Revitalizing Indigenous languages with the help of advanced research computing

Nathan Brinklow (Thanyehténhas), a PhD student at Queen's University, has been deeply involved in both community- and research-based activities designed to reclaim and revitalize indigenous languages, particularly Mohawk. In 2019, he attended a colloquium at Queen's and happened to meet Chris MacPhee, at that time the Director of Operations and Development for Queen's Centre for Advanced Computing (CAC). "Chris said, let's talk about how CAC can help you," said Brinklow. That proved to be a turning point in the way he approaches his work. "Often, the reason Humanities researchers don't take advantage of computational methods is that we don't know about what's available," Brinklow said. "CAC is enabling us to have these conversations and think about these possibilities."

Right: Community data collection: A group of Mohawk first-language speakers using plays and scripts to elicit and record vocabulary.

Photo credit: Tsi Tyónnheht Onkwawén

Left: Tyendinaga Mohawk Territory: Data collection through story-telling and discussion.

Photo credit: Tsi Tyónnheht Onkwawén

One of the possibilities involves solving the 'transcription bottleneck'. Analyzing linguistic data requires spoken and written versions of the same text, both indexed to time. While there is a wealth of language data available, most of it is either in text form with no audio or in audio form with no text. This means hiring people to read and record text or transcribe audio recordings — a time-consuming and expensive process. The lengthy recordings or transcripts that result are usually difficult to index and search. Brinklow and his CAC collaborators plan to start by creating a data management app that volunteers can use to input their work within set parameters that make it possible to more quickly digitize and index the data.

> "Ultimately, the database will be a valuable resource that can be leveraged to develop text-to-speech or speech-to-text applications, and to perform linguistic analyses, such as word frequency, which will aid in building teaching and learning tools."

Given a sufficient quantity of well managed voice and text data, machine learning can be applied to match sounds with written symbols and develop speech recognition algorithms that can then be used to automate the transcription process. Ultimately, the database will be a valuable resource that can be leveraged to develop text-to-speech or speech-to-text applications, and to perform linguistic analyses, such as word frequency, which will aid in building teaching and learning tools.

CAC's role in the project involves providing expert support around data management and security, as well as archival data storage. CAC will also be providing computing cluster support once development of the speech recognition algorithm begins. "CAC is also helping us to solve for policy issues related to Indigenous data sovereignty," Brinklow noted. CAC is helping to enable the data to be securely stored locally at the Tyendinaga Language and Culture Centre, rather than on an open cloud platform.

The project slowed to a pause during the pandemic but is restarting now, with work on data management and planning for new app-based data collection.