/v1/2023.acl-long.26

[121] Celeste Rodríguez Louro and Glenys Collard. 2021. Yarns from the heart: the role of Aboriginal English in Indigenous health communication. https://theconversation.com/yarns-from-the-heart-the-role-of-aboriginal-english-in-indigenous-health-communication-163892. https://theconversation.com/yarns-from-the-heart-the-role-of-aboriginal-english-in-indigenous-health-communication-163892 Accessed: 2024-10-03.

[122] Celeste Rodríguez Louro and Glenys Collard. 2024. The Yarning Corpus: Aboriginal English in Southwest Western Australia. *Australian Journal of Linguistics* (2024), 1–17.

[123] Celeste Rodríguez Louro and Glenys Collard. 2021. Australian Aboriginal English: Linguistic and sociolinguistic perspectives. *Language and Linguistics Compass* 15, 5 (2021), e12415.

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

[125] Celeste Rodríguez Louro and Glenys Collard. 2024. Hearing the Voices: Embracing Diversity in the Study of Language in Society. *Voice and Speech Review* (2024), 1–8.

[126] Jan Sawicki, Maria Ganzha, and Marcin Paprzycki. 2023. The State of the Art of Natural Language Processing—A Systematic Automated Review of NLP Literature Using NLP Techniques. *Data Intelligence* 5, 3 (2023), 707–749.

[127] Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. 2019. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–33.

[128] Eva Schultze-Berndt, Felicity Meakins, and Denise Angelo. 2013. Kriol. In *The Survey of Pidgin and Creole Languages: Volume 1: English-based and Dutch-based Languages*. Oxford University Press, 241–251.

[129] 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*, Vol. 2.

[130] Jeff Siegel. 1999. Creoles and minority dialects in education: An overview. *Journal of Multilingual and Multicultural Development* 20, 6 (1999), 508–531.

[131] Ranjit Singh and Steven J Jackson. 2017. From margins to seams: Imbrication, inclusion, and torque in the Aadhaar Identification Project. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 4776–4824.

[132] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. 2022. Participation is not a design fix for machine learning. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–6.

[133] Andrew Smart, Ben Hutchinson, Lameck Mbangula Amugongo, Suzanne Dikker, Alex Zito, Amber Ebinama, Zara Wudiri, Ding Wang, Erin van Liemt, João Sedoc, et al. 2024. Socially Responsible Data for Large Multilingual Language Models. In *Proceedings of the fourth ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO'24)*.

[134] Linda Tuhiwai Smith. 2021. *Decolonizing methodologies: Research and indigenous peoples.* Bloomsbury Publishing.

[135] Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS nexus* 3, 9 (Sept. 2024), gae346. doi:10.1093/pnasnexus/pgae346

[136] Jennyfer Lawrence Taylor, Wujal Wujal Aboriginal Shire Council, Alessandro Soro, Paul Roe, and Margot Brereton. 2020. 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* (Fremantle, WA, Australia) *(OzCHI '19)*. Association for Computing Machinery, New York, NY, USA, 161–172. doi:10.1145/3369457.3369471

[137] Te Mana Raraunga. 2016. Our Charter. https://www.temanararaunga.maori.nz/tutohinga. https://www.temanararaunga.maori.nz/tutohinga Accessed: 2025-1-10.

[138] Daniela Teodorescu, Josie Matalski, Delaney Lothian, Denilson Barbosa, and Carrie Demmans Epp. 2022. Cree corpus: A collection of nêhiyawêwin resources. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 6354–6364.

[139] Jakelin Fleur Troy. 1994. *Melaleuka: A history and description of New South Wales pidgin.* Ph. D. Dissertation. Australian National University.

[140] Scientific UNESCO (United Nations Educational and Cultural Organization). 2020. Los Pinos Declaration [Chapoltepek]—making a decade of action for Indigenous languages.

[141] Kimi Wenzel, Nitya Devireddy, Cam Davison, and Geoff Kaufman. 2023. Can voice assistants be microaggressors? Cross-race psychological responses to failures of automatic speech recognition. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–14.

[142] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* 3, 1 (2016), 1–9.

[143] Shawn Wilson. 2020. *Research is ceremony: Indigenous research methods.* Fernwood publishing.

[144] Heike Winschiers-Theophilus and Nicola J. Bidwell. 2013. Toward an Afro-Centric Indigenous HCI Paradigm. *International Journal of Human–Computer Interaction* 29, 4 (2013), 243–255. doi:10.1080/10447318.2013.765763 arXiv:https://doi.org/10.1080/10447318.2013.765763

[145] Lesley Woods. 2023. *Something's Gotta Change: Redefining Collaborative Linguistic Research.* ANU Press.

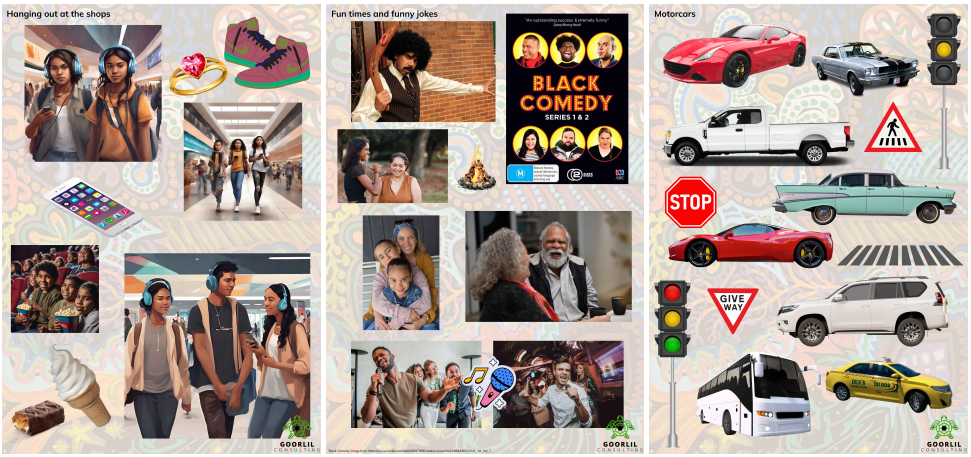
[146] Meg Young, Upol Ehsan, Ranjit Singh, Emnet Tafesse, Michele Gilman, Christina Harrington, and Jacob Metcalf. 2024. Participation versus scale: Tensions in the practical demands on participatory AI. *First Monday* (2024).

[147] Shiyue Zhang, Ben Frey, and Mohit Bansal. 2022. How can NLP Help Revitalize Endangered Languages? A Case Study and Roadmap for the Cherokee Language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 1529–1541. doi:10.18653/v1/2022.acl-long.108

Appendix A: Visual prompts for eliciting speech data collection

The prompts in Figure 1 were designed by Indigenous-owned consultancy Goorlil Consulting.

(a) Food (b) Fishing (c) Sport (d) Visiting family



(e) Hanging out at the shops (f) Fun times and funny jokes (g) Motorcars

Figure 1: Visual prompts for speech data elicitation.

Appendix B: Composition of corpus of Australian Aboriginal English on YouTube

For readers who wish to hear examples of Australian Aboriginal English, we recommend searching on YouTube using queries such as "My art, My culture", "Telling Our Stories, Our Stolen Generations", "Through our Eyes, Series 3", or "Community Spotlight Artists Profile Cairns Indigenous arts fair".

YouTube grants users of the platform 'a worldwide, non-exclusive, royalty-free license to access your Content through the Service, and to use that Content, including to reproduce, distribute, prepare derivative works, display, and perform it' (https://www.youtube.com/static?template=terms). Our use of YouTube videos to create transcription guidelines was consistent with this. These transcription guidelines do not contain any personal identifying information of offensive content.

The YouTube video corpus that we describe in §4 has the composition summarised in Tables 1–3.

Domain	# videos	Total duration (hours)
Arts and Culture	203	23.63
Autos and Vehicles	4	0.47
Comedy	29	2.20
Education	90	14.38
Entertainment	16	2.60
Film and Animation	30	7.82
Gaming	0	0
Health	26	2.53
How to / Style	14	1.79
News and Politics	42	6.25
Nonprofits and Activism	102	14.31
People and Blogs	149	22.29
Pets and Animals	7	0.82
Science and Technology	25	4.00
Sports	17	1.91
Travel and Events	22	4.37
Total	776	111.37

Table 1: Domains of videos.

Scripted?	# videos	Total duration (hours)
Yes	134	18.31
No	510	77.17
Unknown	132	15.89
Total	776	111.37

Table 2: Whether videos were scripted.

Percent in Australian Aboriginal English	# videos	Total duration (hours)
[50, 60)	34	8.43
[60, 70)	34	5.77
[70, 100]	708	97.18
Total	776	111.37

Table 3: Percent of speech in Australian Aboriginal English in videos. Only videos with at least 50% were considered; 70+% was preferred.

Appendix C: Summaries of mitigations of risks and enabling factors of opportunities

Risk	Mitigated by
1. Deploying inappropriate technology	Building on prior work that identified the appropriateness of ASR for this context; Validation through expert partners with expertise working with this community.
2. Low ecological validity	Multi-region data collection to guarantee that regional variations are represented; In-person data collection techniques led by Indigenous employees, so that speakers are comfortable speaking naturally.
3. Psychological harms from ASR errors	Primary goal of the project: reducing ASR errors for Australian Aboriginal English. We also call for more HCI research both on potential harms experienced by Indigenous communities, as well as on potential mitigations in the application layer.
4. Torquing identities	Community involvement in transcription guidelines; We call for more work on the transcription preferences of speakers of non-standardised languages.
5. Harming relationships	Partnering with experts with existing relationships with community; Establishing an Indigenous Advisory Committee.
6. Failing to partner with community	Indigenous researcher on the project; Hiring Indigenous staff; Engaging an Indigenous-owned consultancy; Establishing an Indigenous Advisory Committee.
7. Culturally unsafe methods in data collection	In-person, Indigenous-led data collection; Using culturally safe prompts developed by an Indigenous-owned consultancy; Verifying collected data for cultural safety; Researchers on the project with experience running culturally safe data collections.
8. Not respecting Indigenous data sovereignty	Google not owning the collected speech data; Indigenous Advisory Committee guidance on appropriate Indigenous Data Governance.
9. Not acknowledging Indigenous knowledge and expertise	Recognising contributors as co-authors on research artefacts; Engaging an Indigenous-owned consultancy; Establishing an Indigenous Advisory Committee.

Table 4: Project risks and mitigations.

Opportunity	Enabled by
1–4. Deployed system helpful to community and non-community members in diverse application contexts	Explicitly including support for various application contexts in design desiderata.
5. Project members develop greater understanding of Indigenous contexts and histories	Team discussions on power dynamics, barriers, and social injustices; Team discussions about language use during the writing of this paper and other artefacts.
6. Economic opportunities	Equitable payment for Indigenous data providers; Employing Indigenous staff.
7. Building capabilities	Employing Indigenous staff in roles that enable skill development, including designing elicitation materials and transcription guidelines, and managing voice data collection.
8. Self-determination	Pursuing data governance frameworks which give Indigenous people the power to decide how the Indigenous voice recordings might be used in the future.

Table 5: Project opportunities and enabling factors.