The Indigenous Languages Technology (ILT) project at the National Research Council of Canada, and its context

Kuhn, Roland

Linguapax Review 2021, 9, pp. 85-104, 2021

THE INDIGENOUS LANGUAGES TECHNOLOGY (ILT) PROJECT AT THE NATIONAL RESEARCH COUNCIL OF CANADA, AND ITS CONTEXT

Roland Kuhn

INTRODUCTION

Recent decades have seen the creation of hundreds of Indigenous language revitalization and reclamation projects across Canada. Most are located inside Indigenous communities; a few are centred in universities. They range in scale from the ambitious, multi-language efforts of the First Peoples' Cultural Council (FPCC), which is located in and funded by the Province of British Columbia and is active there and elsewhere in Canada (https://fpcc.ca/), to unfunded one- or two-person volunteer efforts in remote communities.

The project that I have the honour to lead described in this paper – and on this periodically updated website: https://nrc.canada.ca/en/research-development/research-collaboration/programs/canadian-indigenous-languages-technology-project – is managed inside the National Research Council of Canada (NRC), which is the primary research and technology organization of the Government of Canada. However, the reader should not conclude that "settler" (non-Indigenous) governments play a key role in Indigenous language revitalization in Canada. All successful language revitalization projects that I am aware of are Indigenous-run or collaborate closely with Indigenous language activists.

Our Indigenous Languages Technology (ILT) project at NRC (henceforth, "NRC-ILT") is **not** itself a language revitalization project: it builds tools that people working on language revitalization sometimes find useful. We – the NRC-ILT team – are like lighting technicians or stagehands in a theatre: we are not the actors (we do not appear on stage), nor are we in charge of the production. Those roles are filled by Indigenous language activists and their communities. Language revitalization would go on without us. At times, however, our technical help may make it go slightly better. Everything we have accomplished has been done in collaboration with Indigenous stakeholders.

Disclaimer: this paper is one person's partial, subjective perspective on the technologies being applied to language revitalization in Canada, along with a description of the NRC-ILT project. It is by no means an exhaustive survey of the field; it leaves out many, probably most, organizations active in language revitalization, and their accomplishments. This article is an extended version of (Kuhn et al., 2020), which means it is deeply indebted to the contributions of my co-authors on that paper.

INDIGENOUS LANGUAGES IN CANADA: HISTORY AND DEMOGRAPHICS

Figure 1 shows the range of Indigenous language families in what was later to become Canada, at the time of first contact with Europeans. Each family was and is different from the others, in terms of phonetics, vocabulary, and syntax – in some cases, as different as the Germanic, Turkic, and Sino-Tibetan families in Eurasia are from each other. Languages that were geographical neighbours, even if from different families, sometimes borrowed words and aspects of phonetics or syntax from each other.

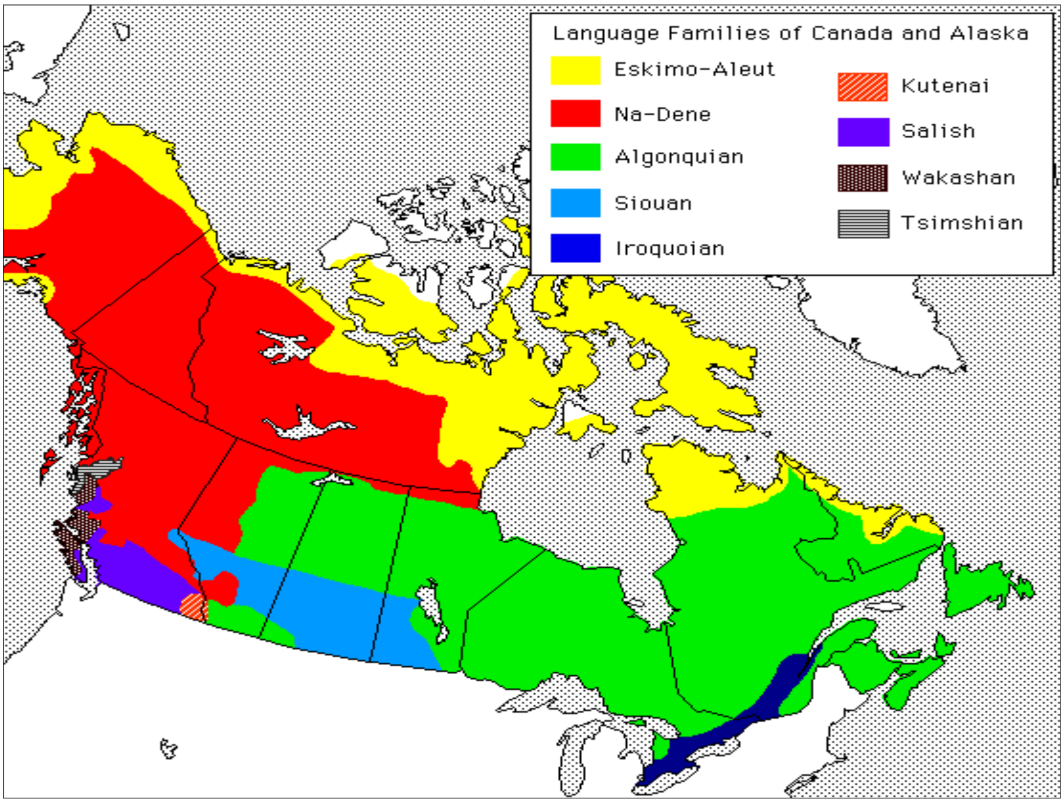

Figure 1. Indigenous languages in Canada during first contact with Europeans.
(Map from Matthew Dryer, Buffalo University)

The terminology for Indigenous communities in Canada often employs three labels: Inuit, Métis, and First Nations. The Inuit are a cohesive ethnic group, as are the Métis. However, "First Nations" groups together all other Indigenous peoples in Canada, a diverse set of ethnicities.

Most Indigenous languages spoken in Canada are polysynthetic: a single word often expresses a complex meaning that would require a full sentence in non-polysynthetic languages. A word is also typically made up of more morphemes than in other languages. An average English word has about 3 morphemes; an average Mohawk word has in the range 5-6. Mohawk (Iroquoian language family), Cree (Algonquian family), and Inuktitut (Eskimo-Aleut family) are all polysynthetic, though they are from otherwise dissimilar families.

After contact between Indigenous peoples and Europeans, creoles with aspects of both Indigenous and European languages arose; most have since disappeared. For instance, contacts between Dutch traders and Mohawks in the Hudson valley gave rise to a Dutch-Mohawk creole. The Chinook Jargon creole in its final form (it survived into the 20th century) had elements from West Coast Indigenous languages, Russian, Hawaiian, and Chinese.

The Métis emerged as a distinct nation of mixed Indigenous and European ancestry in the early 19th century (Bakker, 1997; Rosen & Souter, 2009). Bungi, a creole spoken by the Scottish Red River Métis in present-day Manitoba, had elements from Orkney Scottish English, Scottish Gaelic, French, Cree and Anishinaabemowin (Ojibwe). However, most of the Métis people spoke Michif,

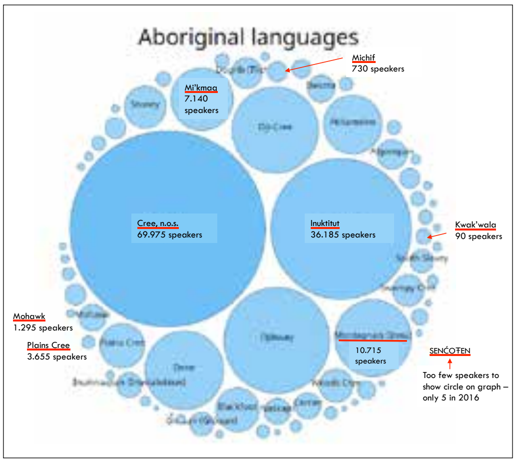

Figure 2. Number of speakers by language (2016 census).
https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/dv-vd/lang/index-eng.cfm

a contact language. Michif survives and is the focus of strong language revitalization efforts. Roughly, it may be viewed as combining mostly French nouns with verbs from Plains Cree (from the Algonquian family); it is mostly polysynthetic (Davies, Santos, and Souter, 2021; Sammons, 2019).

After Canadian Confederation in 1867, Indigenous languages were targeted by government policies that sought to eradicate them and traditional Indigenous customs, with the goal of assimilating Indigenous people into "white" society. The Indian Act discouraged, and often made illegal, gathering for cultural practices and speaking ancestral languages. Many Indigenous children were forcibly removed from their communities and placed in compulsory boarding schools ("residential schools") by the federal government. These schools were run by churches; some by the Catholic Church, others by Protestant denominations. Multiple forms of psychological, physical and sexual abuse that occurred in these schools are well-documented. According to the Truth and Reconciliation Commission of Canada, the residential school system was "created for the purpose of separating Aboriginal children from their families, in order to minimize and weaken family ties and cultural linkages" (Government of Canada (2015), preface). Furthermore, during a period in the mid-20th century, many Indigenous children were taken from their birth families so that they could be adopted by non-Indigenous families, in the notorious Sixties Scoop (Fachinger, 2019).

Despite this history of systematic discrimination, Indigenous communities have resisted assimilation and persisted in preserving and teaching their languages. The benefits of learning an ancestral language have been correlated with improved well-being in youth (Chandler and Lalonde, 1998; Hallett et al., 2007).

On the basis of personal observation in the course of the NRC-ILT project, I believe that many Indigenous languages spoken in Canada are experiencing a renaissance. Young people in Indigenous communities – who often have grown up speaking only English or (in Quebec) French – are enrolling in record numbers in courses that will enable them to speak with each other in their ancestral languages. I've met some extraordinary Indigenous individuals – almost all poorly paid, and some unpaid volunteers – who teach these languages. These language activists form an elite: a self-selected group of people who combine idealism with high intelligence and a capacity for hard work. Teaching a polysynthetic language to people who originally spoke only English or French is not for the faint of heart.

Figure 2 shows the number of speakers of each Indigenous language in 2016, according to the Canadian census (Statistics Canada, 2017). Languages underlined in red are those the ILT-NRC project has interacted with in some way. The numbers shown should be taken with a grain of salt; experts disagree with some of them. However, this figure gives a rough sense of the demographic weights of these languages. Plains Cree (Algonquian family) has the most speakers; Inuktitut (Eskimo-Aleut family) ranks second. Inuktitut is, uniquely, an important government language (in Nunavut Territory). The existence of a government bureaucracy that partly functions in Inuktitut means that far more text and speech resources are available for this language than for any other Indigenous language in Canada. The disparity with Cree, which has far more speakers but far fewer linguistic resources, because no large bureaucracy uses it, is striking.

The reader should not conclude that most of the "mid-sized" or "small" languages shown have dismal long-term prospects. The figure cannot show the relative vigour of revitalization efforts for each language. For instance, the smallest language shown, SENĆOŦEN, which only had 5 fluent, elderly speakers in 2016, has a vigorous, well-managed language revitalization program that is teaching many young people in the community the language – the next census will count far more SENĆOŦEN speakers (Montler, 2018; First Voices Portal, 2021).

Similarly, the Mohawk language (Kanyen'kéha), with a modest 1,295 speakers in **Figure 2**, is experiencing strong growth, thanks to good language schools in several Mohawk communities. The Kanyen'kéha school for adults at the Six Nations of the Grand River community in southwestern Ontario (Onkwawenna Kentyohkwa; see https://onkwawenna.info) is admired by Indigenous educators across Canada for its success in graduating many fluent speakers after these students have undergone two years of intensive immersion (the language had almost died out at Six Nations, though not in all Mohawk communities). Some graduates are marrying each other, speak Kanyen'kéha at home, and are bringing up children for whom it is their first language. Energy, teaching expertise, and community cohesion have a large positive impact on language revitalization – they may be better predictors of a language's future than the current number of fluent speakers.

TECHNOLOGIES FOR LANGUAGE REVITALIZATION

"Indigenous people must take the lead in developing the next wave of responsive and responsible language technologies … Fortunately, Indigenous people around the world are exercising technological leadership as they claim spaces for their languages in the digital realm" (Brinklow, 2021).

Indigenous languages spoken in Canada are **heterogeneous**, in their linguistic properties (10 unrelated language families) and in the number of speakers per language. Indigenous language activists in Canada are keenly aware of the remarkable achievements of their peers in Aotearoa – i.e., New Zealand, where Māori has been an official language since 1987 (King, 2013) – and in Hawai'i (Asensio, 2019). However, language revitalization in Canada is an even more complex challenge. Each language here has different needs, ranging from those of severely endangered languages – where the priority tends to be recording a small number of Elders who still speak the language fluently – to those of Inuktitut, a language of government in Nunavut that is widely spoken, where the priority is to encourage its wider use (e.g., in medical settings) and to improve education in the language.

Two recent articles (Brinklow et al. 2019, Brinklow 2021) discuss language technologies for revitalizing and reclaiming Indigenous languages. Quoting the linguist Francis Tyers, Brinklow points out that because big software companies are profit-hungry, commercial technologies focus on the needs of "small numbers of rich people, or large numbers of poor people." Indigenous languages in Canada are spoken by small numbers (by world standards) of relatively poor people – they don't matter much to big companies. Insofar as companies are interested in Indigenous languages, they often want to retain IP in linguistic data, thus threatening Indigenous data sovereignty – Keegan (2019) gives an example from Aotearoa that illustrates this danger.

There are also technical issues with commercial language technologies:

- None of the "big" languages (English, Chinese, Japanese, Arabic, etc.) for which these technologies have been developed are polysynthetic. Standard technologies applied to polysynthetic languages run into difficulties – for instance, the "out of vocabulary" problem. Even with a vocabulary for automatic speech recognition (ASR) of 1.3 million Inuktitut words, Gupta and Boulianne (2020a) found that more than 60% of the words in held-out Inuktitut stories were **not** in that vocabulary (see below). In a comparable experiment carried out with English texts, one would expect this "out of vocabulary" rate to be much, much lower than 60% – almost certainly less than 5%, possibly lower than 1%.
- Indigenous languages in Canada lack the massive quantities of training data on which machine learning relies. NRC-ILT team members have become familiar with questions from technically unsophisticated, non-Indigenous people like: "Why don't you create software for translating between Cree (or another Indigenous language) and English?"

Our two-fold response:

1. Apart from the Inuit, no Indigenous community we've spoken has shown much interest in machine translation (MT) between their ancestral language and English (or French, in Quebec). Communities are typically more interested in tools to encourage learning and use of their ancestral language.
2. The majority of Indigenous languages in Canada have little parallel data – often, it's limited to a few books of the Bible translated by missionaries. The existence of dialects in most of these languages exacerbates this data sparsity problem. (For Cree, our team and our Indigenous collaborators **might** be able to scrape together at most 100,000 Plains Cree sentences in parallel with English – not nearly enough to train an MT system).

We encounter misunderstanding of this situation even among natural language processing experts. Some experts have hinted that we are technically backward because we use rule-based approaches in most of our work. After all, other teams have successfully applied machine learning to low-resource languages … The term "low-resource language" is a barrier to understanding. It is applied to languages like Tamil, with 75 million speakers, most of them literate in the language, and a history of written texts that goes back thousands of years. It is ridiculous to use the same term to describe the "biggest" Indigenous language in Canada, Plains Cree, with 75,000 speakers and few written texts. Most NRC-ILT subprojects deal with "**extremely** low-resource" languages.

Inuktitut is an exception to this "extremely low-resource" situation (see below).

Littell, Kazantseva et al. (2018) outline technologies for language revitalization in Canada. In the list below taken from this paper, technologies in **bold** were created by NRC-ILT with Indigenous collaborators; underlined technologies were created outside NRC with help from NRC-ILT funding:

- **Implementing digital fonts and layouts for Indigenous languages with uncommon character sets;**
- **Predictive text;**
- **Conversion between different orthographies for the same language;**
- Spell-checking *(no NRC-ILT involvement);*
- **Paradigm generation for teaching complex morphology;**
- **Approximate search;**
- **Machine translation;**
- Speech technologies:
  - · Automatic speech recognition;
  - · Keyword search through speech data;
  - · **Speech-text alignment;**
  - · **Text to speech** *(preliminary NRC-ILT work)*;
- Image technologies;
- Computer-aided language learning (CALL).

Littell, Kazantseva, et al. (2018) allude briefly to four other important technological activities:

- Building tools for recording and annotating the speech of fluent Elders
- Digitizing old magnetic and audio tapes of Elders' speech
- Curating speech archives
- Making digital dictionaries.

HISTORY AND COLLABORATIVE APPROACH OF THE PROJECT

An academic overview of NRC-ILT is (Kuhn, Davis, Désilets, Joanis et al., 2020); a detailed technical report is at https://nrc-publications.canada.ca/eng/view/object/?id=d4f10144-c711-43c5-b80b-5ace7df5e68b. Regular updates are at https://nrc.canada.ca/en/research-development/research-collaboration/programs/canadian-indigenous-languages-technology-project.

The project began with the March 2017 budget of the Government of Canada, which provided $6 million (CAD) to NRC to develop, in collaboration with Indigenous stakeholders, software technologies to support Indigenous languages. Our part of NRC was thought to be qualified for this work because of experience with technologies for major world languages – e.g., we had worked on MT from Chinese and Arabic into English, funded by US DARPA's GALE and BOLT programs. There had also been a small-scale NRC initiative focused on Inuktitut during 2003-2012 (Martin et al., 2003; Farley, 2012).

The NRC-ILT team tries to fulfill its mandate by developing close, respectful relationships with Indigenous communities and trying to break with the long, painful history of extractive research practices (Keegan, 2019; Brinklow et al., 2019). We believe in the "empowerment" philosophy, whereby research is carried out collaboratively, with equal emphasis on the agenda of the researcher and of the community (Czaykowska-Higgins, 2009).

This has meant asking Indigenous language activists which software tools would be useful to **them**, rather than offering technologies based on intriguing research themes. Hence, the rather disjointed collection of technologies described below: we've responded to different needs from different communities. We were guided by an Advisory Committee made up of Indigenous language revitalization experts. Their counsel has been invaluable. NRC-ILT never claimed ownership of Indigenous language data collected with project funding. Where possible, we implemented a "kiln and pottery" model: language-independent open-source software (OSS) (the kiln) creates language-specific software (the pots) which will be owned by the community.

Our initial funding was more generous than anticipated, so we funded work outside as well as inside NRC. ILT-NRC has three "rings":

Owennatékha (Brian Maracle): founder & head, Onkwawenna Kentyohwa immersion school.Suggested **Kawennón:nis,** & collaborated in its creation.

Akwiratékha Martin: language instructor in Kahnawá:ke – collaborating with NRC-ILT team to create Eastern version of **Kawennón:nis.**

Kawennón:nis - verb conjugator for Kanyen'kéha

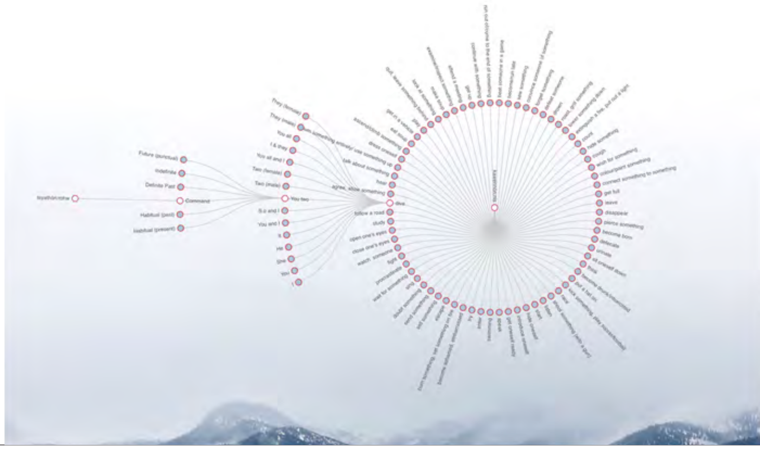

Figure 3. Collaborations on Kanyen'kéha (Mohawk).

- An inner ring of work carried out at NRC in collaboration with Indigenous stakeholders.
- A middle ring of work related to automatic speech recognition (ASR) carried out by the Centre de Recherche Informatique de Montréal (CRIM). CRIM has an illustrious record of ground-breaking research in ASR and related technologies. CRIM delivered experimental ASR results for two important Indigenous languages, and also OSS that speeds up a key stage in annotating recorded speech.
- An outer ring of other subprojects carefully selected by the ILT-NRC team and its Indigenous Advisory Committee for funding. Most of these subprojects were managed by Indigenous organizations; all employed an almost entirely Indigenous work force.

INNER RING: TECHNOLOGIES DEVELOPED INSIDE NRC

VERB CONJUGATORS

The first ILT-NRC subproject arose with a proposal from Owennatékha (Brian Maracle), the director of the Onkwawenna Kentyohkwa (Our Language Society) school in Six Nations of the Grand River in southwestern Ontario: https://onkwawenna.info/. This school is renowned for producing fluent speakers of Kanyen'kéha (Mohawk). Owennatékha suggested that the NRC-ILT team create Kawennón:nis (the WordMaker) – a tool that helps students master the complex verbal system of Kanyen'kéha. There are so many possible verb conjugations for even the most common verb stems in Kanyen'kéha that a physical reference tool for them would be impossibly large; however, it's feasible to compute these conjugations in software (Kazantseva et al, 2018). Kawennón:nis, which is implemented for the Western dialect of Kanyen'kéha, underwent extensive user testing with students and teachers at the school before being released. It has been received enthusiastically by them. It is designed to match the curriculum, and uses the visual metaphors employed by the school's teachers: https://kawennonnis.ca/wordmaker.

Subsequently, ILT-NRC built general frameworks for building verb conjugators for polysynthetic languages. These are being applied to create verb conjugators for Eastern Kanyen'kéha spoken in Kahnawá:ke (Quebec), and for two languages unrelated to Kanyen'kéha: Algonquin Anishinaabemowin as spoken at Kitigan Zibi (near Maniwaki, Quebec), and Michif.

Figure 3 shows two of our Mohawk collaborators. Both teach Kanyen'kéha: Owennatékha teaches the Western dialect, and Akwiratékha the Eastern dialect. (Other Mohawk educators also helped with building and testing of Kawennón:nis). The figure also shows the Kawennón:nis user interface.

Heather Souter of the Prairies to Woodlands Indigenous Language Revitalization Circle is currently encouraging students in her courses to test a beta version of the Michif verb conjugator. Related work is described in (Davies, Santos, and Souter, 2021).

We plan to roll out verb conjugators for several other languages. Furthermore, Dr. Patrick Littell is leading an effort to develop a user-friendly, spreadsheet-based framework called "Gramble" that may make it far easier for language activists to build conjugators **on their own** (without NRC expertise).

READALONG STUDIO

This is the "surprise hit" of NRC-ILT: we didn't anticipate how many teachers would be interested in having us add a key functionality to pre-existing audio books and videos in their languages. This functionality was pioneered by Prof. Marie-Odile Junker's team at Carleton University.

The functionality is illustrated in **Figure 4**; it is very simple. As an audio book or video with Indigenous speech (here, Atikamekw) is played, the word being spoken is highlighted in accompanying text. In the figure, the word "atisokasotcik" is being spoken, and is thus highlighted. If the listener – a student or teacher – wishes to focus on pronunciation of a word in the text, they click on it, and the pronunciation will be played back to them; they can slow the playback down. We (the Carleton U. and NRC-ILT teams) call audio books or videos with this functionality "ReadAlongs".

Many communities have educational books or videos with speech in their languages, with a transcription into text of the spoken (or sung) words. To turn these books or videos into ReadAlongs, words in the speech must be aligned with words in the text. The Carleton team was aligning ReadAlongs by hand when we began to collaborate with them. Our contribution was to automate this process, using software we developed called "ReadAlong Studio". Delasie Torkornoo of the Carleton team has been helping us improve the ReadAlong Studio code. Our two teams have produced and sent back to Indigenous educators ReadAlongs in the Algonquin, Atikamekw, Southern East Cree, Northern East Cree, Gitksan, Inuktitut, Kwak'wala, Kanyen'kéha (Mohawk), Seneca, and SENĆOŦEN languages. Every couple of weeks, we receive new requests to transform teaching materials in this manner. Indigenous

Figure 4. READALONG STUDIO – A SURPRISE HIT!

educators tell us that the ReadAlongs have high pedagogical value.

KEYBOARDS, ORTHOGRAPHY CONVERSION, & PREDICTIVE TEXT

NRC-ILT has released OSS that implements keyboards for some poorly served writing systems, and that converts between writing systems for some languages. Eddie Antonio Santos of the team worked with Saulteaux, Makah, and Plains Cree. Along with University of Alberta linguist Arok Wolvengrey, he also persuaded Google to change its syllabic keyboard for Plains Cree on Chromebooks (the previous version was appropriate for East Cree, but not Western Cree).

NRC-ILT has also released code for text prediction for mobile devices. In theory, the software can implement text prediction for any language. However, it is difficult for non-experts to use. So far, we have only implemented text prediction for the SENĆOŦEN language. However, since SENĆOŦEN orthography is difficult, members of that community have told us they are very happy with this functionality – it makes entering text on their devices much easier.

INUKTUT

All the work in this section benefited from collaboration with the Pirurvik Centre, an Inuit-run culture and language centre (https://www.pirurvik.ca/).

The related languages or dialects spoken by Inuit people in Nunavut are called collectively "Inuktut". Most NRC-ILT work on Inuktut has involved its Inuktitut version, spoken on Baffin Island. This work differs from our work on other languages in two ways: 1. We didn't build educational tools, but **office tools that assist writing or reading Inuktut text** (or translating between Inuktut and English); 2. We worked on **machine translation (MT)** between Inuktut and English, using machine learning. These differences are because there exist a bureaucracy and a legislative body that function partly in Inuktut: these organizations require text to be written in Inuktut, and supply a growing amount of bilingual Inuktut-English text that can be used to train MT systems.

A small-scale effort at NRC during 2003-2012 to build software tools for Inuktut yielded valuable results – most importantly, a morphological analyzer. A key contributor was Benoît Farley, who subsequently retired from NRC but who maintains a repository for these tools (Farley, 2012).

Another NRC output during this period was successive releases of parallel English-Inuktut text corpora based on proceedings of the Nunavut Legislative Assembly, the "Nunavut Hansard" (henceforth, "NH"). The first of these, NH 1.0, was released to the research community in 2003 (Martin et al., 2003); it covered 155 days of proceedings of the Nunavut Assembly, and comprised 3,432,212 English tokens and 1,586,423 Inuktitut tokens.

When NRC-ILT restarted work on Inuktut in 2017, many more years of unaligned NH proceedings were available. We **created and released a new aligned corpus**, NH 3.0. It comprises 8,068,977 Inuktitut tokens and 17,330,271 English tokens; it covers the proceedings of 687 debate days from April 1, 1999 to June 8, 2017 (Joanis et al., 2020).

The size of NH 3.0 makes it feasible to train **machine translation** (MT) systems. We helped the organizers of WMT 2020, an annual, international evaluation of MT systems, to set up evaluations for MT between English and Inuktitut in both directions. We also (along with Microsoft) paid for human evaluators fluent in Inuktitut to assess translations of test data into Inuktitut from the competing systems. The extreme polysynthetic nature of Inuktitut, unusual among the world's languages, attracted many research teams to this MT challenge.

We built our own MT systems that competed in this evaluation (Knowles et al., 2020). The overall results are here: http://www.statmt.org/wmt20/translation-task.html. We hope that continuing interest in MT research for this language pair will eventually yield practical tools for Inuktitut.

We continue to improve the **office tools for Inuktitut** that NRC released in 2003-2012, and to build new ones:

- we are improving Benoît Farley's morphological analyzer.
- we are reviving WeBInuk, a web-based NRC tool that searches for Inuktitut translations of English words. It displays parallel English-Inuktitut sentences where the English sentence contains the search word. WeBInuk was heavily used by Inuktut translators in 2007-2017, then allowed to disappear (NRC management had other priorities).
- we are building a lemmatizer that inputs an Inuktitut word and outputs the most frequent word with the same root. This is a key step in supporting monolingual search in Inuktitut, or search in the Inuktitut —> English direction. It is much harder to build a search engine for users who enter search terms in Inuktitut than for English or French, because of polysynthesis. (Whatever Inuktitut word the user enters has probably never existed before in the history of the language, and will probably never exist again). Fuzzy search is necessary. **Figure 5** illustrates progress we've made on this and related tools.

PRELIMINARY TEXT TO SPEECH (TTS)

The curriculum at Onkwawenna Kentyohkwa focuses on mastering the complex ver-

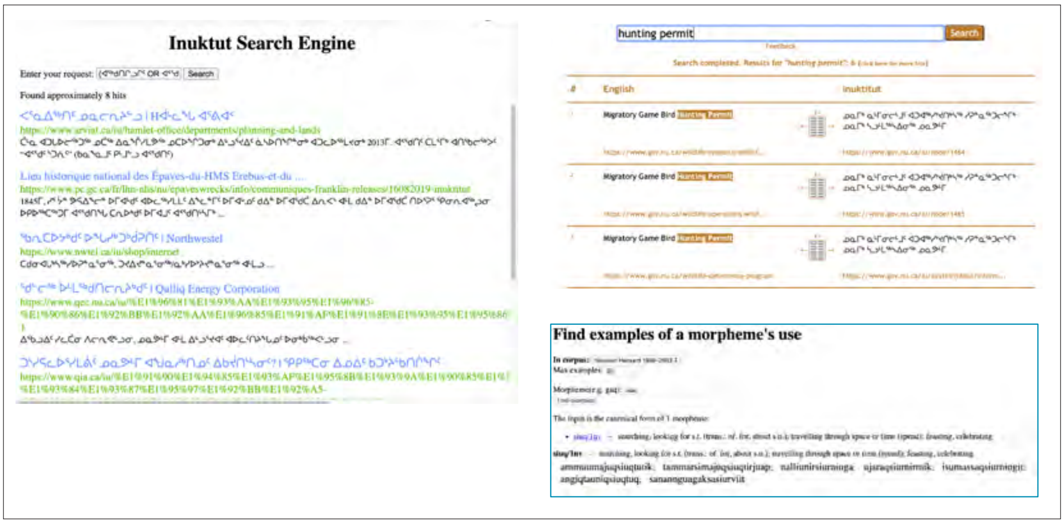

Figure 5. Inuktut Search Engine, WeBInuk (Concordancer), and Morpheme Example Search.

bal morphology of Kanyen'kéha. This is why NRC-ILT and the school jointly developed Kawennón:nis (above).

Kawennón:nis has one limitation: students can generate complex words in text, but do not necessarily know how to pronounce them. The goal of the school is to produce fluent *speakers*, not just *readers/writers*, so there is a gap between what Kawennón:nis produces (text) and what students need most (speech to imitate). Yet just as it would not be feasible to compose and print Kawennón:nis in book form because of its enormous implicit vocabulary, it would likewise not be feasible to **record** all forms. This suggests implementation of TTS – can we train a TTS system on a recorded, phonetically balanced subset of the vocabulary?

We built a concatenative TTS prototype. It rearranges subword units from 852 recordings of conjugations from fluent speakers into complete coverage of the first 122,966 conjugations in Kawennón:nis. The teachers at the school declared the quality of the output speech to be better than acceptable. We have also recently developed a neural TTS model capable of producing acceptable-quality synthesized speech from as few as 30 minutes of audio. Owennatékha recently declared TTS to be in his top three technological priorities for Kanyen'kéha.

Kawennón:nis illustrates the challenges facing Indigenous curriculum developers: complex languages, few speakers to teach the language, few recorded materials, and few speakers to record new materials, and thus large parts of the language that students never get to hear. TTS enables resource-strapped organizations to make the most of their most precious, limited resource – fluent speakers' time – by leveraging a small number of recordings to cover a much larger domain.

MIDDLE RING: SPEECH TECHNOLOGIES DEVELOPED BY CRIM

Annotation and transcription are the biggest part of the workload in most language documentation and conservation projects. Cox et al. (2019) discuss the "transcription bottleneck": speech (especially of Elders) is being recorded in Indigenous languages much faster than it is being transcribed, or even annotated (it should be called the "annotation bottleneck" instead). This is true of minority languages across the world. Annotation and especially transcription take time, and are tedious. It is tempting to record many hours of speech and plan to carry out these tedious steps "later on" – they are often postponed indefinitely …

This has been going on for decades, long before digital recording media were available. All over Canada, there are thousands of hours of speech in Indigenous languages on magnetic tapes or audio cassettes that are potentially valuable resources for language revitalization, but that have become inaccessible in practice because nobody knows what is in them. They are scattered across communities and university departments (I have heard stories of magnetic tapes with important recordings being found, unlabeled, on the shelves of a deceased professor's garage). Nobody has the time to listen to most such recordings; in some tragic cases, nobody is left who still understands the language that was recorded. Older recordings have typically never been copied over into a digital medium.

CORE RESEARCH

We funded the Centre de recherche informatique de Montréal (CRIM) to research automatic speech recognition (ASR) for a small number of Indigenous languages spoken in Canada. Full-vocabulary ASR for such languages is unrealistic in the short term, because the large amounts of parallel speech-transcription data required to train ASR systems do not exist. There was not enough funding available to create such data. Thus, the goal was to create imperfect ASR systems trained on small amounts of parallel data. Such systems have been shown, for other languages, to make "audio search" (also termed "keyword spotting" or "spoken term detection") feasible. When ASR accuracy is low, it may still be possible for users to find certain words or phrases in speech with audio search techniques.

ASR experiments at CRIM have focused on Inuktitut and East Cree. The CRIM researchers arranged for nearly 81 hours of Inuktitut speech, and 102 hours of East Cree speech, to be transcribed; they split these parallel data into training and test sets. Their subsequent work on Inuktitut and East Cree ASR is described in (Gupta and Boulianne, 2020a) and (Gupta and Boulianne, 2020b), respectively. The work benefited from collaboration with Pirurvik Centre (https://www.pirurvik.ca/) and the Canadian Broadcasting Corporation (https://www.cbc.ca/).

A key question for the CRIM team was what units to use for ASR. Inuktitut is highly polysynthetic; thus, the language has a very high "out of vocabulary" rate (60%) – many words are only ever spoken once. An ASR system whose vocabulary is made of words will misrecognize most words in new Inuktitut speech data. The CRIM researchers tried morphemes, syllables and phonemes as the basic ASR unit, and obtained the best results with syllables. But even with syllable units, about 71% of Inuktitut words in the test data were misrecognized.

In unpublished work, these researchers nevertheless managed to get reasonably good keyword-spotting performance for Inuktitut *(personal communication)*. The best-performing approach was conventional: search through a phoneme lattice generated by syllable-based ASR, with a weighted confusion lattice providing robustness to ASR errors.

In CRIM experiments, East Cree had a significantly lower "out of vocabulary" rate than Inuktitut, though much smaller amounts of textual data were available for building an East

Cree vocabulary. With a 30,000-word vocabulary obtained from text in two genres – transcripts of video stories and Bible translations – the “out of vocabulary” rate for held-out text data was 25% for video stories and 9% for Bible data. These numbers made words a plausible unit for East Cree ASR. In ASR experiments on held-out East Cree speech data, 70% of words in the test video stories were misrecognized, and 25% of the words in the test Bible data. No keyword-spotting experiments were conducted for East Cree.

The CRIM researchers obtained extremely promising experimental results (low rates of misrecognition) for both Inuktitut and East Cree when the ASR system was speaker-**dependent**. The poor results quoted above are speaker-**independent.** For Indigenous language revitalization, however, where many hours of speech are often collected from each of a small number of fluent Elders, speaker-**dependent** ASR would be practically useful. Consider a mode of work in which (e.g.) a couple of hours of speech from an Elder are transcribed by hand by a human expert, as is done now. Then these two transcribed hours would be used to train an ASR system adapted to the speech of this particular Elder, which could produce a first-draft transcription of the many remaining hours of speech from the same Elder. This would alleviate the “transcription bottleneck”.

PRODUCTIVITY TOOLS

The CRIM team also released tools that make early stages of processing recorded speech easier. These were packaged as Web services on CRIM’s VESTA platform (http://vesta.crim.ca) and are available there or through an ELAN extension – ELAN is an annotation tool at https://archive.mpi.nl/tla/elan/ (Wittenburg et al., 2006). The tools enable segmenting speech files into speech vs. non-speech (silence, noise, music, etc), language retrieval (finding segments spoken in a particular language, with 32 languages being identifiable), speaker retrieval (finding segments spoken by a particular speaker), multichannel voice activity detection (detecting segments containing speech separately for each track in a multichannel, multi-microphone recording), and other useful capabilities.

Drs. Chris Cox and Olivia Sammons (of Carleton University and First Nations University respectively), used these CRIM tools recently while collecting Michif speech data. They report a 4-to-5-fold speedup for early stages of annotation *(personal communication)*.

OUTER RING: OTHER WORK FUNDED BY NRC-ILT

This section outlines Indigenous language revitalization subprojects to which NRC-ILT funding contributed – at least partially – but which were managed outside NRC.

ONLINE COURSES AND GAMES

Online language learning tools must be updated frequently because the underlying software has become obsolete. For instance, Adobe stopped supporting Flash Player in early 2021. NRC-ILT funded remedial measures for several subprojects.

Learning Platforms at Carleton University:
Prof. Marie-Odile Junker and her Carleton University team have been collaborating for years with Indigenous partners to develop online lessons for Algonquian languages. She uses a participatory-action framework to work with communities. Her team’s contributions are too numerous to list here – see https://www.marieodile-junker.ca/. Prof. Junker holds the Governor General’s Innovation Award (2017).

NRC-ILT funding began to flow to the pre-existing Carleton project in 2018. Software changes had partially stranded educational language tools on Carleton's web-based platform for Innu (not to be confused with Inuktut, an unrelated language) and East Cree. The funding helped pay for a technological update (removal of dependence on Flash and other outdated technologies) and for the expertise of two Indigenous language experts who contributed large amounts of new pedagogical content. Prof. Junker's team continues to develop the platform with other funding (Innu: https://lessons.innu-aimun.ca/; East Cree: https://lessons.eastcree.org/).

7000 Languages:
Some Indigenous communities wished to develop online courses for their languages. We provided modest amounts of funding to 7000 Languages, a non-profit that creates free language learning software, to jointly create courses with these communities (whom we also funded). Three of these courses are online at the 7000 Languages site, www.7000.org: for Kwak'wala, Michif, and Mi'kmaq.

Computer-assisted Language Learning (CALL) at the University of Alberta:
This subproject supports the Y-dialect of Cree (Plains Cree). The U. Alberta team, together with members of Cree-speaking communities, has been creating an adaptive CALL system called "CreeTutor". Cree-language content has also been developed: 13 personal stories from 8 Elders have been recorded, transcribed, and translated. CreeTutor will soon go live. A key element of CreeTutor is *Sound Hunters,* a receptive phonemic awareness activity (Lothian et al., 2020). The main creator of *Sound Hunters,* Delaney Lothian (Cree-Métis), is a U. Alberta MSc student and a part-time NRC-ILT member. She is working on a Michif version of *Sound Hunters*.

FirstVoices:
The First Peoples' Cultural Council (FPCC) of British Columbia has a strong record of providing state-of-the-art technologies, training and technical support to Indigenous language activists (mainly but not exclusively in the province), within its FirstVoices program (http://www.fpcc.ca/about-us/). The Language Tutor, which enables communities to build language lessons, is part of FirstVoices and had become technologically "stranded". ILT-NRC funding helped to redress this situation.

On the Path of the Elders:
"On the Path of the Elders" is a role-playing game designed to acquaint players with historical and cultural topics related to the James Bay region. When this free online game was released in 2007, it was widely praised for its innovative use of historical resources. It suffered loss of functionality due to changes in the underlying software industry. NRC-ILT supplied funding to update the software and add Swampy Cree content. The "On the Path of the Elders" website is now mobile-friendly and compatible with Chrome, FireFox and Internet Explorer browsers: www.pathoftheelders.com.

CAPACITY-BUILDING WORKSHOPS

Two subprojects funded by NRC-ILT provided training for people who document Indigenous languages. NRC-ILT also provided modest financial assistance to the language activist Caroline Running Wolf; she organized four successful online sessions for language activists, the "Community Workshops for Indigenous Language Technology" (CWILTs) in Dec. 2020 – July 2021.

Yukon Native Language Centre (YNLC):
Eight Indigenous languages (Gwich'in, Hän, Kaska, Northern Tutchone, Southern Tutchone, Tagish, Tlingit, and Upper Tanana) are spoken in the Yukon: almost 15% of Indig-

enous languages in Canada. Helped by funding from NRC-ILT, YNLC supported 12 trainees in a 10-month program during which they acquired practical skills in developing, disseminating, and preserving digital materials for these languages in their communities. The Covid-19 pandemic disrupted some of the planned activities in this subproject, which was centred on in-person activities, but competent management by YNLC staff made it very successful nevertheless.

- In Workshop 1 (October 2019), trainees practiced how to use the video equipment and how to carry out interviews with fluent speakers. They were also trained in file management.
- In Workshop 2 (November 2019), attendees were trained to use the ELAN and SayMore software suites for transcribing, annotating, and translating speech recordings.
- Workshop 3 (February 2020), began with review and practice of skills learned earlier. It then covered sharing and repurposing of videos, and using ELAN materials as a resource for language learning and teaching. On the last day, students showed their finished videos.

The 12 trainees created 548 minutes of documentation and mastered skills that will enable them to record Elders in their communities on a continuing basis.

Indigitization:
This is a partnership between the University of British Columbia, the Musqueam Archives, and the Heiltsuk Cultural Education Centre that focuses on digitization of speech data. While the YNLC subproject trained members of Indigenous communities to collect **new** data, the continuing Indigitization subproject trains people to convert **old** data into digital, accessible formats. It helps make Indigenous communities autonomous by familiarizing them with hardware and software options for digitization. For communities that already have digitized content, the subproject offers training in transcription of speech into Indigenous languages, translation of the transcriptions into English or French, and management of the resulting archives.

Indigitization began in January 2020, and underwent re-planning because of the COVID-19 crisis (in-person workshops became inadvisable). More resources were devoted to creation of instructional resources than in the original plan: "We have a unique opportunity to write experience-based, practical guides that will help small, often underfunded, Indigenous organizations to start managing their collections in a structured manner" *(personal communication from Gerry Lawson, Indigitization project lead)*.

The text content for 65 of these new guides has been developed; the guides are currently undergoing editing and graphic design. This subproject plans to lend hardware to communities when necessary.

CWILTS:
CWILTs ("Community Workshops for Indigenous Language Technology") is a series of video sessions for Indigenous language activists organized by Caroline Running Wolf of Apsáalooke Nation (Crow), each featuring lightning talks centered on a language revitalization theme followed by conversations in breakout rooms. Her goal is to build a community of practice. Many CWILTS attendees have kept in touch with each other – there has been progress towards this goal.

- Session 1 (Dec. 7, 2020): 26 attendees. Ms. Running Wolf and Sara Child (Kwakwaka'wakw community) introduced CWILTs; Delaney Lothian (Cree, Métis) presented *Sound Hunters*, her game for training learners to recognize Cree phonemes and syllabics (above); Eddie Antonio Santos presented Western Cree syllabics conversion software (above).

- Session 2 (Jan. 7, 2021): 33 attendees. Stephanie Tenasco (Kitigan Zibi Anishinabe) shared her experience of creating language learning videos; Patrick Littell and Aidan Pine (NRC) presented ReadAlong Studio.
- Session 3 (Jan. 28, 2021): 80 attendees (most were from Canada but there were 13 from USA, 2 from Norway, 1 from Mexico, 1 from Bangladesh, and 2 from Morocco). Theme: "Strengthening Language through Technology: A Context for Capacity Building". Aaron Plahn and Justin Bambrick presented their corpus collection and capacity-building initiative for Tŝilhqot'in: https://www.youtube.com/watch?v=L-eXWFHnCSc&feature=youtu.be .
- Session 4 (July 14, 2021): number of attendees not recorded (at least 40). Sean Smith, Krista Dempster, and Dustin Hill of YNLC described capacity building in their subproject.

SINGLE-LANGUAGE SUBPROJECTS

Cree text digitization at Blue Quills University:
Blue Quills was the first university in Canada to be fully owned and operated by First Nations people. This subproject digitized and indexed the largest known collection of Cree syllabics text: approximately 40,000 pages from monthly newsletters called kihcitwāw miteh (Sacred Heart) produced by the Catholic Church between 1906 and 1978. The corpus will soon be available at http://language-unbq.bluequills.ca.

Kanyen'kéha data collection at TTO:
This effort is led by Nathan Thanyehténhas Brinklow, a language teacher at Queen's University in Ontario, in collaboration with the Tsi Tyónnheht Onkwawén:na Language and Cultural Centre (TTO) on Tyendinaga Mohawk Territory. The goal is creation of text and audio corpora to support research on ASR for Kanyen'kéha. This subproject has collected 112,420 Kanyen'kéha written words and 26 hours of Kanyen'kéha speech, covering genres such as scripts and audio from movies and TV shows, translated books of the Bible, and recordings of Elders.

Kwak'wala data collection:
This effort also involves corpus collection followed by ASR research. Kwak'wala (Wakashan family) is spoken by 18 Kwakwaka'wakw Nations whose traditional territory is on northern Vancouver Island, nearby islands, and the adjacent mainland. NRC-ILT funding supported a partnership among two community-based teams from three different Kwakwaka'wakw communities, a technical team, and three university-based linguists (two of whom are Kwakwaka'wakw). This subproject has compiled 25 hours of machine-readable Kwak'wala audio data, consisting of conversational speech, pedagogical materials, and storytelling by Elders. It has also identified over a hundred hours of further Kwak'wala speech recordings in various locations.

Michif talking dictionary:
This subproject mobilized and made accessible an out-of-print Michif dictionary called "The Michif Dictionary: Turtle Mountain Dictionary Chippewa Cree", first published in 1983. Partly funded by NRC-ILT, and with assistance from Michif first-language speakers, computational linguists, and others, the team developed a digital, spoken version of this resource. The subproject produced over 181 hours of high-quality audio recordings of the dictionary from four speakers; the dictionary includes 15,422 entries. All 350 pages of Michif lexical entries and example sentences have been recorded by at least one speaker, with some entries being recorded by two or more speakers. The link to the dictionary: https://dictionary.michif.org/home.

The final report for this subproject refers to the new CRIM technologies described above: "Each recording was automatically segmented into pause-delimited utterances using a Deep Neural Network (DNN) voice activity detection service that was developed within the VESTA-ELAN project by the Computer Research Institute of Montréal (CRIM). This auto-segmentation saved an immeasurable amount of time in the annotation process" *(communication from Heather Souter and Olivia Sammons, subproject co-leads).*

Nsyilxcn:
The Nsyilxcn language (Interior Salish family) was traditionally spoken in the Okanagan Valley in present-day British Columbia. It is critically endangered, with a dozen highly fluent Elders remaining. This subproject began in December 2019; it was carried out by Syilx Language House (SLH), a community-based organization, with some funding from NRC-ILT. Seven hours of fluent Elder stories in Nsyilxcn were recorded, transcribed (with a glossary provided in English) and archived. All recordings are shared at www.thelanguagehouse.ca. 4 trainees and 10 volunteer learners were trained to fluency as part of the subproject.

SENĆOŦEN:
SENĆOŦEN (Coast Salish family) was spoken just north of present-day Victoria, British Columbia. It is the most severely endangered language NRC-ILT worked with: until recently, there were only five fluent speakers left. However, the community is engaged in a vigorous language revitalization effort, led by the WSÁNEĆ School Board. During 1981-1991, the linguist Dr. Timothy Montler recorded Elsie Claxton, the last monolingual SENĆOŦEN speaker, telling stories (Montler, 2018). NRC-ILT paid for two Elders to work with Dr. Montler to transcribe these recordings, and translate them into English. This material will be an invaluable resource for teaching of the language.

Tŝilhqot'in:
Tŝilhqot'in (Dene, i.e. "Athabaskan" family) was traditionally spoken in the southern interior of present-day British Columbia. Funding from NRC-ILT enabled the Tŝilhqot'in National Government (TNG) to build on an already impressive language revitalization effort. TNG had already recorded 35,000 audio clips, developed a linguistic database, published verb paradigms in a "Digital Verb Book", and created a website with diverse learning tools: www.tsilhqotinlanguage.ca. With our funding, TNG recorded a dozen hours of speech from 20 speakers; 20 hours of speech were digitized, transcribed, and aligned with the transcriptions. Labeling of 46,000 audio clips was completed. A list of 26,200 English words and phrases to be translated has been drawn up; many of these have been translated into Tŝilhqotin. Four community members were trained in IT-oriented roles.

A superb video on this community's language revitalization efforts from CWILTs is at https://www.youtube.com/watch?v=L-eXWFHnCSc&feature=youtu.be .

FUTURE WORK AND CONCLUSION

Brinklow (2021) states that it is possible for communities to collaborate productively with some non-Indigenous partners on Indigenous Language Technology (ILT):

"Many communities in Canada are collaborating with non-profit partners for the development of ILT outside of a profit-driven ecosystem … These innovative partnerships include universities, communities, research institutions, governments, and others … In the Canadian context, these partnerships (ironically, often funded by the original colonizing governments) are producing ground-breaking language technologies that are innovative by any standard. "

This Indigenous author has kind words for an NRC-ILT subproject: "One pioneering partnership has developed between Onkwawenna Kentyohkhwa at Six Nations in Ontario and the National Research Council (NRC) of Canada. The partnership is between the NRC-ILT project team and local teachers to develop a verb-conjugator for Kanyen'kéha … While the initial development was done with a specific language (Mohawk), the underlying tool was created to work with any language in a 'first deep, then broad' approach to design and development … The partnership with the NRC adds value to the project at the national level because a broad approach to development is not the responsibility of an individual language community."

What are the next steps?

Our priorities are:

- Extending technologies we have already developed to new languages.
- Making these technologies much easier to use. It's not enough to release open-source software (OSS); the "empowerment" approach implies that such OSS should be easy to use for people inside Indigenous communities. It should not require an advanced degree in computer science or computational linguistics. An example of our current efforts in this direction: work inside NRC-ILT led by Dr. Patrick Littell on "Gramble" – a spreadsheet-based software framework that, we hope, will make it much easier to create verb conjugators.
- A new focus for NRC-ILT will be text-to-speech (TTS). It is preferable for students to learn a language from fluent human speakers. However, given the shortage of teachers for most Indigenous languages, TTS may sometimes be an acceptable second-best. An NRC-ILT team member, Aidan Pine, already created a prototype TTS system for Kanyen'kéha (see above); recently, he has built experimental TTS systems for two other Indigenous languages spoken in Canada, Gitksan and SENĆOŦEN, that yielded surprisingly good voice quality.

The prospect for many Indigenous languages spoken in Canada is brighter than it has been for decades – thanks to the efforts of Indigenous peoples themselves. The NRC-ILT project shows that it is possible for non-Indigenous organizations, even governmental ones, to develop technology in collaboration with Indigenous stakeholders that is useful to these language revitalization efforts. However, non-Indigenous organizations active in this area must remember that their role is a secondary one, and that their goal should be empowerment of Indigenous communities.

BIBLIOGRAPHY

Rubèn Fernández Asensio (editor). 2019. "Old kava in new gourds: language revitalization and schooling in Hawaii". In *Linguapax Review*. https://www.linguapax.org/wp-content/uploads/2020/02/linguapax19-1-1.pdf

Peter Bakker. 1997. "A language of our own". Oxford & New York: Oxford University Press.

Nathan Thanyehténhas Brinklow, Patrick Littell, Delaney Lothian, Aidan Pine, and Heather Souter. 2019. "Indigenous Language Technologies and Language Reclamation in Canada". In *Collection of research papers of the 1st international conference on language technologies for all*. European Language Resources Association (ELRA). https://lt4all.elra.info/proceedings/lt4all2019/pdf/2019.lt4all-1.100.pdf.

Nathan Thanyehténhas Brinklow. 2021. "Indigenous Language Technologies: Anti-Colonial Oases in a Colonizing (Digital) World". In *WINHEC: International Journal* *of Indigenous Education Scholarship,* pp. 239-266; Special Issue on Indigenous Language Revitalization. Innovation, Reflection and Future Directions. http://dx.doi.org/10.18357/wj1202120288.

Michael J. Chandler and Christopher Lalonde. 1998. "Cultural continuity as a hedge against suicide in Canada's First Nations". In *Transcultural Psychiatry*, 35(4):191–219.

Christopher Cox, Gilles Boulianne, and Jahangir Alam. 2019. "Taking aim at the transcription bottleneck: Integrating speech technology into language documentation and conservation". [Conference Presentation]. *International Conference on Language Documentation & Conservation*. https://scholarspace.manoa.hawaii.edu/bitstream/10125/44841/44841.pdf

Ewa Czaykowska-Higgins. 2009. "Research models, community engagement, and linguistic fieldwork: Reflections on working within Canadian Indigenous communities". In *Language documentation & conservation*, 3(1):182–215.

Fineen Davies, Eddie A. Santos, and Heather Souter. 2021. "On the Computational Modeling of Michif Verbal Morphology". *EACL*. https://aclanthology.org/2021.eacl-main.226.pdf.

Petra Fachinger. 2019. "Colonial violence in sixties scoop narratives: from In Search of April Raintree to A Matter of Conscience". In *Studies in American Indian Literatures*, 31(1-2):115.

Benoît Farley. 2012. "The Uqailaut project". [Website]. http://www. inuktitutcomputing.ca.

First Voices Portal. 2021. "SENĆOŦEN Home Page". Downloaded from https://www.firstvoices.com/explore/FV/sections/Data/THE%20SEN%C4%86O%C5%A6EN%20LANGUAGE/SEN%C4%86O%C5%A6EN/SEN%C4%86O%C5%A6EN on Oct. 1, 2021.

Government of Canada. 2015. "Final report of the Truth and Reconciliation Commission". http://nctr.ca/reports.php.

Vishwa Gupta and Gilles Boulianne. 2020a. "Automatic transcription challenges for Inuktitut, a low-resource polysynthetic language". In *Proceedings of LREC*. https://aclanthology.org/2020.lrec-1.307.pdf.

Vishwa Gupta and Gilles Boulianne. 2020b. "Speech transcription challenges for resource constrained Indigenous language Cree". In *Proceedings of 1st Joint SLTU-CCURL Workshop*. https://aclanthology.org/2020.sltu-1.51/

Darcy Hallett, Michael J. Chandler, and Christopher E. Lalonde. 2007. "Aboriginal language knowledge and youth suicide". In *Cognitive Development*, 22(3):392–399.

Eric Joanis, Rebecca Knowles, Roland Kuhn, Samuel Larkin et al. 2020. "The Nunavut Hansard Inuktitut–English parallel corpus 3.0 with preliminary machine translation results". In *Proceedings of LREC-2020*. https://aclanthology.org/2020.lrec-1.312.pdf .

Anna Kazantseva, Owennatékha Brian Maracle, Ronkwe'tiyohstha Josiah Maracle, and Aidan Pine. 2018. "Kawennón:nis: the wordmaker for Kanyen'kéha". In *COLING Workshop on Computational Modeling of Polysynthetic Languages*. https://aclanthology.org/W18-4806.pdf

Te Taka Keegan. 2019. Issues with Māori sovereignty over Māori language data. [Video Presentation]. http://video.web.gov.bc.ca/public/fpcc/letlanguageslive.html.

Jeannette King. 2013. "Te Kōhanga Reo: Māori Language Revitalization". In L. Hinton and K. Hale (editors), *The Green Book of Language Revitalization in Practice*. Brill.

Rebecca Knowles, Darlene Stewart, Samuel Larkin, and Patrick Littell. 2020. "NRC Systems for the 2020 Inuktitut-English

News Translation Task". In *Proceedings of WMT*. https://aclanthology.org/2020.wmt-1.13/.

Roland Kuhn, Fineen Davis, Alain Désilets, Eric Joanis et al. 2020. "The Indigenous Languages Technology project at NRC Canada: An empowerment-oriented approach to developing language software". In *Proceedings of COLING 2020*. https://www.aclweb.org/anthology/2020.coling-main.516.pdf.

Patrick Littell, Anna Kazantseva, Roland Kuhn, Aidan Pine, Antti Arppe, Christopher Cox, and Marie-Odile Junker. 2018. "Indigenous language technologies in Canada: Assessment, challenges, and successes". In *27th International Conference on Computational Linguistics (COLING 2018)*, Santa Fe, New Mexico, USA. https://www.aclweb.org/anthology/C18-1222.pdf

Delaney Lothian, Gokce Akcayir, Anaka Sparrow, Owen Mcleod, and Carrie Demmans Epp. 2020. "SoundHunters: Increasing Learner Phonological Awareness in Plains Cree". In *Artificial Intelligence in Education*. doi: 10.1007/978-3-030-52237-7_28.

Joel Martin, Howard Johnson, Benoît Farley, and Anna Maclachlan. 2003. "Aligning and using an English-Inuktitut parallel corpus". In *Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts*, Volume 3, pages 115–118. Association for Computational Linguistics.

Timothy Montler. 2018. "SENCOTEN: a dictionary of the Saanich language". University of Washington Press, ISBN: 9780295743851.

Nicole Rosen and Heather Souter. 2009. "Language revitalization in a multilingual community: the case of Michif." In *1st International Conference on Language Documentation and Conservation (ICLDC)*, Honolulu.

Statistics Canada. 2017. "Proportion of mother tongue responses for various regions in Canada, 2016 census". https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/dv-vd/lang/index-eng.cfm.

Olivia N. Sammons. 2019. "Nominal classification in Michif". *Ph.D. thesis, University of Alberta*. https://era.library.ualberta.ca/items/903b7ee2-4f3f-4959-b4f0-8231205d5d67/view/9a967b05-b223-4716-97de-6da32e6931fe/Sammons_Olivia_N_201901_PhD.pdf

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. "ELAN: A professional framework for multimodality research". In *European Language Resources Association (ELRA)*.