NRC·CNRC

NRC Publications Archive
Archives des publications du CNRC

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, Owennatekha Brian; Martin, Akwiratékha'; Cox, Christopher; Junker, Marie-Odile; Sammons, Olivia; Torkornoo, Delasie; Brinklow, Nathan Thanyehténhas; Child, Sara; Farley, Benoît; Huggins-Daines, David; Rosenblum, Daisy; Souter, Heather

For the publisher's version, please access the DOI link below./ Pour consulter la version de l'éditeur, utilisez le lien DOI ci-dessous.

https://doi.org/10.4224/40001304

NRC Publications Archive Record / Notice des Archives des publications du CNRC :
https://nrc-publications.canada.ca/eng/view/object/?id=d4f10144-c711-43c5-b80b-5ace7df5e68b
https://publications-cnrc.canada.ca/fra/voir/objet/?id=d4f10144-c711-43c5-b80b-5ace7df5e68b

Access and use of this website and the material on it are subject to the Terms and Conditions set forth at
https://nrc-publications.canada.ca/eng/copyright
READ THESE TERMS AND CONDITIONS CAREFULLY BEFORE USING THIS WEBSITE.

L'accès à ce site Web et l'utilisation de son contenu sont assujettis aux conditions présentées dans le site
https://publications-cnrc.canada.ca/fra/droits
LISEZ CES CONDITIONS ATTENTIVEMENT AVANT D'UTILISER CE SITE WEB.

Questions? Contact the NRC Publications Archive team at PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca. If you wish to email the authors directly, please see the first page of the publication for their contact information.

Vous avez des questions? Nous pouvons vous aider. Pour communiquer directement avec un auteur, consultez la première page de la revue dans laquelle son article a été publié afin de trouver ses coordonnées. Si vous n'arrivez pas à les repérer, communiquez avec nous à PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca.

National Research Council Canada | Conseil national de recherches Canada

Canada

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

Owennatekha 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 (e.g., ChristopherCox@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
S̲anya̲kola 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

Contents

Abstract

This paper describes a three-year project at the National Research Council of Canada aimed at developing software to assist Indigenous communities in their efforts to preserve their languages and extend their use. The project aimed to work within the empowerment paradigm, where the linguistic goals of communities have at least equal weight with those of the researchers, and where collaboration with communities is central. Because many of the technological directions we took 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 (a verb conjugator for Kanyen'kéha, in the Iroquoian language family, was built in the framework), 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 recordings of speech in Indigenous languages (and other languages), limited-domain text-to-speech synthesis for some Indigenous languages, and several other subprojects.

1 Introduction

This paper describes the Indigenous Languages Technology (ILT) project at the National Research Council of Canada (henceforth NRC). This project came into existence thanks to funding of $6 million over three years, granted by the Canadian government's Treasury Board Secretariat according to the provisions of the March 2017 budget. The project's primary goal was to serve Indigenous communities by producing software that would enhance their efforts to preserve and revitalize their languages. The ILT project did not primarily aim at carrying out academic research, though interesting academic research resulted as a by-product. Some ILT activities were carried out by employees of NRC, others were financially supported by NRC but carried out by other organizations. A second phase of the project began in April 2020.

The ILT project should be viewed in the context of a much larger set of efforts being carried out by a variety of organizations to deploy technologies that promote the revitalization and documentation of Indigenous languages spoken in Canada. Littell et al. (2018a) surveys these efforts. The current paper depicts a tiny corner of a big canvas.

The activities carried out within the ILT project were extremely heterogeneous. Different communities have very different linguistic needs, so, given our primary goal of serving communities, ILT was made up of a diverse set of subprojects. Another aspect of the project was that there is so little textual or speech data for Indigenous languages in Canada (with the partial exception of Inuktut) that they could be characterized in the fields of natural language processing and speech recognition as **extremely low-resource** languages. Googling the term "low-resource language" brings up Sindhi as an example. Sindhi is spoken by about 23 million people. Inuktut is spoken by about 40,000 people—and that is far more speakers than most Indigenous languages in Canada have. Thus, most of the technologies developed within the ILT project have been rule-based, rather than relying on data-driven machine learning.

Drawing on Cameron et al. (1992) and Czaykowska-Higgins (2009), one can list three approaches to linguistic research involving Indigenous languages: 1. the "linguist-focused" or "ethical research" model, in which members of the speech community are the passive subjects of research. They are "informants" from whom the linguist in control collects data. 2. the "advocacy research" model, where the linguist is still in a position of power, but the research is carried out not only by and for the linguist, but also partly for the benefit of the speech community. 3. the "empowerment" model, in which research is carried out with equal emphasis on the agenda of the linguist and of the community. The focus is on collaboration and dialogue.

Most of the work carried out within the ILT project fits into the empowerment model. The most ambitious subproject described in this paper was suggested to us by an Indigenous educator: the creation of a verb conjugator for an Iroquoian language, Kanyen'kéha (Mohawk). Later, when a prototype for the verb conjugator was shown to students of Kanyen'kéha, the first question they asked was whether it would have audio. This led us to research text-to-speech (TTS) capabilities, in order to create a version of the system that can speak hundreds of thousands of unique verb forms out loud. Similarly, the "readalong" subproject for automating word-speech alignment for audio books was in response to strong interest from several communities after a research team at Carleton University had aligned several audio books manually. None of these research themes would have occurred to members of the team at NRC—they were driven by demand from communities. In other cases, members of the NRC team suggested several possible technologies to communities, then focused on the ones that members of the communities thought would be helpful.

Several subprojects were almost purely community-driven, with little direct involvement by members of the NRC team, such as those focusing on speech data collection for the Cree, Kwak'wala, Michif, Nsyilxcn, SENĆOŦEN, and Tŝilhqot'in languages. NRC's role in these subprojects was mainly to provide funding, with little micromanagement (though the team followed their progress carefully).

To ensure that the project did not stray too far from the empowerment model, or inadvertently engage in unethical behaviour, the NRC team enlisted a group of volunteers to guide the project: an Advisory Committee, made up of Indigenous people with expertise in language revitalization. Their counsel has been invaluable. For a list of the members of this committee, scroll down to the end of the project web page.[1]

An important part of our approach was that at no stage did NRC claim ownership of Indigenous language data collected with the project's funding. Furthermore, all software generated by the ILT project has been or will be released as open source. **This refusal by the NRC to retain intellectual property in software or data generated by the project** was one of the project's guiding principles from the beginning. We were determined to break with the unfortunate history of academics and government departments often refusing to return linguistic data to the Indigenous communities from which it was collected.

After outlining the current state of Indigenous languages in Canada, we will discuss the five types of ILT subprojects: text-based subprojects, speech-based subprojects, creation of online courses and educational games, capacity-building subprojects, and subprojects focused on collection of new data.

2 Sociolinguistic Background

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

All Indigenous languages in Canada were targeted by deliberate government policies that sought to eradicate their use in the course of the late 19th century to the end of the 20th century. These policies focused heavily on creating fear around speaking Indigenous languages and ending the generational transmission of such languages. They were implemented through government legislation such as the Indian Act,[2] which discouraged, and often made illegal, gathering for cultural practices and speaking ancestral languages.

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

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

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)). The Truth and Reconciliation Commission of Canada recently led an inquiry into the atrocities, which included physical and sexual abuse, committed during the residential school era in Canada from 1883 to 1996. In 2015, the Commission released a report confirming a reality of these schools that Indigenous people have known all along, that 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). Some children were sent to day schools with the same assimilationist philosophy as the residential schools; these were also damaging.

Although the negative impacts of the effort to eliminate Indigenous languages in Canada are deep and widespread, this effort did not succeed. The resilience of Indigenous language communities can be seen in the myriad of 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). Research in psychology has shown 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).

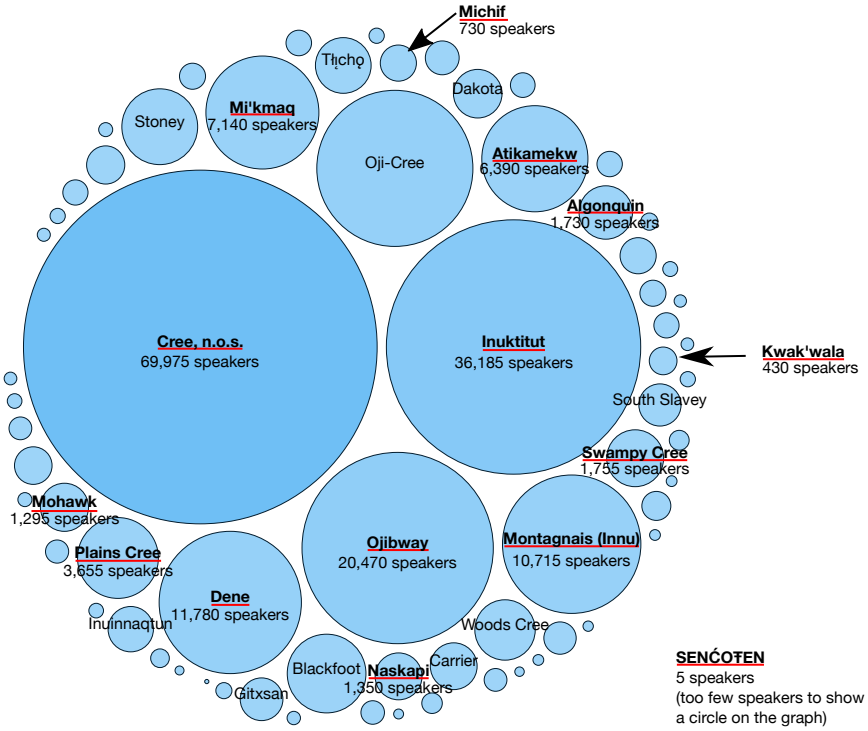

Figure 1: Number of speakers of Indigenous languages spoken in Canada, according to the 2016 census (From Statistics Canada (2017))

Figure 1 (obtained from the Statistics Canada website) shows the number of speakers per language in the second decade of the 21st century (Statistics Canada, 2017). The languages underlined in red are those with which the ILT project has interacted in some way. These census figures are controversial among linguists. For one thing, they imply that each language is a discrete entity, while many of them are more accurately viewed as a collection of dialects that blend into each other as one moves geographically from one community to another (e.g., Cree has been subdivided in different ways by different linguists).

Some of the details about specific languages may be inaccurate, too.[3] 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.

Speakers of the "larger" languages in the figure tend to have different linguistic priorities from those of "smaller" languages. Speakers of Inuit language varieties, such as Inuktitut and Inuinnaqtun, who live in the territory of Nunavut constitute a numerical majority there, and are in the unique position of having these varieties, grouped together under the name Inuktut, designated as official in the territory (along with English and French). Thus, some of their concerns focus on ensuring that the Nunavut bureaucracy is capable of providing services in Inuktut, and that it is taught well by the local school system. (The Northwest Territories recognize three Inuit languages—Inuktitut, Inuinnaqtun, and Inuvialuktun—among the eleven languages recognized there, but outside the court system, there is no guarantee that government services will be provided in these languages.)

A common linguistic phenomenon in many Indigenous communities is a generational split where the majority of fluent speakers, especially those taught as children, largely belong to the eldest generations of grandparents and great grandparents. Today, many communities face decreasing numbers of first language (mother tongue) speakers, due to declining language transmission rates to the younger generations (Norris, 2018), likely due to the assimilation efforts described above. That is why much Indigenous language revitalization work in Canada focuses on preservation of language through recording and transcribing the speech of Elders before their language knowledge is lost with them. Many of the language learning technologies that are being created, and for which these recordings and transcriptions can be a vital resource, are to support the learning of younger generations.

3 Text-based subprojects

In this section, we discuss subprojects based around text. They include work on polysynthetic verb conjugation for language teaching, work on tools for the Inuktut language, and work on text prediction for mobile devices.

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 a community-based 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", that is spoken in territory that spans present-day Ontario, Quebec, and New York State. All Iroquoian languages are highly polysynthetic; that is, their words are composed of many morphemes. Thus, single verbs in Kanyen'kéha are routinely translated as entire sentences in English.

The proposed verb conjugator was a tool that could provide a helpful reference for learners and teachers alike, as learning and teaching verbs in Kanyen'kéha is a formidable task. The equivalent for the French language is the pocket Bescherelle: except that polysynthetic languages may have millions of conjugations for even the most common verbs, far too many to print out. It is therefore only possible to create a verb conjugator with reasonable coverage for Kanyen'kéha in software, not on paper. Verb roots in the language are bound morphemes, meaning that 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

[3] For instance, Dr. Marie-Odile Junker writes: "the East Cree speakers numbering 18,000 are partly bundled under Cree. There was a forest fire in Northern Ontario that year and most people were moved South and did not reply to the Census (affecting Oji-Cree, Moose Cree and Swampy Cree)."

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 significantly larger than in English, French, or other European languages. This complexity adds to the difficulty of teaching the language, as learners cannot be expected to memorize paradigms arbitrarily, and 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).

3.1.2 WordWeaver

Given that many Indigenous languages are polysynthetic and have rich inflectional morphology, the NRC team wanted to take Owennatékha's request and build a tool that could be extended to other languages as well. WordWeaver is a tool designed to do exactly that. 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 built in a way that was tightly coupled to the instance's language model, specifically Foma (Hulden, 2009). However, the language model was later decoupled from WordWeaver architecture in favour of storing all data in a database. The improvements are several-fold: WordWeaver UI can now run offline using PouchDB, it is typically faster (depending on the queries), and the WordWeaver code base has been significantly simplified without the requirement of encoding and serializing HTTP arguments to validate up-side tags of a finite-state transducer (FST) and then decoding and de-serializing them in response. Additionally, this allows verb conjugators to be made in a wide variety of ways without necessarily requiring specialist knowledge of building FSTs (see Sections 3.1.6 and 9.1 for additional discussion).

3.1.3 WordWeaver Instances

The first instance of WordWeaver, called Kawennón:nis, models the Western dialect of Kanyen'kéha that is taught at Onkwawenna Kentyohkwa. 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. Additional sample instances have been made for languages whose data is publicly available, including French.

To design the rules encoded in the first Kawennón:nis FST, we relied on a textbook that describes the Western dialect (Maracle, 2017), along with writings on other dialects of Kanyen'kéha and the closely related language Oneida. Even more important, we benefited from a close, respectful relationship with the staff of Onkwawenna Kentyohkwa, who added hundreds of new verbs to the system through a collaborative development process that allows new verb stems to be declared in a spreadsheet and to then be compiled into a valid FST lexicon formalism (lexc).

Quality control for this Western version of Kawennón:nis was done both by teachers at Onkwawenna Kentyohkwa and by NRC researchers. Members of NRC made several in-person visits to Onkwawenna Kentyohkwa to demonstrate Kawennón:nis to the students and staff, and to participate in multi-day collaborative sessions to design, evaluate, and improve the user interface. Creation of the Eastern (Kahnawà:ke) version relied heavily on the expertise of a single gifted individual, Akwiratékha' Martin.

The current 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 UI

The user interface is of prime importance, if Kawennón:nis is to be useful to students. The process for prototyping, designing, and evaluating the WordWeaver UI was extensive, involving multiple in-person visits for defining the requirements of the UI, hiring in-community designers, and extensive user

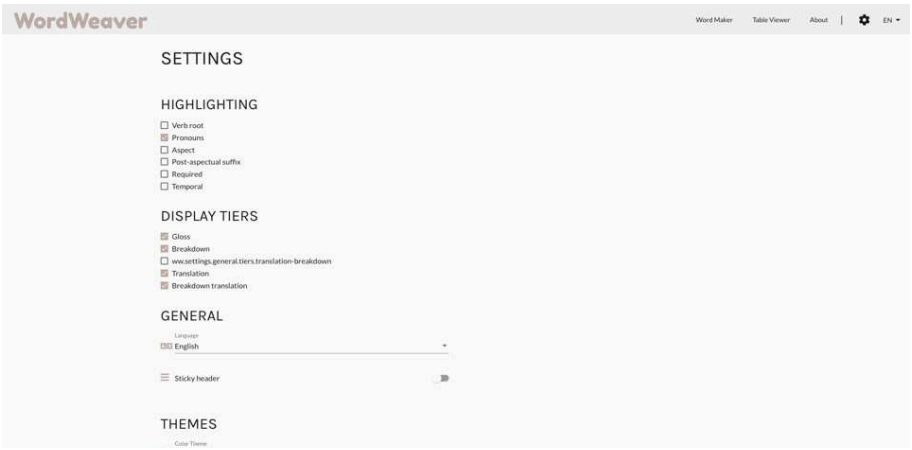

Figure 2: WordWeaver UI application settings: in English using the ‘light theme’

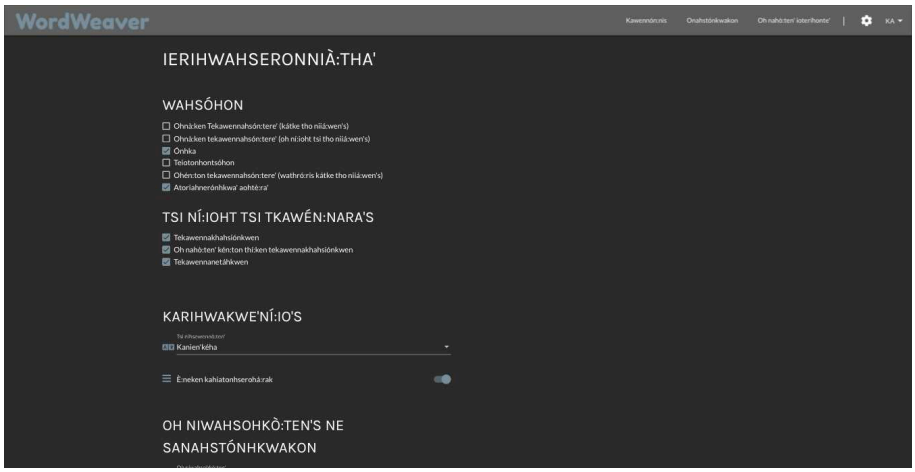

Figure 3: WordWeaver UI application settings: in Kanyen’kéha using the ‘dark theme’

interface and user experience (UI/UX) review. The resulting UI is interactive, highly theme-able, entirely translated in English, French and Kanyen’kéha (see Figures 2 and 3), 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. This reflects three basic categories in WordWeaver generally. 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 suggested by the demonstration version of WordWeaver.

The Tableviewer is the more advanced view, but allows users to create paradigm tables of conjugations instead of single output forms. Here users can non-linearly 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 4) or in a ‘tree’ format (see Figure 5). 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.

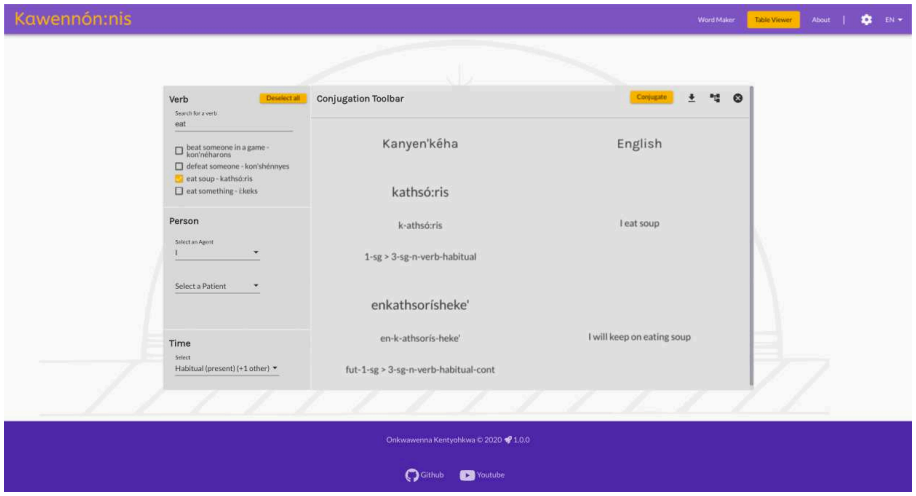

Figure 4: Tabular grid view of the Kawennón:nis Tablemaker using the ‘dark purple theme’

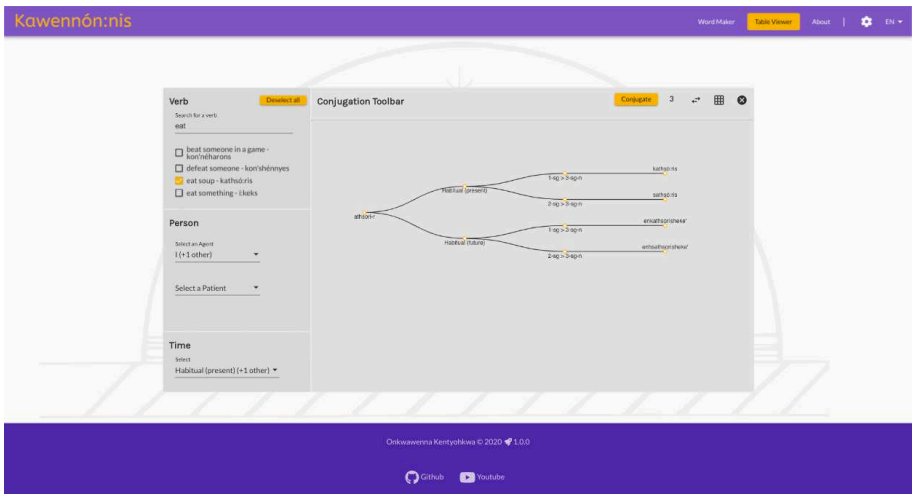

Figure 5: ‘Tree’ view of the Kawennón:nis Tablemaker using the ‘dark purple theme’

3.1.5 Work on Michif

The work on a verb conjugator for Michif began after the NRC team was asked by an influential language revitalization activist, Heather Souter of the Prairies to Woodlands Indigenous Language Revitalization Circle (P2WILRC), whether NRC could implement a system like Kawennón:nis for Michif. Kanyen’kéha and Michif are not related, but they are similar in that they are both highly polysynthetic languages. The same problem is faced by Michif as by other Indigenous languages: there is a shortage of data and formal documentation for the language, which are needed to create teaching tools.

Michif is a mixed language which arose during the 19th century from the intermarriage between French fur traders and Cree and Ojibwe women in the Métis homeland, stretching from present-day north-western Ontario across the Prairies to the north-east of British Columbia and from southern Northwest Territories to North Dakota and Montana in the south (Rosen and Souter, 2009). Their descendants are the Métis people, whose official language is Michif. As a mixed language, Michif takes most of its nominal patterns from French, and its verbal patterns from Cree. There is also a high degree of regional variation (Sammons, 2019).

With Ms. Souter's help, the NRC team has been building `MichifVerb`, a finite-state transducer (FST) which models the verbal morphology of Michif according to the lexc formalism stated in Beesley and Karttunen (2003). The current implementation allows for the conjugation of 22 verb stems, which generates 6791 possible verb forms.[4] Though `MichifVerb` 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 `MichifVerb` 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. We hope these features will promote transmission by giving learners the opportunity to ask about the language directly from speakers in informal Métis community settings instead of limiting learning to classrooms.

3.1.6 Discussion

Work on Kawennón:nis and `MichifVerb` illustrates two of the themes of the ILT project: the use of rule-based approaches and close collaboration with members of an Indigenous community in pursuit of a goal set by that community, rather than by researchers. The goal in both cases has been to develop an assistive reference tool that does not replace learners' experience of acquiring verbal morphology at community schools, but that complements it.

As mentioned, the described reference tools rely on finite-state transducers which are rule-based. In modern-day computational language modelling, rule-based approaches may seem outdated in contrast to more widely used statistical methods. However, as with most polysynthetic languages, existing corpora are not large enough to produce statistical models that are sufficiently accurate. That being said, rule-based methods come with their own set of disadvantages. By definition, they must be built from a standard model. This means imposing a standard onto languages that were never previously standardized and continue to vary from community to community. Through close collaboration with our Indigenous community partners and working with them as experts of the language, we have tried to mitigate these limitations.

Several educators who teach Indigenous languages other than Kanyen'kéha and Michif have expressed a desire for a verb conjugator teaching tool. To facilitate this, WordWeaver has been designed to be as language-independent as possible. We discuss the challenge this poses in terms of human resources in Section 8 and discuss work on tools to lower the barrier for building verb conjugators in Section 9.1.

3.2 Corpus and Tools for Inuktut (including machine translation)

The Inuktut language, a member of the Inuit-Yupik-Unangan language family, is a polysynthetic language spoken across Arctic Canada. The Government of Nunavut uses the term Inuktut to represent all of the Inuit language varieties spoken in Nunavut, including Inuktitut and Inuinnaqtun. Inuktut is an official language of the Territory of Nunavut; Inuvialuktun, Inuinnaqtun, and Inuktitut are official languages of the Northwest Territories,[5] and Inuit languages have recognition in additional regions.

As part of the ILT project and with support from the Government of Nunavut, the NRC team (Joanis et al., 2020) released the sentence-aligned Inuktut–English "Nunavut Hansard" corpus based on the proceedings of the Legislative Assembly of Nunavut, covering sessions from April 1999 to June 2017. To ensure a high quality of alignment, we relied on the expertise of speakers of Inuktut from the Pirurvik Centre. With approximately 1.3 million aligned sentence pairs, this is, to our knowledge, the largest parallel corpus of a polysynthetic language or an Indigenous language of the Americas released to date. The

[4] See version 1.0.0-alpha at https://github.com/finguist/MichifVerb.

[5] https://www.ece.gov.nt.ca/en/services/le-secretariat-de-leducation-et-des-langues-autochtones/languages-overview

corpus is available at the NRC Digital Repository[6] under the CC-BY-4.0 license.

Because of the size of the Nunavut Hansard corpus, this is an ILT subproject in which we were able to apply machine learning techniques to data: Joanis et al. (2020) describe our preliminary experiments on statistical and neural machine translation between Inuktut and English, in both directions. A pre-release version of the same corpus was also used for Inuktut–English and Yupik language machine translation experiments during the 2019 JSALT Workshop on Neural Polysynthetic Language Modelling (Schwartz et al., 2020). The existence of the Nunavut Hansard has made possible a shared task for the Inuktut–English language pair in the Workshop for Machine Translation,[7] with the evaluation taking place in May/June 2020. We hope to make tools based on machine translation available to translators during the second phase of the project.

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 Inuktut–English sentence pairs from a parallel corpus (in practice, the portion of the Nunavut Hansard that was available at the time). Despite its unidirectionality, translators found it extremely useful. Unfortunately, funding for the WeBInuk online service lapsed for several years.

One of the goals of the ILT project was to revive WeBInuk, and to make it bidirectional—i.e., to allow users to enter Inuktut search terms as well as English ones. In addition, the project aimed at providing tools that are currently unavailable for Inuktut: a comprehensive dictionary, a gister, a spell checker, and so on. A member of the NRC team and experts from the Pirurvik Centre began building versions of some of those tools.

They faced a number of challenges. Inuktut words resemble short phrases in English. They are composed by stringing together: a **root** (about 2000 possibilities, versus about 500,000 in English) 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 tend to be invariant, with minor exceptions, many Inuktut morphemes have surface forms that change in different contexts. Furthermore, the same surface form may correspond to different morphemes.

The first phenomenon is illustrated by the word "umiarjualiuqtiujumajunga" *(I want to be a ship builder)*. It is made of

"umiar": from umiaq, *boat*, with final Q voiced to R by following morpheme;
"jua": from juaq, *big*, with final Q deleted by following morpheme;
"liuq": from liuq, *to build*;
"ti": form of ji after consonant, *one whose occupation is to ...*;
"u": *to be*;
"juma": form after vowel, *to want*;
"junga": verb ending after vowel, 1st person non-specific with no indication of the object, in the declarative mood.

Since many, perhaps most, Inuktut words in a given text will not have occurred before, building a word-based dictionary with good coverage is unrealistic. The only realistic method for increasing coverage is to decompose the Inuktut word into a sequence of morphemes, then to 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; we would expect that users would frequently enter words for which exact matches cannot be found in the parallel corpus. Instead, one might prefer to look up Inuktut words in the parallel corpus that share the root of the user-entered word (and perhaps some other key morphemes).

[6] https://doi.org/10.4224/40001819
[7] http://www.statmt.org/wmt20/

Changes in surface forms make decomposing words far from trivial. Nevertheless, by building on a morphological analyzer created earlier by one of them (Benoît Farley), the NRC-Pirurvik collaborators 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 revised version of WeBInuk;
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 search engine that, given an Inuktut word, searches for its five most frequent morphological variants.

All of these need to be sped up. Otherwise, the first three are almost ready to be released. The last two, the spell checker and the search engine, require testing by speakers of the language to see how useful they are in practice.

3.3 Predictive Text

Writers in majority languages have benefited from predictive text suggestions on their smartphone keyboards. Predictive text is the bar on top of smartphone keyboards that suggests words as you type, and makes corrections to mistyped words (Figure 6). The impetus for creating predictive text was to make it more accessible and less error-prone to type on tiny touchscreen 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 in Canada either do not have robust or extensive text corpora, or it is too sensitive to share this data outside of their community.

Thus, 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 language activists, using a spreadsheet of words in their language, to provide predictive text suggestions for their community. We view this as an enabling tool: the intended users of this tool are language activists, who are in charge of their own data, and are entitled to choose how they share their predictive text keyboard. This is in contrast to efforts such as Gboard (van Esch et al., 2019), where a centralized entity mines text corpora and creates a language-specific model, which

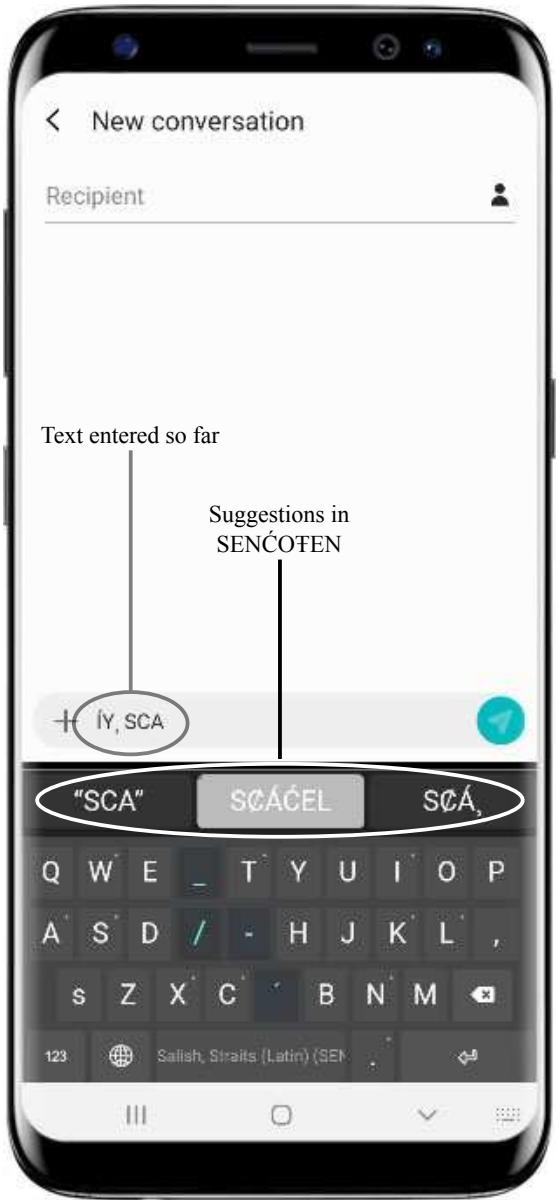

Figure 6: Predictive text for SENĆOŦEN.

is then shared back to the relevant language community. Since we expect data to be sparse, the generated language models are word-level unigram models; anecdotally, this has been enough to make typing in tricky orthographies, such as SENĆOŦEN (Figure 6), dramatically easier.

4 Speech-based subprojects

Traditionally, Indigenous languages in Canada were spoken,[8] not written. It therefore makes sense that several ILT subprojects focused on applications of speech technology, as described below.

4.1 Work at CRIM on Audio Segmentation and Speech Recognition

Transcription and further annotation of speech recordings (often including translation) 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. Even if it's impractical to transcribe all that speech now, could one apply ASR-based audio indexation techniques to carry out such searches?

Originally conceived as a long-term project, the CRIM effort yielded practically useful tools well before the first phase of the ILT project had ended. This was in large part because of collaboration between CRIM's ASR team and Christopher Cox of Carleton University, a field linguist.

The CRIM subproject faced two main challenges. First, very little data for training ASR models was available before the start of the project, even for the two most frequently spoken Indigenous languages in Canada, Inuktut and Cree, with only a few hours of transcribed recordings for each. Second, the speech recognition field was developed over decades with only a small subset of the world's languages. Even the US IARPA BABEL program, which focused on speech recognition for low-resource languages, covered mostly languages with low morphological complexity.[9] Hungarian and Turkish represented the most complex cases in this program. Yet, these two are relatively mid-range among languages. Most Indigenous languages in Canada are in the high range; even among morphologically complex languages, Inuktut is known to be particularly morphologically complex (as measured by mean distance to novel type (Schwartz et al., 2020)).

4.1.1 Data Collection and Transcription

Four transcription activities were carried out as core parts of the CRIM project, as shown in Table 1. Inuktut transcriptions were provided by the Pirurvik Centre; the others were provided by a company called WinTranslation.

In addition, CRIM has partnered with several community-university teams to assist in assessing the usefulness of the new ASR tools in supporting Indigenous language documentation projects, with each partner team providing feedback to CRIM on these services and their integration into common documentation workflows. In one of these partnerships, Elder Bruce Starlight (Language Commissioner of the Tsuut'ina Nation) and Christopher Cox (Carleton University) focused on producing high-quality audio recordings of existing Tsuut'ina language materials (e.g., written resources for which audio was not previously available), applying CRIM's ASR tools to assist in the transcription process, and developing

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

[9] See https://www.superlectures.com/asru2013/the-babel-program-and-low-resource-speech-technology

a selection of the resulting time-aligned transcripts into a collection that could be more readily shared within the community. Revised versions of several of the stories recorded and annotated in this process are now being prepared for publication, accompanied by illustrations produced by graphic artist Emil Starlight.

In another partnership, Michif speaker and Elder Verna DeMontigny (Prairies to Woodlands Language Revitalization Circle) and Olivia Sammons (Carleton University) recorded and transcribed speech in Michif, with an emphasis on lexical resources. As of March 2020, approximately 60 hours of audio recordings have been produced, and will be integrated into the online Michif dictionary discussed in Section 7.

Language	**Target (hours)**	**Current (hours)**	**Planned completion**
Inuktut	100	100	Done at end of March 2020
East Cree	100	100	Done at end of March 2020
Innu	25	3.4	Plan: end of May 2020
Denesuline	10	1.7	Plan: end of August 2020

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

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,[10] and are available for researchers and communities directly on the platform or through an ELAN extension.[11] The tools include:

DNN-VAD: Deep neural net (DNN) voice activity detection, extracts segments containing speech (as opposed to silence, noise, music, etc.);

Diarization: Creates annotations that identify who spoke when in a recording;

Language Retrieval: Finds segments which are spoken in a particular language (among 32 possible languages, including Cree and Inuktut);

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.

Until recently, teams developing speech resources for Indigenous languages had access to a relatively small number of ASR-related services when using popular multimedia annotation tools such as ELAN. The addition of the ASR services developed by CRIM not only expands this list significantly, but also makes recent advancements in speech technology (e.g., deep neural network-based techniques, which are

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

[11] A widely-used annotation tool available at https://archive.mpi.nl/tla/elan/; see (Wittenburg et al., 2006) for an overview.

found in a number of CRIM's ASR tools) much more accessible to current documentation development projects.

In the CRIM partnerships mentioned above, teams noted marked improvements in the efficiency of their annotation workflows when using these services. In many cases, applying the DNN-based voice activity detection services to automatically segment 'raw' recordings into utterance-level units significantly reduced or even eliminated the need for a separate stage of manual segmentation that would otherwise be required before transcription could be undertaken effectively, leaving linguists more time to focus on the actual contents of these recordings. As well, a speech-to-text service able to recognize English, French, and Spanish utterances, coupled with the language retrieval service, has proven to be valuable in providing first-pass automatic transcriptions of annotations that have been flagged as being in a non-Indigenous language (e.g., translations of Indigenous-language utterances that are offered in English). While these services are still relatively new, they have received a warm welcome from the language documentation research community, and are starting to be deployed in the field.

4.1.3 Automatic Speech Recognition (ASR)

This theme was inspired by the possibility of audio indexation for the large backlog of Indigenous speech recordings that have never been transcribed, to be carried out through ASR on the audio files. Even error-prone ASR on a speech file could make it possible to search through the recordings for keywords and phrases of interest. Keyword search looks for phoneme sequences (as opposed to word sequences) with some leeway for phoneme errors. If the ASR error rate is high, one will miss some of the segments one is looking for, and bring up some words or phrases one was not looking for, but this is much better than having no way at all of searching through hundreds or thousands of hours of speech.

To train an ASR system, one requires transcribed speech and ideally, additional monolingual text data, to train the so-called language model (LM). CRIM focused on building ASR systems for Inuktut and East Cree. This work highlighted major differences between polysynthetic languages and those more commonly studied in ASR research.

Inuktut, as was mentioned earlier, is a highly polysynthetic language for which a word dictionary was ineffective in modeling the language. Even with a dictionary containing 300,000 words (from 6 million words of Hansard corpus from Nunavut), 60% of the words in an unseen story text were out of vocabulary (OOV) (Gupta and Boulianne, 2020a). So CRIM resorted to sub-word units in order to reduce the out-of-vocabulary rate. They tried both 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 both morphemes and syllables, word boundaries can be marked by special syllables (and morphemes) with B_ and _E markers; CRIM also experimented with boundaries chosen by deep neural nets (DNNs). The syllable or morpheme LM then predicts the start or end of the words through these markers so that syllable (or morpheme) sequences can be converted back to word sequences. These results are obtained with acoustic models trained on 40 hours of transcribed Inuktut recordings. When training data is increased from 40 hours to 80 hours, the speaker independent (SI) WER goes down from 74.3% to 72.3%.

Though East Cree is also polysynthetic, in this case a word-based LM with a 30,000 word dictionary obtains a much lower OOV rate (Gupta and Boulianne, 2020b). Training a word-based LM for Cree from much less text data than for Inuktut (260,000 words of text from reports from Cree organisations and scriptures text) still yields only 25% OOV rate on video stories and 9% OOV rate on scriptures. Video stories are very different from the LM text, and so they 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

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: Summary of weighted OOV rate and SI WER on the Inuktut development set, for various subword units. The acoustic model is trained on 40 hours of audio.

(80 hours instead of 40 hours) and far more LM training data: 69.0% WER for Cree versus 72.3% for Inuktut.

These experiments are outlined in the publications for East Cree and Inuktut at the LREC 2020 conference and workshops (Gupta and Boulianne, 2020a; Gupta and Boulianne, 2020b). An important detail: training texts written in Inuktut syllabics allowed easy creation of a syllable dictionary, but because the Cree training data were transcribed with Roman characters, these data could be syllabified only after training a relatively complex syllabification model, using other texts in Cree syllabics for supervision. The syllable unit approach is difficult to apply for non-syllabic writing. Byte pair encoding (BPE) of subword units is a more general alternative that can be trained without supervision, for any language. However, for Inuktut, the speaker dependent phoneme error rate (PER) increases to 26.4% when using BPE units, compared to 18.4% PER obtained with syllables.

The SD ASR result for East Cree above is noteworthy: 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 good news: it 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.

ASR research generally does not focus on SD systems, but 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 (as with East Cree in the above experiments), 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. This could be a partial solution to the transcription bottleneck. In the next phase of the project, the CRIM team will look more deeply into this possibility, and also build on work done during the first phase to research audio indexation.

4.2 Read-along audiobooks

An ongoing concern of the project has been "What technologies are feasible for the *least*-resourced languages?" While for a few of the 70-some Indigenous languages spoken in Canada, medium-scale resources for, say, machine translation or ASR can be collected, the great majority are unlikely to pass that threshold in the near future. Littell et al. (2018a) surveyed a variety of text and speech technologies with an eye towards identifying those that could be feasibly rolled out to *any* Indigenous languages in the near term.

High on this list were automatically-aligned text/speech "read-along/sing-along" activities, inspired by the East Cree online activities at eastcree.org. Interactive read-along/sing-along activities, 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 are highly labour-intensive to make and require the maker to be skilled in specialist software like ELAN or Audacity. A collaboration between NRC, Carleton University, 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 (often 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). This can be done by constructing an approximate mapping between target language phonemes and phonemes in a high-resource "donor" language like English, converting the target document into the donor languages' phones, and using an acoustic model trained on the donor language to recognize when each word is spoken. This is widely done on an ad-hoc basis—it is, for example, the assumed default when working with a new language in the Festival toolkit (Black et al., 1998)—but we are seeking to simplify and generalize this process to the point where a non-specialist could use it.

Our ReadAlong Studio[12] combines a custom G2P engine, the PanPhon (Mortensen et al., 2016) phonetic distance library (for automatically determining cross-linguistic phoneme mappings), and the lightweight PocketSphinx (Huggins-Daines et al., 2006) speech recognition library to allow a user to easily create a text/speech alignment system for a new language without being a specialist in speech technology.

ReadAlong Studio currently supports about 22 languages; among Indigenous languages spoken in Canada it supports Anishinaabemowin (Ojibway), Dakelh, Gitxsan, Heiltsuk, Inuktut, Kanyen'kéha, Kwak'wala, SENĆOŦEN, Seneca, Tagish, Tŝilhqot'in, and Tsuut'ina. Adding a new language is the work of only a few hours, depending on the complexity of the language's orthography. For best results, someone (e.g., a linguist) with knowledge of the International Phonetic Alphabet is necessary, but for relatively transparent orthographies an automatic fallback using the `text-unidecode` Python library can suffice.

ReadAlong Studio can already create high-quality interactive webpages (see Figure 7), MP4 movies, and EPUB documents, and can also export in ELAN, TextGrid (PRAAT), and subtitle formats. A user-friendly interface, with the goal of making a read-along as easy to make as a PowerPoint presentation, is currently in development.

4.3 Limited-domain Text-to-Speech (TTS)

Having discussed speech-to-text technologies, we now consider text-to-speech. One of the main suggestions in user evaluations of Kawennón:nis (see Section 3.1.2) was to provide audio for the conjugated verb forms. However, this is unfortunately not a straightforward task, as the Kawennón:nis language model (LM) has hundreds of thousands of unique output forms, which would be entirely infeasible to record by hand. Without any pre-existing data, the task of figuring out the fewest number of recordings required in order to generate the rest emerged.

Typically, text-to-speech (TTS) requires a lot of data. In many use cases, TTS can compromise on the naturalness of the speech so long as it is intelligible (e.g., for speech-generating devices for augmentative and alternative communication). It is also a common requirement that speech synthesizers function in a general or open-domain setting. Our specific use case with Kawennón:nis is slightly different. The domain we need to synthesize is limited to the LM, which is made up entirely of single prosodic words (however, due to the polysynthetic nature of Kanyen'kéha, these are often quite meaningful words).

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

Figure 7: The ReadAlong Web Component displaying Atikamekw story ‘Nikikw’

Additionally, the main goal is not intelligibility, rather it is pedagogical, and so we have to be fairly uncompromising when it comes to the naturalness of speech.

Following (Black and Lenzo, 2001), we implemented a database selection algorithm that analyzes a phonemic representation of the total output of the LM for Kawennón:nis, and runs a greedy search to select the fewest candidates required to cover every unique phoneme trigram in the data. Converting the graphemes of the LM to phonemes for analysis was luckily straightforward, as Kanyen’kéha’s orthography is extremely faithful to the surface representation, and also includes marking of important prosodic features like stress. Given the relatively small phonemic inventory for Kanyen’kéha, this process yields a very high ratio of possible generated forms to required recordings.

We took a sample from the Kawennón:nis LM with 122,966 conjugations and found that we could record at least one sample of each unique phoneme trigram by recording 852 of the forms. We have since built an evaluation tool to randomly sample the data and ask a speaker-evaluator to rate the sample’s quality both in terms of intelligibility and naturalness. Unfortunately, while some samples score nearly perfectly, others are not at all intelligible. Most of these issues have proven to be resolved either by manually fixing errors in the forced alignment or by fixing incorrectly labelled recordings (where the phonemic representation is transcribed incorrectly). This process however is time consuming and we have opted to put this part of the project on hold until either the method or the tools can be improved.

Using the same method with less ambitious datasets has proven to be very effective, however, and we have successfully built synthetic ‘talking clocks’ in Kanyen’kéha using 24 recorded sentences to generate 1440 times of day. The unit of concatenation in this instance is at the word-level, not the phoneme or diphone level, and seems to be less error prone.

5 Online courses and games

Several external subprojects funded by ILT involved online courses and games designed to help people acquire or enhance Indigenous language skills. As will be seen below, efforts of this nature sometimes run into what might be called the “game development trap”: to appeal to learners (especially younger ones),

these courses are often highly interactive and incorporate complex graphics. To enable this, they are built on top of software that will quickly become outdated. Small communities with limited funding may come to regret the need to constantly update their language learning software to ensure that it continues to work.

5.1 Learning Platforms at Carleton University

Prof. Marie-Odile Junker and her Carleton University team have been collaborating for years with Indigenous partners to develop interactive online language lessons that support literacy in two Algonquian languages, East Cree and Innu (not to be confused with the unrelated language of the Inuit, Inuktut). See https://www.marieodilejunker.ca/publications-by-topic/ for a list of the team's publications on this and related topics.

The web-based platform underlying these lessons allows for the creation of multimedia interactive online lessons with auto-generated and author-configured exercises and games. Users are able to listen to a word or phrase in several dialects. They can play computer-generated interactive activities that test and enhance their vocabulary, orthography and grammar acquisition, and also engage in more advanced grammatical and textual activities. Teachers can go online to develop customized lesson plans, and track students' progress. Language experts can access an administrative interface to develop new content.

In this case, the NRC had the privilege of contributing funding to assisting valuable technologies developed long before the ILT project existed. When ILT funding first began to flow to Carleton in the summer of 2018, the rapid pace of change in the software industry had partially stranded these educational language tools: many of the functionalities no longer worked as intended.

In the first phase of this subproject (up to March 2019), the platform was updated to align with current technology. Specifically, loading time and the overall server load were reduced, the database was optimized, the UI was modernized (this included simplifying navigation and coping with different screen sizes), dependence on Flash and other outdated technologies was removed, third-party authentication was integrated, and system administration was simplified.

In the second phase (April 2019–March 2020), the platform—which was originally developed with first-language (L1) speakers in mind—was expanded to help second-language (L2) learners. This expansion involved further software improvements, such as improved ability to import data from the Carleton group's other linguistic resources (dictionaries and databases of inflected verbs) and modification of the Cree modules to accommodate both syllabics and Roman orthography (they had previously been syllabics-only). Above all, this second phase involved incorporation of new pedagogical material, such as new audio recordings (both scripted and spontaneous) and creation of new L2 activities for both languages (e.g., conversations, narratives, structural exercises, verb conjugation exercises, and learning of thematic vocabulary).

Throughout both phases of this subproject, two language experts—one for East Cree, one for Innu—made content corrections and enhancements. User testing was carried out by collaborating with partner organisations (Cree School Board, Institut Tshakapesh). The East Cree and Innu modules have been used in courses at the Université de Montréal, the Native Montreal organization, and in teacher training courses in Indigenous communities. The enhanced bilingual (English and French) platform with its new L2 content is now live at the following URLs: East Cree: https://lessons.eastcree.org/; Innu: https://lessons.innu-aimun.ca/.

5.2 7000 Languages

Several organizations provide online courses for many different world languages. Could one of these organizations apply its expertise to develop courses for Indigenous languages in Canada? There was a stumbling block: several companies that create good online language courses charge large amounts for

doing this, or are only interested in languages spoken by large numbers of people. We decided to provide modest amounts of funding to 7000 Languages (www.7000.org), a non-profit that creates free language-learning software in partnership with Indigenous communities around the world, to create courses with interested communities (who would also be funded).

Four communities ended up working with 7000 Languages on courses for their languages. The languages were Kwak'wala, Michif, Naskapi, and Mi'kmaq. In discussions with 7000 Languages, each community worked out a slightly different plan for data collection (especially of Elder speech recordings and of key vocabulary items) and incorporation of these materials, along with grammar instruction, into the lessons.

The Kwak'wala and Michif courses are almost ready to go online. The Naskapi course requires a few more months to be completed, and work on the Mi'kmaq course has been suspended indefinitely due to unforeseen circumstances. The ILT project was also able to support the Mi'kmaq language in a minor way by providing a different organization, the Mi'kmaq Language Lab at Cape Breton University, with funding to buy videoconferencing equipment that will enable the lab to reach many more learners.

5.3 Computer-assisted Language Learning at the University of Alberta

The University of Alberta subproject has focused on supporting the Y-dialect of Cree (Plains Cree). The U. Alberta team has been creating an adaptive computer-assisted language learning (CALL) system to support learners' ability to recognize and understand oral language. The system will use a combination of games, stories, and other audio materials to teach listening skills. A Cree-language instructor at the university has been developing this system in consultation with members of nearby Cree-speaking communities.

Despite delays caused by contract and hiring difficulties, the project has recently been making good progress. It has been developing Cree-language content: 13 personal stories from 8 Elders have been recorded, and are currently being transcribed. Several key educational activities have been implemented in software: two **phonemic awareness** activities, a game that helps learners of Cree map spoken phonemes onto characters and a "shadowing" activity in which the learner reads along with a fluent speaker; two sets of **morphological awareness** activities, the first set aimed at enhancing recognition of Cree verb forms, the second set asking learners to conjugate Cree verbs themselves by assembling morphemes in the correct order; a **vocabulary knowledge** activity in which learners build and discuss their family trees; and a **local history and culture** augmented reality "walking tour".

These diverse software modules are currently being integrated into the overall system; user testing will follow. Until the COVID-19 outbreak intervened, the system was on track to go live at the end of April (it was to have been shown at the International Conference in Artificial Intelligence in Education, in early July).

5.4 FirstVoices

The First Peoples' Cultural Council (FPCC) of British Columbia has a remarkable record of providing state-of-the-art technologies, training and technical support to Indigenous language champions (mainly but not exclusively within British Columbia), within its FirstVoices program.[13] The Language Tutor was developed as part of FirstVoices; this software allowed communities to build intuitive language lessons that mimic the way a child learns a language. Users were able to listen to a word or phrase, record themselves speaking and then compare the result with a recording of a fluent speaker. They could also match images, video and audio clips from the FirstVoices library with words and phrases.

Unfortunately, as with the original version of the Carleton University lessons described above, changes in the software industry rendered key functionalities of Language Tutor inoperable. FPCC decided to em-

[13] http://www.fpcc.ca/about-us/

bark on a radical rethinking of their whole approach. With ILT funding, they conducted two focus groups to assess what features communities wanted in an online language learning tool, and commissioned a study of the technical literature on online learning of Indigenous languages. Based on the findings of these exercises, they produced a detailed plan for new language learning functionalities.

FPCC has already implemented two useful functionalities, an "immersion portal" that allows communities to input text and audio translations of all user interface controls into FirstVoices.com, thus ensuring learners are fully immersed in the language being studied, and a "flashcard view" that converts any list (e.g., of words) into a printable flashcard format for self-study. Other functionalities are currently being implemented (with funding from sources other than NRC).

5.5 On the Path of the Elders

"On the Path of the Elders" is a role-playing educational game designed to acquaint players with historical and cultural topics related to Treaty 9, signed in 1905 by the Government of Canada and several Indigenous communities in the James Bay region. When this free online game was released in 2007, it was widely praised for its innovative use of historical resources. Though the primary purpose of the game is historical and cultural education, the site has substantial linguistic resources—for instance, audio-enhanced syllabics charts, textbooks, and most important, recordings of almost sixty oral stories recounted by Elders and split among four different dialects of Cree: Swampy Cree, Swampy Cree with N dialect, Moose Cree, and Kashechewan Cree.

During the years since the original launch in 2007, technology has changed and users offered improvement suggestions. The ILT project provided funds to the education branch of the Mushkegowuk Council to have the game converted from Flash to HTML5 format, to make it accessible to mobiles, to add new content, and to make other improvements. While some of these improvements (especially to the teacher's guide part of the site) still need to be made, the website is now fully mobile-friendly and compatible with Chrome, FireFox and Internet Explorer browsers. It can be viewed at www.pathoftheelders.com. https://www.overleaf.com/project/5e4d7f6edc5a70000121500e

6 Capacity-building: Yukon and Indigitization Subprojects

Two subprojects funded by NRC focused on training and support for people who document Indigenous languages. In an initiative run by the **Yukon Native Language Centre** (YNLC) that began in October 2019, 12 speakers of 8 languages (Gwich'in, Hän, Kaska, Northern Tutchone, Southern Tutchone, Tagish, Tlingit, and Upper Tanana) from 14 communities attended three workshops in Whitehorse. Each was five days long, with training starting at 9 am and finishing at 4:30 pm.

In these hands-on YNLC workshops, after being issued with the appropriate equipment, the trainees learned how to make high-quality video and audio recordings of their languages, and how to transcribe, annotate, and translate the speech in the recordings. In workshop 1, held in October, 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, held in November, attendees were trained to use the ELAN and SayMore software suites for transcribing, annotating, and translating speech recordings. Trainees worked with fluent speakers of their languages to practice these skills. A professional video editor from Toronto who happened to be in Whitehorse kindly volunteered his time to teach editing skills. Workshop 3, held in February, began with review and practice of the skills learned earlier. Subsequently, this workshop focused on sharing and repurposing videos (on DVD, on websites, through social media). On the last day, trainees brainstormed on how best to use ELAN-based materials as a resource for language learning and teaching. In the afternoon there was a discussion about the new skills attendees had acquired, and a showing of finished videos.

YNLC's original plan had been for this work to culminate in an Open House on National Aboriginal Language Day (March 31, 2020) where selected videos from the project would have been showcased to the public and to the Grand Chief and other dignitaries. Each trainee would have had the opportunity to show 10 minutes (approximately two videos) from the videos they made. The trainees produced 10 to 20 short videos, each at least 3 to 5 minutes long, each in the trainee's language. Unfortunately, because of the outbreak of COVID-19, the Open House has been postponed to the second phase of the ILT project. This postponement does not diminish the overall accomplishments of this ILT subproject: in the course of 10 months, 12 Indigenous trainees created 548 minutes of documentation and mastered complex skills that will enable them to record Elders in their communities on a continuing basis. It is likely that the YNLC workshops will end up having a strongly positive impact on most languages in the Yukon.

The **Indigitization** subproject, a partnership between the University of British Columbia (UBC), the Musqueam Archives, and the Heiltsuk Cultural Education Centre, focuses on digitization of audio and video language data in non-digital formats. Thus, while the YNLC subproject trains members of Indigenous communities to collect **new** data, Indigitization involves training that will render **old** data in obsolete formats far more accessible. For instance, some communities possess many potentially valuable hours of speech by fluent speakers recorded years ago on cassette tapes.

While digitization can be performed by outside organizations, Indigitization will make Indigenous communities across the country autonomous by developing resources and offering workshops that will train their members in the necessary skills. For communities that already have digitized content, or that move quickly past the digitization stage, the project will also offer training in the next steps of mastering their own data: transcription of speech into text into the relevant Indigenous language and, where desired, translation of that text into another language (e.g., English or French). Manuals and training resources for digitization and transcription/translation methodology are being developed. Training workshops in several different parts of the country are part of the plan. This subproject began in January 2020, and has been hit hard by the COVID-19 crisis. Several key aspects, such as the regional workshops, have been postponed until later in 2020.

7 Single-language subprojects focused on data collection

This section provides brief descriptions of data collection efforts conducted within Indigenous communities. These efforts were carried out by members of those communities and funded by ILT.

- **Cree**: Blue Quills University, known as Nuhelot'įne thaiyots'į nistameyimâkanaks in Dene and Cree, was the first university in Canada to be fully owned and operated by First Nations people. The subproject at Blue Quills focused on digitizing and indexing the largest known collection of Cree text in syllabics: approximately 40,000 pages from monthly newsletters called kihcitwāw miteh (Sacred Heart) produced by the Catholic church between 1906 and 1978. The resulting corpus will cover local and international news, legends, jokes, and so on—a wealth of diverse material that will be of enormous value to students and teachers of Cree, and to scholars. This subproject has a significant training component: students have been carrying out much of the work. Fortunately, the pandemic has caused only a slight delay. At the time of writing, cataloguing and indexation of the corpus are now complete; it is now being formatted, and a historical introduction written. The corpus is likely to be available on the Blue Quills website in late 2020.

- **Kanyen'kéha (Mohawk)**: This effort was made up of two subprojects. The first subproject took place over the summer of 2019; the second started in fall 2019. Both were led by Tsi Tyónnheht Onkwawén:na Language and Cultural Centre (TTO) on the Tyendinaga Mohawk Territory, near Kingston, Ontario; a key figure in both was Nathan Thanyehténhas Brinklow, a language teacher at Queen's University. The long-term goal inspiring them was the creation of a text and audio

corpora to support research on automatic speech recognition (ASR). The medium-term goal is to turn hundreds of hours of unannotated and untranscribed Mohawk audio into a useful resource for language learning and research. The summer 2019 subproject succeeded in collecting a total of 112,420 Kanyen'kéha written words and 26 hours of Kanyen'kéha speech, covering a mix of genres: scripts and audio from movies and TV shows, translated books of the Bible, and recordings of Elders. More Kanyen'kéha material has been collected since then, mostly text from the Internet, including a translation of the United Nations Declaration of Indigenous Peoples, the report of the Mohawk language standardisation committee, blog entries, and an Ontario government document on COVID-19. Nancy Bonvillain, a linguist who worked with the Akwesáhsne Kanyen'kéha-speaking community in the 1970s, is contributing copies of text and audio resources in her collection to TTO. A "Donate Your Voice" website is in development: it's currently running as a test site, with Mohawk translations and text being added, but has not yet gone live.

The pandemic has caused in-person collection from community members to be suspended. This subproject will be extended so that it can meet its original targets. As the "Donate Your Voice" effort is an online activity, social media outreach could enable this portion of the subproject to continue despite physical distancing measures.

- **Kwak'wala**: Like the Kanyen'kéha subproject just described, this effort aims at an initial corpus collection effort followed by work on ASR. The Kwak'wala language (Wakashan) is spoken by 18 Kwakwaka̲'wakw Nations whose traditional territory is on northern Vancouver Island, nearby smaller islands, and the adjacent mainland. NRC ILT funding supported a partnership among the Gwa'sala-'Nakwaxda'xw Language Revitalization Program, the Sa̲nyak̲ola Foundation, and the University of British Columbia (UBC). The partnership includes two community-based teams from three different Kwakwaka̲'wakw communities, a technical team, and three university-based linguists (two of whom are Kwakwaka̲'wakw).

  In assembling a corpus of machine-readable time-aligned transcriptions of recorded Kwak'wala audio for the purpose of automating speech-to-text, the subproject prioritized capacity-building among community-based researchers. The project trained 20 community members from three North Island communities in skills such as audio recording, data management, ear training, Kwak'wala orthographies and transcription, and annotation in ELAN transcription software. The team assessed available resources of analog, digitized, and born-digital audio data in Kwak'wala, created by community members and linguistic researchers and held in personal, local, museum and provincial archives. The team then developed a data pipeline to track data at various stages of progress toward completion, from cassette recordings requiring digitization, to audio that is transcribed but not time aligned, with machine-readable time-aligned ELAN transcriptions using IPA as the desired end product.

  So far, the project has compiled 25 hours and 3 minutes of machine-readable Kwak'wala audio data, consisting of basic daily conversational speech, pedagogical materials, and Elders' storytelling. The project identified over a hundred hours of audio recordings in various locations, at various stages of readiness for machine processing, and has submitted funding requests for community-based efforts to make these recordings more accessible through digitization and transcription. Community-based teams have also prioritized working with fluent Elders to create audio records of text resources to serve language reclamation goals. An additional technical challenge relates to language modelling: while there is a rare wealth of written textual data in the Kwak'wala language, both published and unpublished, there is a significant need to develop optical character recognition (OCR) for legacy orthographies which are not currently machine readable. The corpus creation effort will continue for several more months.

In addition to contributing to the development of ASR tools to lessen the transcription bottleneck and support Indigenous language revitalization, a key outcome of the subproject will be to share these recordings with community members through a graded-access website using the locally-preferred U'mista orthography.

- **Michif**: the Michif language is of great interest for linguists, as a mixed language that incorporates complex elements of its "parent" languages (mainly Plains Cree and French). It is spoken by fewer than 200 individuals, primarily in small communities across western Canada and the northern United States. Intergenerational transmission of the language has ceased, and the majority of speakers are well over sixty years of age. Few print and digital resources for second language acquisition are available to support Métis communities. Thus, Michif is both critically endangered and significantly under-resourced.

  This subproject aims to enhance a particular resource (Laverdure et al., 1983), which is now out of print and largely inaccessible to learners of Michif unless purchased at a high price. This inaccessibility is a significant barrier to Michif language revitalization.

  After an unanticipated delay, formal written permission was granted on January 17, 2020 by the dictionary's copyright holder (Turtle Mountain Community College) to the Prairies to Woodlands Indigenous Language Revitalization Circle (P2WILRC) to create a digital version of the dictionary for online, offline, and mobile use. This digital version of the dictionary will retain the original content, and allow for the inclusion of additional orthographies, grammatical information for lexical entries, and audio recordings of headwords and example sentences. Meanwhile, the P2WILRC team has fully digitized the 350 pages of (Laverdure et al., 1983) using OCR, and is in the process of correcting OCR errors. Approximately 218 pages of the dictionary have been recorded in spoken form. This was accomplished by collaboration between the P2WILRC team and the two-person team of Elder Verna DeMontigny and Olivia Sammons mentioned in Section 4.1.1. This subproject will be extended into the second phase of the ILT project.

- **Nsyilxcn**: The Nsyilxcn language is part of the Interior Salish family. Traditionally spoken in the Okanagan Valley in present-day British Columbia, it is now considered critically endangered, with only 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. The goal of the project was to carry out advanced-level Elder recordings to complement the existing Nsyilxcn curriculum. Syilx community members recorded and archived fluent Elders, while being trained in language skills.

  By the end of this subproject, seven hours of Elder stories in fluent Nsyilxcn had been recorded and archived. The stories are personal narratives, history, and traditional knowledge. They are transcribed in Nsyilxcn, with a glossary provided in English, at an intermediate level. These recordings are shared publicly on www.thelanguagehouse.ca, along with earlier recordings. Furthermore, during this period (Dec. 2019–March 2020) SLH continued to train fourteen learners, delivering sixty hours of Nsyilxcn via live-streaming immersion lessons. Six hours of these lessons can be viewed at http://www.thelanguagehouse.ca/nsyilcxn-1.html.

- **SENĆOŦEN**: SENĆOŦEN belongs to the Coast Salish language family and was traditionally spoken just north of present-day Victoria, British Columbia. It is the most severely endangered language the ILT project became involved with: there are only five fluent speakers left. On the other hand, the community is engaged in a vigorous language revitalization effort, led by the W̱SÁNEĆ School Board. From 1981 to 1991, the linguist Dr. Timothy Montler recorded Elsie Claxton, the last monolingual SENĆOŦEN speaker, telling anecdotes, historical narratives, and traditional tales

of her people (Dr. Montler's contributions to the study of this language include (Montler, 2018)). The NRC was able to assist the School Board's efforts by paying for two Elders to work with Dr. Montler to transcribe the Claxton recordings in SENĆOŦEN, and translate them into English. The Elders worked for a total of 560 hours in 2019 and 2020. This team—Dr. Montler and the two Elders—transcribed and translated all of the Claxton recordings. All this material, together with a list of all the words in the transcriptions, has been entered in a pre-existing database of SENĆOŦEN information. This new material will be an invaluable resource for the School Board's teaching of the language.

- **Tŝilhqot'in**: The traditional territory of the Tŝilhqot'in Nation is in the southern interior part of present-day British Columbia. The Tŝilhqot'in language is the southernmost member of the Dene (Athabaskan) family in British Columbia. Funding from the ILT project enabled the Tŝilhqot'in National Government (TNG) to build on an already impressive language revitalization effort. When the NRC-funded part of this effort began, TNG had already recorded 35,000 audio clips, developed an associated linguistic database, published a set of third person paradigms in a "Digital Verb Book", and created a website with a diverse set of learning tools: www.tsilhqotinlanguage.ca.

  The ILT funding enabled TNG to record over a dozen hours of speech from around 20 speakers. This was somewhat fewer hours than originally targeted, because of the difficulty of interviewing fluent speakers unfamiliar with the process. On the other hand, because the technical workers in the community proved to be proficient at processing existing audio and text data, far more recorded speech than foreseen in the original plan was transcribed, aligned with the transcriptions, and digitized (20 hours of speech). The linguistic database is now much more representative of the broad range of Tŝilhqot'in dialects. Labeling of around 46,000 audio clips has been completed. Furthermore, a list of 26,200 English words and phrases to be translated has been drawn up; many of these have already been translated into Tŝilhqot'in. In the course of this work, four community members were trained in technical roles involving general IT, front-end development, audio processing and business software. This provides a solid basis for the rest of this subproject, which will continue to be funded over the next few months.

8 Discussion

So far, reception by parts of the Natural Language Processing (NLP) research community of much of the work described here has been lukewarm. With the field's current focus on machine learning techniques, rule-based approaches like those deployed in many of the subprojects described in this paper tend to gather less attention than the application of novel deep neural architectures to tasks with larger data sets. New is not always better. We have endeavoured to choose approaches that are most appropriate to the data we have available, attempting to mitigate some of the pitfalls of old-fashioned rule-based techniques through our community collaborations. That being said, we have sometimes been able to import models trained on high-resource language data to tasks involving low-resource languages, as when acoustic models trained on English data turned out to be suitable for aligning speech in several Indigenous languages with the corresponding text to create readalong books.

By contrast, field linguists have been enthusiastic about the audio segmentation tools released by CRIM as part of the ELAN suite of recording and transcription tools (Cox et al., 2019). By incorporating state-of-the-art ASR techniques into the early stages of speech processing, they offer a substantial reduction in the time these linguists must spend on the most tedious part of their task. Though these CRIM tools were developed for use with Indigenous languages in Canada, there is little language-specific about them: they are likely to be useful for field linguists on every continent.

Indigenous educators have welcomed the readalong capability for audio books, which we've already

rolled out for several books in many different languages (as described above). However, making educators dependent on NRC experts to create new readalong books is antithetical to the goals of the ILT project, which seeks to empower communities to meet their own linguistic needs. In the next phase of the project, making the readalong software easier to use will be one of the top priorities.

The same is true for text prediction for mobile devices, which is a capability that is sought after by speakers of several Indigenous languages (especially younger ones). In this case, however, we have only implemented it for one language, SENĆOŦEN. This language community is very happy about this, but one language isn't nearly enough. We have open-sourced the software for creating text prediction capabilities for a new language, but work is needed to make it much more user friendly.

The trickiest case is WordWeaver, because its Kanyen'kéha offspring—Kawennón:nis—has created so much enthusiasm, not merely among teachers of Kanyen'kéha, but also among teachers of other, unrelated, polysynthetic languages, such as Inuktut and Cree. Teachers of several Indigenous languages in Canada have told us they would like to build similar verb conjugators for their languages. The problem is that Kawennón:nis is a labour-intensive software creation, the result of several person-years of hard work by two members of the NRC team and several instructors at Onkwawenna Kentyohkwa (Western dialect) and Kahnawà:ke (Eastern dialect). The NRC does not have the human resources to repeat its part of that effort for over sixty languages, yet specialised expertise in building FSTs (which is rare even among computational linguists) was an essential component of the effort.

Part of the solution may involve producing less ambitious verb conjugators for other languages, by covering fewer stems and forms. This might suffice for some pedagogical purposes. Members of the NRC team have been doing this for Michif, in collaboration with two experts (Heather Souter and Olivia Sammons). Even this approach—of building model T Ford conjugators rather than Rolls-Royce versions—is unsatisfactory, since programming even small FSTs is currently unintuitive. We'd like communities to be able to build verb conjugators for themselves, instead of having to rely on outside experts. We have begun to explore a path to that, as described in the next section.

Three research outcomes of the project have not yet had a practical impact on Indigenous language revitalization. The first is the new tools for Inuktut, including an improved version of the previously available aid for translators, WeBInuk—but these apps are due to be rolled out in the coming months. The second is the creation of an up-to-date Inuktut–English version of the Nunavut Hansard corpus, consisting of 1.3 million sentence pairs, and associated work on machine translation for this language pair (including work by other researchers that will be stimulated by the WMT 2020 shared task for this language pair). The third is CRIM's work on speech recognition for six Indigenous languages. We believe, however, that all three research directions will also soon benefit language revitalization.

The long-term prospects for Indigenous language revitalization in Canada are excellent. All over the country, Indigenous communities are taking their linguistic future into their own hands by collecting speech and text data and enrolling record numbers of students in language courses. Even some of the poorest, most remote communities have activists working to help the ancestral language regain its rightful place in the community. We hope this project will have placed some useful software tools in the hands of these activists and communities.

9 Future Work

NRC has decided to fund a second phase of the project, though it is likely that funding for externally managed projects will be more limited than in the first phase.

9.1 Visual Programming for Building Rule-based Grammars

Since many Indigenous language technologies depend on a morphological FST or other rule-based system (Arppe et al., 2016), development of them is practically limited to a few handfuls of specialist

linguist-programmers. Some of our community collaborators have raised concerns about this: if only a linguist-programmer, typically not a member of the community, can practically write and understand the FST source, then updating the system depends on their availability. This is a barrier to full Indigenous sovereignty over their language technology, and thus broadening the accessibility of rule-based technology development is one of our priorities for future development of Kawennón:nis and similar systems.

Following up on a line of research into making linguistic program development more rapid and intuitive (Hutton and Meijer, 1988; Frost and Launchbury, 1989; Littell et al., 2018c; Littell et al., 2018b), and inspired by visual and tabular programming environments like Blockly (Fraser, 2015; Pasternak et al., 2017), LAPPs/Galaxy (Ide et al., 2016), and Tabular (Gordon et al., 2014), we are currently experimenting with visual linguistic programming intended for language teachers, students, and linguists. Ideally, these interfaces could make it easier for Indigenous language experts to build verb conjugators and other tools for their own languages, without prior experience in technology development and with comparatively little assistance from outsiders.

9.2 Other Priorities

- We plan to continue the NRC–Pirurvik Centre collaboration, which will soon result in the release of a new version of WeBInuk and other tools for Inuktut learners, writers, and translators. In addition, we hope to create a new bilingual Inuktut–English corpus that will extend the parallel resources available to researchers beyond the limited-genre Nunavut Hansard.

- Where appropriate, we will continue to roll out readalong audio books, text prediction for specific languages, and limited-domain TTS.

- However, in accordance with the empowerment goal of the project, our focus will be on making the technologies listed in the previous bullet point easier for communities to deploy themselves without outsider expertise. We hope this will have a multiplier effect.

- CRIM will continue to publicize its open-source tools for audio segmentation among field linguists and Indigenous activists involved in recording speech. The CRIM team will also explore the application of speaker-dependent speech recognition to widening the transcription bottleneck, and to work on audio indexation.

- The NRC team will emphasize transfer of expertise to Indigenous communities. For instance, we plan to hire interns from those communities.

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 report 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 (Jordan Brant), Ronkwe'tiyóhstha (Josiah Maracle), Michael Running Wolf, Lorne Shapiro, Tye Swallow, and Shawn Tsosie.

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.

Antti Arppe, Jordan Lachler, Lene Antonsen, Trond Trosterud, and Sjur N. Moshagen. 2016. Basic language resource kits for endangered languages: A case study of Plains Cree. In *Proceedings of the 2016 CCURL Workshop. Collaboration and Computing for Under-Resourced Languages: Towards and Alliance for Digital Language Diversity, LREC 2016, May 23, 2016*, pages 1–9.

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

Alan Black and Kevin Lenzo. 2001. Optimal data selection for unit selection synthesis. In *4th Speech Synthesis Workshop*, pages 63–67. ICSA.

Alan W. Black, Paul Taylor, and Richard Caley. 1998. The Festival speech synthesis system.

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.

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.

Neil Fraser. 2015. Ten things we've learned from Blockly. In *2015 IEEE Blocks and Beyond Workshop*, pages 49–50.

R. Frost and J. Launchbury. 1989. Constructing natural language interpreters in a lazy functional language. *The Computer Journal*, 32:108–121.

Andy Gordon, Thore Graepel, Nicolas Rolland, Claudio Russo, Johannes Borgström, and John Guiver. 2014. Tabular: A schema-driven probabilistic programming language. In *POPL '14 Proceedings of the 41st ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, pages 321–334. ACM Press, January.

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 *Proc. LREC2020*.

Vishwa Gupta and Gilles Boulianne. 2020b. Speech Transcription Challenges for Resource Constrained Indigenous Language Cree. In *Proc. 1st Joint SLTU and CCURL Workshop*.

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.

Graham Hutton and Erik Meijer. 1988. Monadic parser combinators. *Journal of Functional Programming*, 8:437–444.

Nancy Ide, James Pustejovsky, Christopher Cieri, Eric Nyberg, Denise DiPersio, Chunqi Shi, Keith Suderman, Marc Verhagen, Di Wang, and Jonathan Wright. 2016. The language application grid. In Yohei Murakami and Donghui Lin, editors, *Worldwide Language Service Infrastructure*, pages 51–70, Cham. Springer International Publishing.

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.

Patline Laverdure, Ida Rose Allard, and John C. Crawford (ed). 1983. *The Michif dictionary : Turtle Mountain Chippewa Cree*. Pemmican Publications, https://search.library.utoronto.ca/details?4541064.

Patrick Littell, Anna Kazantseva, Roland Kuhn, Aidan Pine, Antti Arppe, Christopher Cox, and Marie-Odile Junker. 2018a. Indigenous language technologies in Canada: Assessment, challenges, and successes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2620–2632.

Patrick Littell, Tom McCoy, Na-Rae Han, Shruti Rijhwani, Zaid Sheikh, David Mortensen, Teruko Mitamura, and Lori Levin. 2018b. Parser combinators for Tigrinya and Oromo morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Patrick Littell, Tian Tian, Ruochen Xu, Zaid Sheikh, David Mortensen, Lori Levin, Francis Tyers, Hiroaki Hayashi, Graham Horwood, Steve Sloto, Emily Tagtow, Alan Black, Yiming Yang, Teruko Mitamura, and Eduard Hovy. 2018c. The ARIEL-CMU situation frame detection pipeline for LoReHLT16: A model translation approach. *Machine Translation*, 32(1–2):105–126, June.

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.

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

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.

Erik Pasternak, Rachel Fenichel, and Andrew N. Marshall. 2017. Tips for creating a block language with Blockly. In *2017 IEEE Blocks and Beyond Workshop (B&B)*, pages 21–24. IEEE.

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

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.

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.

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).