## Machine Translation Through Cultural Texts: Can Verses and Prose Help **Low-Resource Indigenous Models?**

**Antoine Cadotte** 

### Nathalie André

Université du Ouébec à Montréal

École Kanatamat, Matimekush-Lac John antoine.cadotte@courrier.uqam.ca nathalie.andre@ecolekanatamat.ca

> **Fatiha Sadat** Université du Québec à Montréal sadat.fatiha@ugam.ca

### Abstract

We propose the first MT models for Innu-Aimun, an Indigenous language in Eastern Canada, in an effort to provide assistance tools for translation and language learning. This project is carried out in collaboration with an Innu community school and involves the participation of other participants, within the framework of a meaningful consideration of Indigenous perspectives. Our contributions in this paper result from the three initial stages of this project. First, we aim to align bilingual Innu-Aimun/French texts with collaboration and participation of Innu-Aimun locutors. Second, we present the training and evaluation results of the MT models (both statistical and neural) based on these aligned corpora. And third, we collaboratively analyze some of the translations resulting from the MT models. We also see these developments for Innu-Aimun as a useful case study for answering a larger question: in a context where few aligned bilingual sentences are available for an Indigenous language, can cultural texts such as literature and poetry be used in the development of MT models?

#### 1 Introduction

Innu-Aimun, formerly known as Montagnais (ISO code moe,<sup>1</sup> Glottolog mont $(1268)^2$ , is the language of the Innu, an Indigenous people present in the Quebec and Labrador provinces of Canada. It is a polysynthetic language, member of the Algonquian language family and part of the Cree-Innu-Naskapi dialect continuum (Drapeau, 2014b).

According to the latest Statistics Canada census (2021), an estimated 11,605 locutors speak Innu-Aimun, when including the related Naskapi language.<sup>3</sup> This figure has seen a negative variation compared to the previous census (2016),  $^4$ 

and UNESCO considers Innu-Aimun to be endangered/unsafe.<sup>5</sup> This echoes other assessments made by language specialists (Baraby et al., 2017; Drapeau, 2014a, 2011).

A lack of available services in Innu-Aimun has been documented, with insufficient availability of professional translators and interpreters, for Innu-Aimun and for all Indigenous languages in Quebec.<sup>6</sup> In a revitalisation effort, a new professional Innu-Aimun translation and interpretation program is being offered.<sup>7</sup> In this context and with the aim of helping Innu-Aimun translators as well as language learners and users in general, we are taking the first steps to develop a Machine Translation (MT) system for the Innu-Aimun language and French language pair. This project is carried out in collaboration with the Innu-Aimun teaching staff of Innu community school École Kanatamat. The project also involves the participation of students in Innu-Aimun translation and interpretation from the new college program at Cégep de Sept îles.

Innu-Aimun being originally an oral language, the Innu have a rich tradition of oral storytelling but also, since more recently, an ever growing written, literature, including novels and poetry (St-Gelais, 2022). It is for the most part written in French, with some publications in bilingual Innu-Aimun and French editions. Although working with literary and poetic data presents big challenges, such as artistic and figurative translations or particular writing styles, we take these constraints as an opportunity to answer the following questions: Can these types of texts have a positive impact on the results of Machine Translation models for Innu-Aimun and French languages?

Innu-Aimun, as is generally the case with Indigenous language in Canada, is not covered by

<sup>&</sup>lt;sup>1</sup>ISO 639-3 - moe

<sup>&</sup>lt;sup>2</sup>Glottolog - mont1268

<sup>&</sup>lt;sup>3</sup>Statistics Canada: Indigenous languages in Canada, 2021

<sup>&</sup>lt;sup>4</sup>Indigenous languages across Canada

<sup>&</sup>lt;sup>5</sup>UNESCO World Atlas of Languages - Montagnais

<sup>&</sup>lt;sup>6</sup>Final report of the Viens Commission (in French)

<sup>&</sup>lt;sup>7</sup>Innu-Aimun Interpretation and Translation program at Cégep de Sept-Îles (in French)

currently available MT systems. It is also underrepresented—if not absent—of Large Language Models (LLM) which are increasingly used as translation tools. In the face of recent questioning on the cultural sensitivity and awareness of general NMT models and LLMs, we propose a more culturally-aware approach, based on cultural texts and involving collaboration and participation of Innu-Aimun locutors.

Our paper is structured as follows. We present a brief state-of-the-art related to our topic in Section 2 and our proposed methodology for the present study in Section 3. Section 4 presents results from the first MT models as well as qualitative observations and analyses. Section 5 concludes this paper.

### 2 Related Works

The existing language technologies for Innu-Aimun today are mainly online language tools, such as a bilingual dictionary, a verb conjugation application and learning games (Junker et al., 2016). Similar tools exist for the neighboring language of Eastern Cree.<sup>8</sup> Other important building blocks, such as morphological models, have been developed for related languages among which Plains Cree is notable (Arppe et al., 2016). A word segmentation tool based on deep learning has also been proposed for Innu-Aimun (Tan Le et al., 2022).

The only Indigenous language for which MT currently exists in Canada is Inuktitut. The development of MT for Inuktitut was made possible by the availability of the Nunavut Hansard Corpus, constructed from the bilingual Inuktitut-English debates of the Legislative Assembly (Joanis et al., 2020). This availability of data also made it possible to improve word segmentation (Le and Sadat, 2020) and the study of gender biases in the bilingual corpus (Hansal et al., 2022; Le et al., 2023).

As several computational linguists have noted (notably Bird (2020)), it is primordial, when working on Indigenous language technology or resource development, to adopt a community-based and collaborative approach, to avoid repeating colonial research patterns. An exemplary participatory approach to Neural Machine Translation (NMT) was demonstrated in the Masakhane Project for African languages (Nekoto et al., 2020). Closer to Indigenous languages in Canada, Bontogon (2016) has taken the necessary but often overlooked step of human evaluation of the Indigenous language learning tool, in this case including by native speakers.

In addition to research approaches, closer attention has also been brought to the cultural awareness in machine translation tools themselves, especially with the recent advent and ubiquity of Large Language Models (LLM). For example, Yao et al. (2024) have underlined the limits of NMT models and LLMs when it comes to translating sentences with cultural content and proposed a data curation pipeline and an evaluation metric to better address this issue. Masoud et al. (2024) have benchmarked the performance of several state-of-the-art LLMs on cultural differences, using three of the largest and most represented cultural groups (American, Chinese, and Arab) and have noticed significant challenges even at this level of representation.

### **3** Proposed Methodology

### 3.1 Creation of Innu-Aimun/French alignments with collaboration and participation of Innu-Aimun locutors

Two main reference aligned corpora are used for evaluation as part of this comparative study, based on three bilingual Innu-Aimun and French texts. The first corpus, denonated as **kapesh** is based on literary texts by Innu author An Antane Kapesh.<sup>9</sup> These texts were manually aligned with the participation of Innu-Aimun translation students acting as single annotators.

The second corpus, denonated as **youth**, is based on a collection of poems written by Innu youth,<sup>10</sup> for which a sample representing 25% of the poems from the original text were manually aligned, with verses being treated as short sentences. Part of these the manual alignments for this sample of poems were performed by Innu-Aimun translation students, with every poem aligned by a single annotator. Another part was aligned in collaboration with the Innu-Aimun teaching staff from Kanatamat school. The goal was to allow immediate community benefits from the collaboration, even at the data collection and validation stage. These took the form of the creation of elementary-level exercises based on the poems in Nin Auass and their alignments formed in collaboration with the teaching staff.

<sup>&</sup>lt;sup>9</sup>Eukuan nin matshi-manitu innu-ishkueu (Kapesh, 2019) and Tanite nene etutamin nitassi? (Kapesh, 2020)

<sup>&</sup>lt;sup>8</sup>An example is the bilingual East Cree dictionary

<sup>&</sup>lt;sup>10</sup>The collection is titled *Nin Auass* (Bacon and Morali, 2021)

Based on this sample of manually aligned poems, we evaluated several automated alignment methods, found that, for the type and quantity of text in the **youth** corpus, the best alignment method was Gale and Church (1993). Hence the remaining 75 % was automatically aligned with this method.

Table 1: Studied corpora

Corpus	Domain	Nb sentences
kapesh	Novel/Essay	1280
youth	Youth poems	1907

Table 1 present the aforementioned bilingual corpora with their domain and their total number of sentences (before splitting).

# **3.2** Evaluating MT performance of aligned texts as one corpus

The main objective of this first MT study involving the Innu-Aimun language is to examine its feasibility with the available published bilingual texts (Innu-Aimun and French), mostly literary and poetic texts. Given the small amount of texts, our approach is to assess how these texts behave as a single unified corpus. Similarly to Joanis et al. (2020), we train/validate and test on the same resulting corpus, as the data is too limited to conduct a generalization study (i.e. testing on an entirely different corpus than the one used in training/validation).

We test two types of MT methods: Statistical Machine Translation (SMT), as proposed by Koehn et al. (2003) and Neural Machine Translation (NMT). While neural methods have achieved state of the art for many high-resourced languages, the statistical approach requires less data and testing it is particularly relevant in low-resource contexts such as ours. The model used for NMT is based on the standard Transformer architecture (Vaswani et al., 2017).

For the distribution between the training, validation and test sets, a portion equivalent to 85% of the corpus is reserved for training, then the rest (15%) is divided in two for validation and testing (7.5% each). This split scheme was chosen because of the low availability of bilingual data for training. Datasets were cleaned and then segmented using language-agnostic BPE (Sennrich et al., 2016) with a vocabulary size of 16K.

Metrics used for NMT and SMT results quantitative evaluations are sacrebleu (Post, 2018) and ChrF++ (Popović, 2015). In addition to quantative evaluation, the Innu-Aimun teaching staff from Kanatamat also contributed to a series of qualitative observation and analyses, for a small sample of translations generated by the SMT models.

### 4 Evaluations

### 4.1 Quantitative evaluations of MT Models

Tables 2 and 3 respectively present the results of the NMT and SMT evaluations for different combinations of individual corpora. The **kapesh** model is trained and evaluated solely on the **kapesh** corpus, the **youth** model is trained and evaluated solely and the **youth** corpus and the **kapesh+youth** model is trained and evaluated on a combination of both corpora.

From our NMT results, we can conclude that at this scale of data (i.e. less than four thousand sentences), NMT performance does not seem to be viable for any usage. The SMT results show that, for the Innu-Aimun/French pair at this data scale, the statistical approach to machine translation generally performs much better than the neural approach. This hypothesis of SMT superiority over NMT is statistically significant with p < 0.05 for both corpora and their combinated scores. Statistical significance was tested using Bootstrap Resampling, proposed for MT by Koehn (2004), and suggested by Dror et al. (2018). We used the latters' implementation of the algorithm specified by Berg-Kirkpatrick et al. (2012).

The overall best quantitative results are achieved by the **kapesh**-only SMT model. Its scores are much higher than the **youth**-only scores, but only slightly higher than that of the combined **kapesh+youth** scores.

### 4.2 Qualitative observations

Since the combined evaluation set include sentences/verses from both corpora, it is hard to discriminate impact of the combination on each of the corpora. The following subsection examines in detail and qualitatively a set of SMT-generated translations, from individual and combined models, to better understand this impact.

Those translations were collaboratively analyzed, involving computational linguistics researchers and Innu-Aimun teaching staff. This also allows to gain a more concrete and qualitative appreciation of the generated translations quality.

Tables 4, 5 and 6 present example translations

Table 2: NMT evaluations of Innu-Aimun/French corpora

Corpus	moe-fr (BLEU)	moe-fr (ChrF++)	fr-moe (BLEU)	fr-moe (ChrF++)
kapesh	0.28	13.4	0.62	13.2
youth	0.05	7.0	0.11	5.2
kapesh+youth	0.19	12.3	0.25	13.1

Table 3: SMT evaluations of Innu-Aimun/French corpora

Corpus	moe-fr (BLEU)	moe-fr (ChrF++)	fr-moe (BLEU)	fr-moe (ChrF++)
kapesh	4.41	22.7	4.16	33.6
youth	0.652	11.4	0.249	20.9
kapesh+youth	4.22	20.9	3.27	30.7

from the kapesh corpus, while tables 7, 8 and 9 present example from the youth corpus.

In table 4, the two translations seem quite far from the reference one when we compare the words. However, from the perspective of an Innu-Aimun locutor, the general meaning of the translated sentence from the kapesh+youth model can still be understood. In this translation, it is rather the syntax that is not good. The sentence could have been considered a better translation without the last five words (" à ce qu'il dit ").

In table 5, the translation of the kapesh model is more incomplete: it is partially composed of Innu-Aimun words. We can see in the translation of the kapesh+youth model that the French words " chose " and " je vais " (Kapesh, 2020) are also used in the reference translation. However, for the translation of the kapesh+youth model to be a good representation of the concepts in the source sentence, the concept of "other" should be removed and the concept "say" should be added.

In table 6, the translation of the kapesh model is syntactically incorrect in French. The translation of the combined kapesh+youth model, even if it does not represent a complete sentence, could be considered correct except for its plural (the source sentence is singular only).

In table 7, the youth model uses the Innu word "*kie*" rather than the word "et" (Bacon and Morali, 2021): the presence of the kapesh sentences during training allowed the model to perfect its translation, perhaps by taking advantage of the frequency of the word pair kie/et in this other corpus.

In table 8, we see that training on both youth and kapesh corpora allows to generate a more complete sentence than training on youth alone. In collaboration with the Innu-Aimun teaching staff, we can confirm that "moi je suis le loup" is a not only a correct translation, but it is also closer to the literal meaning that the reference translation, which is a more figurative one.

In table 9, although the kapesh+youth model translation gets none of the words of from the reference translation, it is a correct translation of the source sentence. This hints that the very low BLEU and ChrF++ scores of the youth model might not reflect the actual quality of the translations. Additionally, similarly to the previous example, it is actually closer to the literal meaning of the source sentence.

We can emphasize two key points from the above observations. First, we can see that poetry can contribute to an MT corpus: the youth corpus allows the kapesh corpus to get better, more precise translations for its literary sentences. Second, even though they were considered somewhat imprecise, with artistic quality (e.g. rhymes) often being prioritized over translation accuracy, the youth verses can be properly translated by an SMT model. This is true especially when the training also involves sentences from a literary corpus, as the translations can then become more complete and syntatically correct.

We can also conclude that poetry can be involved in—and be useful for—low-resource MT for an Indigenous language, especially if other quality textual sources are scarce. In that matter, quantitative scores might not be suitable for proper evaluation of the generated translations, since the latter can greatly differ from the reference ones even, if they are correct.

Overall, the results show that while interesting

Table 4: Qualitative comparison of kapesh SMT model, with and without youth: example #1

Source sentence (Innu-Aimun)	« Ne auass tapuetueu nenu etikut. » (Kapesh, 2020)
Translation (Reference)	« L'enfant se laisse convaincre. » (Kapesh, 2020)
Translation (kapesh SMT Model)	l'enfant est qu'il <i>tapuetueu</i> ce que lui dit.
Translation (kapesh+youth SMT Model)	l'enfant est d'accord avec ce que lui dit à ce qu'il.

Table 5: Qualitative comparison of kapesh SMT model, with and without youth: example #2

Source sentence (Innu-Aimun)	« Mak kutak tshekuan tshe uitamatan. » (Kapesh, 2020)
Translation (Reference)	« Je vais te dire encore une chose. » (Kapesh, 2020)
Translation (kapesh SMT Model)	la <i>uitamatan</i> et les autres.
Translation (kapesh+youth SMT Model)	et les autres <b>choses</b> que <b>je vais</b> .

Table 6: Qualitative comparison of kapesh SMT model, with and without youth: example #3

Source sentence (Innu-Aimun)	« Mak kutak tshekuan. » (Kapesh, 2020)
Translation (Reference)	« Et il y a autre chose. » (Kapesh, 2020)
Translation (kapesh SMT Model)	et les autres y.
Translation (kapesh+youth SMT Model)	et les autres choses.

Table 7: Qualitative comparison of youth SMT model, with and without kapesh: example #1

Source sentence (Innu-Aimun)	« kie nutin » (Bacon and Morali, 2021)
Translation (Reference)	« et le vent » (Bacon and Morali, 2021)
Translation (youth SMT Model)	kie <b>le vent</b>
Translation (kapesh+youth SMT Model)	et le vent

Table 8: Qualitative comparison of youth SMT model, with and without kapesh: example #2

Source sentence (Innu-Aimun)	« NIN MAIKAN » (Bacon and Morali, 2021)
Translation (Reference)	« MOI LE LOUP » (Bacon and Morali, 2021)
Translation (youth SMT Model)	je loup
Translation (kapesh+youth SMT Model)	moi je suis un loup

Table 9: Qualitative comparison of youth SMT model, with and without kapesh: example #3

Source sentence (Innu-Aimun)	« shashish shash »(Bacon and Morali, 2021)
Translation (Reference)	« les jours s'éternisent »(Bacon and Morali, 2021)
Translation (youth SMT Model)	la longtemps
Translation (kapesh+youth SMT Model)	il Y A LONGTEMPS déjà

results can be achieved with a few thousand sentences available, fully ready-to-use MT will require more than the current data for Innu-Aimun.

### 5 Conclusion

The current study has shown how difficult it can be to obtain good translations when first developing MT for an Indigenous language for which there are very few sentence pairs, and yet how the literary and poetic texts available, even if written in particular styles, can potentially contribute to building a general corpus for MT. We have also shown how at the present scale and types of available text for Innu-Aimun and French, SMT offers much better results than NMT. We carried this project collaboratively, involving both computational linguistics researchers and Innu-Aimun teaching staff, and with the participation of Indigenous and non-Indigenous speakers of Innu-aimun.

### Acknowledgments

The authors are grateful to all anonymous reviewers for their suggestions and fruitful recommendations. Also, we are very thankful to all Indigenous and non-Indigenous collaborators and partners. This work has been partially supported by the Northern Scientific Training Program (NSTP).

### References

- 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).
- J. Bacon and L. Morali, editors. 2021. *Nin Auass. Moi l'enfant: Poèmes de la jeunesse innue.* Mémoire d'encrier.
- 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.
- Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in NLP. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 995–1005, Jeju Island, Korea. Association for Computational Linguistics.
- Steven Bird. 2020. Decolonising Speech and Language Technology. In Proceedings of the 28th International Conference on Computational Linguistics, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- 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. 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*. PUQ.
- Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing.

In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

- William A. Gale and Kenneth W. Church. 1993. A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics*, 19(1):75–102. Place: Cambridge, MA Publisher: MIT Press.
- Oussama Hansal, Ngoc Tan Le, and Fatiha Sadat. 2022. Indigenous language revitalization and the dilemma of gender bias. In *Proceedings of the 4th Workshop* on Gender Bias in Natural Language Processing (GeBNLP), pages 244–254, Seattle, Washington. Association for Computational Linguistics.
- 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.
- 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.
- A.A Kapesh. 2019. Je suis une maudite Sauvagesse: Eukuan nin matshi-manitu innushkueu. Mémoire d'encrier.
- A.A. Kapesh. 2020. Qu'as-tu fait de mon pays? Tanite nene etutamin nitassi? Mémoire d'encrier.
- Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.
- Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, pages 127–133.
- Ngoc Tan Le, Oussama Hansal, and Fatiha Sadat. 2023. Challenges and issue of gender bias in under-represented languages: An empirical study on Inuktitut-English NMT. In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 89–97, Remote. Association for Computational Linguistics.
- Tan Ngoc Le and Fatiha Sadat. 2020. 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.

- Reem I. Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. 2024. Cultural alignment in large language models: An explanatory analysis based on Hofstede's cultural dimensions. *Preprint*, arXiv:2309.12342.
- Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onvefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2144-2160, Online. Association for Computational Linguistics.
- Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- 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.
- Myriam St-Gelais. 2022. *Une histoire de la littérature innue*. PUQ, Tshakapesh Institute, Montréal, Canada.
- Ngoc Tan Le, Antoine Cadotte, Mathieu Boivin, Fatiha Sadat, and Jimena Terraza. 2022. Deep learningbased morphological segmentation for indigenous languages: A study case on Innu-Aimun. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 146–151, Hybrid. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz

Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Binwei Yao, Ming Jiang, Diyi Yang, and Junjie Hu. 2024. Benchmarking llm-based machine translation on cultural awareness. *Preprint*, arXiv:2305.14328.