Proceedings of the Language Technologies for All (LT4All) , pages 311–314
Paris, UNESCO Headquarters, 5-6 December, 2019.

21st Century Language Technology Tools – 21st Century Challenges vs. 21st Century Opportunities

Antti Arppe[1] & Jordan Lachler[2]
[1]Alberta Language Technology Lab, [2]Canadian Indigenous Languages and Literacy Institute, University of Alberta
arppe@ualberta.ca, lachler@ualberta.ca

Abstract

This paper presents a brief overview of the historical and current circumstances of Indigenous languages spoken in Canada, forming the basis for contemporary needs, challenges as well as opportunities presented in the development of modern language technological tools and applications for these languages in the 21st century.

Keywords: Canadian Indigenous languages, Algonquian languages, Dene languages, Cree, Intelligent on-line dictionaries, Corpora, Spell-checkers, Intelligent Computer-Aided Language Learning Applications

Tiivistelmä (in Finnish)

Tämä artikkeli esittää tiiviin yleiskatsauksen Kanadassa puhuttavien alkuperäiskielten olosuhteiden historiallisista kehityskuluista ja nykytilanteesta. Näiden perusteella artikkelissa kuvataan, mitä tarpeita, haasteita ja mahdollisuuksia on modernin kieliteknologian kehittämisessä näille alkuperäiskielille 21. vuosisadalla.

1. Introduction and Context

1.1 Indigineous Languages Spoken in Canada

Canada has much more linguistic diversity than e.g. Europe. In the 2016 Census of Population in Canada (Statistics Canada, 2017), participating people reported over 70 Indigenous languages, grouped into 8 distinct language families, namely the Algonquian, Inuit, Dene (Athabaskan), Siouan, Salish, Tsimshian, Wakashan and Iroquoian language families, as well as the isolates Kutenai and Haida, plus Michif which combines both Cree and French origins.

The aforementioned language families do not know national boundaries, spanning the Canadian-American and provincial/state borders**.** Dene languages are spoken from Alaska into Canada and hopping into South-Western United States; Algonquian languages are spoken on both sides of the border from the Rockies through the Great Lakes to the Atlantic, as are Iroqoian languages on both sides of Lake Erie and the St. Lawrence River, Siouan languages stretch north across the Western Plains, Salishan languages extend from British Columbia to the states of Washington, Idaho and Montana, and Inuit languages span the entire North-American Arctic from Alaska across Northern Canada to Greenland.

As many as 260,550 people in Canada, a number which has grown since 2006, reported being able to speak an Indigenous language well enough to carry out a conversation, out of 1,673,785 people (4.9% of the entire Canadian population) reporting Indigenous/Métis/Inuit identity/heritage**.** Nearly 213,225 people reported speaking an Indigenous language as a mother tongue, defined as the first language learned at home in childhood and still understood. Thus, the number of people able to speak an Indigenous language exceeded the number who reported an Indigenous mother tongue. According to Statistics Canada, this suggests that many people, especially young people, are learning Indigenous languages as second languages. (Statistics Canada 2017)

Algonquian languages were the family with the largest number of speakers, 175,825, in Canada, of which Cree with 96,575 speakers and Ojibwe with 28,130 speakers were the largest individual languages in this family. The next largest Indigenous language families and languages in terms of speakers were the Inuit languages (42,065 speakers, of which 39,770 spoke Inuktitut), the Dene (Athapaskan) languages (23,455 speakers, of which 13,005 spoke Dene [sųłiné]), followed by the rest, for which the numbers of speakers range from several thousand down to only a few tens. In terms of age groups, older Indigenous people were more likely to be able to speak an Indigenous language than younger generations, so that 35.6% out of seniors (65 years and older) could speak an Indigenous language, with the proportion decreasing with each subsequent age bracket. However, since there are four times more Indigenous children than seniors, in absolute terms there are more Indigenous children (45,135) than seniors (22,125) who could speak an Indigenous language. (Statistics Canada, 2017).

Furthermore, the results of this Census indicate that there were many times more Indigenous languages than Indigenous communities (over six hundred Sovereign First Nation communities) in Canada. Quite often and not surprisingly, most communities consider the language as it is spoken in that community as distinct and as a symbol of identity. In the Province of Alberta alone, there are nine Indigenous languages, which are (Plains or Woods) Cree, Blackfoot and Saulteux in the Algonquian family, Dene [sųłiné], Beaver, Slavey and Tsuut'ina in the Dene family, and Stoney/Nakoda in the Siouan family, plus Michif, spoken in as many as 46 First Nations Communities.

1.2 Historical Context in Canada

One could argue that language technology has hitherto generally proved to be of most use in supporting written language, though recently speech technology has finally also started to show real promise. In this respect, writing systems were developed and used for many Indigenous languages in Canada by the 1800s, for instance Aboriginal Syllabics for Cree, Blackfoot, and Ojibwe as well as Inuktitut (Rogers 2005). Furthermore, there has been a tradition of rapid uptake of Indigenous literacy. For instance, there is evidence of high literacy among Indigenous peoples in the 1800s, sometimes anecdotally claimed to have been higher than the proportion of literate people among the colonizing population (Rogers, 2005). What appears to have changed the general dynamics is Canadian Confederation in 1867 and the notorious Residential School system that it continued, which had started in the 1800s and lasted as long as until 1990s in some parts of Canada. This entailed the forced removal of many Indigenous children from their families to residential schools, often far away from their home communities, even in a different provinces, with children from multiple Indigenous language communities intermixed. In residential schools, the use of Indigenous languages by pupils was forbidden. Nevertheless, many children spent summers and other breaks with their families in their home communities, and returned there after finishing residential school, continuing to retain knowledge of their Indigenous mother tongues.

A possible turning point in the 1970s was the arrival of mass media, in particular TV, in many Indigenous communities. This resulted in significantly expanded exposure to majority languages (i.e. English and French in Canada) in previously isolated communities. Indeed, some members of Indigenous communities have personally noted the advent of mass media as the moment when their Indigenous language went into decline.

All in all, as a result of these aforementioned policies and technological changes, as well as other factors, the use of Indigenous languages was gradually squeezed into the oral realm, and more and more out of the public sphere.

1.3 Characteristics of Indigenous Languages in Canada and Their Resources

Many Indigenous languages spoken in Canada and the rest of North America are quite unlike majority languages such as English, in that they exhibit extensive morphological complexity, to the degree that there is no way to enumerate all possible word forms, or even all likely word forms, as one could do for English. In this respect they are similar to some other Indigenous languages, such as the Sámi languages spoken in Northern Scandinavia, for which extensive language technological tools have been created over the last decade. Table 2 illustrates with a Plains Cree text passage the extent of morphological complexity in practice, in a language with extensive morphological complexity, where only 10 of the 21 Cree words are easily listable non-inflecting forms, and 4 rarer inflected forms that would not normally be included in any core or even extended inflectional paradigm. Thus, without computational modeling of inflection and word structure, less than half of the words in running text could be matched *as is* with a dictionary entry, meaning that over half of the words would remain inaccessible using a software application that would not make use of a computational morphological model (cf. Klavans 2018).

ᐃᐢᐲˣ ᑳ ᑮ ᐊᐚᓯᓰᐏᔮᐣ ᓂᑮ ᓂᑕᐏ
ᑭᐢᑭᓄᐦᐊᒫᑯᓯᐣ ᐅᐦᐱᒣ ᓂᑎᐢᑯᓂᑲᓂˣ ᐅᐦᒋx ᑭᐢᑕᐱᓈᓂˣ ᒫ
ᓇ ᑳ ᑮ ᓂᑕᐏ ᑭᐢᑭᓄᐦᐊᒫᑯᓯᔮᐣ, ᐃᑯᑕ **Residential School**
ᐁ ᑮ ᐃᑐᐦᑕᐦᐃᑲᐏᔮᐣx ᒥᐢᑕᐦᐃ ᒫᓇ
ᓂᑮ ᑭᑎᒫᑭᐦᐃᑲᐏᓈᐣ, ᒫᑲ ᓇᒨᔭ ᐁᐘᑯ ᓂᐑ ᐋᑐᑌᐣ, ᓂᐑ
ᐋᒋᒫᐘᐠ ᓃᒋ ᑭᑎᒫᑭᓴᐠ, ᓃᒋ ᐅᑭᐢᑭᓄᐦᐊᒫᑲᓇᐠx

Table 1. The 1st paragraph of a Plains Cree story, *Dog Biscuits* by Solomon Ratt, in (Western) Canadian Syllabics, courtesy of Cree Literacy Network (URL: https://creeliteracy.org/2014/01/20/dog-biscuits-y-dialect-with-audio/). Words in **red** are non-inflecting words; words in **blue** are complex inflected forms outside any core paradigm, and words in **black** inflected forms in the core paradigm.

Moreover, written corpora, if any exist, are relatively small, with at most several hundred thousand word form tokens representing 10-20 thousand word form types (e.g. Arppe et al., forthcoming, in the case of Plains Cree). Thus, these corpora, in terms of their representativeness of word form types, are not conventionally sufficient for applying machine learning to learn the languages' morphological system. On the other hand, there often exist relatively comprehensive lexical databases as well as complete inflectional paradigms for many of these Indigenous languages, as a result of extensive linguistic documentation work in the 19th and 20th centuries. Importantly, these are resources that are highly amenable for 20th century rule-based computational modeling techniques of morphology using finite-state transducers (cf. e.g. Beesley & Karttunen 2003), see e.g. Harrigan, Schmirler, Arppe et al. (2017) in the case of Plains Cree.

2. Challenges

As a result of the historical context and developments or accidents, there exist few if any established, obvious institutions for collaboration concerning Indigenous languages in Canada, which typically would concern multiple communities speaking the same language or its dialects across several provinces (e.g. Cree or Ojibwe). This has been facilitated by the splintering of the language communities through the reserve system. Furthermore, one can observe a predominance of oral tradition, though we should recall the history of Indigenous literacy in the 1800s. This has been bolstered by the fact that there are few large collections of written texts or literature, as noted above, which can generally be associated with people having less exposure to written materials in their heritage language.

Moreover, orthographical standard forms are sometimes seen as representing a continuation of colonial thinking. Alongside this view, one can also see the success of the principle "The language is written as it is spoken", resulting in substantial variation in how the languages are written, even from one individual to another. On the other hand, this lack of standard also allows for representing the rich variation from individual/community to another. Nevertheless, common orthographical conventions would allow others to more easily both find and understand what has been written by someone else.

Another set of challenges arises from representing appropriately sounds not apparent in the majority languages of Canada, English and French, in a native way. Historical, pre-computer era solutions to this have included using underlining as a diacritic (e.g. Haida), the adoption of apostrophes for glottal stops (e.g. Haida) and other punctuation marks or numerals for non-Western phonemes (in particular for Indigenous languages in British Columbia), and marking e.g. long vowels or geminate consonants as in IPA using a colon (e.g. Mohawk). Furthermore, one can sometimes observe the explicit marking of allophonic, non-phonemic differences in pronunciation, which can result in unnecessarily detailed, complex writing systems. All this poses in the contemporary world the practical challenge of how to to deal with and input some of such non-English characters for non-English sounds on computers and mobile devices. Nevertheless, to some extent, this can be addressed with the flexible expandability of character sets and scripts within the Unicode standard, fuzzy matching algorithms linking what a user types with the properly spelled word form using the standard characters according to a language's writing system, and clever keyboard design.

Moreover, one can observe not only multiple scripts, but also multiple orthographical standards/conventions per these scripts, for writing the very same utterance, as a result of the historical context and developments and accidents. Only for (Plains) Cree (Table 2), one can, or has to, choose between Canadian Aboriginal Syllabics vs. Standard Roman orthography (SRO). Furthermore, there are varying conventions as to whether and how one marks, or does not mark, long vowels (with circumflexes or macrons or not at all) and syllable final aspirations (*-h*) with diacritics. In addition, there is variation, in both Standard Roman Orthography (SRO) and Canadian Syllabics, concerning whether one writes prefixes or suffixes joined with the stems (with or without a hyphen), or not, separating the affixes with spaces.

kâ-kî-awâsisîwiyân
kā-kī-awāsisīwiyān
kâ kî awâsisîwiyân
kakihawasisiwiyan
ᑳ ᑮ ᐊᐋᐧᓯᓰᐃᐧᔮᐣ

Table 2. Varying ways of writing the (Plains) Cree word *kâ-kî-awâsisîwiyân* 'when I was a child'

This poses a general challenge for language technological tools and applications of recognizing as well as generating these multiple standards and conventions in writing. Nevertheless, this is in principle quite straight-forward to address with the computational modeling of both word structure and of orthographical variation, which allows the easy creation of transcriptors converting between all/known, standardized variants.

3. Expectations of Indigenous Communities and Individual Speakers and Learners

Based on our experiences, Indigenous communities in Canada we have worked with have appeared to value recording the richness of their vocabulary, as known by Elders and other fluent native speakers, and also recording how their language is properly spoken by Elders; in contrast, the creation of proofing tools to support the proper writing of texts, while potentially valued, has not always been considered as high in urgency. Indigenous learners/speakers have appeared to value the ability to look up words in an Indigenous language for a concept (i.e. an English word), the ability to find out what a particular word in an Indigenous language means (i.e. a translation into English), and the ability to write, or often importantly say, a particular word or phrase in their Indigenous language. Nevertheless, one needs to recognize here that there are in fact multiple Indigenous speaker/learner subgroups within even a single First Nations community, e.g. Elders, other fluent speakers, language instructors/teachers, school pupils and other language learners, as well as other community members. These different subgroups have different levels of language proficiency, different linguistic community contexts, and consequently different needs and priorities; thus, they also need, expect and would benefit from different language technological solutions. Furthermore, these subgroups within individual First Nations communities may have more in common with similar subgroups in other communities speaking the same Indigenous language, perhaps even more so than with other subgroups in their own community.

As for language technological applications and Indigenous communities, electronic/on-line dictionaries are often considered very desirable, and having a spoken component included in such a resource is considered very important. Morphological intelligence, that is the recognition and generation of complex word-forms is certainly seen as a plus, but this feature is not among the first expectations. Moreover, there is a concern that word-form/paradigm generation produces non-validated unnatural or weird forms. Furthermore, proofing tools such as spell-checking or predictive text do not appear to be the very first expectation, though views of community members seem to change when shown demo mock-ups illustrating such a functionality for their own Indigenous language.

As a language-independent but relevant recent development, the last decade has witnessed a diffusion of mobile/digital devices; Indigenous youth have become digital natives (and increasingly older folks, too). The youth are often considered the "missing generation" in language revitalization, since they did not learn the language at home, but are now too old to benefit from language nests and immersion schools targeting younger children. Indigenous teens and young adults are indeed a vital piece of the puzzle in language revitalization. We are seeing an increasing amount of mobile media use which is based on written communication, and with this the emergence of need of language technology supporting written language, where spell-checking, predictive text, and word-form generation based on computational modeling has a role. Moreover, Indigenous individuals are more and more moving and living outside the reserve in urban centers, without direct access to Elders and other fluent speakers (most of whom are still remaining on the original reserves); thus, they would greatly benefit from written and spoken Indigenous language resources available on-line.

4. Opportunities and Solutions

The development and diffusion of digital devices makes literacy-based language tools useful for a critical generation of Indigenous communities in Canada, as well as others. In this, we in the Alberta Language Technology Lab (ALTLab: altlab.artsrn.ualberta.ca) have been inspired by what the Giellatekno and Divvun research and development teams at UiT – Arctic University of Norway have been able to create for the Indigenous Sámi languages, and have started adapting their work to the Canadian context, encountering both similarities and differences in the circumstances of Indigenous languages in Canada. Following their example, we have aimed at the "low-hanging fruit" that can be created with the existing scarce but rich documentation resources which are amenable to rule-based computational models, and language technological applications and resources that can be created with such models (Arppe et al., 2016). These include (1) **web-based intelligent dictionaries** (I-DICT) presenting both the written and spoken form of words, and that allow for searching with inflected forms and the generation of inflectional paradigms (altlab.ualberta.ca/itwewina); (2) **searchable databases** of both written and spoken usage examples (Arppe et al., forthcoming: altlab.ualberta.ca/korp); (3) **spell-checkers** to support the creation of high-quality texts by speakers and learners; and (4) **intelligent computer-aided language learning** (I-CALL) applications which include training in both the spoken and written forms of the language (Bontogon et al, 2018: oahpa.no/nehiyawetan). To date, we have created our first full demonstration versions of these tools only for Plains Cree, but have started work on similar tools and applications for several other Indigenous languages spoken as well in Canada. Importantly, though such tools are oriented firstly towards supporting literacy, it is worth noting that they can also provide substantial support in creating spoken resources and tools, thus presenting a further significant benefit as oralcy is valued highly by many Indigenous communities.

Alongside our work, we need to note a substantial number of parallel on-going projects on developing language technology for Indigenous languages in Canada, a comprehensive overview of which is presented in Littell et al. (2018).

5. Conclusion

The recent development and diffusion of digital devices makes literacy-based language tools useful for a critical generation of Indigenous language learners and speakers, And indeed, supporting literacy is historically not entirely foreign to Indigenous languages. In all this, we consider serving the needs and expectations of Indigenous communities as paramount, since we want our tools to be of genuine use to these communities, but we also recognize that there are in fact multiple user subgroups whose needs may diverge.

9. Acknowledgements

We acknowledge funding by the Social Sciences and Humanities Council of Canada (awards #890-2013-0047, #611-2016-0207, and #895-2019-1012), and the University of Alberta, Kule Institute of Advanced Study (KIAS Research Cluster grant 2015-2018).

6. Bibliographical References

Arppe, A. Lachler, J., Trosterud, T., Antonsen, L. and Moshagen, S. N. (2016). Basic Language Resource Kits for Endangered Languages: A Case Study of Plains Cree. *CCURL 2016 - Collaboration and Computing for Under-Resourced Languages – Towards an Alliance for Digital Language Diversity*, 1-8, Portorož, Slovenia, 23 May 2016. URL: http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-CCURL2016_Proceedings.pdf

Arppe, A., Schmirler, K., Harrigan, A. G. and Wolvengrey, A. (forthc.). A Morphosyntactically Tagged Corpus for Plains Cree. *Papers of the 49th Algonquian Conference*, Oct. 2017, Montréal, Quebec.

Beesley, K. R. and Karttunen, L. (2003). *Finite-State Morphology*. California: CSLI.

Bontogon, M., Arppe, A., Antonsen, L., Thunder, D. and Lachler, J. (2018). Intelligent Computer Assisted Language Learning (ICALL) for *nêhiyawêwin*: An In-Depth User Experience Evaluation. *Canadian Modern Language Review*, 74(3), 337–362

Harrigan, A. G., Schmirler, K., Arppe, A., Antonsen, L., Trosterud, T. and Wolvengrey, A. (2017). Learning from the computational modelling of Plains Cree verbs. *Morphology*, 27(4), 565–598.

Klavans, J. L. (2018). Computational Modeling of Polysynthetic Languages. Proceedings of Workshop on Polysynthetic Languages, pages 1–11 Santa Fe, New Mexico, USA, August 20-26, 2018.

Littell, P., Kazantseva, A., Kuhn, R., Pine, A., Arppe, A., Cox, C. and Junker, M-O. (2018). Indigenous language technologies in Canada: Assessment, challenges, and successes. In Bender, E., Derczynski, L. and Isabelle P. (eds.), *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, 2620–2632. Santa Fe, New Mexico: ACL.

Rogers, H. (2005). *Writing systems: a linguistic approach*. Blackwell publishing.

Statistics Canada (2017). Census of Population, 2016. URL: https://www12.statcan.gc.ca/census-recensement/2016/as-sa/98-200-x/2016022/98-200-x2016022-eng.cfm