Artificial Intelligence (AI) for Indigenous Language Revitalization: Emerging Concerns

Authors:
The First Peoples' Cultural Council; Endangered Languages Project; with Aidan Pine, Associate Research Officer, National Research Council of Canada

Overview

- We are writing as a group of language revitalization practitioners, engineers and researchers who share a common goal of supporting Indigenous language revitalization in close collaboration with communities.
- Digital tools using artificial intelligence (AI) are being marketed to communities to solve many different challenges that communities face in language revitalization.
- In the rush to adopt the latest technological trend, we believe that key elements of technology development and delivery best practises are being ignored, and core principles of Indigenous data ownership are being violated.
- Further, communities are not being adequately informed about the risks of developing these technologies or of handing over their data for use with these technologies.

Harms and Risks of AI

- Fabricated information
    - AI tools are known to "hallucinate" or confidently deliver false or fabricated information.
    - For example, in the US, federal emergency services have delivered "unintelligible" and garbled emergency response information, believed to be AI-translated, in Alaska Native languages.[1]
    - AI-generated "vocabulary" books for Indigenous languages, filled with false information and attributed to non-existent authors, have been sold online through platforms like Amazon.[2]
    - Revitalization efforts for Indigenous languages, particularly those with few living fluent speakers, can experience disproportionate harm from false or fabricated language information.

[1] https://www.kyuk.org/alaska-state-news/2023-01-07/lost-in-translation-fema-sent-unintelligible-disaster-relief-application-information-to-alaska-natives-impacted-by-typhoon-merbok
[2] https://www.aptnnews.ca/national-news/amazon-removes-questionable-plains-cree-language-book-after-being-contacted-by-aptn-news/

- Exaggerated benefits
    - The challenges of developing AI tools for Indigenous languages are often not clearly communicated. Demonstrations of the abilities of AI might be done in English, but to our knowledge, reaching English-level quality for almost every AI technology is currently not technically feasible for any Indigenous language, and this fact may not be communicated.
    - Importantly, this kind of selective communication exaggerates the potential benefits of the technology, which redirects funding and attention away from the steady, community-led work that actually sustains language use.
- Environmental costs
    - CO_2 and water consumption rates are extremely high for many AI technologies, especially for training large language models.
    - Increased electricity demands can lead to land dispossession, as with hydroelectric dam projects in Quebec (where much of Canada's AI compute infrastructure is located).[3]
    - Indigenous communities are already disproportionately harmed by the climate crisis.[4] Creating AI technologies for Indigenous languages may exacerbate this harm.
- Privacy and consent violations
    - Indigenous languages and intellectual property have already been subject to widespread theft and misappropriation. AI tools are frequently developed in ways that violate the First Nations principles of ownership, control, access, and possession (OCAP)[5] and Indigenous data sovereignty rights.
    - The amount of data required by most AI systems has led to a 'smash-and-grab' approach to data mining where terms of use are disregarded.
    - Meta recently published a paper[6] which built text-to-speech systems for four Indigenous Canadian languages using scraped data from the data provider Faith Comes By Hearing (in apparent violation of the data provider's terms).[7] Meta released the models without the knowledge or permission of the speakers whose voice data had been taken for training.
    - AI models produced by large tech companies such as Meta are often published under open licenses. This allows anyone to easily reuse language data, which was acquired without proper consent, leading to compounding harms.

[3] https://www.newswire.ca/news-releases/export-of-canadian-hydropower-to-the-united-states-first-nations-in-quebec-and-labrador-unite-to-oppose-hydro-quebec-project-845431188.html
[4] https://www.pbssocal.org/shows/tending-nature/the-disproportionate-impact-of-climate-change-on-indigenous-communities
[5] https://fnigc.ca/ocap-training/
[6] https://arxiv.org/pdf/2305.13516
[7] https://www.faithcomesbyhearing.com/terms

- Indigenous communities have expressed concern about AI tools being used to produce offensive or culturally inappropriate materials in their languages, or to impersonate the voices of community members without their consent, especially those who have passed away.

Strategies for Harm Reduction

Given the above concerns, there are a variety of strategies to mitigate possible harm.

- For communities wishing to engage in AI technologies for Indigenous language revitalization:
  - Do not rush into AI partnerships.
  - Before choosing to adopt AI technology, clarify the goals of using the technology. For example, go through each of the questions in First Peoples' Cultural Council's 'Check Before You Tech' document.[8]
  - Practise due diligence in terms of vetting requests for data.
  - Publish terms of use and ensure all language data hosted online has anti-scraping measures in place or consider restricting public access to some or all language data.
  - Develop communication strategies to raise awareness about the risks associated with relinquishing control of data.
- For companies wishing to support Indigenous language revitalization:
  - Data and models must be owned entirely by community collaborators. The use and distribution of synthesized audio must also be determined by them.
  - Develop strong permissions-based access to data and models. Do not release models or data in open-source format without clear consent from a community language authority.
  - Support the development of more effective software tools to protect Indigenous language data from data scraping, theft and misappropriation.
  - Ensure free, prior, informed and continued consent throughout the duration of a project, including prior to deployment or distribution of models, training data, or synthesized speech.

[8] https://fpcc.ca/resource/check-before-you-tech/

Ten Questions to Ask Tech Companies about Indigenous Data Sovereignty

1. Intent and Consent: What are your intentions for the development of the technology? What are your processes for securing free, prior, and informed consent from Indigenous Nations and communities when developing and publishing models of their languages? How are you securing consent to create LLMs or other language models and technologies *before* creating them? What is your current and ongoing commitment to the communities for the use of their data?

2. Data Sources and Consent Processes: What were the sources of the data used to train your models for Indigenous languages? How do you navigate the complexities of using materials that may be considered "public domain" while ensuring you respect Indigenous data sovereignty? What processes do you have in place for securing consent from both data copyright holders and Indigenous communities or organizations?

3. Consultation During Model Development: Pratap et. al. 2024, published by Meta employees, reported that they did not consult with any Indigenous communities nor speakers prior to building and releasing text-to-speech models for their languages (see Section 2.3). The scraping of data from the source named in the paper also appears to be in contravention of the terms of use of the publisher. What are your policies around sourcing data and developing and publishing Indigenous language models?

4. Policy Alignment with Indigenous Data Sovereignty Frameworks: Are you aware of the OCAP principles for Indigenous data, developed by the First Nations Information Governance Centre, the FAIR principles developed by the Research Data Alliance and the CARE principles developed by the Global Indigenous Data Alliance (GIDA)? In what ways do your policies align with these Indigenous data governance frameworks? Do you have any safeguards or processes in place for adhering to OCAP, FAIR, and CARE principles when working with Indigenous languages?

5. Community Engagement: How are you involving the relevant Indigenous communities in the development of tools for their languages? Is your company developing sustainable long-term relationships with these communities or Nations, or are your connections to them more focused on short-term projects and ad hoc consultation? Please describe your processes in detail.

6. Assessing Performance of Indigenous Language Models: How do you assess the performance of models or technologies developed for specific Indigenous languages before public release? Are there review processes involving qualified Indigenous speakers and language experts in evaluating the quality and performance of these models? If so, please describe them. How do you verify that the Indigenous speakers with whom you are engaging have the mandate and authority to share or verify information that belongs to their community?

7. Risk Assessment: Can you describe the process of assessing risk when developing or publishing models? How are decisions made about whether potential benefits outweigh potential risks? Who is invited to be part of those determinations, and how is their input implemented? In what ways do you incorporate or prioritize the perspectives of the relevant Indigenous communities?

8. Preventing and Responding to Harm: What measures do you have in place to prevent the release of models that do not perform well in Indigenous languages, particularly given the potential harm to revitalization efforts and to Indigenous speech communities? What processes are in place to respond to communities' concerns or objections to any models or tools your company releases?

9. Benefits and Equity: Benefits are likely to accrue to your company from the production of models representing Indigenous languages, whether financial or otherwise. Do you compensate Indigenous Nations and communities sufficiently/adequately for their efforts to create and produce the data? In what ways do you ensure equitable distribution of benefits with the communities whose languages are being used?

10. Company Involvement in Indigenous Language Efforts: Can you broadly outline your company's track record of supporting Indigenous languages, particularly in relation to data sovereignty principles, and explain what increases in investment and effort you have made during the International Year of Indigenous Languages (2019) and the International Decade of Indigenous Languages (which started in 2022)? What are your

company's motivations and objectives in developing models or technologies for Indigenous languages?

About us

The First Peoples' Cultural Council (FPCC) is a provincial Crown Corporation formed by the government of British Columbia in 1990 to administer the First Peoples' Heritage, Language and Culture Program. The mandate of FPCC is to assist B.C. First Nations in their efforts to revitalize their languages, arts, cultures and heritage.

The Endangered Languages Project (ELP) is a U.S.-based nonprofit organization supporting the revitalization of Indigenous and endangered languages around the world. ELP brings people together across borders and boundaries to address the urgent issue of language endangerment.

Authors:

The First Peoples' Cultural Council; Endangered Languages Project;, with Aidan Pine, Associate Research Officer, National Research Council of Canada

Contributors:

Pius Akumbu, CNRS-LLACAN and Endangered Languages Project
Bridget Chase, Typotheque
Marie-Odile Junker, Carleton University
Keoni Mahelona, Te Hiku Media
Gerald Roche, La Trobe University
Heather Souter, Prairies to Woodlands Indigenous Language Revitalization Circle
Jeff Ward, Animikii

For more information

Anna Belew, Executive Director, Endangered Languages Project, anna@endangeredlanguages.com
Suzanne Gessner, Research & Development Linguist, First Peoples' Cultural Council, suzanne@fpcc.ca

Last revised: June 2026