Aidan Pine, Nathan Brinklow, Delaney Lothian, Heather Souter
National Research Council Canada, Queen's University, University of Alberta, Prairies to Woodlands Language Revitalization Circle
aidan.pine@nrc-cnrc.gc.ca, nathan.brinklow@queensu.ca, delaney.lothian@gmail.com, hsouter@gmail.com

National Research Council Canada

Ioienawà:ses ne Onkwehón:we Raotiwenna'shón:'a

Indigenous Language Technology Project

Tsi Niwathró:ris ne Kaio'ténhsera (Abstract)

Ne Canada kakó:ra óntka'we ne 2017 ohwísta, $89.9M nè:ne áhsen niiohserà:ke nika'shén:nes nè:ne aiakoié:nawa'se' ne Onkwehón:we raotiwenna'shón:'a tánon tsi nihotirihó:ten. Ne Canadian National Research Council (NRC) $6M tahonwén:non e'tho tonterá:ko. Onkwehón:we rotitióhkwaien ne shakonaten'nikonhrá:wis ne enhonten'nikón:raren, ne NRC enhatirihó'kwate tánon enhonnón:ni ne tsi ní:ioht ne aiontá:ti tánon tsi ní:ioht tsi akahiatón:hake, á:se nahò:ten ronaterien'tatshén:rion tsi ní:ioht tsi akonwaié:nawa'se' kwah tokén:en aiorihwahnírha, aonsontónhnhete tánon taontá:we ne Onkwehonwehnéha owenna'shón:'a. Iakwá:ton ne aioiá:neren'ne' tsi nahatí:iere ne á:se ronaterien'tatshén:rion tsi nakaié:ren wentá:'on ki' kwah tsi tka'nónhkwaien ne owén:na tsi sewatónhnhet, tóhsa ó:ia' nahò:ten ohén:ton watohétstakw. Kí enkahiatonhserohá:rake áhsen niiorì:wake enwentó:ren'ne' ne Onkwehonwehnéha owén:na aorihwà:ke ne tsi nón:we nikanakerahseraké:ron ne Canada, tánon ó:ni ronne'néstha ne áhsen niiorì:wake tsi nahatí:iere' ne NRC.

The Canadian federal budget released in 2017 invested $89.9M over three years to support Indigenous languages and cultures. The Canadian National Research Council (NRC) was granted $6M of this funding. Under the guidance of an Indigenous advisory committee, the NRC is researching and developing speech- and text-based technologies which aim to assist the stabilization, revitalization and reclamation of Indigenous languages. We argue that successful technological solutions must support grass-roots language revitalization efforts and not overshadow them. This poster session describes three separate challenges faced by Indigenous language communities in Canada and highlights three corresponding NRC solutions.

Wa'karihwahnhotón:ko' (Introduction & Background)

There are over 90 Indigenous languages in Canada from 10 distinct language families (See Figure 1). The richness of this diverse linguistic landscape is precious and important, but also poses a challenge for planning and resource management.

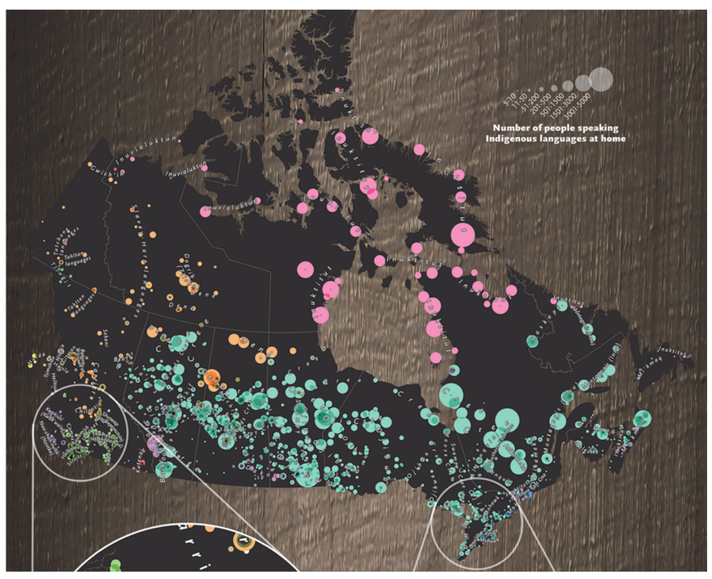

Figure 1. Indigenous languages in Canada
National Geographic Canada

Every one of the languages in Figure 1 endured systemic oppression throughout the residential school period which lasted from 1883 until very recently in the late 1990's which tried to replace Indigenous languages with English & French. The fact that Indigenous people are *still* learning and speaking their languages despite this is proof of the resilience of Indigenous people and the failure of colonial language policy.

Understanding why *re*-vitalization is important to Indigenous communities in Canada, requires understanding the sociohistorical context surrounding their *de*-vitalization. Many view language revitalization as not **just** about learning nouns, verbs, and grammar, but as part of a broader political movement for self-determination. Given the context of oppression, speaking an Indigenous language in Canada is a political act.

At NRC, we believe that technology has a modest, but important role to play in supporting and assisting community-led language revitalization efforts. To guide this process, the NRC has recruited a wide range of Indigenous experts to participate in an advisory committee which has a meaningful role in determining the direction of the ILT project.

Tsi ní:ioht aiakoiénawa'se' (Barriers & Technological Solutions)

Excluding Indigenous languages from the future is the same as relegating them to the past. It is important that as new technologies become commonplace, Indigenous languages are included. The following describes three barriers to learners and teachers of Indigenous languages and three corresponding NRC solutions.

Predictive Text

Typing on mobile devices is slow and for many Indigenous languages, it is difficult to type with the proper keyboard. Additionally, learners often have difficulty remembering the precise spelling of a word. When English or French auto-correct and predictive text incorrectly applies to an Indigenous language, it has the effect of making the user experience (UX) with the Indigenous language feel *foreign.*

NRC has developed a *predictive text platform (*bundled in Keyman 12) which allows users to create their own models (thereby keeping control of their data) and build keyboards with predictive text for their language. This speeds up typing, gives language-appropriate approximate suggestions, and improves the localization of the user experience for an Indigenous language user.

Figure 2. Predictive Text
SENĆOŦEN (Salish)

WordWeaver & Verb Conjugation

Iroquoian languages have some of the longest running and most successful immersion programs in the country. Conjugating verbs is notoriously complex though and learning is a major hurdle for learners. Nothing like a Bescherelle exists, and teachers have to act as references for learners and others which is a burden on their already limited time.

Figure 3. Verb Chart
Maura Abrams

(1) *iah th-a-etsi-te-w-ate-wistohsera-'tarih-á:t-ha-k-e'*
no NOT-WOULD-AGAIN-WE-ALL-OWN-butter-HOT-CAUSE-HABIT-CONTIN-PERF.
'We will no longer keep heating up our butter.' Mohawk (Mithun, 1996, p. 170)

Figure 4. Example of polysynthesis in Kanien'kéha

Founder of the Onkwawenna Kentyohkwa immersion school Brian Maracle had an idea of a sort of Mohawk 'Bescherelle' – a resource for students of his school to be able to look up verbs. NRC has collaborated with Brian, along with two teachers from Onkwawenna Kentyohkwa, as well as two other communities to build software to be able to create these types of applications.

The software allows users to create their own web applications by providing a language model and configuration files. The tool's user interface is designed following ideas from the students.

Given that the language model for the Onkwawenna Kentyohkwa instance has over 400,000 conjugations, it is infeasible to record audio for each of those conjugations. NRC is therefore involved in *speech synthesis* (text-to-speech) research to figure out what the smallest number of recordings needed is, in order to generate audio for all conjugations produced by the language model

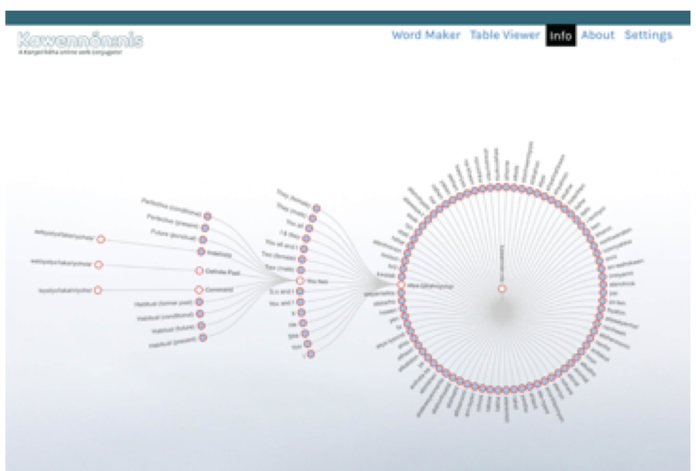

Figure 5. Kawennón:nis
Ohswekenéha (Ohsweken Mohawk) instance of WordWeaver

Automatic Audio/Text Alignment

Many communities have raw audio and text, but reach a content-creation bottleneck when turning these raw materials into meaningful educational materials. Raw materials can be frustrating for learners to use because it is easy to get lost in recordings. Conversely, manually aligning audio and text is a technical and time-consuming skill.

NRC has collaborated with a researcher at Nuance (David Daines) to create a tool called Read Along Studio (RAS). RAS automatically aligns audio and text and exports the resulting alignment to a number of relevant formats for widening the content creation bottleneck. As a primary focus, RAS exports to a web component which can be embedded into any web application or hybrid mobile application.

RAS relies on grapheme-to-phoneme (G2P) models to turn the orthography of a given language into the phonetic alphabet. From there, the closest English equivalents are found, and RAS aligns the data on the English mapping. This allows languages with zero data to use RAS.

RAS can also export to EPUB and various subtitle formats, as well as Praat TextGrids and Elan files to meet a host of different potential users' needs.

RAS can also be used with images, and the interface allows the user various controls over features such as pagination or playback speed.

Figure 6. Read Along Web Component for Gitksan

A visual programming interface has also been created to facilitate the development of the required G2P models for a new language.

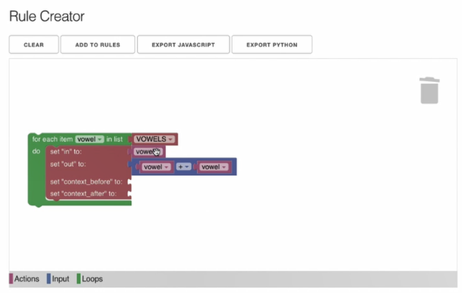

Figure 7. G2P Visual programming interface

RAS currently supports Atikamekw, Western Highland Chatino, Chukchi*, Southern East Cree, Northern East Cree, Danish, English, Gitksan, Hoocąk, Inuktitut, Kwak'wala*, Mohawk, and SENĆOŦEN.

Iako'nikonhraientá:tha Ken'nikakarésha (Conclusion)

We have shown three examples of practical research goals that the NRC has achieved. This is **not** an exhaustive list of the projects we are involved in, but simply a sample to illustrate some of NRC's research specializations.

Through this project, we have learned some key insights. Here are a number of them:

Process is equally as important as **product** in the context of language revitalization in Canada.

Technology efforts should *amplify* rather than **overshadow** community-led revitalizations efforts.

Technology is the icing, not the cake!

Technology should catalyze actual language learning or language use.

Smash the lore! NRC should endeavor to document projects and involve communities to reduce community dependency on a small group of technologists. Releasing tools under open source licenses is an important part of this.

*We would like to sincerely thank and acknowledge Tewateronhíakhwa Mina Beauvais of Kanehsatà:ke for her invaluable help in translating parts of this poster into Kanien'kéha