NRC·CNRC

NRC Publications Archive
Archives des publications du CNRC

El Projecte Indigenous Languages Technology (Ilt) Del Consell Nacional D'investigació De Canadà, I El Seu Context

Kuhn, Roland

This publication could be one of several versions: author's original, accepted manuscript or the publisher's version. / La version de cette publication peut être l'une des suivantes : la version prépublication de l'auteur, la version acceptée du manuscrit ou la version de l'éditeur.

Publisher's version / Version de l'éditeur:

Linguapax Review 2021, 9, pp. 105-126, 2021

NRC Publications Archive Record / Notice des Archives des publications du CNRC :
https://nrc-publications.canada.ca/eng/view/object/?id=46570dae-12b7-4a2b-b12b-613b39ff1e80
https://publications-cnrc.canada.ca/fra/voir/objet/?id=46570dae-12b7-4a2b-b12b-613b39ff1e80

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

EL PROJECTE *INDIGENOUS LANGUAGES TECHNOLOGY (ILT)* DEL CONSELL NACIONAL D'INVESTIGACIÓ DE CANADÀ, I EL SEU CONTEXT

Roland Kuhn

INTRODUCCIÓ

En les últimes dècades, s'han creat centenars de projectes de revitalització i recuperació de llengües indígenes per tot Canadà. La majoria es troben a comunitats indígenes; d'altres se centren a universitats. Varien en dimensió des de les ambicioses iniciatives multilingües del First Peoples' Cultural Council (FPCC)[1] que es troba dins de i és finançat per la Província de Columbia Britànica, i és actiu allà i a altres regions de Canadà (https://fpcc.ca/), fins a iniciatives d'un o dos voluntaris sense finançament a comunitats remotes.

El projecte que tinc l'honor de liderar descrit en aquest article —i a aquesta pàgina web que s'actualitza periòdicament: https://nrc.canada.ca/en/research-development/research-collaboration/programs/canadian-indigenous-languages-technology-project— es gestiona des del Consell Nacional d'Investigació de Canadà (NRC, per les seves sigles en anglès, *National Research Council of Canada*), que és la principal organització d'investigació i tecnologia del Govern de Canadà. Tanmateix, el lector no hauria de concloure que els governs "colons" (no indígenes) juguen un paper clau a la revitalització de les llengües indígenes a Canadà. Tots els projectes exitosos de revitalització lingüística que em consten els gestionen les comunitats indígenes o hi col·laboren estretament activistes de llengües indígenes.

El nostre projecte *Indigenous Languages Technology* (ILT)[2] al NRC (d'ara endavant, "NRC-ILT") **no** constitueix un projecte de revitalització lingüística: construeix eines que a vegades resulten útils per a les persones que treballen en la revitalització lingüística. L'equip del NRC-ILT som com els tècnics d'il·luminació o tramoistes al teatre: no som els actors (no sortim a l'escenari), ni som els encarregats de la producció. Aquests papers els juguen els activistes de llengües indígenes i les seves comunitats. La revitalització lingüística seguiria endavant sense nosaltres. Hi ha moments, però, on la nostra ajuda tècnica pot facilitar lleugerament la feina. Tot el que hem aconseguit ha estat gràcies a la col·laboració amb les parts interessades indígenes.

Avís: aquest article és l'opinió subjectiva i parcial d'un individu sobre les tecnologies que s'estan aplicant a la revitalització lingüística a Canadà, amb una descripció del projecte NRC-ILT. No pretén ser un estudi exhaustiu del camp; exclou moltes, probablement la majoria de les organitzacions actives en la revitalització lingüística, i els seus èxits. Aquest article és una versió ampliada de (Kuhn et al., 2020), per la qual cosa em sento profundament en deute amb les contribucions dels meus coautors en aquell treball.

1.- Consell cultural dels primers pobles.
2.- Projecte de tecnologies per a llengües indígenes.

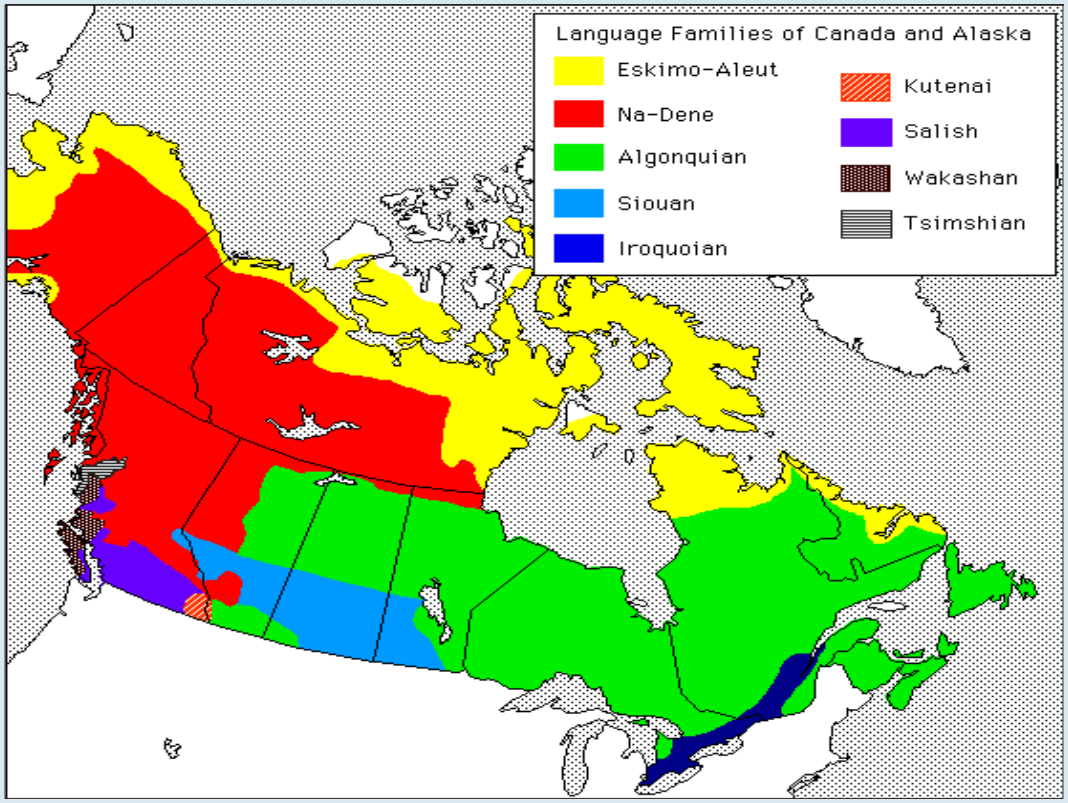

Figura 1. Llengües indígenes a Canadà durant el primer contacte amb europeus .
(Mapa de Matthew Dryer, Universitat de Buffalo)

LES LLENGÜES INDÍGENES A CANADÀ: HISTÒRIA I DEMOGRAFIA

La **Figura 1** mostra la diversitat de famílies de llengües indígenes al que després va esdevenir Canadà, en el moment del primer contacte amb els europeus. Cada família era i és diferent de les altres, en quant a fonètica, vocabulari i sintaxi —en alguns casos, tan diferent com les famílies de llengües germàniques, turques i sinotibetanes d'Euràsia ho són entre elles. Les llengües que eren veïnes geogràfiques, malgrat ser de diferents famílies, a vegades manllevaven paraules i aspectes fonètics o sintàctics entre sí.

Les comunitats indígenes a Canadà sovint es denominen amb tres termes: Inuit, Métis, i Primeres Nacions. Els inuit són un grup ètnic cohesiu, al igual que els métis. No obstant, el terme "Primeres Nacions" agrupa a tots els altres pobles indígenes a Canadà, un conjunt divers d'etnicitats.

La majoria de les llengües indígenes parlades a Canadà són polisintètiques: una sola paraula sovint expressa un significat complex que necessitaria una frase sencera en llengües no polisintètiques. Normalment, una paraula també consta de més morfemes que en altres llengües. Una paraula anglesa consta d'uns 3 morfemes de mitjana; una paraula en mohawk consta d'entre 5-6 morfemes de mitjana. Les llengües mohawk (família iroquesa), cree (família algonquina), i inuktitut (família esquimoaleutiana) són totes polisintètiques, tot i pertànyer a famílies diferents en altres aspectes.

Després del contacte entre els pobles indígenes i els europeus, van sorgir llengües criolles que combinaven aspectes tant de les

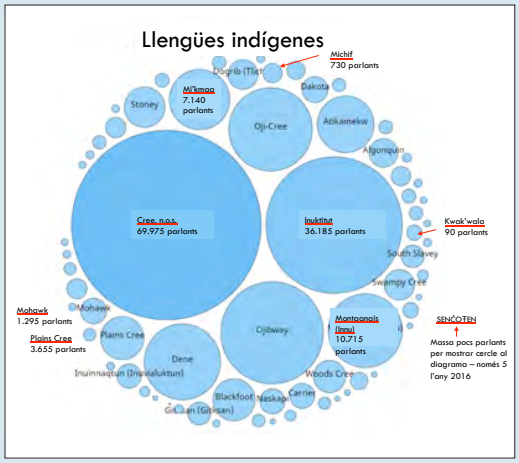

Figura 2. Nombre de parlants per llengua (cens de l'any 2016)
https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/dv-vd/lang/index-eng.cfm

llengües indígenes com europees; la majoria d'elles han desaparegut. Per exemple, el contacte entre comerciants holandesos i els mohawk a la vall del riu Hudson va donar lloc a un crioll holandès-mohawk. L'argot chinook a la seva última forma (va sobreviure fins el segle XX) contenia elements de llengües indígenes de la costa oest, rus, hawaià i xinès.

Els métis van sorgir com a nació pròpia d'una mescla d'avantpassats indígenes i europeus a principis del segle XIX (Bakker, 1997; Rosen & Souter, 2009). El bungi, un crioll parlat pels Métis de Red River, en l'actual Manitoba, constava d'elements de l'anglès escocès d'Orkney, el gaèlic escocès, el francès, el cree i l'anishinaabemowin (ojibwa). Tanmateix, la majoria dels métis parlaven el michif, una llengua de contacte. El michif encara sobreviu i és el protagonista de moltes iniciatives de revitalització lingüística. A grans trets, podríem dir que combina substantius principalment francesos amb verbs en cree de la plana (de la família algonquina); és majoritàriament polisintètic (Davies, Santos, i Souter, 2021; Sammons, 2019).

Després de la Confederació de Canadà l'any 1867, les polítiques governamentals es van enfocar en erradicar les llengües indígenes i les costums indígenes tradicionals, amb l'objectiu d'integrar els pobles indígenes a la societat "blanca". La *Indian Act*, la llei relativa als indígenes, dissuadia i sovint declarava il·legals les trobades culturals i parlar llengües ancestrals. Molts infants indígenes van ésser llevats a la força de les seves comunitats i van ésser col·locats a internats obligatoris ("escoles residencials") pel govern federal. Les esglésies eren les encarregades d'aquestes escoles; algunes les gestionava l'Església catòlica, d'altres, esglésies protestants. S'han documentat múltiples formes d'abús psicològic, físic i sexual que tenien lloc a aquestes escoles. Segons la Comissió canadenca de la Veritat i la Reconciliació[3], el sistema d'escoles residencials es va crear amb l'objectiu de separar els infants aborígens de les seves famílies, per minimitzar i debilitar els lligams familiars i les connexions culturals (Government of Canada (2015), prefaci). A més, durant un temps a mitjans del segle XX, es van separar molts infants indígenes de les seves famílies d'origen per ésser adoptats per famílies no indígenes, denominat *Sixties Scoop* (Fachinger, 2019).

Tot i aquesta història de discriminació sistemàtica, les comunitats indígenes s'han resistit a integrar-se i han persistit en conservar i ensenyar les seves llengües. Els avantatges d'aprendre una llengua ancestral s'han vinculat amb un millor benestar entre els joves (Chandler i Lalonde, 1998; Hallett et al., 2007).

Segons la meva observació personal durant el projecte NRC-ILT, penso que moltes de les llengües indígenes parlades a Canadà

3.- Truth and Reconciliation Commission of Canada.

estan revifant. Quantitats històriques de joves de comunitats indígenes —que sovint s'han criat parlant només anglès o francès (al Quebec)— s'estan inscrivint a cursos que els permetran parlar-se en les seves llengües ancestrals. He conegut a individus indígenes extraordinaris —gairebé tots mal pagats, i alguns voluntaris no retribuïts— que ensenyen aquestes llengües. Aquests activistes lingüístics formen una elit: un grup de persones que combina l'idealisme amb una gran intel·ligència i capacitat per esforçar-se. Ensenyar una llengua polisintètica a persones que originàriament només parlaven anglès o francès no és apte per a dèbils.

La **Figura 2** mostra el nombre de parlants de cada llengua indígena l'any 2016, segons el cens de Canadà (Statistics Canada, 2017). Les llengües subratllades en vermell són amb les quals el projecte NRC-ILT ha interactuat d'alguna manera. Les xifres que es mostren no s'han de prendre al peu de la lletra; hi ha experts que no estan d'acord amb algunes d'elles. Tanmateix, aquesta figura ens dona una idea aproximada del pes demogràfic d'aquestes llengües. La llengua cree de la plana (família algonquina) és la que té més parlants; l'inuktitut (família esquimoaleutiana) ocupa la segona posició. L'inuktitut és, excepcionalment, una important llengua de govern (al territori Nunavut). L'existència de burocràcia governamental que funciona parcialment en inuktitut significa que existeixen molts més recursos escrits i orals per aquesta llengua que per qualsevol altra llengua indígena a Canadà. La disparitat amb la llengua cree, que té molts més parlants però molts menys recursos lingüístics, perquè no hi ha cap òrgan de govern que l'utilitzi, és notable.

El lector no hauria de concloure que la majoria de les llengües "mitjanes" o "petites" mostrades tenen escasses possibilitats a llarg termini. La figura no pot mostrar l'energia relativa dels esforços de revitalització per a cada llengua. Per exemple, la llengua més petita mostrada, saanich (SENĆOŦEN), que només tenia 5 Ancians que dominessin la llengua l'any 2016, compta amb un programa de revitalització lingüística ben gestionat i rigorós que està ensenyant la llengua a molts joves de la comunitat —el següent cens comptarà amb molts més parlants de SENĆOŦEN (Montler, 2018; First Voices Portal, 2021).

De forma similar, la llengua mohawk (kanyen'kéha), amb 1.295 parlants a la **Figura 2**, està experimentant un creixement considerable, gràcies a bones escoles d'idiomes a vàries comunitats mohawk. L'escola de kanyen'kéha per adults a la comunitat Six Nations of the Grand River al sud-oest d'Ontàrio (Onkwawenna Kentyohkwa; vegeu https://onkwawenna.info) és admirada per educadors indígenes a tot Canadà pel seu èxit en graduar molts estudiants que parlen la llengua amb fluïdesa després de dos anys d'immersió intensiva (gairebé s'havia extingit la llengua a Six Nations, tot i que no a totes les comunitats mohawk). Alguns graduats es casen, parlen kanyen'kéha a casa, i crien els seus fills en aquesta llengua, que esdevé la seva llengua materna. L'energia, l'experiència educativa i la cohesió comunitària tenen un gran impacte positiu en la revitalització lingüística —fins i tot poden ser millors indicadors del futur d'una llengua que el nombre actual de parlants que la dominen.

TECNOLOGIES PER A LA REVITALITZACIÓ LINGÜÍSTICA

Cal que els pobles indígenes liderin el desenvolupament de la següent onada de tecnologies lingüístiques responsives i responsables... Per sort, els pobles indígenes de tot el món exerceixen un lideratge tecnològic a mesura que guanyen terreny digital per a les seves llengües (Brinklow, 2021).

Les llengües indígenes parlades a Canadà són **heterogènies** en quant a propietats lingüístiques (10 famílies lingüístiques inconnexes) i en el nombre de parlants per llengua. Els activistes de llengües indígenes a Canadà són molt conscients del notable èxit dels seus iguals a Aotearoa —és a dir, Nova Zelanda, on el maori ha estat llengua oficial des de l'any 1987 (King, 2013)— i a Hawaii (Asensio, 2019). Tanmateix, la revitalització lingüística a Canadà és un repte encara més complex. A Canadà, cada llengua té necessitats diferents, des de les de llengües en perill greu —on la prioritat acostuma a ser documentar un nombre petit d'Ancians que parlen la llengua amb fluïdesa— fins a les de l'inuktitut, una llengua governamental a Nunavut que es parla extensament, on la prioritat és fomentar el seu ús més ampli (p. ex., en entorns mèdics) i millorar l'educació en aquesta llengua.

Dos articles recents (Brinklow et al. 2019, Brinklow 2021) parlen de tecnologies lingüístiques per revitalitzar i recuperar llengües indígenes. Citant al lingüista Francis Tyers, Brinklow senyala que, donat que les grans empreses de programari tenen ànsies de beneficis, les tecnologies comercials se centren en les necessitats "d'un petit nombre de rics, o d'un gran nombre de pobres." Les llengües indígenes a Canadà són parlades per un petit nombre (segons els estàndards mundials) de persones relativament pobres —no els importen gaire a les grans empreses. En la mesura que les empreses sí que estan interessades en les llengües indígenes, sovint volen conservar l'IP de dades lingüístiques i, per tant, amenacen la sobirania de les dades indígenes —Keegan (2019) dona un exemple d'Aotearoa que il·lustra aquest perill.

També existeixen problemes tècnics amb tecnologies lingüístiques comercials:

- Cap de les "grans" llengües (anglès, xinès, japonès, àrab, etc.) per les quals s'han desenvolupat aquestes tecnologies són polisintètiques. Les tecnologies estàndard aplicades a llengües polisintètiques s'enfronten a dificultats —per exemple, el problema de "fora de vocabulari" (OOV, per les seves sigles en anglès, *out of vocabulary*). Fins i tot amb un vocabulari per al reconeixement automàtic de la parla (RAP) d'1,3 milions de paraules inuktitut, Gupta i Boulianne (2020a) van trobar que més del 60% de paraules d'històries inuktitut retingudes **no** estaven incloses en aquell vocabulari (vegeu a continuació). A un experiment comparable realitzat amb texts anglesos, seria d'esperar que aquesta taxa "fora de vocabulari" fos molt més baixa que el 60% —pràcticament segur menys del 5%, possiblement per sota de l'1%.
- Les llengües indígenes a Canadà no compten amb les quantitats massives de dades d'entrenament de les quals depèn l'aprenentatge automàtic. Els membres de l'equip del NRC-ILT estem acostumats a preguntes de persones no indígenes amb pocs coneixements tècnics com: "Perquè no creeu programari per traduir entre el cree (o una altra llengua indígena) i l'anglès?"

La nostra resposta consta de dues vessants:

1. Apart dels inuit, cap altra comunitat indígena amb les quals hem parlat han mostrat gaire interès en la traducció automàtica (TA) entre la seva llengua ancestral i l'anglès (o el francès, al Quebec). Les comunitats solen estar més interessades en eines que fomentin l'aprenentatge i l'ús de la seva llengua ancestral.
2. La majoria de les llengües indígenes a Canadà tenen poques dades paral·leles —sovint es limiten a alguns llibres de la Bíblia traduïts per missioners. L'existència de dialectes a la majoria d'aquestes llengües agreuja aquest problema d'escassetat de dades. (Per la llengua cree, el nostre equip i els nostres col·laboradors indígenes

potser podrien arreplegar com a molt 100.000 frases en cree de la plana en paral·lel amb l'anglès; això no seria suficient per entrenar un sistema de TA).

Ens trobem amb malentesos sobre aquesta situació fins i tot entre experts en processament del llenguatge natural. Alguns experts han insinuat que estem endarrerits tècnicament perquè utilitzem plantejaments basats en regles a la majoria de les nostres tasques. Al cap i a la fi, d'altres equips han aconseguit aplicar exitosament l'aprenentatge automàtic a llengües amb pocs recursos... El terme "llengua amb pocs recursos" pot dificultar la comprensió. S'aplica a llengües com el tàmil, amb 75 milions de parlants, la majoria d'ells alfabetitzats en la seva llengua, amb una història de texts escrits que es remunten milers d'anys. És ridícul emprar el mateix terme per descriure la llengua indígena "més gran" a Canadà, cree de la plana, amb 75.000 parlants i pocs texts escrits. La majoria dels subprojectes del NRC-ILT tracten llengües amb "recursos **extremadament** escassos".

L'inuktitut és una excepció a aquesta situació de "recursos extremadament escassos" (vegeu a continuació).

Littell, Kazantseva et al. (2018) resumeix les tecnologies per a la revitalització lingüística a Canadà. A la següent llista extreta d'aquest article, les tecnologies en **negreta** van ésser creades pel NRC-ILT amb col·laboradors indígenes; les tecnologies subratllades van ésser creades de forma externa al NRC amb l'ajuda de finançament del NRC-ILT:

- **Implementació de fonts i formats digitals per a llengües indígenes amb conjunts de caràcters poc comuns;**
- **Text predictiu;**
- **Conversió entre diferents ortografies per la mateixa llengua;**
- Correcció ortogràfica *(sense participació del NRC-ILT)*;
- **Generació de paradigmes per ensenyar morfologia complexa;**
- **Cerca aproximada;**
- **Traducció automàtica;**
- Tecnologies de veu:
  - Reconeixement automàtic de la parla;
  - Cerca de paraules clau a través de dades de veu;
  - **Alineació de veu a text**;
  - **Conversió de text a veu** *(tasca preliminar del NRC-ILT)*;
- Tecnologies d'imatge;
- Aprenentatge de llengües assistit per ordinador (CALL, per les seves sigles en anglès, *computer-aided language learning*).

Littell, Kazantseva, et al. (2018) al·ludeix breument a quatre altres activitats tecnològiques importants:

- Creació d'eines per documentar i anotar discursos d'Ancians que parlen la llengua amb fluïdesa;
- Digitalització de cintes magnètiques i d'àudio de discursos d'Ancians;
- Conservació d'arxius de discursos;
- Elaboració de diccionaris digitals.

LA HISTÒRIA I L'ENFOCAMENT COL·LABORADOR DEL PROJECTE

Trobareu un resum acadèmic del NRC-ILT a (Kuhn, Davis, Désilets, Joanis et al., 2020), i un informe tècnic detallat a https://nrc-publications.canada.ca/eng/view/object/?id=d4f10144-c711-43c5-b80b-5ace7df5e68b. Les actualitzacions periòdiques es troben a https://nrc.canada.ca/en/research-development/research-collaboration/programs/canadian-indigenous-languages-technology-project.

El projecte va començar amb el pressupost de març de l'any 2017 del Govern de Canadà, que proporcionava 6 milions de dòlars canadencs (CAD) al NRC per

Owennatékha (Brian Maracle): fundador i director, escola d'immersió Onkwawenna Kentyohwa.Va proposar **Kawennón:nis,** i va col·laborar en la seva creació.

Akwiratékha Martin: professor de kanyen'kéha a Kahnawake – col·labora amb l'equip de NRC-ILT per crear la versió oriental de **Kawennón:nis.**

Kawennón:nis - conjugador verbal de kanyen'kéha

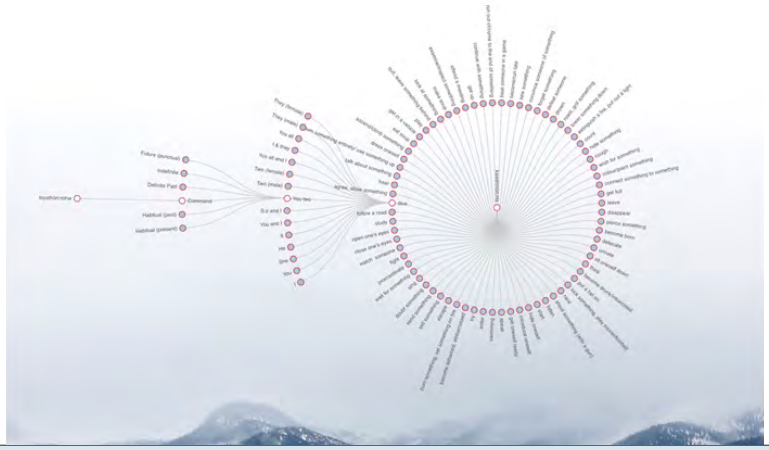

Figura 3. Col·laboracions en kanyen'kéha (mohawk).

desenvolupar, amb la col·laboració de parts interessades indígenes, tecnologies de programari per donar suport a llengües indígenes. Es va considerar que la nostra secció del NRC estava capacitada per fer aquesta tasca gràcies a la nostra experiència amb tecnologies per a grans llengües mundials —p. ex. havíem treballat en la TA del xinès i l'àrab a l'anglès, amb finançament dels programes GALE[4] i BOLT[5] de l'Agència d'Investigacions de Projectes Avançats de Defensa (DARPA, per les seves sigles en anglès, *Defense Advanced Research Projects Agency*) dels Estats Units. El NRC també havia dut a terme una iniciativa a petita escala centrada en l'inuktitut durant els anys 2003-2012 (Martin et al., 2003; Farley, 2012).

L'equip del NRC-ILT procura complir amb el seu mandat desenvolupant relacions estretes i respectuoses amb comunitats indígenes i intentant trencar amb la llarga i dolorosa història de pràctiques de recerca extractives (Keegan, 2019; Brinklow et al., 2019). Creiem en la filosofia de "l'empoderament", on la recerca es fa de manera col·laborativa, amb el mateix èmfasi en les intencions de l'investigador i la comunitat (Czaykowska-Higgins, 2009).

Això ha significat preguntar als activistes de llengües indígenes quines eines de programari trobarien **ells** útils, enlloc d'oferir tecnologies basant-se en temes de recerca interessants. D'aquí la col·lecció

4.- Global Autonomous Language Exploitation.
5.- Broad Operational Language Translation.

una mica inconnexa de tecnologies que es descriuen a continuació: hem donat resposta a les diferents necessitats de diferents comunitats. Ens va guiar un comitè consultiu format per experts en la revitalització de llengües indígenes. El seu consell ha estat de valor incalculable. El NRC-ILT no s'ha declarat mai propietari de les dades de llengües indígenes recollides amb el finançament del projecte. Quan ha estat possible, hem implementat un model de "forn i ceràmica": el programari de codi obert (OSS per les seves sigles en anglès, *open-source software*) independent de la llengua (el forn) crea programari específic per la llengua (la ceràmica) que serà propietat de la comunitat.

El finançament inicial va ésser més generós del que esperàvem, així que vam finançar tasques dins i fora del NRC. El NRC-ILT té tres "cercles":

- Un cercle intern de feina duta a terme al NRC amb la col·laboració de parts interessades indígenes.
- Un cercle mitjà de feina relacionada amb el reconeixement automàtic de la parla (RAP) duta a terme pel Centre de Recerca Informàtica de Montreal (CRIM per les seves sigles en francès, *Centre de Recherche Informatique de Montréal*). El CRIM té una il·lustre trajectòria de recerca pionera en RAP i tecnologies relacionades. El CRIM va entregar resultats experimentals de RAP per a dues llengües indígenes importants, i també l'OSS que accelera una fase clau en l'anotació de discursos gravats.
- Un cercle extern d'altres subprojectes curosament seleccionats per l'equip del NRC-ILT i el seu comitè consultiu indígena per rebre finançament. La majoria d'aquests subprojectes estaven gestionats per organitzacions indígenes; tots ells comptaven amb una plantilla gairebé completament indígena.

CERCLE INTERN: TECNOLOGIES DESENVOLUPADES DINS DEL NRC

CONJUGADORS VERBALS

El primer subprojecte del NRC-ILT va sorgir d'una proposta d'Owennatékha (Brian Maracle), el director de l'escola Onkwawenna Kentyohkwa (Our Language Society) a Six Nations of the Grand River al sud-oest d'Ontàrio: https://onkwawenna.info/. Aquesta escola és coneguda per produir parlants que dominen el kanyen'kéha (mohawk). Owennatékha va proposar a l'equip del NRC-ILT de crear Kawennón:nis (the WordMaker) —una eina que ajuda a estudiants a dominar el complex sistema verbal del kanyen'kéha. Fins i tot pels temes verbals més comuns en kanyen'kéha existeixen tantes possibles conjugacions verbals que una eina de referència física seria impossiblement gran; no obstant això, és factible computar aquestes conjugacions mitjançant un programari (Kazantseva et al, 2018). Kawennón:nis, implementat pel dialecte occidental del kanyen'kéha, es va sotmetre a extenses proves d'usuari amb estudiants i professors a l'escola abans del seu llançament. L'han rebut amb entusiasme. Està dissenyat per complementar el pla d'estudis, i utilitza les metàfores visuals que empren els professors de l'escola: https://kawennonnis.ca/wordmaker.

Posteriorment, el NRC-ILT va elaborar sistemes genèrics per crear conjugadors verbals per a llengües polisintètiques. S'estan aplicant per crear conjugadors verbals pel kanyen'kéha oriental parlat a Kahnawake (Quebec), i per dues llengües no relacionades al kanyen'kéha: La llengua algonquina anishinaabemowin parlada a Kitigan Zibi (a prop de Maniwaki, Quebec), i el michif.

Figura 4. READALONG STUDIO – UN ÈXIT INESPERAT!

La **Figura 3** mostra dos dels nostres col·laboradors mohawk. Tots dos ensenyen kanyen'kéha: Owennatékha ensenya el dialecte occidental, i Akwiratékha el dialecte oriental. (D'altres educadors mohawk també han ajudat a crear i provar el Kawennón:nis). La figura també mostra la interfície d'usuari de Kawennón:nis.

Heather Souter del Prairies to Woodlands Indigenous Language Revitalization Circle actualment anima als estudiants dels seus cursos a provar la versió beta del conjugador verbal de michif. Davies, Santos, i Souter (2021) descriuen obres relacionades.

Tenim previst llançar conjugadors verbals per a vàries altres llengües. A més, el Dr. Patrick Littell dirigeix una iniciativa per desenvolupar un sistema fàcil d'utilitzar basat en fulls de càlcul anomenat "Gramble" que podria facilitar la feina d'activistes lingüístics per crear conjugadors **de forma independent** (sense els coneixements del NRC).

READALONG STUDIO

Aquest és "l'èxit inesperat" del NRC-ILT: no anticipàvem la quantitat de professors que estarien interessats en afegir una funcionalitat clau a audiollibres i vídeos preexistents en les seves llengües. Aquesta funcionalitat ha estat promoguda per l'equip de la Prof. Marie-Odile Junker a la Universitat de Carleton.

La funcionalitat s'il·lustra a la **Figura 4**; és molt senzilla. A mesura que s'escolta un audiollibre o es veu un vídeo amb discurs indígena (aquí, atikamekw), la paraula parlada apareix destacada al text acompanyant. A la figura, s'està escoltant la paraula "atisokasotcik", i per això apareix destacada. Si l'oient —un estudiant o professor— vol centrar-se en la pronunciació d'una paraula dins del text, fa clic a sobre la paraula, i n'escoltarà la pronúncia; també pot alentir la reproducció. Nosaltres (els equips de la Universitat de Carleton i del NRC-ILT)

anomenem els audiollibres o vídeos amb aquesta funcionalitat “ReadAlongs” (llibres o vídeos amb lectura gravada).

Moltes comunitats tenen llibres o vídeos educatius amb discursos en les seves llengües, amb una transcripció en text de les paraules parlades (o cantades). Per poder afegir la lectura gravada a aquests llibres o vídeos, cal que les paraules del discurs s’alineïn amb les paraules del text. L’equip de Carleton estava fent aquestes alineacions a mà quan vam començar a col·laborar junts. La nostra contribució va ésser automatitzar aquest procés, utilitzant un programari que vam crear anomenat “ReadAlong Studio”. Delasie Torkornoo de l’equip de Carleton ens ha estat ajudant a millorar el codi de ReadAlong Studio. Ambdós equips han produït i enviat ReadAlongs a educadors indígenes en les llengües algonquines, atikamekw, cree sud-oriental, cree nord-oriental, gitksan, inuktitut, kwak’wala, kanyen’kéha (mohawk), seneca i SENĆOŦEN. Cada quinze dies rebem noves sol·licituds per convertir materials educatius d’aquesta manera. Els educadors indígenes ens comenten que els ReadAlongs tenen un alt valor pedagògic.

TECLATS, CONVERSIÓ ORTOGRÀFICA, I TEXT PREDICTIU

El NRC-ILT ha llançat un OSS que implementa teclats per a alguns sistemes d’escriptura mal atesos, i que converteix entre sistemes d’escriptura per a algunes llengües. Eddie Antonio Santos de l’equip va treballar amb les llengües saulteaux, makah i cree de la plana. Juntament amb el lingüista Arok Wolvengrey de la Universitat d’Alberta, també va persuadir a Google per canviar el seu teclat sil·làbic pel cree de la plana als Chromebooks (la versió anterior era adequada pel cree oriental, però no pel cree occidental).

El NRC-ILT també ha llançat codi per a la predicció de text per a dispositius mòbils. En teoria, el programari pot implementar la predicció de text per a qualsevol llengua. No obstant això, és difícil d’utilitzar per a usuaris no experts. Fins ara, només hem implementat la predicció de text per a la llengua SENĆOŦEN. Tanmateix, donat que l’ortografia del SENĆOŦEN és difícil, els membres d’aquesta comunitat ens han comentat que estan molt contents amb aquesta funcionalitat —fa que sigui molt més fàcil introduir text als seus dispositius.

INUKTUT

Tota la feina feta a aquesta secció ha estat gràcies a la col·laboració amb el Pirurvik Centre, un centre cultural i lingüístic gestionat pels inuit (https://www.pirurvik.ca/).

Les llengües o dialectes relacionades i parlades pels inuit a Nunavut s’anomenen col·lectivament “inuktut”. La majoria de la feina del NRC-ILT amb l’inuktut ha estat relacionada amb la versió inuktitut, parlada a Baffin Island. Aquesta feina difereix del nostre treball amb altres llengües de dues maneres: 1. No vam crear eines educatives, sinó **eines ofimàtiques que ajuden a escriure o llegir text inuktut** (o traduir entre l’inuktut i l’anglès); 2. Vam treballar en la **traducció automàtica (TA)** entre l’inuktut i l’anglès, mitjançant l’aprenentatge automàtic. Aquestes diferències existeixen perquè hi ha una burocràcia i un òrgan legislatiu que funcionen parcialment en inuktut: aquestes organitzacions necessiten texts escrits en inuktut, i proporcionen una quantitat creixent de texts bilingües inuktut-anglès que es poden utilitzar per entrenar sistemes de TA.

Una iniciativa a petita escala al NRC durant els anys 2003-2012 per crear eines per l’inuktut va donar resultats valuosos —el més important, un analitzador morfològic.

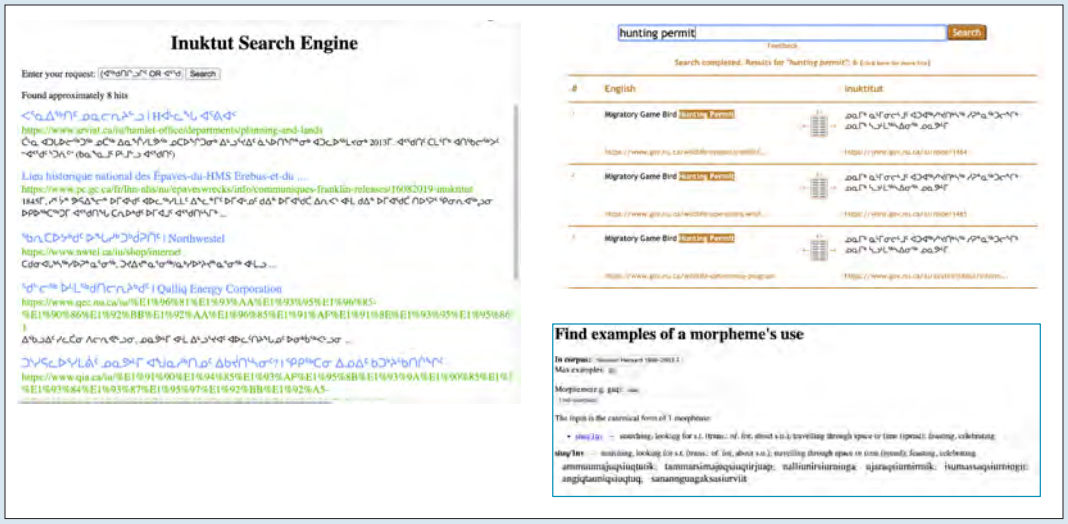

Figura 5. Cercador d'inuktut, WeBInuk (programa de concordança), i exemple de cerca per morfema

Un col·laborador clau va ser Benoît Farley, que tot i haver-se retirat del NRC manté un repositori per a aquestes eines (Farley, 2012).

Una altra producció del NRC durant aquest període va ser la publicació consecutiva de corpus paral·lels de texts anglès-inuktut basat en les actes de l'Assemblea Legislativa de Nunavut, el "Nunavut Hansard" (d'ara endavant, "NH"). El primer d'aquests, el NH 1.0, es va cedir a la comunitat d'investigadors l'any 2003 (Martin et al., 2003); cobria actes de 155 dies de l'Assemblea de Nunavut, i constava de 3.432.212 símbols anglesos i 1.586.423 símbols inuktitut.

Quan el NRC-ILT va reprendre el seu treball amb l'inuktut l'any 2017, es disposava d'actes de molts més anys de NH sense alinear. Vam **crear i publicar un nou corpus alineat**, el NH 3.0. Consta de 8.068.977 símbols inuktitut i 17.330.271 símbols anglesos; comprèn les actes de 687 dies de debat des de l'1 d'abril de 1999 fins el 8 de juny de 2017 (Joanis et al., 2020).

La dimensió del NH 3.0 fa que sigui factible entrenar sistemes de **traducció automàtica** (TA). Vam ajudar als organitzadors del WMT 2020, una avaluació internacional anual de sistemes de TA, a establir avaluacions per a TA entre l'anglès i l'inuktitut en ambdues direccions. Juntament amb Microsoft, també vam invertir en avaluadors humans que dominaven l'inuktitut perquè valoressin les traduccions de dades de prova cap a l'inuktitut de part de sistemes de la competència. La naturalesa extremadament polisintètica de l'inuktitut, insòlit d'entre les llengües del món, va atraure molts equips d'investigació a aquest repte de TA.

Vam crear els nostres propis sistemes de TA que van competir a aquesta avaluació (Knowles et al., 2020). Els resultats globals es troben aquí: http://www.statmt.org/wmt20/translation-task.html. Esperem que es mantingui l'interès per la recerca en TA en aquest parell lingüístic i que finalment resulti en eines pràctiques per l'inuktitut.

Seguim millorant les eines ofimàtiques per l'inuktitut que el NRC va llançar els anys 2003-2012, i seguim creant-ne de noves:

- Estem millorant l'analitzador morfològic de Benoît Farley.
- Estem recuperant el projecte WeBInuk, una eina del NRC basada en la web que cerca

traduccions inuktitut de paraules angleses. Mostra frases en anglès i inuktitut on la frase anglesa conté la paraula clau cercada. Hi havia traductors inuktut que utilitzaven molt WeBInuk els anys 2007-2017, però la direcció del NRC tenia altres prioritats i es va permetre que desaparegués.

- Estem creant un lematitzador que a l'importar una paraula inuktitut exporta la paraula més comuna amb la mateixa arrel. Aquest és un pas clau per donar suport a la cerca monolingüe en inuktitut, o per cercar en la direcció inuktitut —> anglès. És molt més difícil crear un cercador per a usuaris que introdueixen paraules de cerca en inuktitut que en anglès o francès, degut al seu caràcter polisintètic. (És probable que qualsevol paraula inuktitut que introdueixi l'usuari no hagi existit mai en la història de la llengua, i segurament no tornarà a existir mai més). Cal fer cerques difuses. La **Figura 5** il·lustra els avenços que hem fet amb aquesta i d'altres eines relacionades.

CONVERSIÓ PRELIMINAR DE TEXT A VEU (TTS PER LES SEVES SIGLES EN ANGLÈS, *TEXT TO SPEECH*)

El pla d'estudis a Onkwawenna Kentyohkwa se centra en dominar la complexa morfologia verbal del kanyen'kéha. Per aquest motiu, el NRC-ILT i l'escola van desenvolupar conjuntament Kawennón:nis (descrit anteriorment).

Kawennón:nis té una limitació: els estudiants poden generar paraules complexes en format text, però potser no saben com pronunciar-les. L'objectiu de l'escola és produir *parlants* que dominin la llengua, no només *lectors/escriptors*, així que no concorda el que produeix Kawennón:nis (text) amb el que més necessiten els estudiants (parla per imitar-la). Tot i això, de la mateixa manera que no seria factible escriure i publicar Kawennón:nis en format llibre degut al seu enorme vocabulari implícit, seria igual de poc factible **gravar** totes les formes. Això suggereix la implementació de TTS; podem entrenar un sistema de TTS sobre un subconjunt gravat i equilibrat fonèticament del vocabulari?

Vam crear un prototip de TTS concatenatiu. Reordena unitats de subparaules de 852 gravacions de conjugacions de parlants que dominen la llengua per cobrir completament les primeres 122.966 conjugacions a Kawennón:nis. Els professors de l'escola van declarar que la qualitat del discurs de sortida era més que acceptable. Recentment, també hem desenvolupat un model de TSS neuronal capaç de produir parla sintetitzada de qualitat acceptable a partir de tan sols 30 minuts d'àudio. Owennatékha va declarar fa poc que el TTS és una de les seves tres prioritats tecnològiques per al kanyen'kéha.

Kawennón:nis il·lustra els reptes als quals s'enfronten els elaboradors de plans d'estudis indígenes: llengües complexes, pocs parlants per ensenyar la llengua, pocs materials gravats, i pocs parlants per gravar nous materials i, per tant, grans parts de la llengua als quals els estudiants no poden accedir ni escoltar. El TTS permet que organitzacions amb pocs recursos puguin aprofitar al màxim el seu recurs més preciós i limitat —el temps dels parlants que dominen la llengua— al fer ús d'un petit nombre de gravacions per cobrir un domini molt més ampli.

CERCLE MITJÀ: TECNOLOGIES DE VEU DESENVOLUPADES PEL CRIM

L'anotació i la transcripció són la càrrega més gran a la majoria de projectes de documentació i conservació lingüística. Cox et al. (2019) parlen del "coll d'ampolla de la transcripció": s'estan gravant discursos (especialment d'Ancians) en llengües

indígenes de forma més ràpida del que s'estan transcrivint, o fins i tot anotant (s'hauria d'anomenar "el coll d'ampolla de l'anotació"). Això passa a llengües minoritàries arreu del món. L'anotació i especialment la transcripció porten temps, i són tasques tedioses. Resulta temptador gravar moltes hores de discursos i planejar dur a terme aquests passos tediosos "més endavant", però sovint es posposen indefinidament...

Això passa des de fa dècades, molt abans de que es disposés de mitjans de gravació digital. Per tot Canadà, existeixen milers d'hores de discursos en llengües indígenes a cintes magnètiques o cassets d'àudio que són recursos potencialment valuosos per a la revitalització lingüística, però que han esdevingut inaccessibles a la pràctica perquè ningú sap què contenen. Estan repartides per comunitats i departaments universitaris (he sentit històries de cintes magnètiques amb gravacions importants que s'han trobat sense etiquetar a les prestatgeries del garatge d'un professor difunt). Ningú té temps d'escoltar la majoria d'aquest tipus de gravacions; en alguns casos tràgics, no queda ningú que entengui la llengua gravada. Sovint les gravacions més antigues no s'han copiat a mitjans digitals.

RECERCA PRINCIPAL

Vam fundar el Centre de Recerca Informàtica de Montreal (CRIM per les seves sigles en francès, *Centre de Recherche Informatique de Montréal*) per fer recerca sobre el reconeixement automàtic de la parla (RAP) per un petit nombre de llengües indígenes parlades a Canadà. Un RAP amb vocabulari complet per aquestes llengües no és realista a curt termini, perquè no existeixen les grans quantitats de dades de transcripció de veu paral·leles necessàries per entrenar els sistemes de RAP. No disposem de suficient finançament per crear aquestes dades. Per tant, l'objectiu era crear sistemes de RAP imperfectes entrenats amb quantitats petites de dades paral·leles. S'ha demostrat que aquests tipus de sistemes han fet que la "cerca d'àudio" (també anomenada "reconeixement de paraules clau" o "detecció de termes parlats") sigui factible per a d'altres llengües. Quan l'exactitud del RAP és baixa, encara és possible que els usuaris trobin certes paraules o frases en discursos mitjançant tècniques de cerca d'àudio.

Els experiments de RAP al CRIM s'han centrat en l'inuktitut i el cree oriental. Els investigadors del CRIM van gestionar la transcripció de gairebé 81 hores de discurs inuktitut, i 102 hores de discurs cree oriental; van dividir aquestes dades paral·leles en conjunts de prova i d'entrenament. La seva feina posterior en el RAP de inuktitut i cree oriental es descriu a (Gupta i Boulianne, 2020a) i (Gupta i Boulianne, 2020b), respectivament. La feina va ser possible gràcies a la col·laboració amb el Pirurvik Centre (https://www.pirurvik.ca/) i la Canadian Broadcasting Corporation (https://www.cbc.ca/).

Una qüestió clau per l'equip de CRIM era quines unitats utilitzar pel RAP. L'inuktitut és molt polisintètic; per tant, la llengua té una taxa molt alta de "fora de vocabulari" (60%) — moltes paraules només s'expressen un cop. Un sistema de RAP amb un vocabulari format per paraules no reconeixeria bé la majoria de les paraules a dades de veu inuktitut noves. Els investigadors del CRIM van provar-ho amb morfemes, síl·labes i fonemes com a unitat bàsica de RAP, i van obtenir els millors resultats amb síl·labes. Però fins i tot amb unitats sil·làbiques, al voltant d'un 71% de les paraules inuktitut utilitzades a les dades de prova no es van reconèixer correctament.

No obstant això, a obres no publicades, aquests investigadors van aconseguir un rendiment de reconeixement de paraules clau

relativament bo per l'inuktitut (*comunicació personal*). L'enfocament amb el millor rendiment era convencional: cercar a través d'una matriu de fonemes generada per RAP basat en síl·labes, amb una matriu de confusió ponderada que aporta robustesa als errors de RAP.

Als experiments del CRIM, el cree oriental té una taxa de "fora de vocabulari" significativament més baix que l'inuktitut, tot i que es disposava de quantitats molt més petites de dades textuals per crear un vocabulari de cree oriental. Amb un vocabulari de 30.000 paraules obtingut de textos en dos gèneres —transcripcions de vídeos i traduccions de la Bíblia—, la taxa de "fora de vocabulari" per a dades de text retingudes era del 25% pels vídeos i del 9% per les dades bíbliques. Aquestes xifres van fer que les paraules fossin una unitat factible pel RAP en cree oriental. A experiments de RAP sobre dades de veu retingudes de cree oriental, el 70% de les paraules als vídeos de prova no es van reconèixer correctament, així com el 25% de les paraules a les dades bíbliques de prova. No es van realitzar experiments de reconeixement de paraules clau pel cree oriental.

Els investigadors del CRIM van obtenir resultats experimentals molt prometedors (taxes baixes de reconeixement erroni) tant per l'inuktitut com el cree oriental quan el sistema de RAP **depenia** del parlant. Els pobres resultats anteriorment citats s**ó**n **independents** del parlant. No obstant, per a la revitalitzaci**ó** de lleng**ü**es ind**í**genes, on sovint es recullen moltes hores de discurs d'un petit nombre d'Ancians que parlen la llengua amb flu**ï**desa, un sistema de RAP que **depèn** del parlant seria **ú**til a nivell pr**à**ctic. Considerem una manera de treballar on (p.ex.) un expert hum**à** transcriu un parell d'hores del discurs d'un Anci**à**, com es fa actualment. Llavors, aquestes dues hores transcrites s'utilitzarien per entrenar un sistema de RAP adaptat al discurs d'aquest Anci**à** en particular, la qual cosa podria produir un esborrany de la transcripci**ó** de les nombroses hores restants de discurs de la mateixa persona. Aix**ò** alleujaria el "coll d'ampolla de la transcripci**ó**".

EINES DE PRODUCTIVITAT

L'equip del CRIM també va publicar eines que faciliten les fases inicials del processament de discursos gravats. Aquestes eines estaven empaquetades com a serveis web a la plataforma VESTA del CRIM (http://vesta.crim.ca) i estan disponibles allà o a través d'una extensió ELAN —ELAN és una eina d'anotació a https://archive.mpi.nl/tla/elan/ (Wittenburg et al., 2006). Les eines permeten la segmentació d'arxius de veu, en veu i no veu (silenci, soroll, música, etc.), la recuperació lingüística (trobar segments parlats en una llengua concreta, podent-se identificar 32 llengües), la recuperació per parlant (trobar segments parlats per un parlant concret), la detecció d'activitat de veu multicanal (detectar segments que contenen veu de forma separada per a cada pista d'una gravació multicanal amb més d'un micròfon), i d'altres capacitats útils.

Els Drs. Chris Cox i Olivia Sammons (de la Universitat de Carleton i la Universitat First Nations, respectivament), van fer servir aquestes eines CRIM recentment mentre recollien dades de veu michif. Informen d'una acceleració quadruplicada o quintuplicada per a les fases inicials d'anotació (*comunicació personal*).

CERCLE EXTERN: ALTRES TASQUES FINANÇADES PEL NRC-ILT

Aquesta secció defineix els subprojectes de revitalització de llengües indígenes a les

quals ha col·laborat el finançament del NRC-ILT, almenys parcialment, però que han estat gestionats des de fora del NRC.

CURSOS I JOCS EN LÍNIA

S'han d'actualitzar periòdicament les eines d'aprenentatge lingüístic en línia perquè el programari subjacent ha esdevingut obsolet. Per exemple, Adobe va deixar d'oferir suport pel Flash Player a principis de l'any 2021. El NRC-ILT va finançar mesures correctives per a diversos subprojectes.

Plataformes educatives a la Universitat de Carleton:
La Prof. Marie-Odile Junker i el seu equip a la Universitat de Carleton porten anys col·laborant amb companys indígenes per desenvolupar lliçons en línia per a llengües algonquines. Utilitza un sistema d'acció participativa per treballar amb les comunitats. Les contribucions del seu equip són massa nombroses per enumerar-les aquí —vegeu https://www.marieodilejunker.ca/. La Prof. Junker va rebre el *Governor General's Innovation Award* (2017).

El finançament del NRC-ILT va començar a ajudar el projecte ja existent a Carleton l'any 2018. Els canvis de programari havien deixat parcialment penjades les eines educatives lingüístiques a la plataforma web de Carleton per l'innu (no confondre amb l'inuktut, una llengua independent) i el cree oriental. El finançament va ajudar a pagar una actualització tecnològica (eliminació de la dependència de Flash i d'altres tecnologies obsoletes) i l'experiència de dos experts en llengües indígenes que van contribuir amb grans quantitats de continguts pedagògics. L'equip de la Prof. Junker segueix desenvolupant la plataforma amb altres fonts de finançament (innu: https://lessons.innu-aimun.ca/; cree oriental: https://lessons.eastcree.org/).

7000 Languages:
Algunes comunitats indígenes volien desenvolupar cursos en línia per a les seves llengües. Vam proporcionar un finançament modest a 7000 Languages, una organització sense ànim de lucre que crea programari educatiu lingüístic, per crear cursos juntament amb aquestes comunitats (a les quals també vam finançar). Tres d'aquests cursos es troben en línia a la web de 7000 Languages, www.7000.org: per al kwak'wala, el michif, i el mi'kmaq.

Aprenentatge de llengües assistit per ordinador (CALL) a la Universitat d'Alberta:
Aquest subprojecte dona suport al dialecte Y del cree (cree de la plana). L'equip de la Universitat d'Alberta, juntament amb membres de comunitats que parlen cree, ha estat creant un sistema CALL adaptatiu anomenat "CreeTutor". També s'han creat continguts en la llengua cree: s'han gravat, transcrit i traduït 13 històries personals de 8 Ancians. CreeTutor es publicarà aviat. Un element clau de CreeTutor és *Sound Hunters* [caçadors de sons], un exercici de consciència fonèmica receptiva (Lothian et al., 2020). La principal creadora de *Sound Hunters,* Delaney Lothian (cree-métis), és una estudiant de llicenciatura en ciències a la Universitat d'Alberta i membre a temps parcial del NRC-ILT. Està treballant en la versió michif de *Sound Hunters*.

FirstVoices:
El First Peoples' Cultural Council (FPCC) a la Columbia Britànica té una sòlida trajectòria proporcionant tecnologies d'última generació, formació i suport tècnic a activistes de llengües indígenes (principalment però no exclusivament a la província), dins del programa FirstVoices (http://www.fpcc.ca/about-us/). Language Tutor [tutor lingüístic], que permet a les comunitats crear lliçons de llengua, forma

part de FirstVoices i s'havia quedat "obsolet" a nivell tecnològic. El finançament del NRC-ILT va ajudar a reconduir la situació.

On the Path of the Elders:
"On the Path of the Elders" [Al camí dels Ancians] és un joc de rol dissenyat per donar a conèixer temes històrics i culturals relacionats amb la regió de James Bay. Quan aquest joc en línia gratuït es va llançar l'any 2007, va rebre molts elogis pel seu ús innovador de recursos històrics. Va perdre funcionalitats degut a canvis en la indústria del programari subjacent. El NRC-ILT va proporcionar finançament per actualitzar el programari i afegir continguts en cree swampy. La pàgina web de "On the Path of the Elders" ara és compatible amb mòbils i amb els navegadors Chrome, Firefox i Internet Explorer: www.pathoftheelders.com.

TALLERS DE DESENVOLUPAMENT D'HABILITATS

Dos subprojectes finançats pel NRC-ILT van proporcionar formació per a persones que documenten llengües indígenes. El NRC-ILT també va proporcionar un finançament modest a l'activista lingüística Caroline Running Wolf; va organitzar quatre exitoses sessions en línia per a activistes lingüístics, els "Community Workshops for Indigenous Language Technology" (CWILTs, tallers comunitaris de tecnologies per a llengües indígenes) del desembre de 2020 al juliol de 2021.

Yukon Native Language Centre (YNLC)[6]:
A Yukon es parlen vuit llengües indígenes (gwich'in, hän, kaska, tutchone septentrional, tutchone meridional, tagish, tlingit, i alt tanana): gairebé el 15% de les llengües indígenes a Canadà. Amb l'ajuda del finançament del NRC-ILT, el YNLC va donar suport a 12 estudiants en un programa de 10 mesos durant el qual van adquirir habilitats pràctiques per elaborar, difondre i conservar materials digitals per aquestes llengües a les seves comunitats. La pandèmia del COVID-19 va interrompre algunes de les activitats planejades a aquest subprojecte, que estava centrat en activitats presencials, però la gestió competent del personal del YNLC va fer que fos un èxit igualment.

- Al Taller 1 (octubre de 2019), els estudiants van practicar com utilitzar equips de vídeo i com fer entrevistes amb parlants que dominen la llengua. També van rebre formació en gestió d'arxius.
- Al Taller 2 (novembre de 2019), van rebre formació sobre l'ús dels paquets de programari d'ELAN i SayMore per transcriure, anotar i traduir gravacions de veu.
- El Taller 3 (febrer de 2020) va començar amb una revisió i posada en pràctica de les habilitats apreses fins al moment. També va incloure compartir i reutilitzar vídeos, i utilitzar materials ELAN com a recurs per a l'aprenentatge i l'ensenyament de llengües. L'últim dia, els estudiants van mostrar els seus vídeos acabats.

Els 12 estudiants van crear 548 minuts de documentació i van arribar a dominar habilitats que els permetran gravar als Ancians dins de les seves comunitats de forma continuada.

Indigitization[7]:
Aquesta és una col·laboració entre la Universitat de la Columbia Britànica, Musqueam Archives, i el centre d'educació

6.- Centre de llengües natives de Yukon.
7.- El nom del projecte és un joc de paraules que combina els conceptes de *indigenization* (indigenització) i *digitization* (digitalització).

cultural Heiltsuk Cultural Education Centre que se centra en la digitalització de dades de veu. Mentre que el subprojecte YNLC formava a membres de les comunitats indígenes per recollir dades **noves**, el subprojecte continuat Indigitization forma a persones per convertir dades **antigues** a formats digitals accessibles. Ajuda a que les comunitats indígenes guanyin autonomia al familiaritzar-los amb les opcions de programari i maquinari per a la digitalització. Per a comunitats que ja tenen continguts digitalitzats, el subprojecte ofereix formació de transcripció de discursos a llengües indígenes, la traducció de transcripcions a l'anglès i al francès, i la gestió dels arxius resultants.

El projecte Indigitization va començar al gener de 2020, i es va haver de reorganitzar degut a la crisis del COVID-19 (els tallers presencials van esdevenir poc aconsellables). Es va dedicar més recursos a la creació de recursos instructius que al pla original: "Tenim una oportunitat única per redactar guies pràctiques basades en l'experiència que ajudaran a organitzacions indígenes petites, sovint poc finançades, a començar a gestionar les seves col·leccions de forma estructurada" (*comunicació personal amb Gerry Lawson, cap del projecte Indigitization*).

S'han elaborat els continguts de text de 65 d'aquestes noves guies; actualment s'està realitzant l'edició i el disseny gràfic. Aquest subprojecte pensa prestar maquinari a les comunitats quan sigui necessari.

CWILTs:

Els "Community Workshops for Indigenous Language Technology" (CWILTs, tallers comunitaris de tecnologies per a llengües indígenes) són una sèrie de sessions en vídeo per a activistes de llengües indígenes organitzat per Caroline Running Wolf de Apsáalooke Nation (crow), cadascuna de les quals inclou xerrades ràpides centrades en el tema de la revitalització lingüística, seguit per converses en grups reduïts. El seu objectiu és crear una comunitat de pràctica. Molts participants als CWILTs han mantingut el contacte entre ells —s'ha avançat cap a aquest objectiu.

- Sessió 1 (7 de desembre de 2020): 26 participants. Running Wolf i Sara Child (comunitat kwakwaka'wakw) van presentar els CWILTs; Delaney Lothian (cree, métis) va presentar *Sound Hunters*, el seu joc per a estudiants per reconèixer fonemes i síl·labes cree (esmentat anteriorment); Eddie Antonio Santos va presentar el programa de conversió sil·làbica de cree occidental (esmentat anteriorment).
- Sessió 2 (7 de gener de 2021): 33 participants. Stephanie Tenasco (Kitigan Zibi Anishinabe) va compartir la seva experiència creant vídeos d'aprenentatge lingüístic; Patrick Littell i Aidan Pine (NRC) van presentar ReadAlong Studio.
- Sessió 3 (28 de gener de 2021): 80 participants (la majoria eren de Canadà, però n'hi havia 13 d'EUA, 2 de Noruega, 1 de Mèxic, 1 de Bangladesh, i 2 del Marroc). Tema: *"Strengthening Language through Technology: A Context for Capacity Building"* [Enfortir la llengua mitjançant la tecnologia: un context per al desenvolupament d'habilitats]. Aaron Plahn i Justin Bambrick van presentar la seva iniciativa de col·lecció de corpus i desenvolupament d'habilitats per al tŝilhqot'in:
- https://www.youtube.com/watch?v=L-eXWFHnCSc&feature=youtu.be.
- Sessió 4 (14 de juliol de 2021): no es va gravar el nombre de participants (almenys 40). Sean Smith, Krista Dempster, i Dustin Hill de YNLC van descriure el desenvolupament d'habilitats al seu subprojecte.

SUBPROJECTES MONOLINGÜES

La digitalització de text cree a la Blue Quills University:
La Blue Quills va ésser la primera universitat a Canadà propietat de i dirigida per pobles de Primeres Nacions. Aquest subprojecte va digitalitzar i indexar la col·lecció més gran de textos sil·làbics cree: aproximadament 40.000 pàgines de butlletins mensuals, anomenats kihcitwāw miteh (Sacred Heart [sagrat cor]) produïts per l'església catòlica entre els anys 1906 i 1978. El corpus estarà disponible aviat a http://language-unbq.bluequills.ca.

Col·lecció de dades kanyen'kéha a TTO:
Aquesta tasca està liderada per Nathan Thanyehténhas Brinklow, un professor de llengua a la Queen's University a Ontàrio, amb la col·laboració del centre cultural i lingüístic Tsi Tyónnheht Onkwawén:na Language and Cultural Centre (TTO) al territori Tyendinaga Mohawk. L'objectiu és la creació de corpus de text i àudio per donar suport a la recerca de RAP per al kanyen'kéha. Aquest subprojecte ha recollit 112.420 paraules escrites en kanyen'kéha i 26 hores d'àudio, i inclou gèneres com guions i àudio de pel·lícules i sèries de televisió, llibres traduïts de la Bíblia, i gravacions d'Ancians.

Col·lecció de dades kwak'wala:
Aquesta iniciativa també implica la col·lecció de corpus, seguit de recerca de RAP. La llengua kwak'wala (família wakash) es parla a 18 nacions kwakwaka'wakw, el territori tradicional dels quals es troba al nord de la illa de Vancouver, a illes properes i als territoris continentals adjacents. El finançament del NRC-ILT va donar suport a una col·laboració entre dos equips comunitaris de tres comunitats kwakwaka'wakw diferents, un equip tècnic, i tres lingüistes universitaris (dos dels quals són kwakwaka'wakw). Aquest subprojecte ha acumulat 25 hores de dades d'àudio kwak'wala llegibles a través de l'ordinador, que consisteixen en converses, materials pedagògics, i narracions d'Ancians. També ha identificat més de cent hores d'altres gravacions de veu kwak'wala a diverses ubicacions.

Diccionari oral michif:
Aquest subprojecte va mobilitzar i fer accessible un diccionari michif descatalogat anomenat "The Michif Dictionary: Turtle Mountain Dictionary Chippewa Cree", publicat per primer cop l'any 1983. Finançat parcialment pel NRC-ILT, i amb l'assistència de parlants nadius de michif, lingüistes computacionals, i d'altres, l'equip va desenvolupar una versió digital parlada d'aquest recurs. El subprojecte va produir més de 181 hores de gravacions d'àudio d'alta qualitat del diccionari de la mà de quatre parlants; el diccionari inclou 15.422 entrades. Almenys un parlant ha gravat totes les 350 pàgines d'entrades lèxiques michif amb frases d'exemple, i algunes entrades les han gravat dos o més parlants. L'enllaç al diccionari: https://dictionary.michif.org/home.

L'informe final d'aquest subprojecte fa referència a les noves tecnologies CRIM descrites anteriorment:

Cada gravació es va segmentar automàticament en expressions delimitades per pauses mitjançant un servei de detecció d'activitat de veu a través d'una xarxa neuronal profunda (DNN per les seves sigles en anglès, *deep neural network*) desenvolupat dins del projecte VESTA-ELAN pel CRIM. Aquesta auto-segmentació va estalviar una quantitat de temps incalculable al procés d'anotació (*comunicació de Heather Souter i Olivia Sammons, coordinadores del subprojecte*).

Nsyilxcn:
La llengua okanagan (nsyilxcn, família salish interior) era parlada tradicionalment a la vall d'Okanagan en l'actual Columbia Britànica.

És una llengua en perill crític, amb només una dotzena d'Ancians que la parlen amb fluïdesa. Aquest subprojecte va començar al desembre de 2019; el va dur a terme Syilx Language House (SLH), una organització comunitària, amb finançament parcial del NRC-ILT. Es van gravar, transcriure (amb un glossari en anglès) i arxivar set hores de narracions d'Ancians que dominen l'okanagan. Totes les gravacions es troben a www.thelanguagehouse.ca. Dins del subprojecte, es va formar a 4 persones en període de pràctiques i 10 estudiants voluntaris per adquirir fluïdesa.

SENĆOŦEN:

La llengua saanich (SENĆOŦEN, família salish de la costa) es parlava al nord de l'actual Victoria, Columbia Britànica. És la llengua en perill més greu amb la qual ha treballat el NRC-ILT: fins fa poc, només quedaven cinc persones que parlessin la llengua amb fluïdesa. No obstant això, la comunitat està compromesa amb una iniciativa vigorosa de revitalització lingüística, liderada pel WSÁNEĆ School Board. Durant els anys 1981-1991, el lingüista Dr. Timothy Montler va gravar l'Elsie Claxton, la última parlant monolingüe de SENĆOŦEN, narrant històries (Montler, 2018). El NRC-ILT va pagar a dos Ancians perquè treballessin amb el Dr. Montler per transcriure aquestes gravacions, i traduir-les a l'anglès. Aquest material serà un recurs de valor incalculable per ensenyar la llengua.

Tŝilhqot'in:

La llengua tŝilhqot'in (família atapascana o dené) es parlava tradicionalment a la zona interior meridional de l'actual Columbia Britànica. El finançament per part del NRC-ILT va permetre al Govern Nacional de Tŝilhqot'in (TNG per les seves sigles en anglès, *Tŝilhqot'in National Government*) ampliar una iniciativa de revitalització lingüística impressionant. El TNG ja havia gravat 35.000 fragments d'àudio, elaborat una base de dades lingüístiques, publicat paradigmes verbals a un "llibre verbal digital", i creat una pàgina web amb diverses eines educatives: www.tsilhqotinlanguage.ca. Amb el nostre finançament, el TNG va gravar una dotzena d'hores de discursos de 20 parlants; es van digitalitzar, transcriure i alinear les transcripcions de 20 hores de discursos. Es va completar l'etiquetatge de 46.000 fragments d'àudio. S'ha elaborat una llista de 26.200 paraules i frases angleses a traduir; moltes d'aquestes s'han traduït al tŝilhqot'in. Es va formar a quatre membres de la comunitat en rols informàtics.

Es pot trobar un excel·lent vídeo dels esforços de revitalització lingüística d'aquesta comunitat als CWILTs a https://www.youtube.com/watch?v=L-eXWFHnCSc&feature=youtu.be.

FEINA FUTURA I CONCLUSIONS

Brinklow (2021) declara que la col·laboració productiva amb col·laboradors no indígenes en tecnologies per a llengües indígenes (ILT) és possible:

"Moltes comunitats a Canadà estan col·laborant amb cooperants sense ànim de lucre per desenvolupar ILT fora d'un ecosistema lucratiu... Aquestes col·laboracions innovadores inclouen universitats, comunitats, institucions de recerca, governs, i d'altres... Al context canadenc, aquestes col·laboracions (irònicament, sovint finançades pels mateixos governs colonitzadors) estan produint tecnologies lingüístiques pioneres que són innovadores des de qualsevol punt de vista."

Aquest autor indígena té elogis per un subprojecte del NRC-ILT: "Ha sorgit una

col·laboració pionera entre Onkwawenna Kentyohkhwa a Six Nations, a Ontàrio, i el NRC. La col·laboració és entre l'equip del projecte NRC-ILT i professors locals per desenvolupar un conjugador verbal pel kanyen'kéha... Tot i que el desenvolupament inicial es va fer amb una llengua concreta (mohawk), l'eina subjacent es va crear per treballar amb qualsevol llengua amb una perspectiva 'primer profunda i després àmplia' al disseny i al desenvolupament... La col·laboració amb el NRC aporta valor al projecte a nivell nacional perquè un enfocament ampli no és responsabilitat d'una comunitat lingüística individual."

Quins són els següents passos?

Les nostres prioritats són:

- Ampliar les tecnologies que ja hem desenvolupat a llengües noves.
- Facilitar l'ús d'aquestes tecnologies. No és suficient llançar programari de codi obert (OSS); la perspectiva "d'empoderament" significa que aquest OSS hauria de ser fàcil d'utilitzar per a persones dins de comunitats indígenes. No hauria de requerir un màster en informàtica o lingüística computacional. Un exemple dels nostres esforços actuals en aquesta direcció: feina dins del NRC-ILT liderat pel Dr. Patrick Littell sobre "Gramble", un sistema de programari basat en fulls de càlcul que, esperem, facilitarà la creació de conjugadors de verbs.
- Un nou enfocament del NRC-ILT serà la conversió de text a veu (TTS). És preferible que els estudiants aprenguin una llengua a partir de parlants humans que la dominin. Tanmateix, donada l'escassetat de professors per a la majoria de llengües indígenes, la TTS podria a vegades esdevenir la segona millor opció acceptable. Un membre de l'equip del NRC-ILT, Aidan Pine, ja va crear un sistema de TTS prototip pel kanyen'kéha (mencionat anteriorment); recentment, ha creat sistemes de TTS experimentals per a dues altres llengües indígenes parlades a Canadà, el gitksan i el SENĆOŦEN, que van obtenir una qualitat de veu sorprenentment bona.

Les perspectives de futur per a moltes llengües indígenes parlades a Canadà són més bones que en les últimes dècades, gràcies als esforços dels mateixos pobles indígenes. El projecte NRC-ILT mostra que és possible que organitzacions no indígenes, fins i tot governamentals, desenvolupin tecnologies en col·laboració amb les parts interessades indígenes que resultin útils per aquestes iniciatives de revitalització lingüística. Tanmateix, cal que les organitzacions no indígenes que hi actuen recordin que el seu paper és secundari, i que el seu objectiu hauria de ser empoderar les comunitats indígenes.

REFERÈNCIES

Rubèn Fernández Asensio (editor). 2019. "*Kava* vell en carbasses noves: revitalització lingüística i escolarització a Hawaii". *Linguapax Review*. https://www.linguapax.org/wp-content/uploads/2020/02/linguapax19-1-1.pdf

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

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

Nathan Thanyehténhas Brinklow. 2021. "Indigenous Language Technologies: Anti-

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

Michael J. Chandler and Christopher Lalonde. 1998. "Cultural continuity as a hedge against suicide in Canada's First Nations". *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". [Presentació de conferència]. *International Conference on Language Documentation & Conservation*. https://scholarspace.manoa.hawaii.edu/bitstream/10125/44841/44841.pdf

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

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

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

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

First Voices Portal. 2021. "SENĆOŦEN Home Page". Descarregat de https://www.firstvoices.com/explore/FV/sections/Data/THE%20SEN%C4%86O%C5%A6EN%20LANGUAGESEN%C4%86O%C5%A6EN/SEN%C4%86O%C5%A6EN l'1 d'octubre del 2021.

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

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

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

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

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

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

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

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

Rebecca Knowles, Darlene Stewart, Samuel Larkin, and Patrick Littell. 2020. "NRC

Systems for the 2020 Inuktitut-English News Translation Task". *Proceedings of WMT*. https://aclanthology.org/2020.wmt-1.13/.

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

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

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

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

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

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

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

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

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