The Indigenous Languages Technology project at NRC Canada: an empowerment-oriented approach to developing language software

Kuhn, Roland; Davis, Fineen; Désilets, Alain; Joanis, Eric; Kazantseva, Anna; Knowles, Rebecca; Littell, Patrick; Lothian, Delaney; Pine, Aidan; Running Wolf, Caroline; Santos, Eddie; Stewart, Darlene; Boulianne, Gilles; Gupta, Vishwa; Maracle, Owennatékha Brian; Martin, Akwiratékha'; Cox, Christopher; Junker, Marie-Odile; Sammons, Olivia; Torkornoo, Delasie; Thanyehténhas Brinklow, Nathan; Child, Sara; Farley, Benoît; Huggins-Daines, David; Rosenblum, Daisy; Souter, Heather

Publisher's version / Version de l'éditeur:
https://doi.org/10.18653/v1/2020.coling-main.516
COLING 2020: Proceedings of the 28th International Conference on Computational Linguistics, pp. 5866-5878, 2020-12-13

The Indigenous Languages Technology project at NRC Canada: An empowerment-oriented approach to developing language software

Roland Kuhn, Fineen Davis, Alain Désilets, Eric Joanis,
Anna Kazantseva, Rebecca Knowles, Patrick Littell, Delaney Lothian,
Aidan Pine, Caroline Running Wolf, Eddie Santos, Darlene Stewart
National Research Council Canada (NRC)
Firstname.Secondname@nrc-cnrc.gc.ca (e.g., Roland.Kuhn@nrc-cnrc.gc.ca)

Gilles Boulianne, Vishwa Gupta
Centre de recherche informatique de Montréal (CRIM)
gilles.boulianne@crim.ca, vishwa.gupta@crim.ca

Owennatékha Brian Maracle
Onkwawenna Kentyohkwa (Our Language Society)
owennatekha@gmail.com

Akwiratékha' Martin
Kahnawà:ke Kanien'kehá:ka Territory
tekhaluvsyou@hotmail.com

Christopher Cox, Marie-Odile Junker, Olivia Sammons, Delasie Torkornoo
Carleton University
FirstnameSecondname@cunet.carleton.ca

Nathan Thanyehténhas Brinklow
Queen's University and
Tsi Tyonnheht Onkwawen:na Language and Cultural Centre (TTO)
nathan.brinklow@queensu.ca

Sara Child
S̲anyak̲ola Foundation
sanyakola2018@gmail.com

Benoît Farley
Pirurvik Centre
benoitfarley@videotron.ca

David Huggins-Daines
Nuance Communications
dhdaines@gmail.com

Daisy Rosenblum
University of British Columbia
daisy.rosenblum@ubc.ca

Heather Souter
Prairies to Woodlands Indigenous Language Revitalization Circle
p2wilrc@gmail.com

Proceedings of the 28th International Conference on Computational Linguistics, pages 5866–5878
Barcelona, Spain (Online), December 8-13, 2020

Abstract

This paper describes the first, three-year phase of a project at the National Research Council of Canada that creates software to assist Indigenous communities in preserving their languages and extending their use. The project aimed to work within the empowerment paradigm, where collaboration with communities and fulfillment of their goals is central. Since many of the technologies we developed were in response to community needs, the project ended up as a collection of diverse subprojects, including the creation of a sophisticated framework for building verb conjugators for highly inflectional polysynthetic languages (such as Kanyen'kéha, in the Iroquoian language family), release of what is probably the largest available corpus of sentences in a polysynthetic language (Inuktut) aligned with English sentences and experiments with machine translation (MT) systems trained on this corpus, free online services based on automatic speech recognition (ASR) for easing the transcription bottleneck for speech recordings, software for implementing text prediction and read-along audiobooks for Indigenous languages, and several other subprojects.

1 Introduction

This paper describes the Indigenous Languages Technology (ILT) project at the National Research Council of Canada (NRC). Phase I of this project received funding of $6 million over three years in the March 2017 budget of the Government of Canada; phase II is ongoing. The project's goal is to produce software that will enhance the efforts of Indigenous communities in Canada to preserve and revitalize their languages. Littell et al. (2018) surveys many different efforts by a variety of organizations that have the same goal. This paper depicts a tiny corner of a big canvas; for space reasons, it does not even cover all aspects of our project. For a detailed technical report on the entire ILT project, see Kuhn et al. (2020).

Different communities have very different linguistic needs, so the ILT project was made up of a diverse set of subprojects. Because there is relatively little textual or speech data for Indigenous languages in Canada (with the partial exception of Inuktut), most of the technologies developed within the project described here have been rule-based, rather than relying on data-driven machine learning.

Different approaches have been taken to linguistic research involving Indigenous languages (Cameron et al., 1992; Czaykowska-Higgins, 2009). Most of the work carried out within this project aims to fit into the "empowerment" approach, in which research is carried out collaboratively, with equal emphasis on the agenda of the linguist and of the community. The most ambitious subproject described here was suggested to the NRC team by an Indigenous educator: the creation of a verb conjugator for Kanyen'kéha (Mohawk) (which inspired the creation of a software framework for building verb conjugators for other languages). Similarly, the "readalong" subproject for automating word–speech alignment for audio books was in response to strong interest from several communities.

The project was guided by an Advisory Committee made up of Indigenous language revitalization experts. Their counsel has been invaluable; they are listed at the end of the project web page[1] and in the Acknowledgements. At no stage did the NRC claim ownership of Indigenous language data collected with the project's funding. We were determined to break with the long, painful history of extractive research practices, in which non-Indigenous researchers have collected language data from Indigenous communities for their own purposes, sometimes even resulting in communities losing access to their language data (see Pool (2016), Keegan (2019), Brinklow et al. (forthcoming), and Lewis et al. (2020)).

2 Sociolinguistic Background

There are about 70 Indigenous languages from 10 distinct language families currently spoken in Canada (Rice, 2008). Most of these languages have complex morphology; they are polysynthetic or agglutinative. Commonly, a single word carries the meaning of an entire clause in Indo-European languages.

[1] https://nrc.canada.ca/en/research-development/research-collaboration/programs/canadian-indigenous-languages-technology-project

All Indigenous languages in Canada were targeted by government policies that sought to eradicate them. These policies were implemented through legislation, such as the Indian Act,[2] which 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 (known as Residential Schools) or adopted by non-Indigenous families (known as the Sixties Scoop (Fachinger, 2019)). 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).

The resilience of Indigenous communities can be seen in the many ways that they have resisted assimilation and continued to teach, learn, and speak their languages (Pine and Turin, 2017). The benefits associated with the use of these languages are wide-ranging (Whalen et al., 2016; Reyhner, 2010; Oster et al., 2014; Marmion et al., 2014). For instance, there is a correlation between Indigenous language use and a decrease in youth suicide rates on reserves in British Columbia (Chandler and Lalonde, 1998; Hallett et al., 2007). However, many communities face decreasing numbers of first language (mother tongue) speakers, due to declining language transmission rates (Norris, 2018). Much Indigenous language revitalization work in Canada focuses on preservation of language through recording the speech of Elders: recordings and transcriptions are a vital resource for language learning by younger generations.

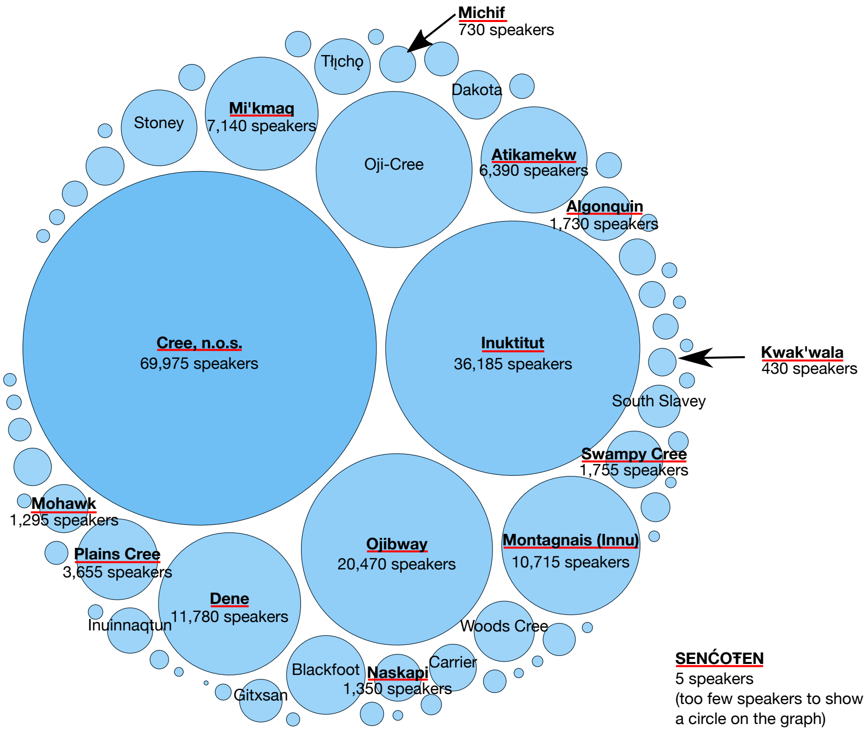


Figure 1: Number of speakers of Indigenous languages spoken in Canada, according to the 2016 census (From Statistics Canada (2017))

Figure 1 (obtained from the Statistics Canada website) shows the number of speakers per language in 2016 (Statistics Canada, 2017). The languages underlined in red are those with which the ILT project has interacted in some way. These census numbers are controversial. This figure is included to give the non-expert reader a general idea of the number of speakers of each language; for detailed demographic information, one should consult experts on each individual language.

[2] See https://www.thecanadianencyclopedia.ca/en/article/indian-act, https://indigenousfoundations.arts.ubc.ca/the_indian_act/.

3 Text-based subprojects

3.1 Polysynthetic Verb Conjugation

3.1.1 Background

Early in the ILT project, Owennatékha (Brian Maracle), the director of Onkwawenna Kentyohkwa (Our Language Society), asked whether NRC would be able to create a software-based verb conjugator. Onkwawenna Kentyohkwa is an adult immersion school in the Six Nations Grand River Territory in Ontario that takes students through 2000 hours of Kanyen'kéha immersion over two years. Kanyen'kéha is an Iroquoian language, commonly known as "Mohawk", spoken in territory that spans present-day Ontario, Quebec, and New York State. Iroquoian languages are highly polysynthetic.

Learning and teaching verbs in Kanyen'kéha is a formidable task. There are millions of conjugations for even the most common verbs. It is therefore only possible to create a verb conjugator with reasonable coverage in software, not on paper. Verb roots in Kanyen'kéha are bound morphemes: they do not, on their own, constitute words. A pronominal prefix, a verb root and an aspectual ending are always present (for commands the aspectual ending is null). A verb can contain pre- and post-pronominal prefixes and pre-aspectual suffixes. Kanyen'kéha has 14 stand-alone or 'free' pronouns, and 72 bound pronouns, meaning the combinatorial inflectional possibilities are much larger than in English, French, or other European languages. This complexity adds to the difficulty of teaching the language: even learning how to properly conjugate a modest number of verbs requires a significant amount of work. For detailed information on Kanyen'kéha syntax, see Kanatawakhon (2002).

In 2017, we began building the verb conjugator for Kanyen'kéha described below (Kazantseva et al., 2018). We are now building a verb conjugator for Michif, an unrelated polysynthetic language. For each language, we have been working with members of the relevant community to develop a conjugator that does not replace learners' experience of studying verbal morphology at schools in the community, but that complements it. These conjugators rely on finite-state transducers (FSTs) which are rule-based. Rule-based approaches may seem outdated in contrast to statistical or neural methods. However, with most Indigenous languages, existing corpora are not large enough to produce accurate statistical models.

3.1.2 WordWeaver

Since many Indigenous languages have rich inflectional morphology, we decided to carry out Owennatékha's request by building a software tool that could be extended beyond Kanyen'kéha to other languages. The structure of the WordWeaver ecosystem consists of two main parts: a front-end interface (WordWeaver UI) implemented in Angular, and a back-end database and API implemented in Python (Fastapi & CouchDB). Initially, WordWeaver was tightly coupled to the instance's language model, specifically Foma (Hulden, 2009). However, all data is now stored in a database. This architecture allows verb conjugators to be made without requiring knowledge about FSTs.

3.1.3 WordWeaver Instances

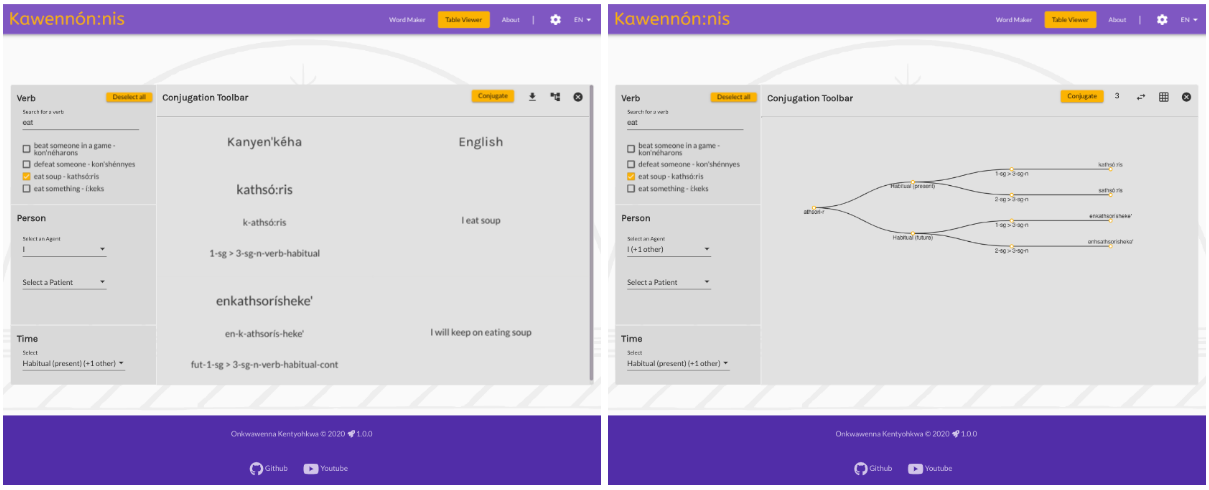
The first instance of WordWeaver, Kawennón:nis, models the Western dialect of Kanyen'kéha that is taught at the Onkwawenna Kentyohkwa school. Kawennón:nis means "It Makes Words" in the language. We have since also created an instance for the Eastern dialect, spoken in the Kahnawà:ke community in Quebec. To design rules for the first Kawennón:nis FST, we relied on a textbook that describes the Western dialect (Maracle, 2017), along with writings on other Kanyen'kéha dialects and the related language Oneida. The NRC team benefited from a close relationship with the staff of the school, who added hundreds of new verbs to the system through a collaborative development process whereby new verb stems are entered in a spreadsheet and are then compiled into an FST lexicon formalism (lexc).

Quality control for this Western version of Kawennón:nis was done by teachers at the school and by members of the research team. Team members made several in-person visits to the school to participate in multi-day collaborative sessions to design, evaluate, and improve the user interface (UI). Creation of the Eastern (Kahnawà:ke) version relied heavily on the expertise of Akwiratékha' Martin, who is familiar with that dialect. The Western version of Kawennón:nis contains over 250 verb stems while the Eastern version has approximately 600. Both versions contain all bound pronouns and 12 tense/aspect combinations (command, habitual forms, perfective forms and punctual forms including the definite past, conditional and future forms) and are capable of generating over 100,000 conjugated forms.

3.1.4 WordWeaver User Interface (UI)

The UI is of prime importance, if WordWeaver is to be useful to students. The process for prototyping, designing, and evaluating the WordWeaver UI involved multiple in-person visits for defining the requirements, hiring in-community designers, and extensive user interface and user experience (UI/UX) review. The resulting UI is interactive, highly theme-able, translated into English, French and Kanyen'kéha, available on the web and mobile as a progressive web application, and available offline.

There are two main views within the application, the **'Wordmaker'** and the **'Tableviewer'**. The **'Wordmaker'** is the simplest view and guides the user linearly through three questions to create a single conjugation, *what* the action is, *who* is doing it, and *when* it's happening. It is assumed that each conjugation will minimally require a root and some sort of pronominal inflection. The third category is the most open and could be used for other 'options' beyond the temporal options currently suggested by WordWeaver. The **'Tableviewer'** is the more advanced view. It allows users to create paradigm tables of conjugations instead of single output forms. Here users can select multiple options from the three categories to create a query for many conjugated forms. The user can then either interact with the conjugations in a tabular grid format (see Figure 2, left) or in a 'tree' format (see Figure 2, right). Users can also download the conjugations as a Microsoft Word Document, CSV file, or a formatted LaTeX file. This functionality allows users to create and print out tables for making their own flashcards or study tools.



Search for a verb: eat

□ beat someone in a game - kon'néharons
□ defeat someone - kon'shénnyes
☑ eat soup - kathsó:ris
□ eat something - í:keks

Select an Agent: I

Select a Patient

Select: Habitual (present) (+1 other)

Kanyen'kéha	English
kathsó:ris	
k-athsó:ris	I eat soup
1-sg > 3-sg-n-verb-habitual	
enkathsorísheke'	
en-k-athsorís-heke'	I will keep on eating soup
fut-1-sg > 3-sg-n-verb-habitual-cont	

Search for a verb: eat

□ beat someone in a game - kon'néharons
□ defeat someone - kon'shénnyes
☑ eat soup - kathsó:ris
□ eat something - í:keks

Select an Agent: I (+1 other)

Select a Patient

Select: Habitual (present) (+1 other)

Figure 2: Tabular grid view (left) and 'Tree' view (right) of the Kawennón:nis Tablemaker

3.1.5 Work on Michif

The work on a verb conjugator for Michif began after the NRC team was asked by Heather Souter of the Prairies to Woodlands Indigenous Language Revitalization Circle (P2WILRC), whether they could implement a system like Kawennón:nis for Michif. Michif is a mixed, highly polysynthetic language which arose during the 19th century from the intermarriage between French fur traders and Cree and Ojibwe women (Rosen and Souter, 2009). Their descendants are the Métis people, whose official language is Michif. Michif takes most of its nominal patterns from French, and its verbal patterns from Cree. There is a high degree of regional variation (Sammons, 2019). As is the case for many Indigenous languages, there is a shortage of data and formal documentation for the language.

With Ms. Souter's help, the NRC team has been building `Li Verb kaa-Ooshitahk di Michif` (Davis and Santos, under review), a finite-state transducer (FST) which models the verbal morphology of Michif according to the lexc formalism (Beesley and Karttunen, 2003). The current implementation allows for the conjugation of 22 verb stems, which generates 6791 possible verb forms. Though `Li Verb kaa-Ooshitahk di Michif` relies on FSTs, just as Kawennón:nis does, it is not yet integrated into the

WordWeaver code base; we have started working on this integration. The `Li Verb kaa-Ooshitahk di Michif` FST serves as the back-end to an app that will allow users to conjugate verbs in Michif without having any previous training relating to FSTs or linguistics. The app contains a simple interface that walks a user through building a verb conjugation in Michif. This interface will be available for both Android and iOS, and as a web application. The application will be available for use offline (after initial download) and deliberately avoids the use of over-technical linguistic terminology, to make it easier for learners to ask about the language directly from speakers in informal Métis community settings.

3.2 Corpus and Tools for Inuktut (including machine translation)

Inuit languages, which are polysynthetic, are spoken across Arctic Canada. The Government of Nunavut uses the term Inuktut to represent all of the Inuit language varieties spoken in Nunavut. With help from the Government of Nunavut and the Pirurvik Centre, the NRC project team created and released the sentence-aligned Inuktut–English “Nunavut Hansard” corpus based on the proceedings of the Legislative Assembly of Nunavut from April 1999 to June 2017. We believe this to be the largest parallel corpus for an Indigenous language of the Americas or a polysynthetic language released to date (1.3 million aligned sentence pairs). The corpus is available at the NRC Digital Repository[3] under the CC-BY-4.0 license; creation of the corpus and preliminary machine translation experiments are described in Joanis et al. (2020).[4] The existence of the Nunavut Hansard has made possible a shared task for the Inuktut–English language pair in the 2020 Workshop for Machine Translation.[5]

Several years ago, researchers at NRC built a search engine for English-to-Inuktut translators, WeBInuk (Désilets et al., 2008). Given an English word or phrase, it would return matching Inuktut–English sentence pairs from a parallel corpus (the portion of the Nunavut Hansard that was then available). This online service lapsed for several years. One of the goals of the ILT project was to create a new bidirectional version of WeBInuk—i.e., to allow users to enter Inuktut search terms as well as English ones. We also aimed at providing other tools for Inuktut: a dictionary, a gister (i.e. a service that provides rough English renderings of the component morphemes of an Inuktut word), a spell checker, etc.

Inuktut words resemble short phrases in English. They are composed by stringing together: a **root** (about 2000 possibilities) and a fairly long **sequence of morphemes** (typically 4-5 morphemes, but potentially up to 9) taken from a set of roughly 450 affixes (verbs, adjectives, etc.), 1300 verb endings, and 320 noun endings. Unlike Kanyen’kéha, where the surface forms of morphemes are usually invariant, many Inuktut morphemes have surface forms that change in different contexts. The same surface form may correspond to different morphemes.

Many, perhaps most, Inuktut words in a given text will not have occurred before. Thus, building a word-based dictionary with good coverage is unrealistic. However, one can decompose the Inuktut word into a sequence of morphemes, then look up the meanings of the morphemes. Similarly, one could easily build a version of WeBInuk that would allow users to look up whole Inuktut words, but it would not have high coverage; users would often enter words for which exact matches cannot be found in the parallel corpus. It would be more helpful to supply them with Inuktut words in the corpus that share the root and a few other morphemes with the word they are interested in.

Nevertheless, by building on a morphological analyzer created earlier (Farley, 2012), Benoît Farley of the Pirurvik Centre and Alain Désilets of NRC were able to create five prototype apps: 1. a **“morpheme example search”** app that, given a morpheme, shows examples of its use; 2. a **gister** that, given Inuktut text, returns not only the meanings of the morphemes in the words composing it, but also sentence pairs that may have related meanings; 3. a **search tool for Inuktut–English sentence pairs** similar to WeBInuk (but bidirectional); 4. an **Inuktut spell checker** that, if it can’t find a word, returns the most similar words which are known to be correctly spelled and the longest correctly spelled head and tail in the word; and 5. a **morphological search engine** that, given an Inuktut word, searches for its five most frequent morphological variants. These tools are being tested and improved; they will soon be released.

[3] https://doi.org/10.4224/40001819

[4] A pre-release version of the same corpus was used for machine translation experiments during a 2019 JSALT workshop (Schwartz et al., 2020)

[5] http://www.statmt.org/wmt20/translation-task.html

3.3 Predictive Text

Writers in majority languages benefit from predictive text suggestions on their smartphone keyboards. Predictive text requires data—typically mined from large text corpora—to generate an n-gram language model (van Esch et al., 2019); however, many Indigenous language communities either do not have extensive text corpora, or have data too culturally sensitive to share outside their community. We have collaborated with Keyman (keyman.com), an open-source keyboard creation platform, to add a predictive text platform to their suite of smartphone keyboards and their keyboard development tool. We have "decentralized" the creation of predictive text models, by making it possible for Indigenous language activists, using a spreadsheet of words in their language, to provide predictive text suggestions for their community. Thus, the language activists are in charge of their own data, and can choose how they share their predictive text keyboard. This contrasts with Gboard (van Esch et al., 2019) where a centralized entity mines text corpora and creates a language-specific model. Since we expect data to be sparse, the language models our software generates are word-level unigrams. We have implemented predictive text for one language so far, SENĆOŦEN. Members of that community report that it makes typing in their language, which has an unusual orthography, dramatically easier. We are working to make the software easier to use, so that communities can implement predictive text for their own languages.

4 Speech-based subprojects

Traditionally, Indigenous languages in Canada were spoken,[6] not written. Thus, several ILT subprojects focused on applications of speech technology.

4.1 Work at CRIM on Audio Segmentation and Speech Recognition

Transcription and further annotation of speech recordings are the biggest part of the workload in most language documentation and conservation projects. The pace at which speech is being recorded in Indigenous languages in Canada and indeed, in minority languages across the world, greatly outstrips the pace at which field linguists and Indigenous language activists can transcribe these recordings. This is the "transcription bottleneck" (Cox et al., 2019). NRC provided funding to the Centre de recherche informatique de Montréal (CRIM) to develop tools based on automatic speech recognition (ASR) to relieve both this bottleneck, and a related one that could be termed the "indexation bottleneck".

Some communities have thousands of hours of recordings of speech in their languages made years earlier, with no means of searching through them for relevant words or phrases. Though it may be impractical to transcribe all that speech now, ASR-based audio indexation might make search possible.

This subproject faced two main challenges. First, very little data for training ASR systems for Indigenous languages in Canada was available. Second, ASR research has focused on languages that are in the low to medium range of morphological complexity. Most Indigenous languages in Canada are in the high range. Inuktut is known to be particularly morphologically complex, as measured by mean distance to novel type (Schwartz et al., 2020). To address these issues, this subproject had three themes: data collection and transcription, creation of tools to facilitate transcription, and ASR experiments.

4.1.1 Data Collection and Transcription

Four transcription activities were carried out - see Table 1 (details in Kuhn et al. (2020)).

4.1.2 Audio Segmentation

This theme yielded a set of tools to make the early stages of processing recorded speech easier, prior to transcription. These were packaged as Web services on CRIM's VESTA platform,[7] and are available on the platform or through an ELAN extension.[8] We have already received reports of significant productivity improvements due to their use in speech preprocessing in the field. The services include: **DNN-VAD:** deep neural net (DNN) voice activity detection, which extracts segments containing speech (versus silence, noise, etc.); **Diarization:** identifies who spoke when in a recording; **Language Retrieval:** finds

[6] Or signed, though our current work does not touch on Indigenous sign languages of Canada.

[7] http://vesta.crim.ca

[8] ELAN is a widely-used annotation tool available at https://archive.mpi.nl/tla/elan/; see (Wittenburg et al., 2006).

Language	Target (hours)	Current (hours)	Status
Inuktut	100	80.9	Complete (original estimate of available hours was wrong)
East Cree	100	102	Complete
Innu	25	3.4	To be completed by December 2020
Denesuline	10	4.35	Suspended due to COVID-19 situation

Table 1: Amount of transcription done or planned for core CRIM research

segments which are spoken in a particular language (among 32 languages); **Speaker Retrieval:** finds segments spoken by a particular speaker, given a short sample of the speaker's voice; **Multichannel Voice Activity Detection:** detects segments containing speech separately for each track in a multichannel recording with multiple microphones; **Language Independent Text-to-Audio Alignment:** works with any grapheme-to-IPA phoneme table.

4.1.3 Automatic Speech Recognition (ASR)

ASR experiments at CRIM have focused on Inuktut and East Cree. They have highlighted differences between these two polysynthetic languages (Gupta and Boulianne (2020a), Gupta and Boulianne (2020b)). Inuktut is highly polysynthetic, causing a word-based language model (LM) to be ineffective. Even with a dictionary containing 300,000 words (from 6 million words of the Nunavut Hansard), 60% of the words in an unseen story text were out of vocabulary (OOV). So the CRIM experimenters tried syllables and morphemes as sub-word units and found that syllables gave the lowest word error rate (WER).

The WER for the reconstituted words obtained from different sub-word units is shown in Table 2 (for acoustic models trained on 40 hours of transcribed Inuktut speech). B_ and _E markers are used to show boundaries for morphemes and syllables; experiments with boundaries chosen by deep neural nets (DNNs) were also carried out. The LM predicts the start or end of the words through these markers so that syllable or morpheme sequences can be converted back to word sequences. When training data is increased from 40 hours to 80 hours, the speaker independent (SI) WER decreases from 74.3% to 72.3%.

Units	N. of units	OOV rate	WER
Words	129 k	62.6%	108.7%
Unsupervised morphemes	35.1 k	0.8%	80.7%
Semisupervised morphemes	23.2 k	0.4%	79.4%
Syllables + B_, _E	3.2 k	0.1%	**74.3%**
Syllables + DNN boundaries	3.2 k	0.1%	75.6%

Table 2: Weighted OOV rate and SI WER on Inuktut; acoustic model trained on 40 hours of audio.

For East Cree, a word-based LM with a 30,000 word dictionary obtains a much lower OOV rate. Training a word-based LM for Cree from much less text data than for Inuktut (260,000 words of text from Cree organisation reports scriptures) yields only 25% OOV rate on video stories and 9% OOV rate on scriptures. Video stories differ from the LM text, and thus have a higher OOV rate than the scriptures development text. Decoding Cree speech using this LM yields a 69.0% WER on video stories (speaker independent = SI WER) and 24.6% on scriptures (speaker dependent = SD WER). Note that SI WER is lower for Cree than for Inuktut, even when Inuktut benefits from twice as much acoustic training data (80 hours instead of 40 hours) and far more LM training data: 69.0% WER for Cree vs 72.3% for Inuktut.

The speaker-dependent (SD) ASR result for East Cree above is good news: with a few hours of transcribed audio from a single speaker, the WER on new data from the same speaker was 24.6%. Phoneme error rate (PER) in this SD condition was 8.7%. This is well below the 30% PER considered good enough to significantly speed up the manual transcription process (if transcription is done by a non-native speaker, as is often the case in field linguistics) (Adams et al., 2018). It took only 3 hours of transcribed audio to achieve this result. A similar SD experiment was run with Inuktut, with 3 hours of training from one speaker, and the same syllable LM as for the SI case. This resulted in 67.3% WER and 18.4% PER. So even for the SD case, Inuktut has higher WER and PER than East Cree. Experiments on other Indigenous polysynthetic languages are needed to see where they fall along the East Cree to Inuktut spectrum, in terms of SI and SD ASR difficulty.

Field linguists often record many hours of speech from each of a very small number of fluent speakers: the Elders of an Indigenous community. If it turns out that good SD ASR is possible for several Indigenous languages, one could imagine a common mode of work in which an SD system is trained on the first few hours of speech from an Elder, then used to produce a first-draft transcription of the remaining hours.

4.2 Read-along audiobooks

A subproject on software that supports "read-along/sing-along" activities was inspired by the East Cree online activities pioneered at Carleton University (see eastcree.org). Interactive read-along/sing-along audiobooks that highlight words as they are spoken and allow students to click on words to hear them aloud are well liked by both students and teachers, but their creation has been labour intensive, requiring expertise with specialized software like ELAN or Audacity. A collaboration between Carleton University, NRC, and David Huggins-Daines (Nuance Communications Montreal) seeks to make the creation of such activities quick and easy, without requiring the creator to manually align each word.

Fortunately, text/audio alignment (also called "forced alignment") is feasible to perform in a "zero-shot" scenario (that is, where there is *no* data available in the language in question). One constructs an approximate mapping between target language phonemes and phonemes in a high-resource "donor" language (in our case, English), converts the target document into phones in the donor language, and trains an acoustic model on the donor language to recognize when each word is spoken. This is the assumed default when working with a new language in the Festival toolkit (Black et al., 1998). Based on this concept, our ReadAlong Studio[9] combines a custom G2P engine, the PanPhon (Mortensen et al., 2016) phonetic distance library, and the lightweight PocketSphinx (Huggins-Daines et al., 2006) speech recognition library to allow a non-expert user to create a text/speech alignment system for a new language.

ReadAlong Studio currently supports 22 languages; among Indigenous languages spoken in Canada it supports Anishinaabemowin (Ojibway), Atikamekw, Dakelh, East Cree, Gitxsan, Heiltsuk, Inuktut, Kanyen'kéha, Kwak'wala, SENĆOŦEN, Seneca, Tagish, Tŝilhqot'in, and Tsuut'ina (the other eight are not spoken in Canada). Adding a new language is the work of only a few hours, depending on the complexity of the language's orthography. ReadAlong Studio can already create high-quality interactive webpages (see Figure 3), MP4 movies, and EPUB documents, and can export in ELAN, TextGrid (PRAAT), and subtitle formats. A user-friendly interface is currently in development.

5 Other Subprojects

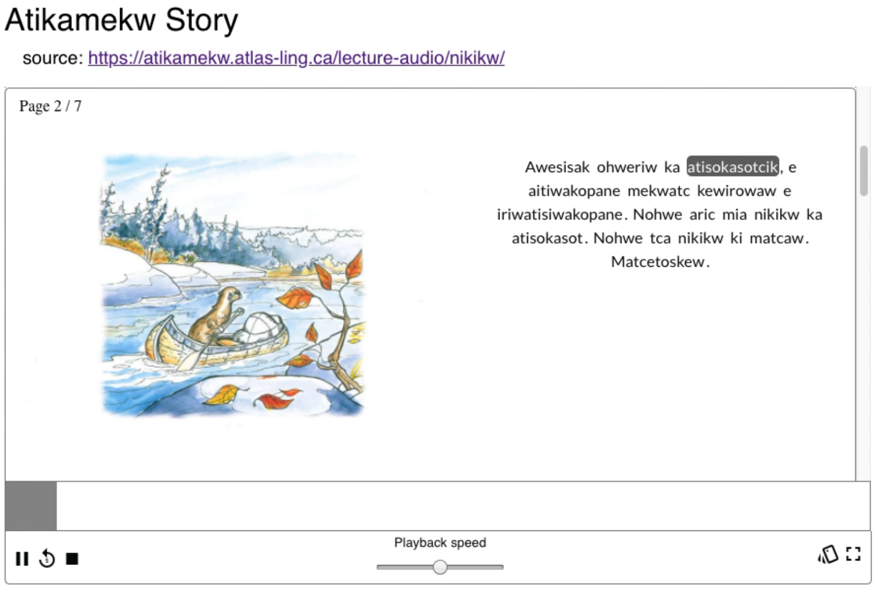
Space is lacking to describe several other subprojects of the NRC project (for these, see Kuhn et al. (2020)). They include subprojects carried out by external organizations (mostly Indigenous-run) but funded by NRC, such as enhancement of online language courses for East Cree and Innu previously developed at Carleton University, development of online courses for Plains Cree, Kwak'wala, Michif, and Naskapi, improvements to a role-playing game with Swampy Cree content, two subprojects for training Indigenous language activists in data collection methodologies, and data collection efforts for Plains Cree, Kanyen'kéha, Kwak'wala, Michif, Nsyilxcn, SENĆOŦEN, Tŝilhqot'in, and Tsuut'ina.

6 Conclusions and Future Work

One of the reviewers of this paper pointed out that its main scientific finding is that "money works!" Strategically deployed, funding for research into technologies for neglected languages can have a significant impact. This reviewer even suggested that the project described here might have lessons for governments in other parts of the world which seek to help minority languages.

There is a caveat: the research must be done respectfully, in collaboration with the communities who are the custodians of these languages. There should be no room in this field for scholars whose main goal

[9] https://github.com/ReadAlongs/Studio



Atikamekw Story

source: https://atikamekw.atlas-ling.ca/lecture-audio/nikikw/

Page 2 / 7

Awesisak ohweriw ka atisokasotcik, e aitiwakopane mekwatc kewirowaw e iriwatisiwakopane. Nohwe aric mia nikikw ka atisokasot. Nohwe tca nikikw ki matcaw. Matcetoskew.

Figure 3: The ReadAlong Web Component displaying the Atikamekw story ‘Nikikw’

is extracting “interesting” research without considering the linguistic needs of communities. Our greatest challenge in this project was doing our best to build respectful relationships with communities.

The two goals do not necessarily conflict. There are many interesting research themes that have the potential to help under-resourced languages. Examples in this project include implementing verb conjugators, machine translation, and speech recognition for polysynthetic languages. But there are also low-hanging fruit, challenges that do not require advanced research — they can be tackled with today’s technologies — but which, when overcome, can have linguistic benefits for communities. An example is the production of Read-along audio books in Indigenous languages. Automating this process posed a technical challenge of only moderate difficulty, but educators find the end product very useful.

Our focus in Phase II is on making the technologies described above easier for Indigenous communities to deploy themselves. For instance, we would like to make the Read-along interface so user-friendly that there is no need to consult the NRC team to produce Read-along versions of audio books (the team is having trouble keeping up with demand). We have a similar goal for text prediction on mobile devices. It will be harder to develop a framework that enables Indigenous educators to build software-based verb conjugators for their languages without requiring the educators to have specialized IT knowledge, but we are working on this as well. Ideally, we will make all the technologies described above more user-friendly. All software developed by NRC’s ILT project team will be released as open-source.

Acknowledgements

We would like to acknowledge the wise counsel of the hard-working volunteers on the ILT project’s Indigenous Advisory Committee (IAC): Nathan Thanyehténhas Brinklow, Tessa Erickson, Amanda Evic-Kuluguqtuq, Blaire Gould, Gerry Lawson, Delaney Lothian, Megan Lukaniec, Onowa McIvor, Glenn Karonhiio Morrison, Marilyn Shirt, Tina Jules Skayda.û, and Heather Souter.

We’re also grateful to several other people who are neither authors of this paper nor members of the IAC but helped the project in other ways: these include Stéphane Cloutier, Ryan DeCaire, Krista Dempster, Sarah Dupont, Carrie Demmans Epp, Riel Gallant, Lucy Hemphill, Bill Jancewicz, Michele Johnson, Gerry Lawson, Alexa Little, Jodi Lynn Maracle, Jeffrey Micher, Yvette Mollen, Timothy Montler, Mimie Neacappo, Christina Neilsen, Gavin Nesbitt, Aliana Parker, Aaron Plahn, Laura Prosper, Rohahí:yo, Ronkwe’tiyóhstha, Michael Running Wolf, Lorne Shapiro, Tye Swallow, and Shawn Tsosie.

Finally, we would like to thank the three reviewers of this paper, whose comments made it stronger.

References

Oliver Adams, Trevor Cohn, Graham Neubig, Steven Bird, and Alexis Michaud. 2018. Evaluating phonemic transcription of low-resource tonal languages for language documentation. In *Proc. LREC*, pages 3356–3365.

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Publications.

Alan W. Black, Paul Taylor, and Richard Caley. 1998. The Festival speech synthesis system. http://www.festvox.org/festival.

Nathan Thanyehténhas Brinklow, Patrick Littell, Delaney Lothian, Aidan Pine, and Heather Souter. forthcoming. Indigenous language technologies & language reclamation in Canada. In *LT4ALL: Enabling Linguistic Diversity and Multilingualism Worldwide*. UNESCO.

Deborah Cameron, Elizabeth Frazer, Penelope Harvey, M. B. H. Rampton, and Kay Richardson. 1992. *Researching language: Issues of power and method*. Routledge.

Michael J. Chandler and Christopher Lalonde. 1998. Cultural continuity as a hedge against suicide in Canada's First Nations. *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. http://hdl.handle.net/10125/44841. Presentation at the 6th International Conference on Language Documentation and Conservation (ICLDC), University of Hawai'i at Mānoa, Honolulu, HI.

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

Fineen Davis and Eddie R. Santos. under review. On the computational modelling of Michif verbal morphology.

Alain Désilets, Benoît Farley, Geneviève Patenaude, and Marta Stojanovic. 2008. WeBiText: Building large heterogeneous translation memories from parallel web content. In *Proceedings of Translating and the Computer*, volume 30. International Association for Advancement in Language Technology.

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

Benoît Farley. 2012. The Uqailaut project. http://www. inuktitutcomputing.ca.

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 *LREC*, pages 2521–2527. European Language Resources Association (ELRA).

Vishwa Gupta and Gilles Boulianne. 2020b. Speech transcription challenges for resource constrained Indigenous language Cree. In *Proceedings of 1st Joint SLTU-CCURL Workshop*, pages 362–367.

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

David Huggins-Daines, Mohit Kumar, Arthur Chan, Alan W. Black, Mosur Ravishankar, and Alexander I. Rudnicky. 2006. Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I. IEEE.

Mans Hulden. 2009. Foma: a finite-state compiler and library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 29–32. Association for Computational Linguistics.

Eric Joanis, Rebecca Knowles, Roland Kuhn, Samuel Larkin, Patrick Littell, Chi kiu Lo, Darlene Stewart, and Jeffrey Micher. 2020. The Nunavut Hansard Inuktitut–English parallel corpus 3.0 with preliminary machine translation results. In *Proceedings of LREC-2020*.

David Kanatawakhon. 2002. *Yonteweyenhstahkwa Kanyen'kéha: a Mohawk Teaching Dictionary*. Centre for Research and Teaching of Canadian Native Languages, University of Western Ontario.

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.

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

Roland Kuhn, Fineen Davis, Alain Désilets, Eric Joanis, Anna Kazantseva, Rebecca Knowles, Patrick Littell, Delaney Lothian, Aidan Pine, Caroline Running Wolf, Eddie Santos, Darlene Stewart, Gilles Boulianne, Vishwa Gupta, Owennatekha Brian Maracle, Akwiratékha' Martin, Christopher Cox, Marie-Odile Junker, Olivia Sammons, Delasie Torkornoo, Nathan Thanyehténhas Brinklow, Sara Child, Benoît Farley, David Huggins-Daines, Daisy Rosenblum, and Heather Souter. 2020. The Indigenous languages technology project at NRC Canada: an empowerment-oriented approach to developing language software. https://nrc-publications.canada.ca/eng/view/object/?id=d4f10144-c711-43c5-b80b-5ace7df5e68b. Technical report on the publications archive of the National Research Council of Canada.

Jason Edward Lewis, Angie Abdilla, Noelani Arista, Kaipulaumakaniolono Baker, Scott Benesiinaabandan, Michelle Brown, Melanie Cheung, Meredith Coleman, Ashley Cordes, Joel Davison, Kūpono Duncan, Sergio Garzon, D. Fox Harrell, Peter-Lucas Jones, Kekuhi Kealiikanakaoleohaililani, Megan Kelleher, Suzanne Kite, Olin Lagon, Jason Leigh, Maroussia Levesque, Keoni Mahelona, Caleb Moses, Isaac ('Ika'aka) Nahuewai, Kari Noe, Danielle Olson, 'Ōiwi Parker Jones, Caroline Running Wolf, Michael Running Wolf, Marlee Silva, Skawennati Fragnito, and Hēmi Whaanga. 2020. Indigenous protocol and artificial intelligence position paper. Project Report 10.11573/spectrum.library.concordia.ca.00986506, Aboriginal Territories in Cyberspace, Honolulu, HI. Edited by Jason Edward Lewis.

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 *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2620–2632.

Brian Maracle. 2017. *Anonymous 1st Year Adult Immersion Program 2017-18*. Onkwawenna Kentyohkwa, Ohsweken, ON, Canada. The book was co-written by several other staff members over the years. Brian Maracle is the author of the latest, 2017 edition.

Doug Marmion, Kazuko Obata, and Jakelin Troy. 2014. *Community, identity, wellbeing: the report of the Second National Indigenous Languages Survey*. Australian Institute of Aboriginal and Torres Strait Islander Studies Canberra.

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. Panphon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484.

Mary Jane Norris. 2018. The state of Indigenous languages in Canada: Trends and prospects in language retention, revitalization and revival. *Canadian Diversity*, 15(1):22–31.

Richard T. Oster, Angela Grier, Rick Lightning, Maria J. Mayan, and Ellen L. Toth. 2014. Cultural continuity, traditional Indigenous language, and diabetes in Alberta First Nations: A mixed methods study. *International journal for equity in health*, 13(1):92.

Aidan Pine and Mark Turin. 2017. Language revitalization. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.

Ian Pool. 2016. Colonialism's and postcolonialism's fellow traveller: the collection, use and misuse of data on indigenous people. *Indigenous Data Sovereignty*, pages 57–76.

Jon Reyhner. 2010. Indigenous language immersion schools for strong Indigenous identities. *Heritage Language Journal*, 7(2):138–152.

Keren Rice. 2008. Indigenous languages in Canada. In *The Canadian Encyclopedia*. Historica Canada. https://www.thecanadianencyclopedia.ca/en/article/aboriginal-people-languages (Accessed on May 6, 2020.).

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.

Olivia N. Sammons. 2019. *Nominal classification in Michif.* Ph.D. thesis, University of Alberta, DOI: 10.7939/r3-b8sq-xz05.

Lane Schwartz, Francis Tyers, Lori Levin, Christo Kirov, Patrick Littell, Chi-kiu Lo, Emily Prud'hommeaux, Hyunji Hayley Park, Kenneth Steimel, Rebecca Knowles, Jeffrey Micher, Lonny Strunk, Han Liu, Coleman Haley, Katherine J. Zhang, Robbie Jimmerson, Vasilisa Andriyanets, Aldrian Obaja Muis, Naoki Otani, Jong Hyuk Park, and Zhisong Zhang. 2020. Neural polysynthetic language modelling. https://arxiv.org/abs/2005.05477.

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.

Daan van Esch, Elnaz Sarbar, Tamar Lucassen, Jeremy O'Brien, Theresa Breiner, Manasa Prasad, Evan Crew, Chieu Nguyen, and Françoise Beaufays. 2019. Writing across the world's languages: Deep internationalization for Gboard, the Google keyboard. https://arxiv.org/abs/1912.01218.

Douglas H. Whalen, Margaret Moss, and Daryl Baldwin. 2016. Healing through language: Positive physical health effects of indigenous language use. *F1000Research*, 5.

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