"It's *how* you do things that matters": Attending to Process to Better Serve Indigenous Communities with Language Technologies

Anonymous ACL submission

Abstract

Indigenous languages are historically underserved by NLP technologies, but this is changing for some languages with the recent scaling of large multilingual models and an increased focus by the NLP community on endangered languages. This position paper explores ethical considerations in building NLP technologies for Indigenous languages, based on the premise that such projects should primarily serve Indigenous communities. We report on interviews with 17 researchers working in or with Aboriginal and Torres Strait Islander communities on language technology projects in Australia. Drawing on insights from the interviews, we recommend practices for NLP researchers to increase attention to the process of engagements with Indigenous communities, rather than focusing only on decontextualised artefacts.

1 Introduction

In this position paper, we discuss how to ethically build Natural Language Processing (NLP) technologies for Indigenous languages, which have historically been poorly served by NLP. This is a timely question, as we are in the UNESCO International Decade of Indigenous Languages (2022–2032), and there has been a recent trend towards more NLP technologies processing Indigenous languages. One thread of recent projects has been motivated by scaling large multilingual models to include Indigenous languages, including Māori, Zulu, Igbo, Southern Quechua, Hawai'ian, Querétaro Otomi, Navajo, and more (e.g., Pratap et al., 2023; ImaniGooghari et al., 2023; Kudugunta et al., 2023). Another thread of recent projects is driven by threats of language extinction, for example, the six Workshops on Computational Methods for Endangered Languages held between 2014–2023, and the ACL 2022 Theme Track on low-resource and endangered languages. Both threads of research are typically based on assumptions that language technology should be accessible to everyone in their native language(s), and that the availability of those language technologies will promote language use and preservation (Bird, 2020).

We start with the premise that NLP for Indigenous languages should primarily serve Indigenous communities. If this is indeed the goal of the NLP community, then we need NLP to be accountable to and benefit Indigenous communities (Schwartz, 2022), and to consider communities' values and experiences with respect to NLP projects. This includes considering the context of Indigenous communities within colonised societies (Schwartz, 2022; Bird, 2020) and the expressed opinions of those communities around data governance (e.g. Liu et al., 2022; Mager et al., 2023). The overarching question for this paper, then, is: *how can NLP better serve Indigenous communities*?

To consider this question, we first review the developing discourse around decolonisation of language technology along with principles for Indigenous data governance. We then report on interviews with researchers working in or with Indigenous communities on language technology projects in Australia, the country in which the authors live. Drawing on insights from the interviews, we recommend practices for NLP researchers working with Indigenous languages. Overall, we encourage NLP researchers to increase attention to the process of engagements with Indigenous communities, rather than focusing only on decontextualised artefacts.

2 Background

Languages can be marginalised in different ways. The NLP research community describes a language as '*low-resource*' when there is insufficient data in that language to train and evaluate statistical and machine learning models (Liu et al., 2022). The poverty-conscious framing of 'low-resource' has been criticised by Bird (2022), however, for being colonial and Eurocentric. We prefer the term *underserved* in this paper (echoing, for example, Bender and Friedman, 2018; Kaffee et al., 2018; Armstrong et al., 2022; Forbes et al., 2022), as we recognise that a language may be fully constituted in its own ways, while it may not be serviced by dominant NLP tools or techniques. Guided by scholars of marginalisation processes (e.g., Bagga-Gupta, 2017), we seek to pivot the discussion from 'low-resource' languages to how technology communities are *under-serving* language communities.

Languages spoken by few people may additionally be defined as *endangered*—at risk of disappearing due to a lack of speakers (Bromham et al., 2022). However, having few living speakers does not necessarily mean a language is low-resource (e.g., Latin has enough data to support Google Translate).

The majority of Indigenous languages—languages native to a particular region and spoken by Indigenous peoples—are forecast to disappear by the end of this century (Bromham et al., 2022). In practice, most Indigenous languages are endangered due not to any inherent linguistic inferiority, but rather due to the global economic, ideological, military, and nationalistic practices that are constitutive of *colonialism*.

2.1 Decolonisation and Language Technology

Decolonial approaches to addressing marginalisation in technology are motivated by social justice and self-determination (Smith, 1999), rather than by data efficiency. These approaches encourage researchers to embrace perspectives from and at the margins in order to surface and critique the persistence of colonial relationships in present-day society (Maldonado-Torres, 2007; Quijano, 2007; Escobar, 2018). The decolonisation literature suggests there are three broad strategies to enact decolonial agendas in language technology work.

Firstly, decolonial agendas require that we *consider whose interests are served by NLP*. Language technologies are laden with cultural perspectives and assumptions (Awori et al., 2016), and NLP has a "habit of … technological colonisation" along with making assumptions about goals and methods (Bird, 2020). Research on languages of Indigenous communities must be conducted on their terms (Dourish et al., 2020) and research outputs must be primarily relevant to those communities, not only to research communities (Alvarado Garcia et al., 2021).

Secondly, decolonial agendas encourage us to *question the universality of values* (Mignolo, 2011; Grosfoguel, 2007), in particular, the primacy of Western values over others. This includes questioning methods and utility functions of NLP projects. Assuming all communities want the same language technologies disempowers local communities (Bird, 2020). Instead, we must critique the universalising logic of our methods, along with technologies (Dourish et al., 2020; Irani et al., 2010).

Thirdly, decolonial agendas direct us to *interrogate power dynamics embedded in NLP projects*. Approaches from the Global North are often disconnected from the life experiences of those in the Global South (Alvarado Garcia et al., 2021). In addition, power asymmetries exist between users and platforms (Couldry and Mejias, 2018), and between different regions of the world (Kwet, 2019).

2.2 Principles for Indigenous Data Governance

We believe it is critical to consider Indigenous perspectives on language data management. Examples of such perspectives are reflected in the CARE principles of the Global Indigenous Data Alliance (Carroll et al., 2020), the Maiam nayri Wingara (2018) Indigenous Data Sovereignty Principles, and the Te Mana Raraunga (2016) principles of Māori data sovereignty. These principles grapple with an ongoing tension for Indigenous communities when engaging with language technologists—between maintaining sovereignty over their language data and engaging with technological developments that could benefit language revitalisation efforts. Although each set of principles is distinct, a thematic analysis by the authors revealed some common areas of concern. a) *Respect*: Acknowledge and support the rights of people and communities to hold and express different values, norms and aspirations regarding data and technology. This requires listening and understanding culture. b) *Relationships*: Act cooperatively. Build positive, long-term relationships. c) *Shared control*: Support data governance and control. Support the exercise of data guardianship using traditional protocols. d) *Benefits*: Understand disparate benefits and ensure equitable distribution of benefits. Provide evidence of individual and collective benefits.

3 Insights from Interviews

Building on the previous section, our focus here narrows to Australia as a case study. Australian

Aboriginal and Torres Strait Islander languages are marginalised in multiple ways. There is a scarcity of language technologies, which reflects a much broader technological under-serving of these communities. Indeed, many communities struggle to even get reliable and affordable access to the internet (Featherstone et al., 2023). Prior to colonisation, there were more than 250 local languages spoken in Australia, though today just over 120 languages are in use or being revitalised and more than 90% of those are considered endangered (Australian Government et al., 2020). However, it is not for a lack of internet, data, or NLP technologies that many local languages are endangered or extinct. We cannot ignore the impacts of colonialism—in many cases, language loss is the byproduct of oppression. Local languages were often the target of colonial oppression as those languages sustained identities and connection to Country.

Specifically, two research questions guided a series of interviews with researchers who work in or with Aboriginal and Torres Strait Islander communities on speech and language technology projects. Firstly, how should language technologists work with local communities to develop speech and language technologies? Secondly, what is the role of speech and language technologies in sustaining language use by local communities? We conducted in-depth, 60-minute interviews with 17 researchers from academia and community-based organisations (see Appendix). Their reflections, detailed in the following sections, shed light on strategies and challenges to enact decolonial agendas and Indigenous data governance principles at the project level.

3.1 *How* to work with Aboriginal and Torres Strait Islander communities

We first asked interviewees how they decide what to work on and who to work with. All interviewees strongly emphasised that speech and language technology projects *"must start with a community need"*, and that recognising such needs requires long-term relationships. The need for translation, for example, often arises where communities or researchers observe something happening across cultures over time. Many interviewees also argued that projects shouldn't start with technology, or solutions. Instead, interviewees encouraged other technologists to demo existing technology and facilitate experimentation with the tools by communities for their languages.

We also asked researchers how they manage relationships with the people they work with. All interviewees emphasised that researchers must clarify to partner communities the mutual benefits of a project at the outset, with some interviewees explicitly mentioning the negotiation of data access rights. Several interviewees noted that community-based work requires researchers to question universal assumptions about the social or cultural factors relevant to technology, and that personal relationships are key to managing those complexities.

Finally, we asked about finishing projects. Most interviewees noted that, though it is important for projects to have an end date, personal relationships between researchers and communities persist. Several interviewees encouraged translating documentation into an accessible form that communities can access ongoing (rather than locking up data in bespoke, single purpose tools). Those same interviewees argued that repositories and archives support the sustainability of project outcomes: *"Apps and websites are disposable . . . store the data in an archival format that is going to persist."*.

3.2 *What* to work on with Aboriginal and Torres Strait Islander communities

Most interviewees stated that the primary motivation of Aboriginal and Torres Strait Islander communities for building speech and technologies is the transmission of culture via language: *"Tap into the intrinsic motivation of transmitting life and knowledge down the generations."* Several interviewees encouraged a *"design for one, then scale"* approach, where researchers collaborate with one community, then scale a *"digital shell"*—a technological template tailored for one community, yet adaptable enough to be customised by others, streamlining early development stages for each new engagement. Others urged technologists to consider the benefits of the production process to communities, to facilitate capacity building in technology development, not only focusing on project outputs like datasets or publications.

In terms of application domains, several interviewees advocated for improving accessibility to archival materials using front-end tools for metadata tagging and information retrieval, especially for audio. Others emphasised the importance of vehicular languages like Aboriginal English and Kriol. Interviewees noted that many communities use vehicular languages to participate in the national economy and access education and health systems. Finally, some interviewees encouraged multi-modal work to support signed Aboriginal and Torres Strait Islander languages, alongside text and audio.

4 Recommendations and Conclusion

To conclude, we propose a set of practices building on the insights from our interviews. These practices grapple with a tension for NLP researchers working with Indigenous languages—between producing work that is relevant to local partner communities and the demands of research communities for projects that scale across many languages. We intend to contribute to the discourse about decolonisation of language technology, not by resolving this tension, but by recommending a cyclical process of engagement to assist researchers to navigate it (Figure 1). As Escobar (2018) suggests for design, we argue that the NLP community can engage with marginalisation and dispossession through a greater focus on the process of engagements rather than on artefacts alone.

An ethical process starts by *seeking out community needs*. This means asking communities we wish to partner with about their goals for their languages, and ensuring our efforts are aligned with those goals (Liu et al., 2022). This approach may lead us to focus more on supporting the transmission of cultural knowledge across generations, not only expanding access to products and services. Focusing solely on data collection by communities to develop products and services risks disenfranchising communities. Instead, one approach might be to demo existing technology at community events (e.g., the PULiiMA Indigenous Languages and Technology Conference) and asking how communities can appropriate it for their needs.

Engaging with community representative bodies can help researchers establish long-term relationships with community members. While personal relationships between researchers and community members are crucial, engaging through representative bodies offers a distinct advantage in balancing power dynamics. Additionally, these bodies already have established relationships within their communities, allowing researchers to build trust and credibility more rapidly.

Relatedly, we must consider how to *negotiate control over project resources and ongoing relationships*. At the start of community-engaged language technology projects, this means scheduling time to interrogate power dynamics (Blodgett et al., 2020) and considering how to share power with community partners by recognising Indigenous (co)-ownership of outcomes of data collection efforts (e.g., community ownership of datasets or other intellectual property, and joint publications (Janke, 2021)).

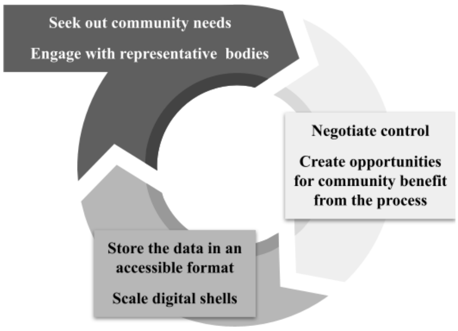
Seek out community needs
Engage with representative bodies
Negotiate control
Create opportunities for community benefit from the process
Store the data in an accessible format
Scale digital shells

Figure 1: Recommended process for engagement.

Where data collection is a component of a project with an Indigenous community, we must consider how *the process of engagement might be an opportunity for community benefit*. In practice, this may involve designing experiences for community members to learn about language technology as part of the process of generating or collecting data, and creating outputs from data collection that are accessible by community members, not only usable by language technologists.

In addition, it is critical to *store and maintain data produced from the project* in a format that community partners can access beyond the project (e.g., archives or repositories). Where researchers also intend to scale projects across languages, we recommend starting small—focusing on one to two communities, then *scaling digital shells* to other contexts (see, for example, Richards, 2023; Foley et al., 2018).

Finally, we also urge the NLP research environment to pay more attention to the process of engaging with Indigenous communities, rather than focusing on de-contextualised model accuracy benchmarks as proxies for utility to communities (Hutchinson et al., 2022). This means including the process of engagement as a core reviewing criterion when processing Indigenous languages, and fostering forums where Indigenous voices can articulate their needs to the NLP community. Let the process of engagement with Indigenous communities and their voices be the pillars of our research.

References

Adriana Alvarado Garcia, Juan F Maestre, Manuhuia Barcham, Marilyn Iriarte, Marisol Wong-Villacres, Oscar A Lemus, Palak Dudani, Pedro Reynolds-Cuéllar, Ruotong Wang, and Teresa Cerratto Pargman. 2021. Decolonial pathways: Our manifesto for a decolonizing agenda in HCI research and design. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, pages 1–9, New York, NY, USA. Association for Computing Machinery.

Ruth-Ann Armstrong, John Hewitt, and Christopher Manning. 2022. JamPatoisNLI: A jamaican patois natural language inference dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5307–5320, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Kagonya Awori, Nicola J Bidwell, Tigist Sherwaga Hussan, Satinder Gill, and Silvia Lindtner. 2016. Decolonising technology design. In *Proceedings of the First African Conference on Human Computer Interaction*, AfriCHI'16, pages 226–228, New York, NY, USA. Association for Computing Machinery.

Sangeeta Bagga-Gupta. 2017. *Marginalization Processes Across Different Settings: Going Beyond the Mainstream*. Cambridge Scholars Publishing.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Steven Bird. 2022. Local Languages, Third Spaces, and other High-Resource Scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7817–7829, Dublin, Ireland. Association for Computational Linguistics.

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*, pages 5454–5476, Online. Association for Computational Linguistics.

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):163–173.

Stephanie Carroll, Ibrahim Garba, Oscar Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, Rodrigo Sara, Jennifer Walker, Jane Anderson, and Maui Hudson. 2020. The CARE principles for indigenous data governance. *Data Science Journal*, 19:1–12.

Nick Couldry and Ulises A Mejias. 2018. Data colonialism: Rethinking big data's relation to the contemporary subject. *Television & New Media*, 20(4).

Paul Dourish, Christopher Lawrence, Tuck Wah Leong, and Greg Wadley. 2020. On being iterated: The affective demands of design participation. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–11, New York, NY, USA. Association for Computing Machinery.

Arturo Escobar. 2018. *Designs for the Pluriverse: Radical Interdependence, Autonomy, and the Making of Worlds*. New Ecologies for the Twenty-First Century. Duke University Press, Durham.

Daniel Featherstone, Lyndon Ormond-Parker, Lauren Ganley, Julian Thomas, Sharon Parkinson, Kieran Hegarty, Jenny Kennedy, and Indigo Holcombe-James. 2023. Mapping the digital gap: 2023 outcomes report. Technical report, ARC Centre of Excellence for Automated Decision-Making and Society.

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

Clarissa Forbes, Farhan Samir, Bruce Oliver, Changbing Yang, Edith Coates, Garrett Nicolai, and Miikka Silfverberg. 2022. Dim wihl gat tun: The case for linguistic expertise in NLP for Under-Documented languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2116–2130, Dublin, Ireland. Association for Computational Linguistics.

Ramón Grosfoguel. 2007. The epistemic decolonial turn: Beyond political-economy paradigms. *Cultural Studies*, 21(2-3):211–223.

Ben Hutchinson, Negar Rostamzadeh, Christina Greer, Katherine Heller, and Vinodkumar Prabhakaran. 2022. Evaluation Gaps in Machine Learning Practice. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1859–1876, Seoul Republic of Korea. Association for Computing Machinery.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Lilly Irani, Janet Vertesi, Paul Dourish, Kavita Philip, and Rebecca E Grinter. 2010. Postcolonial computing: a lens on design and development. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1311–1320, New York, NY, USA. Association for Computing Machinery.

Terri Janke. 2021. *True Tracks: Indigenous cultural and intellectual property principles for putting self-determination into practice*. University of New South Wales Press.

Lucie-Aimée Kaffee, Hady Elsahar, Pavlos Vougiouklis, Christophe Gravier, Frédérique Laforest, Jonathon Hare, and Elena Simperl. 2018. Learning to generate wikipedia summaries for underserved languages from wikidata. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 640–645, New Orleans, Louisiana. Association for Computational Linguistics.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A multilingual and Document-Level large audited dataset. *arXiv preprint arXiv:2309.04662*.

Michael Kwet. 2019. Digital colonialism: US empire and the new imperialism in the global south. *Race & Class*, 60(4):3–26.

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)*, pages 3933–3944, Dublin, Ireland. Association for Computational Linguistics.

Manuel Mager, Elisabeth Mager, Katharina Kann, and Ngoc Thang Vu. 2023. Ethical considerations for machine translation of indigenous languages: Giving a voice to the speakers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4871–4897, Toronto, Canada. Association for Computational Linguistics.

Maiam nayri Wingara. 2018. Indigenous data sovereignty communique. `https://www.maiamnayriwingara.org/mnw-principles`. Accessed: 2023-6-16.

Nelson Maldonado-Torres. 2007. On the Coloniality of Being. *Cultural Studies*, 21(2-3):240–270.

Walter D Mignolo. 2011. *Global Futures, Decolonial Options*. Duke University Press.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1,000+ languages. *arXiv preprint arXiv:2305.13516*.

Aníbal Quijano. 2007. Coloniality and Modernity/Rationality. *Cultural Studies*, 21(2-3):168–178.

Mark Richards. 2023. Listen N talk. `https://westernsydney.edu.au/marcs/impact/case_studies/listen_n_talk`. Accessed: 2023-2-16.

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 (Volume 2: Short Papers)*, pages 724–731, Dublin, Ireland. Association for Computational Linguistics.

Linda Tuhiwai Smith. 1999. *Decolonizing Methodologies: Research and Indigenous Peoples*, 2nd edition. Zed Books, London, UK.

Te Mana Raraunga. 2016. Our charter. `https://www.temanararaunga.maori.nz/tutohinga`. Accessed: 2023-6-16.

A Summary of interviewees

Indigenous status	Count
Non-Indigenous	12
Aboriginal and/or Torres Strait Islander	5

Table 1: Indigenous status of interviewees.

Field of Expertise	Count
Linguistics	7
Computing	7
Community-based research	3

Table 2: Primary field of expertise of interviewees.

Australian State or Territory	Count
Queensland	4
New South Wales	4
Victoria	4
Northern Territory	3
Western Australia	2

Table 3: Location of interviewees.