Challenges and Perspectives for Innu-Aimun within Indigenous Language Technologies

Antoine Cadotte
Université du Québec à Montréal
cadotte.antoine@courrier.uqam.ca

Ngoc Tan Le
Université du Québec à Montréal
le.ngoc_tan@uqam.ca

Mathieu Boivin
Université de Montréal
mathieu.boivin.2@umontreal.ca

Fatiha Sadat
Université du Québec à Montréal
sadat.fatiha@uqam.ca

Abstract

Innu-Aimun is an Algonquian language spoken in Eastern Canada. It is the language of the Innu, an indigenous people that now lives for the most part in a dozen communities across Quebec and Labrador. Although it is alive, Innu-Aimun sees important preservation and revitalization challenges and issues. The state of its technology is still nascent, with very few existing applications. This paper proposes a first survey of the available linguistic resources and existing technology for Innu-Aimun. Considering the existing linguistic and textual resources, we argue that developing language technology is feasible and propose first steps towards NLP applications like machine translation. The goal of developing such technologies is first and foremost to help efforts in improving language transmission and cultural safety and preservation for Innu-Aimun speakers, as those are considered urgent and vital issues. Finally, we discuss the importance of close collaboration and consultation with the Innu community in order to ensure that language technologies are developed respectfully and in accordance with that goal.

1 Introduction

In 2016, there were over 70 different indigenous languages in Canada, which together cumulated 260,550 speakers[1], for a total country population of 35,151,728[2]. This number of indigenous language speakers shows how light their demographic weight is, considering the length of indigenous people's presence throughout the country. Yet this number also hides important disparities between indigenous languages. While Cree, spoken in four of the country's provinces, counts over 96,000 speakers, other languages like Haida are only spoken by a few hundreds[3].

In this paper, we examine one specific indigenous language, spoken in Quebec and Labrador, Innu-Aimun (ISO code moe [4], Glottolog mont1268[5]). Like the aforementioned languages, its presence in the country's linguistic landscape is fragile, especially compared to the official languages (English and French). While linguistic resources do exist for Innu-Aimun, the development of its language technology and NLP applications is almost inexistent.

Our contributions to the current research consist in two main parts: (1) a survey of Innu-Aimun linguistic resources and technology, followed by (2) discussions and perspectives concerning how to develop Innu-Aimun language technology and the role it could have for the community.

The structure of this paper is described as follows: Section 2 describes Innu-Aimun and its linguistic situation, and surveys the available linguistic resources and technologies. Section 3 addresses the question of how Innu-Aimun language technology should be developed, including the importance of collaboration with the community, examining language-related social issues and discussing what role technology could have to help on these issues. Section 4 provides some ideas for short term and longer term developments, focusing on short term development perspectives and how they could be carried out. Finally, Section 5 concludes this paper and suggests future directions for further research.

2 Language Description and Survey

2.1 Innu-Aimun language

Innu-Aimun is a language spoken by the Innu, an indigenous people of Canada, formerly known as Montagnais (Mollen, 2006). It is a polysynthetic

[1] Statistics Canada: The Aboriginal languages of First Nations people, Métis and Inuit

[2] Statistics Canada: Census Profile, 2016 Census

[3] Statistics Canada: The Aboriginal languages of First Nations people, Métis and Inuit

[4] ISO 639-3 - moe

[5] Glottolog - mont1268

Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 99 - 108
May 26-27, 2022 ©2022 Association for Computational Linguistics

language and part of the Algonquian language family and of the Cree-Innu-Naskapi dialect continuum (Drapeau, 2014b). In 2017, the number of speakers was estimated at 12,000, spread over a dozen communities (Baraby et al., 2017).

As noted by Baraby et al. (2017), Innu-Aimun is "[...] alive but still fragile". Its state of preservation can be seen as part of the broader situation of indigenous languages in Canada. Generally speaking there is a transfer to the majority language (English in general, French mostly in the case of Innu) and indigenous language fluency is lower in younger age groups than in older ones (Drapeau, 2011). In some Innu communities, lexical erosion has been observed due to the high rate of bilingualism among speakers (Drapeau, 2014a).

Originally an oral language with several dialects, Innu-Aimun had its orthography standardized in 1989 (Mollen, 2006). This standardization work, done through a consultation between representatives of the different dialects, concerns only the written language; the differences in pronunciation between dialects remain (Mollen, 2006). The standard Innu-Aimun orthography is based on the Latin alphabet and includes a special character: the "superscript-u". This character has its own unicode code point [6], which has been incorporated in a keyboard developed specifically for Innu-Aimun [7].

2.2 Existing Innu-Aimun Resources and Technological Applications

While several linguistics resources are available, there are very few technological applications for Innu-Aimun. These were developed primarily for educational and language preservation purposes. This section describes existing linguistic resources for Innu-Aimun and applications that are part of a joint development effort with Cree language.

2.2.1 Primary linguistic resources

Despite its significant preservation challenges, Innu-Aimun is one of the best documented indigenous languages in Canada, and its documentation has become more technology-based in recent years (Baraby et al., 2017). Among the primary linguistic resources is the *Innu Grammar* (*Grammaire de la langue innue*) by Drapeau (2014b), which describes in detail many aspects of the Innu-Aimun grammatical structures.

The 1991 *Montagnais-French Dictionnary* (*Dictionnaire Montagnais-français*) by Drapeau (1991) was the first to use the standardized orthography (Mollen, 2006). The most up-to-date published dictionary available is the *Innu-French Dictionary* (2016, second edition) which includes more than 28,000 Innu-Aimun words [8]. However, an online version of this dictionary is available[9] and it is regularly expanded with new words (this tool is further discussed in the following section).

For conjugation, the guide *Conjugation of Innu verbs* (*Conjugaison des verbes innus*) by Baraby and Junker (2011) is available as a Website [10]. This guide is based on the work started with *Guide pratique des principales conjugaisons en montagnais* (Baraby, 1998) which has since been updated.

2.2.2 Integrated Web tools and Search Engine

With *Integrated Web tools for Innu language maintenance*, Junker et al. (2016) presented a series of Web tools intended primarily for bilingual speakers of Innu-Aimun and whose main goal was the preservation of the language. These included language learning games, several basic language resources (grammars, lexicons, etc.), a catalog of works in Innu-Aimun (including educational books, children's stories, etc.), an online dictionary and a verb conjugation application.

The verb conjugation application, developed by Baraby and Junker (2011), organizes verbs with respect to Innu-Aimun conjugation structure and includes audio clips for the pronunciation in the eastern and western dialects. The trilingual, pan-dialectal online dictionary (mentioned in the previous section) structures search results with respect to the Innu-Aimun morphology. Table 1 shows examples of entries from the Innu-Aimun-English part of dictionary. The dictionary uses the orthographic flexibility of the Innu-Aimun search engine developed by Junker and Stewart (2008).

In *Building search engines for Algonquian languages*, Junker and Stewart (2008) developed a search engine for East Cree (an Algonquian language related to Innu-Aimun) and then adapted it to Innu-Aimun. The authors' work consists of two parts: flexible orthographic search and verb search. The flexible orthographic search aims to solve different spellings problem for the same word, as the recent standardization of spelling, the existence of

[6] *Innu-Aimun.ca* - Writing and Technology (in French)

[7] Keyboard layout for Ilnu/Innu Aimun

[8] Tshakapesh Institute - *Dictionnaire Innu-Français*

[9] `https://dictionary.innu-aimun.ca/`

[10] `https://verbe.innu-aimun.ca`

Main verb (English)	Examples of bilingual entries	
	Innu-Aimun	**English**
see	eukuan	oh yeah!, I see!
	tepapameu	s/he has seen enough of him/her
	unapatam^{u}	s/he is mistaken about what s/he sees; s/he loses sight of it
make	tutam^{u}	s/he makes it
	tutamueu	s/he makes it for someone

Table 1: Example entries from the Innu-English online dictionary (Junker et al., 2016)

several dialects and the predominance of oral language mean that users will often look for a word with a different orthography than the standard one. The verb search component consists of a flexible search in a database of verbal paradigms, which identifies the most likely root and reconstructs the verb in its standard form. This aims to solve a challenge arising from the different forms of verb inflections in Cree and Innu-Aimun, as users can search for verbs in their non-canonical form.

Hasler et al. (2018) proposed an online terminology forum for multiple Algonquian languages, including Innu-Aimun, in order to provide translations and definitions for specialized terms in several fields such as healthcare, justice or environment. The forum is a tool for collaborative terminology development, with participation from communities and review from translators.

2.3 Existing Innu-Aimun digitized resources

To our knowledge, there is no publicly available annotated or aligned Innu-Aimun corpora, and few research works report on this subject. There exists however a certain amount of publicly available monolingual, bilingual and trilingual Innu-Aimun digitized texts.

2.3.1 Transcription and linguistic annotation

Citing the lack of linguistic documentation despite its importance to the preservation of the language, Drapeau and Lambert-Brétière (2013) presented a project to create and make available a corpus of linguistically analyzed Innu-Aimun texts. The corpus was built through the segmentation and transcription of oral recordings in standard orthography. The text analysis includes morphological segmentation and translation into French and English. The result is a multimodal, multilingual annotated Innu-Aimun corpus.

Kuhn et al. (2020) presented the language technology project by NRC Canada and its collaborators. This project aims to transcribe oral recordings for several indigenous languages in Canada, including Innu-Aimun.

2.3.2 Multilingual textual resources

The Tshakapesh Institute has an online catalog offering many texts in Innu-Aimun. This includes pedagogical books (primary and preschool), stories for children, novels, poetry, non-fiction and other types of works[11]. However, many of these texts are available only in non-digitized versions.

The publishing house *Mémoire d'encrier*, publishes bilingual Innu-Aimun-French works, such as novels (notably reeditions of works by Innu author An Antane Kapesh[12]) and collections of poems (notably by the Innu poet Joséphine Bacon[13]). Some of those titles are also published in bilingual Innu-Aimun-English versions[14]; those can thus be considered as trilingual Innu-Aimun-French-English texts.

On rarer occasions, texts in multiple indigenous languages are made available. The FNQLSDI (First Nations of Quebec and Labrador Sustainable Development Institute) produces documents in 6 languages including English, French, Innu-Aimun and other indigenous languages such as East Cree, and sometimes up to 12 languages [15].

[11] Tshakapesh Institute - Catalogue

[12] Mémoire d'encrier - An Antane Kapesh

[13] Mémoire d'encrier - Joséphine Bacon

[14] For example: Mawenzi House - *Message Sticks (Tshissinuatshitakana)*

[15] FNQLSDI - Multilingual books

3 Discussion: How Should Technology Be Developed for Innu-Aimun?

3.1 The imperative of respecting and collaborating with the community

Social and ethical aspects are of particularly great importance when it comes to practicing research involving indigenous languages in Canada. This should be emphasized not only considering the precarious situation of these languages, but also—and most importantly—in light of the well documented historical prejudices and subsisting societal issues indigenous communities have been subjected to. This includes the appalling legacy of the Indian Residential Schools system, as documented by the Truth and Reconciliation Commission of Canada[16]. Such considerations are crucial for indigenous languages in Canada in general and they should absolutely be kept in mind for Innu-Aimun language technology development.

As per the directives of the Social Sciences and Humanities Research Council (SSHRC) of Canada, "Whatever the methodologies or perspectives that apply in a given context, researchers who conduct indigenous research, whether they are indigenous or non-indigenous themselves, commit to respectful relationships with all indigenous peoples and communities."[17]

Indigenous research should as much as possible be done *by and for* the community. In the case of indigenous language technology development, this takes an even greater significance as such research aims first and foremost to have a concrete positive impact indigenous communities. Language technologies must address in their development the needs as well as the concerns of the community they serve.

If the developed technologies result in tools intended as applications with end users, evaluation of the technologies by members of the community should be a key component of a collaborative development. In the case of indigenous languages in Canada, an example of such an evaluation for a precise language is the one carried by Bontogon (2016) for a Plains Cree computer-aided language learning tool (CALL).

3.2 Language and social issues

Among the most urgent linguistic issues expressed by some community members[18] is the need for better and safer interactions between Innu and health and social workers, as well as in the educational and justice systems. In the latter, ensuring the clarity of interactions with an indigenous person, by using an interpreter if need be, is not only important but a legal obligation, as described by Newashish and Boivin (2019).

Language plays an important role in culturally safe communications with health workers, as discussed by Møller (2016) in their study of language for nursing in Nunavut and Greenland. Cultural safety overall has been identified as important to ensure safe interactions with health workers: if interactions between indigenous patients and health workers are not adequate, this can lead to potentially disastrous situations like death, as has been recently concluded by a coroner inquiry following the death of an indigenous patient in Canada[19].

The Viens Commission final report[20] mentions that 54% of indigenous people in Quebec live in cities rather than in indigenous communities and that this makes access to services in their language all the more difficult. The need to improve the relation and interactions between Innu and non-indigenous in urban context has also been identified as an important matter by Leroux (2014) when examining cohabitation within Sept-Îles: difference in native language between non-indigenous and Innu is considered to play a role in the divide between the two.

The Innu-Aimun language is an integral part of Innu identity and this makes language preservation all the more important. As highlighted by Leroux (2014) through her interviews with Innu community members, the attempted assimilation of the Innu people to the dominant non-indigenous society is still profoundly felt and has had an impact on transmission of the language.

According to one Innu-Aimun teacher from the Uashat mak Mani-utenam community, with whom we exchanged, the language is highly endangered. Rare are the students that properly master their mother tongue and French dominates in day-to-day interactions. Not enough time in the curriculum,

[16] Truth and Reconciliation Commission of Canada

[17] Social Sciences and Humanities Research Council - Definition of Terms, indigenous Research

[18] ITUM (Innu Takuaikan Uashat mak Mani-utenam) - Council of Uashat mak Mani-utenam

[19] Investigation Report on the Death of Joyce Echaquan (in French)

[20] Final report of the Viens Commission (in French)

she says, is allocated to teaching Innu-Aimun and preserving the language should overall be considered as a more pressing societal concern.

3.3 NLP and Innu-Aimun revitalization

As stated earlier, research in Innu-Aimun language technologies should first address the priorities and needs of the Innu communities, as expressed by them. For that matter, consultation with the community is a key part of such research. In this section, we offer ideas of roles Innu-Aimun language technology could play, as a first step towards further consulting the community—should it be to validate these ideas or to stimulate discussion on the matter and encourage other ideas.

Language preservation is a role commonly projected onto indigenous language technologies. We indeed believe language technologies could help preserve Innu-Aimun by acting as educational tools to native speakers and by acting as technologically-oriented language documentation. From the existing tools like online bilingual dictionaries to potential developments like machine translation, conversational agents and learning assistants, we think language technology could help support native speakers learn their language or improve their knowledge of it, and especially so in a context of prevailing bilingualism.

Some community members have said in discussions we held with them that in their view, an even more important role language technology could play is that of raising awareness and understanding within non-indigenous people. It is believed that gaining better knowledge of Innu-Aimun could help better raise awareness and understand Innu realities, which is of great importance for reconciliation. This becomes even more crucial for non-indigenous workers that interact with the community, as is often the case in the health and education sectors. When it comes to interactions in the context of health and education services, ensuring language knowledge becomes a matter of cultural safety. This need has already been recognized and some steps have been taken, like the recent creation of a program for translation and interpretation to and from Innu-Aimun [21]. We think the development of cross-lingual Innu-Aimun technologies is in line with those efforts and could be of great help to ensure cultural safety.

4 Perspectives: Innu-Aimun and NLP

In light of the discussed roles for Innu-Aimun language technology, we present here our proposed vision for potential technological developments in collaboration with the ITUM group[22]. This vision is divided into two more accessible developments in the short term and two longer term developments.

4.1 Short term

4.1.1 Towards a first machine translation system

We consider machine translation could be a useful tool to the Innu community, both for language learning and to assist professional translators and teachers. On the language learning side, machine translation could serve as an extension or an enhancement of bilingual dictionaries. When it comes to forming Innu-Aimun words that correctly grasp the desired context, automated sentence translation could prove useful and machine translation can help reach that goal. On translation assistance, we concur with the view brought by Littell et al. (2018): that a general-purpose system like Google Translate is probably not achievable with the current state of resources and that translation assistance is a more accessible goal. Such an approach would also be more empowering for the community as it would aim to assist rather than replace Innu translators.

4.1.1.1 Parallel corpora

With the publicly available bilingual and trilingual Innu-Aimun texts, it is certainly possible to create experimental Innu-Aimun-French and Innu-Aimun-English parallel corpora. Some of the bilingual works mentioned earlier are only available in paper, while some are available in ebook and PDF formats. Naturally, books available in paper only would require a significant amount of work in order to be rendered usable as parallel data, as it would involve scanning the documents and using OCR (Optical Character Recognition) methods in order to obtain workable text. Considering only Innu-Aimun-French bilingual texts that are easily obtainable as ebook or PDF, we estimate that at least 3000 parallel Innu-Aimun-French sentences could be collectable with minimal effort. Such a

[21] Sept-Îles Cégep launches an Innu language translation program (in French)

[22] ITUM (Innu Takuaikan Uashat mak Mani-utenam) - the council of Uashat mak Mani-utenam

small number of examples might not be enough to train an Innu-Aimun-specific translation model, but it could be put to contribution using machine learning techniques that are better adapted to low-resource or zero-shot settings and that harness data from other languages, as discussed in the following section.

Table 2 examines three bilingual corpora for the Innu-Aimun and French language pair: FNQLSDI books[23], Mémoire d'encrier poetry[24] and Mémoire d'encrier novels & essays[25]. The number of parallel sentences is an approximation and it might vary following proper alignment. Also in this table is the vocabulary size for each corpus and the percentage of words from this vocabulary that are absent from the most complete Innu-Aimun dictionary available [26]. We can observe that in all three cases, a very high proportion of the words found on the Innu-Aimun side (82-87%) are out-of-dictionary. Some of the out-of-dictionary words are simply proper nouns or words borrowed from other languages (e.g. French). But the main explanation probably resides in the polysynthetic nature of the language: as morphemes agglutinate to form longer words, a high proportion of the words will be in fact found in an inflected form that is not present in the dictionary. This observation also serves as a reminder of the importance of segmentation for the development of machine translation for Innu-Aimun.

The Innu-Aimun-French dictionary itself could be put to use as parallel data for Machine Translation. The 28K+ words and definitions found in the dictionary could probably not be counted as so many parallel sentences. But since many of these words are provided with translations that are as long as a sentence due to their high morphology, the parallel data found in the Innu-Aimun dictionary is certainly more useful to machine translation than that found in traditional bilingual dictionaries (where translation is usually provided as single corresponding words).

Since recent NMT (Neural Machine Translation) methods using auxiliary, higher-resourced languages have shown positive results for low-resource language pairs, even when the languages are unrelated (see Section 4.1.1.2), it is appropriate to survey other indigenous languages in Canada, with regard to their proximity and the availability of training data.

While the availability of open parallel corpora (and training data in general) is a major challenge for most indigenous languages in Canada, such a corpus has been published for the Inuktitut-English language pair and has made possible the development of machine translation models (Littell et al., 2018). This corpus contains in its third and latest version over 1.4 million pairs of aligned Inuktitut-English sentences, all collected from the proceedings of the Nunavut Hansard which is published in both languages (Joanis et al., 2020). This, according to the authors, constitutes the largest publicly available parallel corpus for a polysynthetic language.

Atikamekw, another indigenous language of Canada that belongs to the same family as Innu-Aimun (Algonquian languages) has a wikimedia project[27], which could help construct comparable corpora for these category of languages and thus enrich a multilingual neural machine translation framework. In addition, some of the FNQLSDI books available in Innu-Aimun are also available in Atikamekw, as well as in East Cree .

4.1.1.2 Applying methods for extremely low-resource language pairs

A large variety of methods have been proposed in different contexts to improve neural machine translation results for low-resource language pairs, as recent surveys show (Wang et al., 2021; Haddow et al., 2021). Several of these methods involve making use of data from auxiliary languages (such as transfer learning, multilingual modelling and others). Keeping in mind that the main goal of first experiments in machine translation is to assess feasibility, we focus here on the methods that allow the direct use of the parallel resources that have been mentioned so far.

Multilingual modelling, which aims at harnessing the most of many languages by sharing parameters between them, has shown good results for low-resource language pairs. This is done by Johnson et al. (2017), which show that it is possible to improve results for lower-resourced languages by combining them into a single model along with higher-resourced language pairs (with a total of 12 pairs). More recent contributions push

[23] FNQLSDI - Multilingual books

[24] Mémoire d'encrier - Joséphine Bacon

[25] Mémoire d'encrier - An Antane Kapesh

[26] Innu-Aimun online dictionary

[27] https://atj.wikipedia.org/wiki/Otitikowin

Corpus	Number of parallel sentences	Innu-Aimun vocabulary size	% of out-of-dictionary words
FNQLSDI books	1,450	4,453	87%
Mémoire d'encrier poetry	110	1,558	82%
Mémoire d'encrier novels & essays	1,670	4,170	87%

Table 2: Parallel Innu-Aimun - French corpora

the number of languages much higher and label the method "massively multilingual modelling". Aharoni et al. (2019) for example train models in one-to-many and many-to-one settings combining 102 languages and many-to-many models combining 59 languages. They improve previous results for low-resource pairs and show that adding more languages improves zero-shot performances, but pointout there exists a trade-off between the number of languages and overall translation performance, especially for higher-resourced pairs.

The goal of keeping a better performance for all language pairs in a multilingual model does not apply to our proposed experiments for Innu-Aimun translation: if adding more languages helps improve results for our low-resourced target language pairs, then there is no incentive not do so. Furthermore, recent results suggest that transfer learning can even be beneficial to unrelated languages that have different alphabets (Kocmi and Bojar, 2018).

Another method of interest is meta-learning, which Gu et al. (2018) applied on low-resource machine translation. They show meta-learning can significantly improve translation results for lower-resourced pairs, especially in zero-shot situations or when training with few examples.

4.1.2 Developing a morphological segmenter

We take the view that one of the first steps to take for Innu-Aimun language technology development is to consolidate and expand building blocks that are considered part of a basic language toolkit, as described for Plains Cree by Arppe et al. (2016). These building blocks, like lexical databases, written corpora and transcriptors, are not only important technological tools for languages but also form the basis for more advanced developments.

A fundamental building block to develop for Innu-Aimun is automated morphological segmentation. As stated earlier, the polysynthetic nature of Innu-Aimun makes some of its words equivalent to whole sentences in Indo-European languages. This trait makes the use of morphological segmentation almost inevitable for applications like machine translation, as described for Inuktitut and Inuinnaqtun, other indigenous languages in Canada (Le and Sadat, 2020b, 2021).

In the absence of a language-specific segmentation model, some unsupervised methods (i.e. learning methods that do not require annotated data) allow the training of a model using solely monolingual data from the target language. This is the case of the BPE (*Byte Pair Encoding*) method proposed for segmentation by Sennrich et al. (2016), which merges most frequent pairs of characters or n-grams (i.e. sequences of items like words, syllables, letters, etc.) found in the text to construct a subword vocabulary for the targeted language. However, such a method does not replace a language-specific segmentation model. For example, Le and Sadat (2020a) improve the translation results obtained by Joanis et al. (2020) on their Nunavut Hansard Inuktitut-English corpus by proposing their own Inuktitut-specific segmentation model.

A logical step to develop an Innu-specific segmentation model is to adapt to Innu-Aimun the Plains Cree FST model proposed by Snoek et al. (2014). The authors consider their model to be adaptable to other Algonquian languages, since the language structure would be similar. Another similar approach would be to use the same development method used by Snoek et al. (2014) and Harrigan et al. (2017) for Plains Cree or by Arppe et al. (2017) for East Cree, in order to develop an FST model specific to Innu.

Another possible approach for Innu-Aimun automated segmentation is the semi-supervised method, as used by Le and Sadat (2021) to develop their segmentation model for Inuinnaqtun (an endangered indigenous language of Canada). Semi-supervised methods are usually hybrid approaches that combine unsupervised methods with the use of available annotated data. In the case of Le and Sadat (2021), the proposed approach uses the Adaptor Grammars based framework by Eskander et al. (2020), which can learn a model based on a list of unsegmented words using grammar rules. These rules can also include a list of morphemes from the target language.

The semi-supervised approach is promising in the Innu-Aimun context, since enough linguistic documentation exists to define general grammar patterns as done by Le and Sadat (2021) and since a list of Innu-Aimun words and morphemes could be collected from the available dictionaries and verb conjugators. Counting the number of unique Innu-Aimun words currently found in the Innu-Aimun online dictionnary, combined with the words found the corpora analyzed in Table 2, a vocabulary size of 34K can be obtained and put to use in semi-supervised automated segmentation methods.

4.2 Longer term

4.2.1 Cross-lingual conversational agent

We propose the longer term development of a cross-lingual conversational agent whose primary purpose would be to act as an intelligent language tutor. This can be seen as being in line with the existing interactive learning games, proposed as part of the integrated web tools by Junker et al. (2016). First steps in the construction of the agent would involve collecting a Question-Answering dataset within educational and health groups/centers in Innu communities. Such a tool would assist not only native speakers in their learning of Innu-Aimun, but also non-native speakers in their communication and understanding of the communities' culture and realities. It could play a positive role especially for beginner-level learners and in contexts where access to an Innu-Aimun teacher is problematic—which is the case especially outside communities.

4.2.2 Automatic Innu-Aimun multimodal machine translation

As stated earlier, standardization of Innu-Aimun orthography is relatively recent (since 1989) (Mollen, 2006). This means many community members learned the language before the standardization occurred. Automated transcription could bridge the gap between how speakers use their language and how orthography-based tools function.

Among indigenous languages in Canada, an attempt was made with the development of an ASR (Automatic Speech Recognition) system for Inuktitut (Gupta and Boulianne, 2020). This project aimed to automatically transcribe Inuktitut and used 23 hours of transcribed Inuktitut oral stories to build an acoustic model.

However, relying on voice-based technologies brings significant challenges. Due to lack of data, dialect variances, and other restrictions, it is difficult to create strong ASR systems for indigenous languages (Jimerson and Prud'hommeaux, 2018). In the case of Innu-Aimun, not only is the writing far from its pronunciation, but the existence of multiple dialects means there are multiple ways to pronounce, depending on the region or community, as mentionned by Mollen (2006).

Despite significant challenges, developing multimodal systems would help to better represent cultural and ancestral data through voice—considering that Innu-Aimun is traditionally an oral language Mollen (2006). Fortunately, in the last few years, there have been efforts to digitise content in Innu-Aimun, both in text and in audio format, as stated in section 2.2.

5 Conclusion

Despite substantial challenges ahead, like the limited amount of resources available or the complexity of the language, we consider the development of more advanced Innu-Aimun technology to be feasible. We also consider such a development to be important, in view of the very real social issues related to Innu-Aimun. We believe language technologies like machine translation could be useful in the efforts to ensure language transmission and improve cultural safety in services. The first steps we proposed in this article, besides their goal of demonstrating feasibility, will help better understand the difficulties in processing Innu-Aimun texts and building technological modules like morphological and translation models. This will allow defining further steps towards the longer term goals like intelligent tutors, conversational agents and automatic transcription.

Acknowledgements

The authors are grateful to the community of Uashat Mak Mani-Utenam, Samuel Marticotte for his role in initiating this project, Mrs. Denise Jourdain for sharing her knowledge and experience, and Prof. Yvette Mollen for her precious advices and feedbacks. We also thank the anonymous reviewers for their valuable comments. This work has been partially supported by the The Centre for Research on Brain, Language and Music (CRBLM).

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Antti Arppe, Marie-Odile Junker, and Delasie Torkornoo. 2017. 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*, pages 52–56, Honolulu. Association for Computational Linguistics.

Antti Arppe, Jordan Lachler, Trond Trosterud, Lene Antonsen, and Sjur N. Moshagen. 2016. Basic Language Resource Kits for Endangered Languages: A Case Study of Plains Cree. In *Proceedings of the LREC 2016 Workshop "CCURL 2016 – Towards an Alliance for Digital Language Diversity"*, pages 1–8, Portorož (Slovenia).

Anne-Marie Baraby. 1998. Guide pratique des principales conjugaisons en Montagnais. *Sept-Iles: Institut culturel et éducatif montagnais*.

Anne-Marie Baraby and Marie-Odile Junker. 2011. Conjugaisons des verbes innus.

Anne-Marie Baraby, Marie-Odile Junker, and Yvette Mollen. 2017. A 45-year old language documentation program first aimed at speakers: the case of the Innu.

Megan A. Bontogon. 2016. *Evaluating nêhiyawêtân: A computer assisted language learning (CALL) application for Plains Cree*. Ph.D. thesis, University of Alberta.

L. Drapeau. 2011. *Les langues autochtones du Québec: Un patrimoine en danger*. Presses de l'Université du Québec.

Lynn Drapeau. 1991. *Dictionnaire montagnais-français*. Presses de l'Université du Québec.

Lynn Drapeau. 2014a. Bilinguisme et érosion lexicale dans une communauté montagnaise. In Pierre Martel and Jacques Maurais, editors, *Langues et sociétés en contact: Mélanges offerts à Jean-Claude Corbeil*, pages 363–376. Max Niemeyer Verlag.

Lynn Drapeau. 2014b. *Grammaire de la langue innue*. Presses de l'Université du Québec.

Lynn Drapeau and Renée Lambert-Brétière. 2013. The innu language documentation project. In *Proceedings of the 17th Foundation for Endangered Languages Conference*.

Ramy Eskander, Francesca Callejas, Elizabeth Nichols, Judith Klavans, and Smaranda Muresan. 2020. MorphAGram, evaluation and framework for unsupervised morphological segmentation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7112–7122, Marseille, France. European Language Resources Association.

Jiatao Gu, Yong Wang, Yun Chen, Victor O. K. Li, and Kyunghyun Cho. 2018. Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631, Brussels, Belgium. Association for Computational Linguistics.

Vishwa Gupta and Gilles Boulianne. 2020. Automatic transcription challenges for inuktitut, a low-resource polysynthetic language. In *Proceedings of the 12th language resources and evaluation conference*, pages 2521–2527.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2021. Survey of low-resource machine translation.

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.

Laurel Anne Hasler, Marie-Odile Junker, Marguerite MacKenzie, Mimie Neacappo, and Delasie Torkornoo. 2018. The Online Terminology Forum for East Cree and Innu: A collaborative approach to multi-format terminology development. In *LD&C Special Publication No. 20: Collaborative Approaches to the Challenges of Language Documentation and Conservation*. University of Hawai'i Press.

Robbie Jimerson and Emily Prud'hommeaux. 2018. Asr for documenting acutely under-resourced indigenous languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Eric Joanis, Rebecca Knowles, Roland Kuhn, Samuel Larkin, Patrick Littell, Chi-kiu Lo, Darlene Stewart, and Jeffrey Micher. 2020. The Nunavut Hansard Inuktitut–English parallel corpus 3.0 with preliminary machine translation results. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2562–2572, Marseille, France. European Language Resources Association.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Marie-Odile Junker, Yvette Mollen, Hélène St-Onge, and Delasie Torkornoo. 2016. Integrated web tools for Innu language maintenance. In *Papers of the 44th Algonquian Conference*, pages 192–210.

Marie-Odile Junker and Terry Stewart. 2008. Building search engines for Algonquian languages. *Algonquian Papers-Archive*, 39.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252, Brussels, Belgium. Association for Computational Linguistics.

Roland Kuhn, Fineen Davis, Alain Désilets, Eric Joanis, Anna Kazantseva, Rebecca Knowles, Patrick Littell, Delaney Lothian, Aidan Pine, Caroline Running Wolf, Eddie Santos, Darlene Stewart, Gilles Boulianne, Vishwa Gupta, Brian Maracle Owennatékha, Akwiratékha' Martin, Christopher Cox, Marie-Odile Junker, Olivia Sammons, Delasie Torkornoo, Nathan Thanyehténhas Brinklow, Sara Child, Benoît Farley, David Huggins-Daines, Daisy Rosenblum, and Heather Souter. 2020. The Indigenous Languages Technology project at NRC Canada: An empowerment-oriented approach to developing language software. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5866–5878, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Tan Ngoc Le and Fatiha Sadat. 2020a. Low-resource NMT: an empirical study on the effect of rich morphological word segmentation on Inuktitut. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 165–172, Virtual. Association for Machine Translation in the Americas (AMTA 2020).

Tan Ngoc Le and Fatiha Sadat. 2020b. Revitalization of indigenous languages through pre-processing and neural machine translation: The case of Inuktitut. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4661–4666, Barcelona, Spain (Online). International Committee on Computational Linguistics (COLING 2020).

Tan Ngoc Le and Fatiha Sadat. 2021. Towards a first automatic unsupervised morphological segmentation for Inuinnaqtun. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 159–162, Online. Association for Computational Linguistics (NAACL 2021).

Shanie Leroux. 2014. Le point de vue des innus de sept-Îles, uashat et maliotenam sur les relations entre autochtones et allochtones en milieu urbain : vers une concitoyenneté. *Nouvelles pratiques sociales*, 27(1):64–77.

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

Yvette Mollen. 2006. Transmettre un héritage: la langue innue. *Cap-aux-Diamants: la revue d'histoire du Québec*, 1(85):21–25. Publisher: Les Éditions Cap-aux-Diamants inc.

Helle Møller. 2016. Culturally safe communication and the power of language in arctic nursing. *Études/Inuit/Studies*, 40(1):85–104.

Maggie Newashish and Mathieu Boivin. 2019. Interprétation judiciaire atikamekw : ce que c'est; ce qu'il reste à faire... *Histoire Québec*, 24(4):12–14.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

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.

Rui Wang, Xu Tan, Renqian Luo, Tao Qin, and Tie-Yan Liu. 2021. A survey on low-resource neural machine translation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4636–4643. International Joint Conferences on Artificial Intelligence Organization. Survey Track.