Indigenous language technologies in Canada: Assessment, challenges, and successes

Patrick Littell
National Research Council of Canada
1200 Montreal Road
Ottawa, ON, K1A 0R6
patrick.littell@nrc.gc.ca

Anna Kazantseva
National Research Council of Canada
5071 West Saanich Road
Victoria, BC, V9E 2E7
anna.kazantseva@nrc.gc.ca

Roland Kuhn
National Research Council of Canada
1200 Montreal Road
Ottawa, ON, K1A 0R6
roland.kuhn@nrc.gc.ca

Aidan Pine
National Research Council of Canada
5071 West Saanich Road
Victoria, BC, V9E 2E7
aidan.pine@nrc.gc.ca

Antti Arppe
University of Alberta
4-32 Assiniboia Hall
Edmonton, AB, T6G 2E7
arppe@ualberta.ca

Christopher Cox
Carleton University
1125 Colonel By Drive
Ottawa, ON, K1S 5B6
christopher.cox@carleton.ca

Marie-Odile Junker
Carleton University
1125 Colonel By Drive
Ottawa, ON, K1S 5B6
MarieOdile.Junker@carleton.ca

Abstract

In this article, we discuss which text, speech, and image technologies have been developed, and would be feasible to develop, for the approximately 60 Indigenous languages spoken in Canada. In particular, we concentrate on technologies that may be feasible to develop for most or all of these languages, not just those that may be feasible for the few most-resourced of these. We assess past achievements and consider future horizons for Indigenous language transliteration, text prediction, spell-checking, approximate search, machine translation, speech recognition, speaker diarization, speech synthesis, optical character recognition, and computer-aided language learning.

1 Introduction

There are approximately 60 Indigenous[1] languages from 10 distinct language families (Rice, 2008) currently spoken in Canada. Several of these languages have tens of thousands of speakers and are still acquired by most children, whereas others have a few tens or hundreds of mostly-elderly speakers; in all, 260,550 report speaking an Indigenous language at at least a conversational level (Statistics Canada, 2016). All of these languages are under significant pressure from English and French, but also have many young people interested in learning. The resurgent strength of community-driven Indigenous linguistic and cultural reclamation in Canada is at the heart of the growing demand for Indigenous language courses, materials and technology.

Indigenous languages are of paramount importance to the nations that speak them and the benefits associated with their use are wide-ranging (Whalen et al., 2016; Reyhner, 2010; Oster et al., 2014; Marmion et al., 2014). As a specific example, research in psychology has shown a compelling correlation

[1] In this document, "Indigenous languages" will specifically refer to Indigenous languages spoken in Canada.

Proceedings of the 27th International Conference on Computational Linguistics, pages 2620–2632
Santa Fe, New Mexico, USA, August 20-26, 2018.

between Indigenous language use and a decrease in youth suicide rates on reserves in British Columbia (Chandler, 1998; Hallett et al., 2007). With increased awareness of these benefits has come increased interest by both Indigenous communities and federal and provincial governments in language technology development, to promote the revitalization and documentation of these languages.

However, the development of Indigenous language technologies faces many challenges: most of these languages are highly morphologically complex, there is relatively little text and speech data available, and there can be significant differences in dialects and orthographies that make it difficult to develop applications that work for all users. Well-known "flagship" language technologies that require large amounts of training data, like machine translation and automatic speech recognition, are therefore probably only feasible to develop for a few of the better-resourced languages such as Inuktitut.

There are nonetheless many practical language technologies that would be feasible to develop for a large number of these languages, and in some cases already have been developed. In this document we assess the feasibility of text (§4), speech (§5), image (§6), and educational (§7) technologies for Indigenous languages, based on past efforts in developing them for these and other low-resource languages.[2]

Disclaimer This document represents the personal opinions of the authors regarding the feasibility of certain technologies, and is not a statement of Government of Canada policy or priorities.

2 Scope and organization

This document will primarily assess user-level applications like search and spell-checking, rather than software that primarily exists to enable linguistic research. This delineation is very approximate, however, as many such applications will have benefits for both kinds of users.

This document also will not be a general inventory of digital resources such as online corpora and lexica. The collection and dissemination of these resources is, of course, highly important and is often the foundational work that makes these technologies possible; it is just that such an inventory would be outside the scope of a document of this size.[3]

In terms of organization, technologies will be clustered from the point-of-view of the practical application of a technology (e.g., spell-checking or text prediction), rather than be organized by the computational model that makes the application possible (e.g., a finite-state transducer or statistical language model). A finite-state grammar of a language (e.g. Snoek et al. (2014), Dunham (2014), Arppe et al. (2017a), Bowers et al. (2017), Harrigan et al. (2017)) can power a number of practical applications from spell-checking, to morphologically-aware search and browsing of dictionaries and corpora, to computer-aided language learning (Arppe et al., 2016).

This document will categorize technologies into five groups, according to the feasibility of developing these for a wide range of Indigenous languages, ranging from "already available" to "infeasible for most languages" (§8). It should be emphasized that these are not ratings of desirability, impact, worthiness for funding, or the relative importance of these technologies to language communities. By contrast, Krauwer (2003) proposed BLARKs - Basic Languages Resource Kits that list basic language technologies and resources needed for successful support and further research of under-represented languages in the European context. Arppe et al. (2016) extend the model to define resource and application priorities for the endangered languages of Canada - EL-BLARK (BLARK for Endangered Languages). This survey finds many similarities with the applications proposed by Arppe et al. (2016).

3 General challenges

There are many challenges that are commonly encountered during the development of Indigenous language technologies, and are encountered in almost all Indigenous languages.

[2] The inventory of existing technologies presented here is likely incomplete, as many language technologies are not published academically or publicized outside of their communities.

[3] An extensive inventory of open-source resources, in both Indigenous and other languages, is available at `github.com/RichardLitt/endangered-languages`. There are also a number of Indigenous language education and reference apps on the iOS App Store and Google Play; Animikii (`www.animikii.com`) maintains a growing list of these at `www.animikii.com/blog/apps-for-learning-an-Indigenous-language`.

3.1 Morphological complexity

Indigenous languages are typically very morphologically complex, with most being polysynthetic or agglutinative. It is commonly the case that a single word carries the meaning of what would be an entire clause in English and French.

(1) *iah th-a-etsi-te-w-ate-wistohsera-'tarih-á:t-ha-k-e'*
no NOT-WOULD-AGAIN-WE-ALL-OWN-butter-HOT-CAUSE-HABIT-CONTIN-PERF.

'We will no longer keep heating up our butter.' Mohawk (Mithun, 1996, p. 170)

(2) *Qanniqlaunngikkalauqtuqlu,* *aninngittunga*
qanniq-lak-uq-nngit-galauq-tuq-lu, ani-nngit-junga
snow-a.little-frequently-NOT-although-3.IND.S-and go.out-NOT-1.IND.S

'And even though it's not snowing a great deal, I'm not going out.'
Inuktitut (Micher, 2017, p. 102)

This complexity presents a challenge for many applications and algorithms, especially those that encode assumptions about the atomic word being the basic unit of meaning/structure, or even the assumption that concatenative morphological analysis is sufficient for finding sub-word units (Arppe et al., 2017a).

3.2 Limited training data

For most languages, there is little to no digitized text or audio available for use as training data, at least not at the scale required for modern statistical or neural NLP. Existing technologies for Indigenous languages have therefore, with a few exceptions, been exclusively rule-based.

A further problem related to the lack of training data is that available training data often comes from only a single domain. For example, the bulk of Inuktitut parallel text comes from Nunavut Legislative Assembly transcripts, but this genre is highly self-similar, and the application of a machine translation system trained on this corpus will likely have difficulties translating other genres like conversation or literature.

A promising research frontier to address limited training data is multilingual modeling; many of the least-resourced Indigenous languages are reasonably closely related to a more-resourced language. For example, automatic speech recognition in Seneca could be trained in part on other Iroquoian languages like Mohawk and Oneida (Jimerson and Prud'hommeaux, 2018), since they have similar phonetic inventories but more available speech data.

3.3 Dialectal and orthographic variation

Most Indigenous languages have a variety of dialects, but often sources and research articles only represent one dialect, or the orthographic standards were developed for a particular dialect and it is unclear how they apply to related dialects. It is sometimes the case that the roadblock to providing technology more widely in a language community is that the dialectal situation is poorly understood, and more basic research on dialectal differences is needed.

Furthermore, even within a single dialect, published works can use a variety of orthographies, and even works using the same orthography often differ in the details such as the encoding of particular diacritics or which morphemes/enclitics are written as separate words or joined. This variety can even be seen in single works, such as those with multiple contributors or transcribers.

Dialectal and orthographic variation pose a particular problem to rule-based text processing systems, since these are usually based on one relatively-well-studied dialect and use particular writing conventions that user-contributed data do not always share. A promising frontier of research to address this, as seen in Micher (2017), is to begin with an existing rule-based system and use it to bootstrap a statistical or neural system (in this case a recurrent neural network) that is more robust when faced with noisy data and unknown morphemes.

4 Text technologies

4.1 Fonts and keyboard layouts

Given the widespread adoption of Unicode and a substantial expansion of character coverage in standard Windows and MacOS fonts like Times New Roman, font coverage of Indigenous languages is currently very good so far as desktop operating systems are concerned. Both Windows and MacOS ship with the Euphemia font for Canadian Aboriginal Syllabics (§6.2), although for some languages Euphemia can display incorrect character orientations[4].

Special Roman characters, diacritics, and Syllabics are not always supported by system-installed keyboard layouts, necessitating the development of custom keyboard layouts. Fortunately, keyboard layout coverage of Indigenous languages is extensive; LanguageGeek[5] provides Windows and MacOS keyboards in almost every Indigenous language, while Tavultesoft Keyman[6] and FirstVoices[7] have developed keyboards for iOS and Android that offer complete coverage of Indigenous languages as well as support for other non-Indigenous languages.

4.2 Predictive text

One common request concerning keyboards (particularly mobile keyboards) is "predictive text" or "autocomplete", in which the keyboard offers shortcut buttons that suggest probable next words to the user depending on what they have already typed. This technology is especially desirable because it appeals to young users as well as to advanced second language learners.

The Multiling O[8] keyboard app for Android offers dictionary-based predictive text in the SENĆOŦEN language.

Maheshwari et al. (2018) examine word and character-based language models for text prediction of Mi'kmaq, based on a small web corpus.

Given the relative paucity of digital text corpora for many languages, it is likely that most predictive text systems will not be able to rely entirely on statistical models, and will instead be built on rule-based (e.g. finite state) or hybrid statistical/rule-based systems.

4.3 Orthography conversion

Almost all Indigenous languages have been written in several different orthographies. While there is a general trend towards orthographic unification in most communities, it is still common to find geographical or generational differences in how languages are written.

Conversion between modern orthographies is generally straightforward, and there exist many applications that manage these conversions[9,10,11,12,13,14]. Conversion between historical and modern orthographies can be more difficult, as historical orthographies often made different assumptions about the vowel and consonant inventories of these languages. There exists a rule-based transliterator between historical and modern Kwak'wala text[15], but the correspondences are somewhat irregular and thus the results are not completely reliable.

4.4 Spell-checking

Although Indigenous languages of Canada have a relatively short tradition of writing, it is quickly gaining steam, especially among young users and learners. However, writing—especially writing "correctly"

[4] www.eastcree.org/cree/en/resources/how-to/cree-fonts/syllabic-font-orientation/
[5] www.languagegeek.com/keyboard_general/all_keyboards.html
[6] keyman.com
[7] firstvoices.com
[8] play.google.com/store/apps/details?id=kl.ime.oh
[9] syllabics.atlas-ling.ca/
[10] www.creedictionary.com/converter/
[11] inuktitutcomputing.ca/Transcoder/
[12] mothertongues.org/#convertextract (Pine and Turin, 2018)
[13] www.eastcree.org/cree/en/resources/syllabic-convertor/
[14] www.giellatekno.uit.no/index.eng.html
[15] orth.nfshost.com/?lang=kwk&input=umista&output=boas

according to official community standards—can be particularly difficult for English- or French-dominant writers, since it requires making sound distinctions that English and French lack. Spell-checking is therefore a frequently-requested technology.

Since all Indigenous languages are morphologically complex, a purely word-list based spell-checking system is typically infeasible; a given stem can have hundreds or even millions of possible derivations/inflections. Corpus-based spell-checkers would have a similar problem; even when a digital corpus is available, only a small fraction of possible derivations/inflections will occur it. Therefore, efforts to develop spell-checkers in Indigenous languages typically concentrate on finite-state technology, since this allows the specification of very large lexicons in an efficient and succinct manner.

A Plains Cree spell-checker based on FST technology is available for system-wide use in recent versions of MacOS, and versions for Microsoft Office and Libre Office are in development (Arppe et al., 2016). The Giella infrastructure (Moshagen et al., 2013) offers an easy way to create FST-based spell-checkers that can be integrated into LibreOffice and, to a limited extent, into Microsoft Office. The spell-checkers use finite-state transducers as a backend, but it is possible to specify spelling relaxations as well as to include modules for likely or common errors. Theoretically the framework allows other types of language models as well, but they have been relatively untested.

An unexpected problem with integrating spell-checkers into mainstream office software is tokenization, since some Indigenous languages use commas, colons, and apostrophes to indicate phonetic differences, whereas many text processing systems assume internally that these are token boundaries. This points to a need for more flexible tokenization within mainstream office software to accommodate these languages.

4.5 Paradigm generation

It has generally been acknowledged that effectively teaching polysynthetic languages requires teaching morphology (Kell, 2014). Since all Indigenous languages have complex verb morphology, one frequent educational need is verb conjugators (Junker and MacKenzie, 2010; Junker and MacKenzie, 2011; Baraby and Junker, 2011; Arppe et al., 2017b), either stand-alone or integrated into an online dictionary.

For most Indigenous languages, learning morphology automatically from corpora is not a viable option. However, symbolic systems, especially those based on finite-state transducers (FSTs) have been successfully implemented for a number of languages. For example, Arppe et al. (2017b) developed an FST for East Cree by leveraging a lexical database. Arppe et al. (2016) and Arppe et al. (2017a) do not release stand-alone verb conjugators, but make verb conjugations available as a part of morphologically-aware online dictionaries for Plains Cree and Tsuut'ina languages respectively.

4.6 Approximate search

Approximate (or "fuzzy") search is a key language technology in situations where the language has not been widely written, or where a large proportion of technology users are learners. Moreover, whole-word search is problematic in highly polysynthetic/agglutinative languages, since the user's query may not use the inflectional form that appears in the dictionary or corpus. Both of these situations are common for Indigenous languages, and therefore the incorporation of approximate search is appropriate in nearly any text technology for these languages.

In general, approximate search can be done in a language-independent way—i.e., by simply counting the number of deletions, insertions, changes, and transpositions (Damerau, 1964; Levenshtein, 1966), without consideration of any language-specific properties—and can be done efficiently even on a large lexicon (Schulz and Mihov, 2002). There are three ways the user experience can be further improved for a particular language: by adapting to actual user queries, by building phonetic knowledge into the system, by making the search aware of morpheme breakdowns.

The East Cree[16] and Innu[17] dictionaries utilize relaxed search rules based on users' habits (Junker and Stewart, 2008).

[16] dictionary.eastcree.org
[17] dictionnaire.innu-aimun.ca

Mother Tongues Dictionaries[18] incorporates phonological background knowledge (e.g., that two sounds are similar and likely to be confused by users) in a finite-state approximate phonological search algorithm (Littell et al., 2017). It concentrates on Pacific Northwest languages where, due to the extensive consonant inventories and phonological complexity of these languages, approximate search is particularly important. This algorithm powers the search function in e-dictionaries for 17 Indigenous languages spoken in British Columbia, including FirstVoices'[19] mobile dictionary applications for iOS and Android, with dictionaries for 11 more languages currently in development.

Morphologically-aware search allows the user to find instances of their search query that may differ in one or more morphemes(Johnson et al., 2013). The Giella infrastructure offers morphology-aware search in dictionaries that are generated by linking a morphological model with lexical resources (and possibly with text corpora). A user can search with any inflected word form of a lemma (or root), possibly taking into account common spelling errors and spelling relaxations (Moshagen et al., 2013). Snoek et al. (2014) and Harrigan et al. (2017) use this technology to allow searching a dictionary of Plains Cree for specific lemmas. Similar capabilities exist for East Cree (Arppe et al., 2017b), Tsuut'ina (Arppe et al., 2017a), Northern Haida (Lachler et al., 2018) and Odawa (Bowers et al., 2017).[20]

4.7 Machine translation

Machine translation is one of the best-known language technologies, and receives significant attention from academia, industry, and the general public, so one of the more common queries from Indigenous groups is whether machine translation would be feasible for their languages.

The current state-of-the-art of machine translation is relatively language neutral, but requires very large amounts of parallel text, which is currently unavailable in most Indigenous languages save Inuktitut. Even then, given the complexity of Inuktitut morphology and the limited corpus available, it is probable that such systems will be, at best, aides to human translators working within that domain, rather than a general-purpose consumer technology like Google Translate.

Several prerequisite steps for Inuktitut machine translation have been achieved, including morphological segmentation (the Uqailaut analyzer[21] and its neural generalization (Micher, 2017)), and sentence and word-level alignment (Martin et al., 2003; Langlais et al., 2005). There are several Inuktitut-English machine translation systems currently under development.

The prerequisite steps can themselves power practical technology. For example, the WeBInuk translation memory system, an adaptation of the WeBiText system (Désilets et al., 2008) mines Inuktitut-English text and uses word alignments to suggest translations to Inuktitut translators.

5 Speech technologies

There has been little development of Indigenous language speech technology so far, but consultation with language communities has suggested that speech technologies are greatly desired, as these languages and cultures are traditionally oral. Text technologies typically expect the user to be able to write their language using the same conventions that the developer expects, which is a problematic expectation in languages without widespread agreement about written conventions. Speech technologies therefore offer an attractive proposition for users more accustomed to speaking and hearing their language than writing and reading it.

5.1 Automatic speech recognition

Full-vocabulary automatic speech recognition (ASR) currently requires large amounts of transcribed audio, and is therefore unlikely to be feasible in most Indigenous languages for the foreseeable future, at least not at a high degree of accuracy. However, even a low degree of accuracy can significantly assist human transcription; this technology, sometimes called Transcription Acceleration (TA), would probably be feasible for at least some languages now.

[18] mothertongues.org
[19] firstvoices.com
[20] altlab.artsrn.ualberta.ca/tools-applications/
[21] www.inuktitutcomputing.ca/Uqailaut/info.php

Jimerson and Prud'hommeaux (2018) has developed a preliminary ASR system for the Seneca language, and the Persephone ASR (Adams et al., 2018) system is being adapted to provide transcription acceleration within the Dative Online Linguistic Database interface (Dunham, 2014), which currently powers dozens of Indigenous language documentation efforts in Canada.

The frontier in speech recognition that is most promising for low-resource languages is multilingual recognition, in which a model trained on a large variety of languages can help compensate for a lack of transcribed speech data in the target language. A challenge for multilingual speech recognition is that some Indigenous languages, particularly in the Pacific Northwest, are global outliers in terms of phonological complexity, with large consonant inventories, rare consonants such as [t͡ɬ], and sometimes long sequences of consonants without the need for intervening vowels. At the very least, the development of practical multilingual recognition models would allow such languages to pool their resources, even if the difference between these and languages outside the region is too great to use a "universal" model.

5.2 Audio keyword search

The primary challenge of ASR in any language is the wide range of inputs the system might encounter—basically, anything that a person might talk about. On the other hand, ASR can also be used to find a *particular* word in an audio recording: the decision is not "what words are these?" but the simpler decision "is this part of the recording an instance of this word?".

This problem, of audio keyword search, is more tractable, but still potentially very useful for making un-transcribed speech recordings more accessible to the public, allowing users to search more quickly through long audio recordings in search of particular words or topics. The National Research Council of Canada (NRC), is collaborating with the Computer Research Institute of Montréal (CRIM) and the Pirurvik Centre on an audio keyword search project for Canadian Broadcasting Company (CBC) radio broadcasts in the Inuktitut and Cree languages.

5.3 Speech/text alignment

Even when resources are too limited to allow full, "open-vocabulary" ASR, prerequisite steps to ASR can be valuable in their own right. One of the prerequisite steps to both ASR and speech synthesis is speech/text alignment (sometimes called "forced alignment"), which involves taking a speech recording and a transcription of it and determining which segments of audio correspond to words and/or phonemes in the transcription.

This intermediate step can itself be of value for education, in creating time-aligned closed-captions from transcribed recordings, and read-along activities such as those available on the East Cree language portal[22] (Luchian and Junker, 2004; Junker et al., 2016) and in the Inuktitut-language Uqalimaarluk ("Read To Me") app for iPad[23].

5.4 Audio segmentation and speaker diarization

Even if automatic speech recognition ("what was said?") is beyond the means of current technology, speaker diarization ("who spoke when?") can be of great value, helping users to more quickly comb through large amounts of audio data in search of examples by a particular speaker or in a particular language.

The NRC-CRIM collaboration mentioned above (§5.2) will also be developing tools for automatic segmentation and speaker diarization. These are relatively language-neutral technologies that could be used in other languages as well.

5.5 Speech synthesis

The converse of automatic speech recognition, text-to-speech (TTS) is somewhat more feasible in low-resource situations. A limited-domain text-to-speech system (such as a talking clock or public transit announcement system) can be trained with just minutes or hours of total recordings, so long as the samples are adequately representative of the target domain.

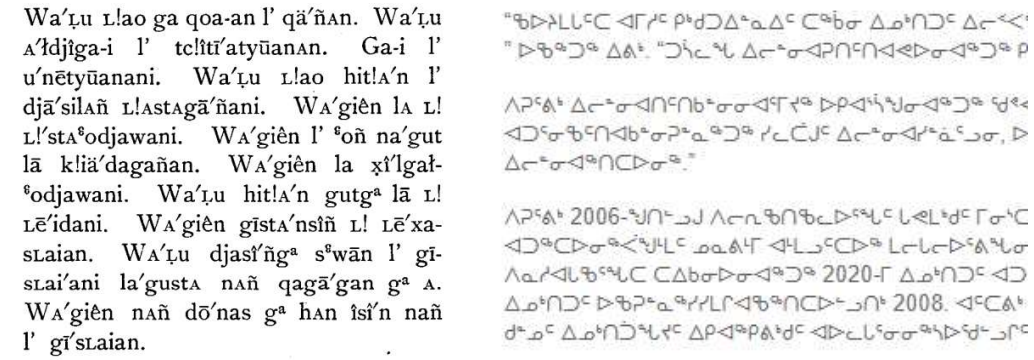
[22] www.eastcree.org/cree/en/lessons/sing-along/

[23] itunes.apple.com/ca/app/uqalimaarluk/id1348117314

Interest in text-to-speech has come from communities where interested learners outnumber fluent speakers, such that the learner might want to know how a word is pronounced but does not currently have access to a speaker to model pronunciation for them. Interest has also come from communities working on projects such as talking online dictionaries, in which the inflectional complexity of the language (§3.1) has meant that it is not feasible to record every possible inflection of a word. In such projects, TTS could allow the user to hear the pronunciation of any inflected form of the word, rather than just uninflected stems.

To our knowledge there are not yet any complete speech-synthesis project in an Indigenous language spoken in Canada, but Synscenter Refsnæs and Oqaasileriffik (the Language Secretariat of Greenland) have developed a general-purpose text-to-speech system for Kalaallisut[24] (West Greenlandic), which is closely related to Inuktitut.

6 Image technologies


Wa′ɻu ʟ!ao ga qoa-an l′ qä′ñan. Wa′ɻu ᴀ′ɬdjîga-i l′ tc!ît!′atyūanan. Ga-i l′ u′nētyūanani. Wa′ɻu ʟ!ao hit!ᴀ′n l′ djā′silᴀñ ʟ!ᴀstᴀgā′ñani. Wᴀ′giên lᴀ ʟ! ʟ!′stᴀˢodjawani. Wᴀ′giên l′ ˢoñ na′gut lā k!iä′dagañan. Wᴀ′giên la x̣î′lgał-ˢodjawani. Wa′ɻu hit!ᴀ′n gutgᵃ lā ʟ! ʟē′idani. Wᴀ′giên gīstᴀ′nsîñ ʟ! ʟē′xa-sʟaian. Wᴀ′ɻu djasî′ñgᵃ sˢwān l′ gī-sʟai′ani la′gustᴀ nᴀñ qagā′gan gᵃ ᴀ. Wᴀ′giên nᴀñ dō′nas gᵃ hᴀn îsî′n nañ l′ gī′sʟaian.

ᖃᐅᔨᒪᒪᑦᑕ ᐊᒥᓯᑦ ᑭᖅᑯᑐᐃᓐᓇᐃᑦ ᑕᖃᐅᓂ ᐃᓚᖅᑎᑐᑦ ᐃᓕᑦᐸᑦ
" ᐅᖃᖓᑐᖅ ᐃᕆᖅ. "ᑐᕋᓪᓚᖏᑦ ᐃᓕᓴᓂᐊᑭᑎᑦᑎᐊᕕᐅᓂᐊᖓᑐᖅ ᑭ

ᐱᑭᑦᕆᖅ ᐃᓕᓴᓂᐊᑎᑦᑎᒃᑲᓂᓂᐊᖏᓪᓗ ᐅᑭᐊᖏᔪᖅᓯᓂᐊᖓᑐᖅ ᖃᑯᑦᓂ
ᐊᑐᑦᓂᖃᑎᒌᐊᒃᑲᓂᑭᓐᓇᖓᑐᖅ ᓯᓚᑖᔪᑦ ᐃᓕᓴᓂᐊᓯᓴᓇᑦᓗᓂᓪ, ᐅ
ᐃᓕᓴᓂᐊᖓᑎᑕᐅᓂᖓ."

ᐱᑭᑦᕆᖅ 2006-ᖃᔪᑎᓴᓗᔭ ᐱᓕᓀᖃᑎᖃᓚᐅᑦᖏᑦᑕ ᒪᓪᒪᒃᑯᑦ ᒥᓂᑦ
ᐊᑐᖓᑕᐅᓂᖓᑦᖃᖏᒪᑦᑕ ᓄᓇᕆᒥ ᐊᒻᒪᓗᑦᑕᐅᖓᑐᖅ ᒪᓕᒪᓕᐅᑦᕆᕆᓂ
ᐱᓇᓯᐊᒪᖃᖏᒪᑦᑕ ᑕᐃᑲᓂᐅᓂᐊᖓᑐᖅ 2020-ᒥ ᐃᓄᖅᑎᑐᑦ ᐊᑐ
ᐃᓄᖅᑎᑐᑦ ᐅᖃᖓᑕᓇᓯᓯᒪᑎᐊᖃᖓᑎᑕᐅᓕᔭᓂᖅ 2008. ᐊᑦᑕᕆᒃ
ᑯᓕᓂ ᐃᓄᖅᑎᑐᖓᒪᓯᑦ ᐃᑭᐊᖓᑭᕆᒃᑯᑦ ᐊᐅᓪᒪᑦᓂᓂᖓᐅᖃᑯᓕᓪᓕᑦᑎᓂᑦ

Figure 1: Example of a historical printed document in the Northern Haida language (Hubert et al., 2016), and Canadian Aboriginal Syllabics representing the Inuktitut language, courtesy of `tusaalanga.ca`.

6.1 Optical character recognition for Roman orthographies

Optical character recognition (OCR) has been successfully been applied to Indigenous language documents, including historical manuscripts printed using pre-digital presses (Fig. 1). Hubert et al. (2016) report a high degree of success on OCR with only a few pages of training data, suggesting that OCR would be feasible to implement for a wide range of Indigenous languages.

The challenge for OCR on many Roman orthographies for Indigenous languages is the proliferation of diacritics and superscript letters, particularly in languages with extensive consonant inventories. Diacritics and superscripts are difficult to differentiate from punctuation and from each other, and depending on the font resources, some letter-diacritic combinations may be very hard to distinguish. For example, British Columbian orthographies based on the Royal B.C. Museum recommendations often use underlined ⟨g⟩ for a uvular voiced plosive, and in some fonts (or in typewritten documents), this underline can overlap the descender on the ⟨g⟩.

6.2 Optical character recognition for Canadian Aboriginal Syllabics

While most Indigenous languages are written using a Roman orthography, several varieties of Inuktut, Cree, and Ojibwe[25] use a system called Canadian Aboriginal Syllabics (Fig. 1).

Canadian Aboriginal Syllabics (often simply called "Syllabics") is a "rotational syllabary" in which the shape of the glyph indicates the syllable's consonant and its rotational orientation to the vowel. There are also smaller, superscript characters that indicate consonants where no vowel follows (e.g. in a syllable

[24] `oqaasileriffik.gl/langtech/martha/`

[25] Syllabics have also historically been used for Blackfoot and some Athabascan languages such as Dakelh (Carrier), but have fallen out of use in favor of Roman orthographies.

coda). Like superscript characters in Roman orthographies, superscript characters in Syllabics can pose a problem for OCR, since due to their size and position they are easily confused for punctuation marks or with each other. Nonetheless, there have been several successful Syllabics OCR projects (e.g. Posgate and Leekam (2014)), and Inuktitut has since been included in the Tesseract OCR project[26] (Smith, 2007).

7 Computer-aided language learning

7.1 Course modules

Computer-aided language learning (CALL) course modules are widely available for Indigenous languages, particularly through the FirstVoices Language Tutor (FVLT) portal[27], which offers approximately 50 online courses covering many Indigenous languages, with exercises on listening, speaking, reading, and vocabulary development, as well as online language-learning games.

There are also language-specific CALL portals, including but not limited to:

- Tusaalanga[28] from the Pirurvik Centre, which offers exercises in five varieties of Inuktut.
- The Institut Tshakapesh learning portal[29], which offers educational games in the Innu language (Junker and Torkornoo, 2012; Junker et al., 2016). These were modeled after the eastcree.org lessons[30] for teaching syllabics, vocabulary and literacy.
- The *nêhiyawêtân* CALL portal for Plains Cree[31] fuses CALL and text technologies, in which students receive targeted feedback made possible by the integration of a finite-state morphology model (Arppe et al., 2016; Bontogon et al., 2018).
- The Yukon Native Language Centre[32] offers online audiovisual adaptations of language courses in eight Indigenous languages.

Several international CALL products have been adapted for Indigenous languages. 7000 Languages[33] adapts the Transparent Language software for low-resource and endangered languages, and offers courses on Denesuline, Dakota, and several varieties of Cree, Ojibwe, and Oji-Cree through partnerships with Grassroots Indigenous Multimedia[34] and the Manitoba First Nations Education Resource Centre[35]. Rosetta Stone[36] has also developed courses for Labrador Inuttitut and Kahnawá:ke Mohawk.

A forthcoming project headed by Dr. Marianne Ignace at Simon Fraser University presents an innovative "chat-based" UI (what is sometimes called "No-interface UI") for CALL apps, in which an AI tutor interacts with the student in a web-chat-like interface.

7.2 Phonetic tutorial

Some education applications focus more narrowly on the acquisition of speech sounds. Phonetic tutorials are particularly important in languages with rarer sounds, like lateral fricatives or ejective plosives.

The Yukon Native Language Centre[37] has developed a phonetic learning game in which players must count the number of instances of a particular sound (e.g., [t͡ɬ]) in a recording to mush a dog sled.

The Tiga Talk[38] app for iOS, originally a collection of speech-language pathology games for English, is currently being adapted to Cree to help support child acquisition.

[26] github.com/tesseract-ocr
[27] tutor.firstvoices.com
[28] tusaalanga.ca
[29] jeux.tsakapesh.ca
[30] lessons.eastcree.org
[31] oahpa.no/nehiyawetan/
[32] www.ynlc.ca
[33] 7000languages.org
[34] gim-ojibwe.org
[35] mfnerc.org
[36] www.rosettastone.com/endangered/projects
[37] ynlc.ca
[38] tigatalk.com/app/

The UBC eNunciate[39] tools use ultrasound to illustrate to students the articulatory gestures of the tongue and vocal tract that cannot ordinarily be seen, in the Upriver Halqemeylem, SENĆOŦEN, Secwepemc, and Blackfoot languages (Bliss et al., 2016).

7.3 Augmented reality and virtual reality

Augmented and virtual reality technologies are not specifically language or learning technologies, but there is a growing amount of interest in their application to Indigenous language education primarily due to their ability to be naturally integrated into popular "place-based eduction" (Sobel, 2004) practices.

The feasibility of implementing augmented and virtual reality projects is aided by the widespread interest in the technology and 3D game engines like Unity and Unreal. However, there are still very few implementations for Indigenous languages in Canada. Some examples include *Tuwitames*, an augmented reality story-telling app narrated in Secwepemctsín (Lacho, 2018), an augmented reality companion app to a Blackfoot card game (Goff et al., 2017), and *Schoolû*, a virtual reality application for teaching Cree syllabics. Yet another augmented reality app, *Wikiup*[40], is designed to take users on tours throughout Canadian cities, transforming landmarks by telling AR-enhanced Indigenous stories and histories, but not necessarily involving Indigenous languages.

8 Summary

Widely available Successful technologies that are already available for many Indigenous languages *Keyboard layouts* (§4.1), *Approximate search* (§4.6), *Computer-aided language learning* (§7).

Ready for wider implementation Technologies that have been developed for some languages, and that could feasibly be developed for most or all Indigenous languages: *Orthography conversion* (§4.3), *Optical character recognition* (§6.1, §6.2).

Awaiting implementation Technologies for which there is already a technological basis in a few languages (e.g., a finite-state analyzer has been written), or for which there exists a language-neutral technological basis, but for which practical user interfaces or language-specific implementations are not yet developed or widely available: *Spell-checking* (§4.4), *Paradigm generation* (§4.5), *Speech/text alignment* (§5.3).

Experimental Technologies that have not yet proven to be successful on Indigenous languages, but show promise in other low-resource language situations: *Predictive text* (§4.2), *Transcription acceleration* (§5.1), *Audio keyword search* (§5.2), *Audio segmentation and speaker diarization* (§5.4), *Text-to-speech* (§5.5).

Restricted feasibility Technologies that will likely be feasible only in more-resourced languages (e.g. Inuktitut, Cree): *Machine translation* (§4.7), *Automatic speech recognition* (§5.1).

From the above, it is clear that there are a number of text, speech, image, and CALL technologies that are either already available, or could be made more widely available, in many cases with relatively reasonable further investment. The boundary between the first three categories at various stages of implementability and the two last experimental and restricted ones appears to determined by the existence of technological solutions that work with typically quite sparse data resources that can be reasonably expected for Indigenous languages. Meanwhile, new developments (particularly in multilingual and finite-state/neural hybrid modeling) may make technologies possible that until recently seemed infeasible for Indigenous languages.

[39] `enunciate.arts.ubc.ca`
[40] `wikiupedia.com`

References

Oliver Adams, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird, and Alexis Michaud. 2018. Evaluating phonemic transcription of low-resource tonal languages for language documentation. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hlne Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, May. European Language Resources Association (ELRA).

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.

Antti Arppe, Christopher Cox, Mans Hulden, Jordan Lachler, Sjur N. Moshagen, Miikka Silfverberg, and Trond Trosterud. 2017a. Computational modeling of verbs in Dene languages: The case of Tsuut'ina. In Alessandro Jaker, editor, *Working Papers in Athabaskan (Dene) Languages*, pages 51–69.

Antti Arppe, Marie-Odile Junker, and Delasie Torkornoo. 2017b. Converting a comprehensive lexical database into a computational model: The case of East Cree verb inflection. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-2)*, pages 43–47, Honolulu, March. Association for Computational Linguistics.

Anne-Marie Baraby and Marie-Odile Junker. 2011. Conjugaisons des verbes innus, (3e d.). http://verbe.innu-aimun.ca.

Heather Bliss, Strang Burton, and Bryan Gick. 2016. Ultrasound overlay videos and their application in indigenous language learning and revitalization. *Canadian Acoustics*, 44:136–37.

Megan Bontogon, Antti Arppe, Lene Antonsen, Dorothy Thunder, and Jordan Lachler. 2018. Intelligent computer assisted language learning (ICALL) for nêhiyawêwin: An in-depth user experience evaluation. *Canadian Modern Language Review*.

Dustin Bowers, Antti Arppe, Jordan Lachler, Sjur Moshagen, and Trond Trosterud. 2017. A morphological parser for Odawa. In *Proceedings of 2nd Workshop on Computational Methods for Endangered Languages (ComputEL-2)*, pages 2326–2330.

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

Fred J. Damerau. 1964. A technique for computer detection and correction of spelling errors. *Commun. ACM*, 7(3):171–176, March.

Alain Désilets, Benoit Farley, Geneviève Patenaude, and Marta Stojanovic. 2008. WeBiText: Building large heterogeneous translation memories from parallel web content. *Proc. of Translating and the Computer*, 30:27–28.

Joel Robert William Dunham. 2014. *The online linguistic database: Software for linguistic fieldwork*. Ph.D. thesis, University of British Columbia.

Rebecca Goff, Brandon Goff, Caroline Running Wolf, Michael Running Wolf, Jesse Desrosier, Naatosi Fish, and Mizuki Miyashita. 2017. Playfully revitalizing languages and traditional knowledge through collaboration. http://hdl.handle.net/10125/41929.

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

Atticus G. Harrigan, Katherine Schmirler, Antti Arppe, Lene Antonsen, Trond Trosterud, and Arok Wolvengrey. 2017. Learning from the computational modelling of Plains Cree verbs. *Morphology*, 27(4):565–598.

Isabell Hubert, Antti Arppe, Jordan Lachler, and Eddie Antonio Santos. 2016. Training & quality assessment of an optical character recognition model for Northern Haida. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3227–3234, Paris, France, May. European Language Resources Association (ELRA).

Robert Jimerson and Emily Prud'hommeaux. 2018. ASR for documenting acutely under-resourced indigenous languages. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hlne Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, May. European Language Resources Association (ELRA).

Ryan Johnson, Lene Antonsen, and Trond Trosterud. 2013. Using finite state transducers for making efficient reading comprehension dictionaries. In Stephan Oepen, Kristin Hagen, and Janne Bondi Johannessen, editors, *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013), NEALT Proceedings Series 16*, pages 59–71, Oslo University, Norway, May.

Marie-Odile Junker and Marguerite MacKenzie. 2010. East Cree (Northern dialect) verb conjugation. http://verbn.eastcree.org/.

Marie-Odile Junker and Marguerite MacKenzie. 2011. East Cree (Southern dialect) verb conjugation. http://verbs.eastcree.org/.

Marie-Odile Junker and Terry Stewart. 2008. Building search engines for Algonquian languages. In Karl S. Hele and Regna Darnell, editors, *Papers of the 39th Algonquian Conference*, pages 378–411, London, ON. University of Western Ontario Press.

Marie-Odile Junker and Delasie Torkornoo. 2012. Online language games for endangered languages: jeux.tshakapesh.ca, www.eastcree.org/lessons/. In *Proceedings of EDULEARN 12: International Conference on Education and New Learning Technologies*.

Marie-Odile Junker, Yvette Mollen, Hélène St-Onge, and Delasie Torkornoo. 2016. Integrated web tools for Innu language maintenance. In J. R. Valentine and M. MacCauley, editors, *Papers of the 44st Algonquian Conference*.

Sarah Kell. 2014. *Polysynthetic Language Structures and their Role in Pedagogy and Curriculum for BC Indigenous Languages*. BC Ministry of Education.

Steven Krauwer. 2003. The basic language resource kit (BLARK) as the first milestone for the language resources roadmap. In *Proceedings of the International Workshop Śpeech and Computer, SPECOM 2003, Moscow*, pages 8–15.

Jordan Lachler, Lene Antonsen, Trond Trosterud, Sjur N. Moshagen, and Antti Arppe. 2018. Modeling Northern Haida morphology. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2018)*, May.

David Lacho. 2018. *Developing an augmented reality app in Secwepemctsín in collaboration with the Splatsin Tsm7aksaltn (Splatsin Teaching Centre) Society*. Ph.D. thesis, University of British Columbia.

Philippe Langlais, Fabrizio Gotti, and Guihong Cao. 2005. Nukti: English-Inuktitut word alignment system description. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 75–78. Association for Computational Linguistics.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10:707–710.

Patrick Littell, Aidan Pine, and Henry Davis. 2017. Waldayu and Waldayu Mobile: Modern digital dictionary interfaces for endangered languages. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 141–150.

Radu Luchian and Marie-Odile Junker. 2004. Developing an on-line Cree read-along with syllabics. *Carleton University Cognitive Science Technical Report*, 2006-01.

Anant Maheshwari, Leo Bouscarrat, and Paul Cook. 2018. Towards language technology for Mi'kmaq. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hlne Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, May. European Language Resources Association (ELRA).

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.

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

Jeffrey Micher. 2017. Improving coverage of an Inuktitut morphological analyzer using a segmental recurrent neural network. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 101–106. Association for Computational Linguistics.

Marianne Mithun. 1996. The Mohawk language. *MULTILINGUAL MATTERS*, pages 159–173.

Sjur N. Moshagen, Tommi A. Pirinen, and Trond Trosterud. 2013. Building an open-source development infrastructure for language technology projects. In *Proceedings of the 19th Nordic Conference of Computational Linguistics, NODALIDA 2013, May 22-24, 2013, Oslo University, Norway*, pages 343–352.

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

Aidan Pine and Mark Turin. 2018. Seeing the Heiltsuk orthography from font encoding through to Unicode: A case study using convertextract. In Claudia Soria, Laurent Besacier, and Laurette Pretorius, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, May. European Language Resources Association (ELRA).

Jess Posgate and Cathy Leekam. 2014. Digitizing Ontario's community memory: Bringing multicultural history online. In *Ontario Library Association Super Conference*, Toronto, ON, Jan. 30.

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*.

Klaus U. Schulz and Stoyan Mihov. 2002. Fast string correction with Levenshtein-automata. *International Journal of Document Analysis and Recognition*, 5(1):67–85.

Ray Smith. 2007. An overview of the Tesseract OCR engine. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition - Volume 02*, ICDAR '07, pages 629–633, Washington, DC, USA. IEEE Computer Society.

Conor Snoek, Dorothy Thunder, Kaidi Lõo, Antti Arppe, Jordan Lachler, Sjur Moshagen, and Trond Trosterud. 2014. Modeling the noun morphology of Plains Cree. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 34–42, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

David Sobel. 2004. *Place-based education: Connecting classroom and community*. Orion Society.

Statistics Canada. 2016. Aboriginal languages in Canada, 2016 census of population, catalogue no. 98-200-X.

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