UNIVERSITÉ DU QUÉBEC À MONTRÉAL

LA TRADUCTION AUTOMATIQUE AU SERVICE DE LA REVITALISATION DE L'INNU-AIMUN : UNE APPROCHE COLLABORATIVE POUR DÉVELOPPER DES OUTILS D'ASSISTANCE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN INFORMATIQUE

PAR

ANTOINE CADOTTE

DÉCEMBRE 2023

REMERCIEMENTS

Je remercie d'abord ma directrice de recherche, Fatiha Sadat, sans qui je n'aurais pas travaillé sur un projet qui m'a tant tenu à coeur, qui m'a offert d'uniques opportunités qui me resteront, et qui par son audace m'a encouragé à me dépasser, à oser, à déployer ma recherche plus loin que je ne l'aurais imaginé.

Je remercie grandement Mathieu Boivin qui, autant avec ses bottines et qu'avec ses paroles et réflexions, m'a généreusement guidé sur le terrain, dans les communautés, et à travers ce sujet à la fois singulier, sensible et passionnant.

Je remercie Nathalie André et Donat Jean-Pierre, de l'école Kanatamat, pour leur inestimable collaboration, ainsi que la générosité avec laquelle ils m'ont offert leur confiance. Je remercie aussi Conrad André Kapesh, Daouda Cissé et tout le personnel de l'école pour leur aide précieuse et leur rôle dans un travail de terrain qui m'a fait énormément grandir.

Je remercie Anne-Christina Thernish, Rosalie Malleck, Sabine Mestokosho et Judith McKenzie, étudiantes en traduction d'innu-aimun du Cégep de Sept-Îles, maintenant diplômées, qui ont eu un apport indispensable à ce mémoire et à notre recherche. Je remercie chaleureusement Monique Durand, qui a cru en notre projet et sans qui cette superbe collaboration n'aurait pas été possible.

Je remercie aussi les Innus que j'ai pu croiser sur mon chemin et qui ont eu la générosité de m'accueillir et de partager avec moi leur culture, leur langue et leur histoire : Denis Vachon et Marie-Luce Jourdain, Dolorès André, et plusieurs autres.

Je remercie Ngoc Tan Le pour ses enseignements, ainsi que mes collègues de laboratoire, Ikram, Soheila, Habiba, Thierno et Ahmed, qui m'ont chacun aidé à leur façon.

Je remercie mes parents, Dominique Côté et François Cadotte, dont le propre parcours m'a inspiré et qui ont su m'écouter, me conseiller et m'encourager. Leur présence m'a, comme toujours, amené force et apaisement.

Je dédie ce mémoire à Ivona Yordanova, celle dont l'amour m'a porté au quotidien, qui m'a soutenu dès les débuts de cette aventure, et qui en cours de route est devenue ma fiancée, puis mon épouse.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

RÉSUMÉ

L'innu-aimun est une langue autochtone membre de la famille algonquienne, présente dans une douzaine de communautés au Québec et au Labrador. L'une des plus parlées au Canada, son nombre de locuteurs est néanmoins en diminution, dans un contexte de minorisation face au français principalement. Vu les défis de transmission de la langue, et le manque de services et de communications dans celle-ci, nous proposons le développement d'outils d'assistance à l'apprentissage et à la traduction. Si des outils informatisés de base existent pour l'innu-aimun, aucun outil n'utilise actuellement de techniques plus avancées de traitement automatique du langage (TAL). Pourtant, notamment en raison d'une riche culture littéraire innue, plusieurs textes bilingues sont disponibles. Nous postulons ainsi qu'il est peut-être possible de développer la traduction automatique pour la langue, dans le cadre du développement d'outils d'assistance.

Pour apporter une réelle utilité et être respectueux des communautés innues, un tel développement doit s'effectuer en collaboration avec elles. Notre première contribution est donc la conception, en collaboration avec une école primaire-secondaire innue, d'une méthode de recherche collaborative et participative en milieu éducatif autochtone. Cette approche que nous suivons se résume à l'arrimage des besoins de la recherche TAL et informatique à ceux de la communauté avec laquelle nous collaborons, ainsi que l'identification de bénéfices immédiats pour cette dernière. Dans notre cas, nous mettons cette approche à l'oeuvre en élaborant deux activités, collaborative et participative, pour la collecte et la validation de textes bilingues alignés en innu-aimun et français. L'une se fait en collaboration avec le personnel enseignant de l'école primaire-secondaire innue et l'autre avec la participation d'étudiants en traduction d'innu-aimun au Cégep. Dans les deux cas, nous adaptons notre approche au contexte et au milieu, pour que la recherche soit faite par et pour les membres de la communauté, tout en s'assurant que ceux-ci bénéficient immédiatement de cette collaboration.

La seconde contribution est une étude comparative des méthodes d'alignement automatique de phrases pour l'innu-aimun et le français. Se basant sur les textes bilingues que nous avons pu aligner de manière collaborative et participative avec des personnes maîtrisant la langue, nous évaluons la performance de différentes méthodes d'alignement existantes. Cette étude a pour but de déterminer quelles sont les méthodes les plus utiles pour aligner les textes bilingues qui ne sont pas alignés phrase à phrase, dans une optique de développement de la traduction automatique ou d'autres outils TAL multilingues. Elle nous permet aussi de tester les limites des méthodes d'alignement et de constituer une étude de cas pour d'autres langues qui seraient dans un contexte de faible quantité de données comme celui qu'est le nôtre.

La troisième contribution est une étude de faisabilité pour la traduction automatique d'innu-aimun. Nous cherchons à savoir quelles performances de traduction sont atteignables avec les données bilingues actuellement disponibles pour la langue en utilisant les meilleures méthodes à ce jour. Ce faisant, nous obtenons les tous premiers résultats connus de traduction automatique neuronale et statistique pour l'innu-aimun. Nous procédons par ailleurs à des comparaisons des deux méthodes testées et en analysons les résultats au regard de l'apport des différents textes utilisés, à savoir des textes littéraires et poétiques innus. Dans une comparaison de significativité statistique, nous déterminons que la meilleure des deux méthodes, avec la taille et le types de textes que nous utilisons, est la traduction statistique. Grâce à ces tests, nous pouvons aussi identifier les avenues les plus prometteuses pour les améliorations des performances de traduction.

Globalement, notre étude nous permet de conclure qu'un modèle de traduction automatique assez performant pour une utilisation immédiate n'est pas actuellement atteignable. Toutefois, la qualité des résultats

obtenus avec aussi peu de texte et les améliorations que nous parvenons à tirer de nos expérimentations nous permettent de croire qu'un tel développement serait possible à moyen terme, et ce dans le respect des meilleures pratiques et des protocoles de recherche autochtone. À court terme, nous croyons qu'il serait possible de mettre à contribution les modèles que nous avons créés dans des systèmes de recherche interlingue où les usagers n'utiliseraient la traduction automatique qu'indirectement. Nous concluons sur des perspectives pour les prochaines étapes dans le développement collaboratif d'outils d'assistance à l'apprentissage et à la traduction de l'innu-aimun, un projet plus large dans lequel s'inscrit notre mémoire.

Mots-clés : traduction automatique, traitement automatique du langage, linguistique computationelle, alignement de phrases, apprentissage automatique, intelligence artificielle, corpus, langues autochtones, innu-aimun, collaboration

INTRODUCTION

En innu-aimun, l'expression « innu-aimun » signifie littéralement « la langue innue » [1]. Dans cette même langue, « innu » signifie « être humain » [2]. L'innu-aimun est la langue des Innus, un peuple autochtone aujourd'hui présent surtout dans la région de la Côte-Nord du Québec, ainsi qu'au Labrador. On a estimé en 2021 le nombre de locuteurs d'innu-aimun à 10 745 [3], un nombre qui est en diminution par rapport à celui comptabilisé en 2016 (11 440) [4]. La langue est en situation de minorisation face aux langues officielles du Canada : surtout le français au Québec, où l'on trouve la plus grande partie des locuteurs d'innu-aimun, et face à l'anglais au Labrador. Au Québec, on a documenté des difficultés de transmission de la langue, tout comme un manque de services dans cette dernière. Face à cette situation, et pour contribuer à la revitalisation de l'innu-aimun, nous proposons le développement d'outils d'assistance à l'apprentissage et à la traduction.

Du point de vue du traitement automatique du langage (TAL) et de la linguistique computationelle, l'innu-aimun serait considéré comme une langue dite peu dotée, c'est-à-dire une langue pour laquelle il existe une faible quantité de données textuelles existantes. Cette caractérisation des langues selon la quantité de données, une caractérisation académique qui a évidemment ses limites, vient surtout du fait que la plupart des techniques TAL les plus performantes se basent sur l'apprentissage automatique, qui nécessite d'importantes quantités de données. Langue polysynthétique, membre de la famille des langues algonquiennes, l'innu-aimun compte à la fois des caractéristiques linguistiques qui rendent plus difficile son traitement automatique (par exemple, une morphologie riche), et d'autres qui pourraient le faciliter (par exemple, une proximité avec d'autres langues autochtones au Canada). Elle peut aussi compter sur une littérature grandissante, qui tire en partie son origine d'une riche tradition orale. Face à ces atouts et ces défis, il nous est permis de croire qu'il est possible, à moyen ou long terme, de développer la traduction automatique pour l'innu-aimun, dans l'objectif plus global du développement des outils linguistiques d'assistance pour revitaliser la langue.

Le développement de technologies linguistiques ici proposé pour l'innu-aimun cadre d'ailleurs dans une mouvance et un intérêt grandissant pour la revitalisation, assistée par les technologies, des langues autochtones. Alors que l'UNESCO a déclaré la période 2022-2032 comme *Décennie internationale des langues au-*

[1] Définition du mot « innu-aimun » tirée du dictionnaire innu en ligne (Ambroise *et al.*, 2023)

[2] Définition du mot « innu » tirée du dictionnaire innu en ligne. (Ambroise *et al.*, 2023)

[3] Statistiques Canada: Profil du recensement, Recensement de 2021

[4] Statistiques Canada: Profil du recensement, Recensement de 2016

tochtones [5], la conférence ACL (*Association of Computational Linguistics*) a fait de sa thématique en 2022 la diversité des langues (*Language Diversity : from Low-Resource to Endangered Languages*) [6], en y incluant les langues peu dotées, mais aussi les langues en danger, desquelles font typiquement partie les langues autochtones. À cela s'ajoute la conférence LREC (*International Conference on Language Resources and Evaluation*), qui a fait des langues peu dotées et en danger l'un de ses sujets majeurs en 2020 [7]. Également en 2020, le groupe de travail pour un protocole autochtone en intelligence artificielle (*Indigenous Protocol and Artificial Intelligence Working Group*) [8] a proposé des directives pour s'assurer du développement d'une intelligence artificielle respectueuse des enjeux et de l'identité autochtone. En proposant le développement collaboratif d'outils d'assistance technologiques pour l'enseignement et la traduction de l'innu-aimun, nous souhaitons nous inscrire dans cet esprit : adapter le traitement automatique du langage à la réalité autochtone, et ce dans le respect des meilleures pratiques et des protocoles de la recherche autochtone.

Le présent mémoire rapporte les résultats de ce qui peut être vu comme la première étape de ce développement : celle de la première collecte et validation de données, menant à la première étude de faisabilité. L'étude tente, en substance, de répondre à la question suivante. Est-il possible, avec les textes disponibles et les techniques existantes, de développer des outils d'aide à l'apprentissage et à la traduction de l'innu-aimun basés sur le TAL ? Ou, plus précisémment, quelle est l'efficacité des techniques TAL existantes sur les textes aujourd'hui disponibles pour l'innu-aimun et comment pourrait-on améliorer cette efficacité ? Pour l'innu-aimun et, de manière plus large la communauté innue, cette étude peut renseigner sur ce qu'il est possible de faire dès lors pour l'innu-aimun avec l'aide du TAL et ce que devraient être les prochaines étapes pour pousser un tel développement collaboratif plus loin. Pour les domaines du TAL et de la linguistique computationelle, cette étude sur l'innu-aimun constitue une étude de cas qui peut renseigner plus largement sur les questions suivantes. Quelles sont les limites des méthodes TAL constituant l'état de l'art, lorsqu'on a affaire à une langue autochtone, qui est de nature polysynthétique, et pour laquelle peu de textes existent ? Quelles sont les perspectives les plus prometteuses pour en améliorer les résultats ? Et comment peut-on travailler à l'étude, à l'application de techniques TAL en contexte autochtone ?

Le chapitre 1 définira plus en profondeur la problématique qui nous occupe. Dans ce chapitre, nous pro-

[5] UNESCO - Décennie internationale des langues autochtones (2022-2032)

[6] ACL 2022 Theme Track: "Language Diversity: from Low-Resource to Endangered Languages"

[7] Hot Topics for LREC 2020

[8] `https://www.indigenous-ai.net/`

céderons à une description linguistique de l'innu-aimun, et examinerons sa situation et les enjeux sociolinguistiques qui s'y rattachent. Ces enjeux sont ce qui motive l'objectif sur le long terme de notre recherche, c'est-à-dire le développement d'outils d'aide à l'apprentissage et à la traduction l'innu-aimun. Ils motivent aussi en bonne partie l'approche que nous proposons pour y parvenir, une approche qui repose sur la collaboration avec la communauté. Nous élaborerons notre proposition de recherche dans ce même chapitre, en expliquant pourquoi nous choisissons une étude de faisabilité sur la traduction automatique comme première étape dans le développement d'outils. Nous procéderons ensuite à la définition des concepts et connaissances de base en traduction automatique. Ceci inclut la notion d'alignement de phrases, puisque nous devrons aligner des textes bilingues qui ne le sont pas encore. Nous aborderons aussi la notion de langues ou paires de langues peu dotées, c'est-à-dire celles pour lesquelles peu de données sont à notre disposition, comme c'est le cas pour l'innu-aimun. L'alignement et la traduction automatique de langues peu dotées sont les sujets informatiques centraux de ce mémoire.

Le chapitre 2 dresse l'état de l'art des principaux sujets touchés par notre recherche. Nous nous intéressons d'abord aux ressources et technologies de la langues existantes pour l'innu-aimun, puis à celles qui existent pour des langues reliées, ainsi que d'autres langues autochtones du Canada. Nous examinerons ensuite les récentes approches et avancées en traduction automatique pour les paires de langues peu dotées. Nous ferons enfin un survol des principales méthodes d'alignement automatique des phrases.

Le chapitre 3 présente une approche collaborative, co-conçue avec l'école primaire-secondaire Kanatamat, pour le développement d'outils d'assistance et présente sa mise en oeuvre pour de premières collectes et validations de données dans le cadre de notre projet. Cette approche en est une qui vise à arrimer les besoins de la recherche en traitement automatique du langage à ceux des milieux avec lesquels nous travaillons. Ces milieux sont deux milieux scolaires, soit l'école primaire-secondaire Kanatamat elle-même, à Matimekush - Lac John, ainsi que les étudiants de traduction et interprétation d'innu-aimun du Cégep de Sept-Îles. Le chapitre décrit le processus d'élaboration de l'entente de collaboration avec l'école, et la mise en oeuvre de deux activités, l'une en collaboration avec le personnel de l'école Kanatamat et l'autre avec la participation des étudiants en traduction d'innu-aimun. Ces activités visent, tel que le propose notre approche, à permettre à la fois des impacts bénéfiques immédiats pour les milieux en question et une certaine collecte et validation de données dans le cadre du projet. Ce dernier est par ailleurs conçu pour être réalisé par et pour les membres de la communauté, ce qui est l'un des objectifs de notre approche. Nous discutons enfin des succès et limites de l'approche, et de sa potentielle application dans d'autres cas que celui du présent mémoire.

Le chapitre 4 présente une étude comparative de différentes méthodes d'alignement. Cette étape vise à identifier quelle est la meilleure méthode pour aligner les textes bilingues qui ne le sont pas déjà, surtout considérant notre contexte, où une faible quantité de données est disponible. La performance des différentes méthodes sera évaluée face aux alignements de référence construits au cours des activités collaborative et participative décrites au chapitre 3. Avoir de tels alignements de référence à notre disposition est une rare occasion qui permet d'évaluer plus en détail la performance de méthodes d'alignement déjà éprouvées, dans le contexte en quelque sorte exigeant d'une langue autochtone polysynthétique en manque de données. Ce chapitre peut donc être vu comme une étude de cas qui pourra être utile à d'autres contextes similaires. Cela constituera aussi la première étude de ce type pour l'innu-aimun.

Le chapitre 5 présente notre étude de faisabilité pour la traduction automatique de l'innu-aimun. Nous y présenterons les tous premiers résultats pour cette langue, qui serviront de référence à de futures expérimentations et améliorations. Nous présenterons différents essais de traduction automatique pour tenter d'améliorer ces résultats et examiner les pistes les plus prometteuses dans le développement collaboratif d'outils futurs. Notamment, nous étudierons l'impact de l'apport de différents types de textes, incluant des textes littéraires et poétiques innus. Nous expérimenterons avec la traduction neuronale et la traduction statistique, analyserons et comparerons entre eux les résultats de ces méthodes. Enfin, nous discuterons des différents constats que l'ensemble de ces essais nous permettent de dresser, surtout au regard de la question de la faisabilité et des avenues à prioriser dans le développement d'outils d'assistance basés sur la traduction automatique.

CHAPITRE 1
DÉFINITION DE LA PROBLÉMATIQUE

Le présent chapitre a pour but d'introduire la problématique qui nous occupe, soit la situation de l'innu-aimun en tant que langue, et ce que ce contexte nous amène à proposer comme approche de recherche. Ces deux aspects de la problématique sont décrits respectivement aux sections 1.1 et 1.2, alors que la section 1.3 apporte des définitions conceptuelles pour les deux principaux sujets de traitement automatique du langage qui sont abordés dans le cadre de ce mémoire, soit l'alignement automatique de phrases et la traduction automatique.

1.1 Contexte : description de l'innu-aimun et situation sociolinguistique

La section qui suit présente une description linguistique de l'innu-aimun, suivie d'un portrait de sa situation actuelle et des enjeux sociolinguistiques qui s'y rapportent.

1.1.1 Description de la langue

L'innu-aimun, autrefois dénommée langue montagnaise (code ISO moe [1], Glottolog mont1268 [2]) est une langue algonquienne faisant partie du continuum dialectal cri-innu-naskapi (Drapeau, 2014). La carte à la figure 1.1 montre l'emplacement de l'innu-aimun (illustré ici en tant que Montagnais du Centre et Montagnais de l'Ouest) au sein de ce continuum, qui s'entend au-delà du Québec et du Labrador, jusque dans les provinces de l'ouest.

Parlée principalement dans onze communautés innues au Québec et au Labrador (Baraby *et al.*, 2017), elle comprend plusieurs dialectes. Ceux-ci sont typiquement regroupés sous le dialecte de l'ouest, le dialecte du centre et le dialecte de l'est (qui comprend les sous-dialectes du Labrador et le sous-dialecte de Mamit, au Québec) [3]. Au Québec, les dialectes sont aussi parfois regroupés sous le « dialecte de l'Ouest » (qui comprend les dialectes dits « en l » et ceux de Uashat mak Mani-utenam et Matimekush) et le « dialecte de Mamit » (pour les communautés de la Basse Côte-Nord) (Drapeau, 2014). La carte à la figure 1.2 montre les

[1] ISO 639-3 - moe

[2] Glottolog - mont1268

[3] innu-aimun.ca - À propos de l'innu

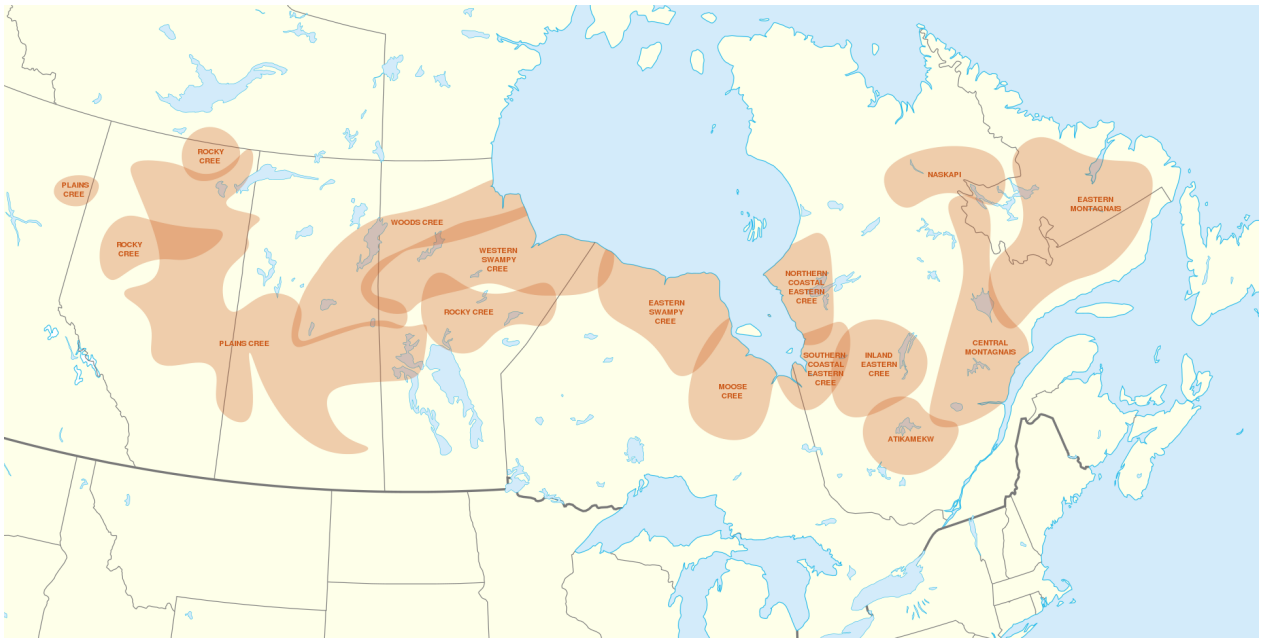

FIGURE 1.1 – Continuum linguistique cri. Image Wikimedia : *Cree Map* (`https://en.wikipedia.org/wiki/Cree#/media/File:Cree_map.svg`), par l'auteur Noahedits (`https://commons.wikimedia.org/wiki/User:Noahedits`) sous licence Creative Commons 4.0 (`https://creativecommons.org/licenses/by-sa/4.0/`).

différentes communautés innues au Québec et au Labrador.

Tel qu'expliqué par Drapeau (2014) dans sa grammaire de la langue innue, comme beaucoup de langues autochtones au Canada et en Amérique du Nord, l'innu-aimun est une langue polysynthétique : ses mots sont formés par combinaison de plusieurs segments de mots (ou morphèmes), résultant en de longs mots qui, en comparaison à des langues non-polysynthétiques comme le français ou l'anglais, peuvent être l'équivalent de phrases complètes. Ces longs mots sont le plus souvent construits sur la base d'un verbe (Drapeau, 2014). Les tableaux 1.1 et 1.2 montrent des exemples de traduction de l'innu-aimun vers le français, tous deux tirés du dictionnaire trilingue innu Ambroise *et al.* (2023), pour illustrer certains aspects caractéristiques de la langue. Le premier exemple montre bien la nature polysynthétique de l'innu-aimun et sa riche morphologie, alors qu'un mot innu doit être traduit par une phrase complète en français. Ce mot innu, *neumitashumitannuetipapekaikaneshu*, est constitué par agglutinement de morphèmes, à commencer par *neumitashumitannu* qui signifie quatre cents. Le second exemple montre que certains mots innus, même courts, peuvent représenter un certain savoir ancestral innu, qui n'a pas son pareil dans la langue française. Ici, le mot *ashteueu* ne peut être traduit que par une longue description en français, non pas parce qu'il est composé de nombreux

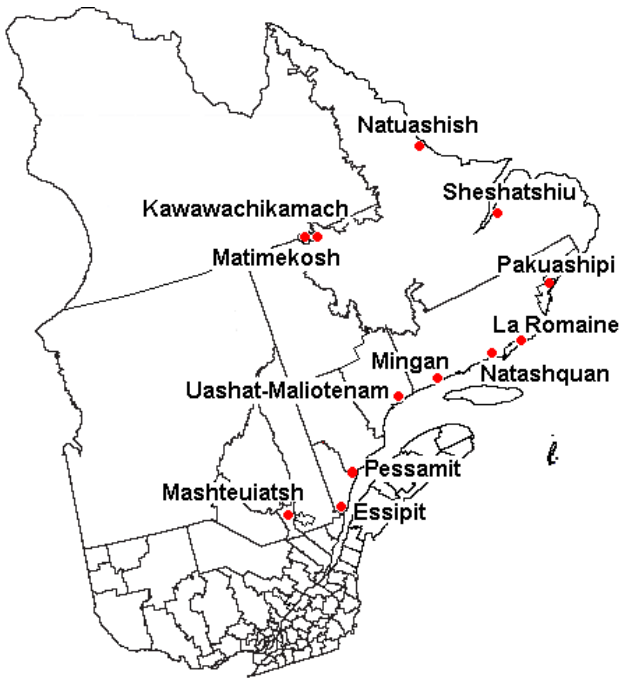

FIGURE 1.2 – Communautés innues du Québec et du Labrador. Image Wikimedia du domaine public : *Villages innus du Québec et du Labrador* (`https://fr.wikipedia.org/wiki/Innus#/media/Fichier:Innus.png`)

morphèmes, mais bien parce qu'il représente une notion très précise et très propre au contexte culturel et au territoire innu, qui n'a pas de vocabulaire propre en français. À eux deux, ces exemples montrent bien les caractéristiques singulières de l'innu-aimun, qui posent un improtant défi pour le développement des techniques de traitement automatique du langage (TAL) pour la langue.

TABLE 1.1 – Illustration de la morphologie : exemple #1, tiré du dictionnaire trilingue innu (Ambroise *et al.*, 2023). La notation (A) indique que le sujet est de type *animé*, ce qui modifie sa conjugaison.

Mot en innu-aimun	« neumitashumitannuetipapekaikaneshu » (Ambroise *et al.*, 2023)
Définition en français	« il ou qqch (A) pèse quatre cents livres » (Ambroise *et al.*, 2023)

L'innu-aimun est à l'origine une langue orale. Ce sont les missionnaires français qui l'on pour la première fois transposée à l'écrit, notamment pour écrire des recueils de prières Mollen (2006). Par exemple, au 18e siècle, le Père Jean-Baptise de la Brosse a écrit abécédaire, dictionnaire et grammaire pour la langue dite

TABLE 1.2 – Illustration de la morphologie : exemple #2, tiré du dictionnaire trilingue innu (Ambroise *et al.*, 2023).

Mot en innu-aimun	« ashteueu » (Ambroise *et al.*, 2023)
Définition en français	« il passe sans voir qqn qu'il allait rejoindre en canot ou en chaloupe » (Ambroise *et al.*, 2023)

« montagnaise » à l'époque et en a fait imprimer de nombreuses copies dans un but d'enseignement religieux (Hébert, 1988).

En 1989, un processus d'uniformisation de l'orthographe innue a été mis en oeuvre, notamment afin d'en faciliter l'enseignement (Mollen, 2006). Un premier dictionnaire utilisant l'orthographe uniformisée a été publié en 1991 : le *Dictionnaire montagnais-français* (Drapeau, 1991). Le processus d'uniformisation a été fait de manière participative, à travers une négociation entre représentants des différents dialectes, et avait pour but d'uniformiser uniquement la langue écrite, en laissant intacts les mots et prononciations propres à chaque communauté (Mollen, 2006). Depuis, le dictionnaire a été mis à jour et republié par l'institut Tshakapesh (Mailhot *et al.*, 2012), pour finalement être transformé en dictionnaire trilingue et multi-dialectal en ligne (Ambroise *et al.*, 2023). Ce dernier continue d'être mis à jour et comprend plus de 28 000 mots innus.

Prenant ses racines notamment dans une littérature traditionelle orale, il existe une littérature innue écrite. Celle-ci comporte plusieurs styles tels l'essai et la poésie et est souvent disponible dans des éditions bilingues en innu-aimun et en français (St-Gelais, 2022). De manière notable, on compte l'autrice innue An Antane Kapesh, qui a publié dans les années 1970 deux livres en innu-aimun et en français, soit *Je suis une maudite Sauvagesse / Eukuan nin matshimanitu innu-iskueu* et *Qu'as-tu fait de mon pays ? / Tanite nene etutamin nitassi ?*. Ces deux oeuvres, qui ont ultérieurement été transcrites en orthographe uniformisée, continuent d'être rééditées (Kapesh, 2019, 2020). Certaines oeuvres, en plus d'être traduites en français, ont été aussi traduites et publiées en anglais. C'est le cas des deux oeuvres susmentionnées, ainsi que du recueil de poésie *Bâtons à messages / Tshissinnuatshitakana* de la poétesse innue Joséphine Bacon.

1.1.2 Situation de la langue et enjeux sociolinguistiques

En 2021, selon Statistiques Canada, 10 745 personnes pouvaient parler innu-aimun [4]. En 2016, ce chiffre était de 11 440 [5], représentant un déclin de 6,1%. Au Québec, l'innu-aimun et le naskapi, mis ensemble, ont subi la plus grande diminution parmi les langues autochtones de la province en nombre absolu (perte de 1 095 locuteurs) depuis 2016 [6]. L'UNESCO considère officiellement que la langue est en danger [7], alors que des chercheurs l'ont décrite comme « vivante, mais encore fragile » (Baraby *et al.*, 2017).

Au-delà du déclin statistique, certains constatent sur le terrain une situation difficile pour la transmission de l'innu-aimun. Dans la communauté innue de Pessamit, une enquête de Drapeau et Lambert-Brétière (2013) montrait déjà en 1994 que, puisque la transmission de la langue se faisait en contexte de fort bilinguisme, il y avait une érosion des connaissances lexicale en innu-aimun. Plus récemment, selon une enseignante d'innu-aimun de Uashat mak Mani-utenam, les élèves innus de cette communauté parlent surtout français en classe et ils maîtrisent rarement la langue (Cadotte *et al.*, 2022).

Historiquement les Innus savaient lire et écrire leur langue. Toutefois, suite à l'imposition de l'éducation dans la langue majoritaire, voire carrément l'interdiction de l'innu-aimun dans le système d'écoles résidentielles, cette capacité a largement reculé en quelques décennies (Baraby, 2000). Dans son enquête dans la région de Sept-Îles, Leroux (2014) rapporte que les Innus sont nombreux à mentionner que les récentes tentatives d'assimilation ont eu un impact très négatif sur la transmission de la langue, tout comme de la culture innue en général et sur « l'image de soi des Innus ». Leroux (2014) rapportait aussi par ailleurs, au moment de l'enquête, d'importantes tensions entre les allochtones, majoitaires à Sept-Île, et les Innus de la région (qui comprend la communauté de Uashat mak Mani-utenam). Selon les personnes interviewées, ces tensions étaient aussi une conséquence du lourd héritage historique et des pratiques assimilatrices. Pour rappel, le système d'écoles résidentielles au Canada et ses conséquences dévastatrices sur les peuples autochtones partout au Canada ont été largement documentés par la Commission de vérité et réconciliation du Canada [8].

[4] Statistiques Canada: Profil du recensement, Recensement de 2021

[5] Statistiques Canada: Profil du recensement, Recensement de 2016

[6] Statistiques Canada: Les langues autochtones au Canada

[7] UNESCO Atlas mondial des langues - Montagnais (en Anglais)

[8] Commission de vérité et réconciliation du Canada

Au Québec, la Commission Viens [9] a documenté le manque de services pour toutes les langues autochtones dans la province, incluant l'innu-aimun. Notamment, le manque de documents disponibles en innu-aimun a été indiqué, ainsi que le manque de traducteurs connaissant l'innu-aimun dans le cadre des services de santé, ce qui permettrait un consentement plus libre et éclairé avant des traitements. Entre autres pour pallier au manque de traducteurs d'innu-aimun, un nouveau programme de traduction et d'interprétation d'innu-aimun a récemment été créé [10] au Cégep de Sept-Îles [11].

De manière générale, la langue joue un rôle important pour la sécurisation culturelle dans le cadre des soins de santé pour les autochtones ; cela a été noté par exemple lors d'une étude portant sur les soins infirmiers en Inuktitut au Groenland et au Nunavut (Møller, 2016). Il y a quelques années, le rapport du Coroner suite à la mort tragique de Joyce Echaquan dans un hôpital à Joliette a conclu à l'importance de la communication et de la sécurisation culturelle pour éviter ce type de drame [12].

1.2 Approche proposée

Face aux défis de revitalisation connus par l'innu-aimun, nous proposons d'entamer le développement collaboratif d'outils d'assistance à la traduction, basés sur des méthodes de traduction automatique et utilisant les textes bilingues disponibles en innu-aimun en français. De tels outils n'existent pas encore pour la langue (Cadotte *et al.*, 2022). L'objectif serait d'assister les traducteurs professionnels dans leur travail, mais ces outils pourraient aussi servir aux personnes apprenant l'innu-aimun ou souhaitant être assistées dans son usage, qu'elles soient jeunes, adultes, autochtones ou allochtones.

Les sous-sections qui suivent établissent les trois principaux aspects de l'approche proposée, soit : l'approche d'assistance, l'étude de la faisabilité, l'innu-aimun comme cas d'étude en TAL. L'approche de collaboration avec la communauté, elle, est définie au chapitre 3.

[9] Rapport final - Commission d'enquête sur les relations entre les Autochtones et certains services publics : écoute, réconciliation et progrès

[10] Le Cégep de Sept-Îles lance un programme de traduction en langue innue

[11] Cégep de Sept-Îles - Programme AEC Traducteur / Interprète en langue innue

[12] Rapport d'enquête, Loi sur la recherche des causes et des circonstances des décès pour la protection de la vie humaine, concernant le décès de Joyce Echaquan

1.2.1 Assistance à la traduction

L'approche d'assistance se justifie par le fait que le but n'est pas de remplacer les traducteurs, mais bien de les aider dans leur tâche afin qu'ils puissent mieux subvenir aux importants besoins. De toute manière, étant donné la grande quantité de données nécessaires pour entraîner des modèles de traduction automatique atteignant des performances proches de l'humain (voir section 1.3) et considérant que le nombre de phrases disponibles en innu-aimun ne se compte qu'en milliers (Cadotte *et al.*, 2022), une automatisation complète n'est probablement pas atteignable dans le court ou moyen terme.

Pour apporter une assistance aux traducteurs d'innu-aimun et assurer une réelle utilité à ces derniers, il est primordial que les outils développés répondent d'abord aux besoins des traducteurs d'innu-aimun tels qu'exprimés par eux-mêmes et impliquent ces derniers à toutes les étapes du développement. Nous avons eu l'occasion de faire un tel postulat, en présentant une vision développement conjointement avec une traductrice d'innu-aimun de la communauté de Uashat mak Mani-utenam, dans le cadre de la conférence *Human-informed Translation and Interpreting Technology* (Cadotte *et al.*, 2023). Cette vision fait notamment part du manque de documents traduits en innu-aimun dans les communautés et du fait que traduire davantage de documents permettrait d'encourager l'usage et l'apprentissage de la langue. Elle postule qu'aider les traducteurs innus devrait être considéré comme une des priorités dans le développement d'outils technologiques pour la langue.

1.2.2 Étude de la faisabilité

Puisqu'il n'existe pas encore de modèle de traduction automatique pour l'innu-aimun, la première étape vers un tel développement est d'en étudier la faisabilité. Cela veut dire répondre à la question suivante : avec les données accessibles et les méthodes existantes, peut-on arriver à développer un premier système de traduction automatique pour l'innu-aimun ? Et quelle en seraient l'efficacité et les limitations ? Cette question sert moins à savoir si un tel système est possible que de savoir ce qu'il est possible de faire avec les données aujourd'hui disponibles et d'identifier les avenues d'amélioration pour les prochaines étapes. En outre, une telle approche permet d'établir de premiers résultats de référence qui pourront ensuite servir de point de comparaison pour les contributions subséquentes.

Deux étapes principales sont nécessaires pour développer de premiers modèles de traduction automatique. La première étape est la création de corpus alignés à partir des textes bilingues existants. Cette étape est

nécessaire puisque les phrases des textes bilingues publiés ne sont pas alignés phrase à phrase, comme l'exige l'entraînement de modèles de traduction automatique. En l'absence de connaissance sur la langue ou de personnes-ressources pouvant effectuer l'alignement manuellement, les phrases peuvent être alignées de manière automatique, en utilisant des techniques existantes (voir prochaine section). La deuxième étape est, ayant en main un jeu des paires de phrases en innu-aimun et en français, l'entraînement et l'évaluation de modèles de traduction automatique en utilisant des méthodes existantes.

1.2.3 L'innu-aimun comme cas d'étude en TAL

Au-delà de l'étude de faisabilité, il ressort du présent projet une opportunité d'étudier l'application de méthodes existantes en traitement automatique du langage naturel à une langue à laquelle elles n'ont pas encore été appliquées. Cette opportunité constitue donc une étude de cas qui, au-delà de renseigner sur les possibilité qu'offrent ces méthodes pour la langue étudiée, renseigne sur les limites de ces méthodes elles-mêmes. Comment se comportent ces méthodes lorsqu'elles sont appliquées sur une langue qui est autochtone et polysynthétique, et pour laquelle il y a très peu de données disponibles ? D'une part, ces caractéristique font de l'innu-aimun un cas d'étude plus exigeant que beaucoup d'autres langues auxquelles ces méthodes ont déjà été appliquées (des langues en contexte majoritaire, avec de plus grandes quantités de textes, faisant partie d'une famille avec d'autres langues déjà étudiées en TAL, etc.). D'autre part, l'innu-aimun n'est pas la seule langue qui correspond à ces caractéristiques. En ce sens, les conclusions du présent mémoire pourront possiblement servir de référence à d'autres langues.

1.3 Traduction automatique : définitions et connaissances de base

Afin d'appuyer la suite de ce mémoire, la présente section apporte des définitions et connaissances de base sur la traduction automatique et les sujets connexes en TAL. Ces définitions et connaissances, lorsqu'elle ne sont pas directement citées à la source, sont tirées de ce qui est considéré comme la base générale du domaine, largement décrite par de nombreux manuels et notes de cours. Nous référons le lecteur aux ouvrages suivants pour des définitions et explications plus étendues : *Speech and Language Processing* (Jurafsky et Martin, 2009), *Foundations of Statistical Natural Language Processing* (Manning et Schutze, 1999) et *Introduction to Natural Language Processing* (Eisenstein, 2019).

1.3.1 Définition et fonctionnement de la traduction automatique

Dit le plus simplement, la traduction automatique (ou *Machine Translation*, en anglais) est « [...] l'utilisation d'ordinateurs pour traduire d'une langue à une autre. »(Jurafsky et Martin, 2009)

La traduction automatique aborde typiquement la traduction phrase par phrase : celle-ci traduit chaque mot en considérant le contexte formé par les autres mots de la phrase, mais ne considérera pas le contexte du document entier.

1.3.1.1 Modèles neuronaux

Basée sur des modèles statistiques à ses débuts, de grandes améliorations de performance ont été possibles grâce aux réseaux de neurones profonds LeCun *et al.* (2015), qui sont à la base de tous les outils de traduction automatique contemporains les plus performants.

Ayant au départ des architectures neuronales de types RNN (*Recurrent Neural Network* ou réseau de neurones récurrents), puis de type LSTM (*Long Short-Term Memory* ou réseau récurrent à mémoire court et long terme), les modèles obtenant les meilleures performances sont ceux qui ont adopté une architecture de type Transformeur (Vaswani *et al.*, 2017), basée entre autres sur un processus nommé mécanisme d'attention (Bahdanau *et al.*, 2016). Cette dernière architecture de base est celle qui est restée dominante depuis, bien qu'ayant fait l'objet de différentes variations et avancées (voir la section **??** au chapitre prochain pour un état de l'art plus complet).

L'architecture générale des modèles de traduction automatique neuronale (TAN) est l'encodeur-décodeur. Cette architecture vise à apprendre les paramètres d'un modèle qui doit transformer une phrase en langue source à son entrée en une phrase en langue cible à sa sortie.

1.3.1.2 Plongements

Pour leur entraînement, les modèles de TAN se réfèrent non pas directement au mots eux-mêmes des phrases en entrée et en sortie, mais sur des représentations sémantiques de ces mots, appelées plongements, qu'ils apprennent à l'entraînement. Les plongements de mots sont des vecteurs sensés représenter sémantiquement les mots en se basant sur leur contexte habituel. Plus spécifiquement, tel qu'expliqué par Jurafsky et Martin

(2009), « Les mots qui apparaissent dans des contextes similaires ont tendance à avoir des significations similaires. Ce lien entre la similarité dans la façon dont les mots sont distribués et la similarité dans ce qu'ils signifient est appelé l'hypothèse distributionnelle. [...] la sémantique vectorielle [...] instancie cette hypothèse linguistique en apprenant des représentations du sens des mots, appelées plongements, directement à partir de leurs distributions dans les textes. »

1.3.1.3 Corpus et alignement

Pour entraîner un modèle de TAN sur une paire de langues, d'importants jeux de données constitués de paires de phrases parallèles (des phrases qui sont des traductions l'une de l'autre). Ces collections de paires de phrases sont appelées corpus parallèles. L'un des corpus parallèles les plus importants et les plus connus est le corpus Europarl (Koehn, 2005), qui dans sa deuxième version réunit pour 11 langues de l'Union Européenne environ 1 million de paires de phrases, tirées des compte-rendus du parlement européen. Un autre corpus parallèle notable est celui de l'ONU (Ziemski *et al.*, 2016), qui procède à un exercice similaire pour les six langues officielles de l'ONU.

Suivant la disponibilité de grandes quantités de textes traduits, une étape fondamentale afin de permettre l'entraînement des modèles de TAN est l'alignement des phrases au sein des textes. Cet alignement est nécessaire pour fournir des paires de phrases équivalentes, alors qu'elles n'ont pas nécessairement une correspondance 1-pour-1 dans les textes d'origine.

Pour les corpus dits parallèles, comme l'Europarl, cet alignement est déjà effectué, mais ce n'est pas le cas lorsqu'on construit de nouveaux corpus (soit pour augmenter la quantité de données ou parce qu'il n'en existe pas pour la paire de langue à laquelle on s'intéresse). Pour ce faire, des méthodes d'alignement automatique sont typiquement utilisées. Ces méthodes considèrent habituellement qu'une phrase ne peut être scindée en parties, mais qu'elle peut être fusionnée à des phrases voisines pour mieux correspondre à une phrase équivalente dans l'autre langue. Un état de l'art des méthodes d'alignement est fourni à la section **??** du prochain chapitre.

1.3.1.4 Évaluation

Pour évaluer la performance des modèles de traduction automatique, l'approche la plus courante est une évaluation automatisée et quantitative, basée sur une comparaison à des traductions de référence. Pour un jeu

de phrases données, on mesure la similarité entre les phrases produites par le modèle et celles considérées commes les traduction de références, provenenant d'un jeu d'évaluation (ou *test set*, en anglais). Ce jeu d'évaluation doit idéalement être constitué de traductions de qualité, produites par des traducteurs humains.

Il existe plusieurs manières de mesurer la similarité entre deux traductions. Peu importe la manière, la principale difficulté dans une telle évaluation quantitative est que, la traduction étant une activité fondamentalement humaine, deux traductions pour une même phrase peuvent être à la fois différentes et valides.

La mesure la plus courante est le score BLEU, introduit par Papineni *et al.* (2002). Celui-ci est basé sur le nombre de fois que les mots de la traduction de référence apparaissent dans la traduction hypothétisée. Cela doit être calculé pour l'ensemble des phrases du jeu d'évaluation (le score BLEU correspond toujours à une mesure pour un jeu d'évaluation complet). D'autres mesures courantes existent, tel le ChrF++ (Popović, 2015), qui compare les phrases non pas mot par mot, mais caractère par caractère.

1.3.1.5 Langues dites peu dotées

Plus une grande quantité de paires de phrases est disponible pour une paire de langues donnée, plus les performances d'un modèle de traduction automatique pour cette dernière seront bonnes. Plusieurs paires de langues très bien représentées sur le web sont des langues officielles d'un ou de plusieurs pays, ou sont parlées par un très grand nombre de locuteurs. C'est le cas pour l'anglais, le français et les autres langues représentées dans les corpus Europarl et de l'ONU.

Toutefois, toutes les langues n'ont pas autant de données disponibles pour la traduction automatique. Selon l'UNESCO, il existe environ 7000 langues en usage dans le monde [13]. C'est une évidence que beaucoup d'entre elles n'ont ni les ressources, ni le statut officiel leur permettant de produire un million de paires de phrases parallèles.

Les paires de langues peuvent être catégorisées selon leur niveau de ressources. Les paires de langues pour lesquelles une faible quantité de données est disponible sont dites « peu dotées » (ou *low-resource*, en anglais), alors que celles pour lesquelles une grande quantité de données est disponibles sont dites « bien dotées » (ou *high-resource*, en anglais). Le tableau 1.3 donne en exemple une estimation du nombre de phrases

[13] UNESCO - Atlas mondial des langues (en Anglais)

parallèles disponibles pour quelques paires de langues ayant différents niveaux de ressources (tableau tiré de l'article de Haddow *et al.* (2021)).

TABLE 1.3 – Niveaux de ressources et nombre de phrases parallèles, pour différentes paires de langues (tiré de l'article de Haddow *et al.* (2021))

Niveau de ressources	**Paire de langues**	**Locuteurs**	**Phrases parallèles**
Élevé	anglais-français	267M	280M
Moyen	anglais-birman	30M	0.7M
Bas	anglais-fon	2M	0.035M

CHAPITRE 2
ÉTAT DE L'ART

Le chapitre qui suit dresse l'état de l'art des principaux sujets touchés par ce mémoire. D'abord, la section 2.1 présente l'état de l'art des ressources et technologies de la langue pour l'innu-aimun, ainsi que pour les langues qui y sont reliés et d'autres langues autochtones du Canada. La section **??** présente l'état de l'art de la traduction automatique pour les langues dites peu dotées. Enfin, la section **??** présente l'état de l'art des méthodes d'alignement automatique des phrases.

2.1 Ressources et technologies pour l'Innu-aimun et autres langues autochtones du Canada

2.1.1 Innu-aimun

Il existe à l'heure actuelle peu d'applications d'outils technologiques de la langue pour l'innu-aimun. Ceux-ci ont principalement un but pédagogique et de préservation de la langue. On peut citer deux principales applications, le dictionnaire en ligne et l'application de conjugaison des verbes, qui s'inscrivent dans un effort commun de développement avec les langues cries, de la même famille linguistique. Les deux applications en question pour l'innu-aimun ont en partie été adaptées d'applications destinées à ces dernières.

Ces deux applications ont été présentées par Junker *et al.* (2016), dans le cadre d'une série d'outils web destinés d'abord aux locuteurs bilingues de l'innu et dont le but principal est la préservation de la langue. Ces ressources en ligne peuvent être vues comme la continuité des ressources linguistiques non-technologiques déjà existantes. Outre le dictionnaire en ligne et l'application de conjugaison des verbes, on y retrouve des jeux intéractifs en ligne pour l'apprentissage de l'innu-aimun, ainsi que plusieurs ressources linguistiques de référence (guides de grammaire, lexiques spécialisés, etc.).

L'application de conjugaison des verbes [1] organise les verbes par classe et radical. Pour chaque verbe on fournit la conjugaison selon les 20 temps et les 9 personnes verbales. Chaque conjugaison est accompagnée de deux extraits audios, un pour la prononciation dans le dialecte de l'est et l'autre pour le dialecte de l'ouest.

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

Le dictionnaire en ligne [2], qui est trilingue et pan-dialectal, organise les résultats de recherche suivant une structure adaptée à la langue innue. Y sont d'abord présentés les gloses innues associées au mot recherché en français ou en anglais, puis les différentes déclinaisons de chacunes de ces gloses (c'est-à-dire les mots formés selon le contexte).

Le dictionnaire en ligne est basé en partie sur le moteur de recherche développé par Junker et Stewart (2008). Parmi les apports de ces derniers, on retrouve la flexibilité orthographique du dictionnaire et la suggestion de l'orthographe correcte (le dictionnaire informera l'usager si un mot n'est pas écrit correctement et lui suggérera l'orthographe appropriée).

Ces fonctions démontrent qu'il est possible d'avoir des outils avec un certain niveau d'avancement ou d'automatisation avec les bases de données et modèles linguistiques existants, sur lesquels le dictionnaire est basé. Ces bases et modèles ne sont toutefois pas disponibles publiquement, on n'y a accès que par l'intermédiaire du dictionnaire. Ce type de modèle ouvre aussi la porte au développement d'autres outils qui ne sont pas encore disponibles pour l'innu-aimun, tel un analyseur morphologique.

Au meilleur de nos connaissances, il n'existe aujourd'hui aucun outil pour l'innu-aimun qui utilise des méthodes plus récentes ou avancées de traitement automatique du langage (TAL) basées sur l'apprentissage profond, tel que la traduction automatique neuronale (TAN). Unique exception à cette observation, on peut noter les résultats préliminaires obtenus par Tan Le *et al.* (2022) pour une approche basée sur l'apprentissage profond de la segmentation morphologique de l'innu-aimun.

Par ailleurs, il est pertinent de mentionner que le site web Glosbe, qui développe de la traduction automatique, des mémoires de traduction et des dictionnaires de manière collaborative pour nombre de paires de langues, propose un dictionnaire pour la paire innu-aimun/français [3]. Il ne semblerait pas toutefois que cet outil comprenne de la traduction automatique et ce développement n'a jusqu'ici pas fait l'objet de publication, au meilleur de nos connaissances. Sur le même plan collaboratif, le site web Wiktionnaire comprend des listes de noms communs [4]

[2] https://dictionary.innu-aimun.ca/

[3] https://fr.glosbe.com/fr/moe

CHAPITRE 3
MÉTHODE DE COLLABORATION EN MILIEU ÉDUCATIF AUTOCHTONE

Comment la recherche en traitement automatique du langage (TAL), incluant au premier chef la collecte et la validation de données, peut-elle s'adapter à la réalité du terrain d'une école primaire-secondaire innue, ou d'un programme de traduction d'innu-aimun dans un Cégep ? Les exigences de la recherche en informatique et en TAL peuvent-elles s'arrimer aux besoins d'une communauté scolaire et de ses acteurs, non seulement au regard des objectifs ultimes de la recherche mais aussi tout au long du processus de cette dernière ? Et si tel est le cas, quelles seraient les limites d'une méthode qui aurait comme objectif premier de respecter ces besoins, dans le but d'un développement par et pour la communauté ? Le présent chapitre a pour but de répondre à ces questions, à travers l'expérience de collaboration vécue avec l'école primaire-secondaire Kanatamat et celle de participation avec les étudiants du programme de traduction et interprétation d'innu-aimun au Cégep de Sept-Îles.

Pourquoi s'intéresser de premier chef au milieu éducatif innu ? Parce que c'est à la fois là où se transmettent les compétences d'innu-aimun (hors des foyers familiaux) et là où l'on peut constater, sur le terrain, les besoins et les difficultés vécus par les acteurs de cette transmission. Tel que l'exprimera la description des processus de collaboration du présent chapitre, s'immerger dans le milieu éducatif permet de mettre à l'épreuve du terrain des hypothèses de recherche qui auraient été élaborées purement à travers le prisme des domaines de l'informatique et du TAL. Puisque notre projet a la prétention d'aider à la revitalisation de l'innu-aimun par le développement d'outils informatiques, il s'agit d'une étape primordiale dans l'initiation d'un tel développement.

Nous examinerons à travers ce chapitre le processus de collaboration et de participation d'acteurs des deux différents milieux scolaires susmentionnés et comment celui-ci a permis d'aboutir à des objectifs communs et à une méthode de travail bénéfique à la fois pour la recherche et pour le milieu éducatif. Nous ferons de cette expérience une étude de cas et chercherons à savoir comment elle peut renseigner plus largement la recherche académique en informatique quant aux méthodes de collaboration dans une communauté (autochtone ou autre).

La section 3.1 établit les principes et le cadre de méthodologique qui guide et structure notre approche collaborative. La section 3.2 présente l'approche générale de collaboration proposée. Cette approche a été

co-conçue avec l'école primaire-secondaire Kanatamat, principalement lors de la conclusion d'une entente collaborative avec cette dernière. La section 3.3 rapporte le processus de conclusion de l'entente collaborative avec l'école Kanatamat. À noter que, des deux activités décrites dans le cadre de l'entente, seule la première a été mise en oeuvre, dû à des contraintes de calendrier scolaire et de disponibilité du personnel de l'école. Cette activité est le sujet de la section 3.4, qui présente la collaboration avec le personnel enseignant de l'école Kanatamat pour la validation d'alignements conjointement à la préparation d'exercices de niveau primaire. La section 3.5 présente la participation des étudiants en traduction et interprétation d'innu-aimun au Cégep de Sept-Îles pour la création d'alignements via une activité parascolaire. On peut considérer cette dernière collaboration comme l'enclenchement général du développement d'outils d'assistance à la traduction d'innu-aimun, puisqu'elle aura été une occasion d'échanger avec de futures usager potentiels de ces outils. La section 3.6 analyse plus en profondeur et de manière plus critique les processus et résultats de collaboration et participation rapportés au cours du chapitre, tout en explorant les implications plus large pour la recherche en TAL et en informatique.

3.1 Principes et cadre méthodologique

L'impératif de collaboration avec la communauté, lorsqu'il est question de recherche sur des sujets autochtones, a été établi de maintes manières, que ce soit par le milieu de la recherche ou par les instances gouvernementales et les représentants autochtones.

Dans les domaines qui nous intéressent, soit le TAL ou la linguistique computationelle, plusieurs chercheurs ont écrit sur l'importance de travailler en partenariat avec les communautés autochtones concernées, dans une approche de développement se voulant « par et pour » ces dernières. De façon notable Steven Bird, un chercheur en linguistique computationelle avec une longue expérience et d'importantes contributions notamment sur les langues autochtones d'Australie, a dans un article de positionnement à la conférence internationale sur la linguistique computationelle (*International Conference on Computational Linguistics* ou COLING) appelé à décoloniser le développement des technologies du langage (Bird, 2020). Il y décrie entre autres la tendance traditionelle en linguistique computationelle à voir les langues autochtones comme une ressource à exploiter, et pour lesquelles il y a un potentiel de développement, sans pour autant considérer les volontées des communautés autochtones concernées.

En présentant leur approche de boîte à outil pour langues en danger, Arppe *et al.* (2016) soutiennent qu'en

s'engageant auprès des communautés autochtones, il est important de prioriser les projets de recherche qui leur sont vraiment utiles compte tenu de la rareté des locuteurs parlant couramment et de la valeur de leur temps. Afin de sélectionner de tels projets, les auteurs suggèrent de caractériser la situation linguistique propre à la communauté cible et proposent à cet effet plusieurs questions à se poser : des questions portant sur le nombre de locuteurs parlant couramment et en situation d'apprentissage, les domaines d'utilisation de la langue, etc.

Mager *et al.* (2023) ont examiné la question du développement éthique de la traduction automatique pour les langues autochtones dans les Amériques, à travers une enquête sur les différentes pratiques ainsi que plusieurs entretiens avec des leaders communautaires et divers spécialistes du langage et de l'enseignement. Ils ont pu réaffirmer, à travers cet exercice, l'importance crucial d'impliquer les locuteurs et membres de la communauté lors du développement de la traduction automatique pour une langue autochtone.

Plusieurs autres ont élaboré des principes de bonnes pratiques plus généralement pour la recherche en intelligence artificielle impliquant les communautés autochtones. Par exemple, Lewis (2020) définit un protocole pour ce faire, qui propose plusieurs lignes directrices pour la conception de l'IA centrée sur les autochtones, en demandant notamment que la conception de l'IA soit réalisée en collaboration avec des communautés spécifiques, indiquant que les connaissances autochtones ont une forte dimension locale et territoriale et que cela devrait se refléter dans les systèmes d'IA.

Au-delà des souhaits, des recommandations ou des appels aux meilleures pratiques, le gouvernement canadien depuis 2018 fait de la collaboration avec la communauté une *obligation* pour pouvoir faire de la recherche sur un sujet impactant les autochtones. Le chapitre 9 de l'énoncé des trois organismes de recherche fédéraux sur l'éthique de la recherche avec des êtres humains [1] indique : « Si le projet de recherche est susceptible d'avoir une incidence sur le bien-être d'une ou de plusieurs communautés autochtones auxquelles appartiennent les participants éventuels, les chercheurs doivent solliciter la participation de la communauté ou des communautés visées. »

Enfin, plus localement, le Protocole de recherche des Premières Nations au Québec et au Labrador [2] établit des normes d'éthiques, incluant les principes dits de PCAP (propriété, contrôle, accès et possession des

[1] EPTC 2 (2018) – Chapitre 9 : Recherche impliquant les Premières Nations, les Inuits ou les Métis du Canada

[2] Protocole de recherche des Premières Nations au Québec et au Labrador

données). Tel que le spécificie le protocole, ces derniers principes visent la « protection du patrimoine informationnel et des connaissances des Premières Nations ».

Ces normes ne permettent pas seulement le respect de l'intégrité des communautés autochtones. Par leur exigences de collaboration, elles poussent à mener une recherche ayant au final une plus grande utilité réelle pour les communautés visées. Et une communauté qui est non seulement consultée, mais qui participe activement à la recherche aura bien plus de chances de s'en approprier les résultats, évitant que ceux-ci tombent dans l'oubli.

Des exemples intéressants d'approche collaborative avec des langues autochtones ont été présentés dans les domaines du TAL et de la linguistique computationelle. Notamment, Bontogon (2016) démontre une démarche exemplaire dans laquelle, après avoir développé un système d'apprentissage assisté par ordinateur pour le nêhiyawêtân, ce dernier est évalué directement auprès d'utilisateurs apprenant la langue, incluant par des personnes dont c'est la langue maternelle. Dans un autre exemple d'approche de collaboration, Nekoto *et al.* (2020) qui démontrent la faisabilité d'une approche participative à grande échelle, en apportant des résultats de références en traductions automatiques pour une trentaines de langues africaines peu dotées.

3.2 Approche suivie

L'essence de l'approche que nous présentons ici, et qui a été élaborée en collaboration avec l'école Kanatamat, est de trouver des objectifs communs à chaque étape du projet. Ces objectifs doivent se baser sur les besoins immédiats des membres de la communauté (ici, l'une ou l'autre des communautés scolaires innues) et les objectifs plus long terme du développement collaboratif des outils informatisé. Nous insistons sur l'importance d'assurer un impact positif pour la communauté à chaque étape du processus de recherche, dès les débuts. Pour s'assurer que la collecte de données servant à la recherche et au développement d'outils n'est pas extractive, la communauté ne devrait pas avoir à attendre la fin de la recherche pour en bénéficier. Que ce soit dès les débuts de la recherche, aux premières collectes de données, où à la fin dans le développement et l'évaluation des outils à destination de la communauté (par exemple, une application), chaque étape doit comporter des avantages pour les membres de la communauté et les participants à la recherche. Pour que le développement soit fait par et pour la communauté, les résultats de recherche doivent aussi être présentés à la communauté et faire l'objet de rétroactions, à chaque étape du projet.

La figure 3.1 schématise l'approche suivie, incluant les intéractions entre la recherche et la communauté tout

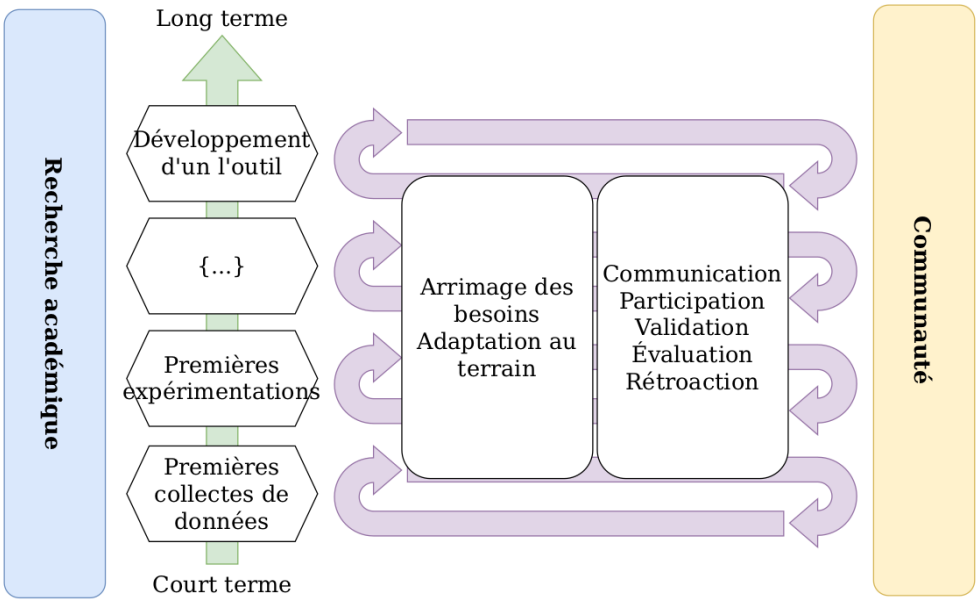

FIGURE 3.1 – Schéma illustrant l'approche suivie

le long du processus.

Dans un contexte où l'objectif est de favoriser l'apprentissage et l'usage de la langue et dans le cas d'une collaboration en milieu éducatif ou d'une participation de ce dernier, où les enseignants d'innu-aimun et où il manque de temps dédié à l'enseignement de la langue, il est crucial que la recherche n'accapare ni le temps des enseignants, ni le temps d'enseignement aux élèves. De ce fait, la collaboration de recherche doit s'arrimer parfaitement avec les objectifs habituels des personnes sollicitées et de la communauté, afin de limiter les impacts négatifs de leur participation sur eux-même et sur leur communauté.

Ainsi, par exemple, si une activité de collecte de données peut s'effectuer à même les activités habituelles d'enseignement, cette collecte de données sera faite sans impact négatif dans le court terme et avec des impacts potentiellement positifs dans le long terme (par exemple, disponibilité d'une application d'aide à l'apprentissage). Si, de surcroît, l'activité de recherche en question peut amener des impacts positifs dès sa mise en oeuvre, et non seulement une fois les développements terminés, la recherche sera d'autant plus respectueuse des membres de la communauté et de leur temps. Par exemple, si la création ou la validation de données peut se faire dans l'esprit de créer à la fois des données de recherche et du matériel pédagogique, alors il y a non seulement des effets positifs potentiels à long terme, mais aussi des effets positifs immédiats dans le milieu éducatif.

La participation des personnes-ressources (enseignants ou élèves) à même leurs occupations permet par ailleurs d'ancrer le projet dans la réalité de la communauté. Cette exposition directe au milieu permet de repenser constatement les objectifs de recherche tout au long du processus et d'éviter d'aboutir au final à un outil mésadapté.

3.3 Élaboration de l'approche et d'une entente de collaboration avec l'école primaire-secondaire Kanatamat

La section qui suit décrit l'entente de collaboration conclue avec l'école Kanatamat, l'école primaire-secondaire de la communauté de Matimekush - Lac John, élaborée dans le respect des principes présentés à la section 3.1 et dans l'esprit de l'approche générale proposée à la section 3.2.

La sous-section 3.3.1 présente la communauté de Matimekush - Lac John, relate le déplacement que nous y avons fait, ainsi que l'immersion et la participation aux activités dans cette communauté et son école, préalablement à la conclusion de l'entente. Bien que difficilement formalisable dans un mémoire d'informatique, on peut considérer cette étape comme critique à la co-conception de l'approche collaborative, avec les acteurs de l'école, ainsi que de l'élaboration de l'entente de collaboration. La lecture de cette sous-section est toutefois optionnelle pour comprendre le contenu de l'entente elle-même. Le lecteur qui souhaite précisément accéder à l'information concernant l'entente peut donc se référer à la sous-section 3.3.2, qui présente l'entente finale et son élaboration.

3.3.1 Matimekush - Lac John et école Kanatamat : allée et immersion dans la communauté

La communauté de Matimekush - Lac John se situe au 54e parallèle nord, tout près de la frontière entre le Québec et le Labrador. Cette frontière n'existait toutefois évidemment sur le territoire ancestral innu, traditionnellement. Ce territoire est en partie partagé avec les Naskapis, dont l'unique communauté, Kawawachikamach est voisine de Matimekush - Lac John. Ce territoire est celui où vivait traditionnellement la famille de l'autrice innue An Antane Kapesh. Celle-ci relate dans son roman *Eukuan nin matshi-manitu innushkueu / Je suis une maudite sauvagesse* la dépossession vécue par les Innus et les injustices vécues avec le début de l'exploitation du minerai de fer sur leur territoire, concomitant avec la création de la ville de Schefferville, qui a amené la sédentarisation des Innus de la région (Kapesh, 2019).

La communauté est située quelques 575 km au nord de Sept-Îles, à mi-chemin entre cette dernière et Kuuj-

juaq. Cette distance est celle parcourue par le train qui fait le trajet entre Sept-Îles et Schefferville (voir trajet à la figure 3.2). Avec l'avion, ce train dénommé *Tshiuetin* (qui signifie « Vent du nord », en innu-aimun) est le seul moyen de transport permettant de se rendre dans la région, car le réseau routier n'y est pas connecté.

Le train entre Sept-Îles et Schefferville, autrefois opéré par la minière IOC pour transporter le minerai de fer, est désormais co-propriété des Innus et des Naskapis. Tel que raconté par l'autrice Innue Naomi Fontaine dans son roman *Manikanetish* (Fontaine, 2017), et tel que nous avons pu le constater nous-même en l'empruntant pour se rendre à Matimekush, ce train qui sillonne la rivière Moisie tel l'ancien chemin de portage est très utilisé par les Innus et les Naskapis pour accéder au territoire ancestral. L'opérateur permet l'arrêt à de nombreuses bornes kilométriques le long du trajet, donnant accès à des campements de chasse et de pêche, le territoire entre Sept-Îles et Schefferville étant constitué de nombreux lots de trappe attitrés traditionnellement à des familles innues ou naskapies.

Notre immersion pendant une semaine complète lors de ce premier séjour à Matimekush - Lac John a permis non seulement de donner le temps à une réflexion plus poussée pour l'élaboration de l'approche, elle a aussi permis de bâtir une relation plus solide avec la communauté de l'école. Sans ces deux choses, il est probable que l'élaboration de notre approche collaborative n'ait pas été possible.

Dès notre arrivée à Matimekush, nous avons pu contribuer à la préparation d'activités reliées à la semaine de l'environnement de Matimekush - Lac John. En effet, les activités de cet évènement, comme de nombreux autres qui se produisent dans la communauté, avaient lieu à l'école elle-même. Cela témoigne du rôle de centre communautaire que joue l'école. Nous avons donc pu nous immerger dans les activités de la communauté en participant à ces événements [3].

Durant cette semaine complète l'auteur du présent mémoire a été amené à remplacer l'enseignante absente d'une classe de 3e année du primaire, l'école n'ayant pas d'autres personnes disponibles à ce moment-là. Cette expérience a permis de mieux comprendre la dynamique des classes de niveau primaire à cette école, tout en assistant l'école, qui était dans le besoin à ce moment.

Globalement, nous avons pu constater à quel point participer aux activités générales de la communauté a permis à la fois de mieux comprendre la réalité de l'école et de la communauté, et de bâtir une relation de

[3] Rapport annuel de l'IDDPNQL - Page 27 (semaine de l'environnement à Matimekush – Lac John)

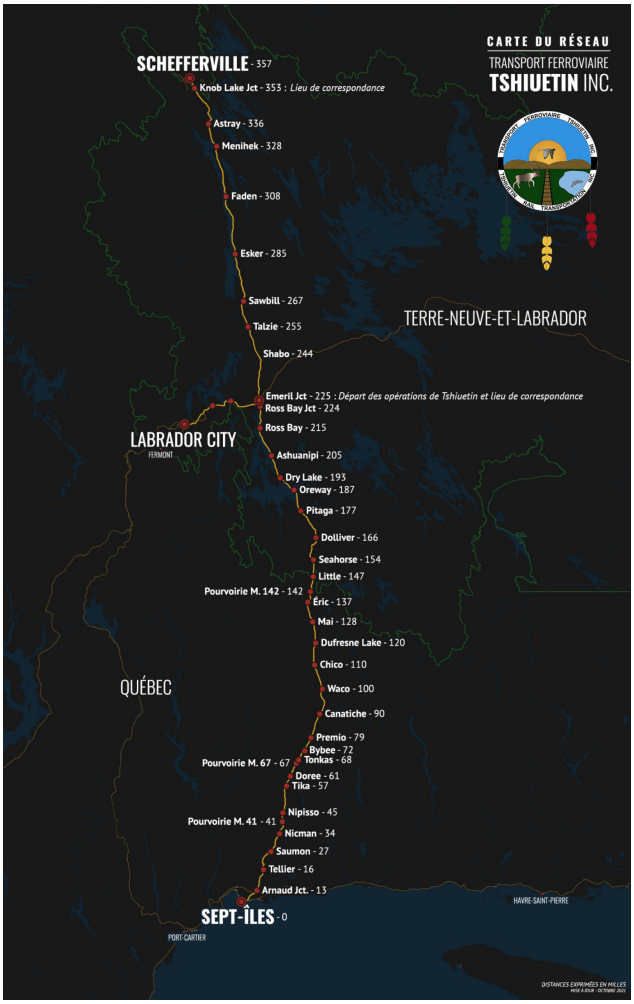

FIGURE 3.2 – Itinéraire du train *Tshiuetin* entre Sept-Îles et Schefferville (Image reproduite depuis `https://tshiuetin.net/`, avec la permission de Tshiuetin)

(a) (b)

FIGURE 3.3 – À gauche, l'auteur du présent mémoire près du Lac Houston. À droite, une des nombreuses résidences permanentes que se sont construits les membres de la communauté de Matimekush - Lac John près de ce lac, qui fait partie du territoire ancestral.

confiance mutuelle. Cette relation a été primordiale pour la co-conception d'une approche de collaboration ainsi que pour la mise en oeuvre des activités avec le personnel enseignant de l'école.

En parallèle du séjour et de l'immersion dans la communauté de l'école, nous avons eu l'occasion d'accorder un entretien à Radio-Canada Côte-Nord [4], afin d'expliquer en quoi consistait le projet de recherche et de proposer un appel à la participation aux Innus qui le désirent. Cet entretien a été partagé sur les réseaux sociaux dans la communauté, et quelques-uns de ses membres nous ont abordé à ce sujet, en témoignant de leur opinion positive du projet. Il est possible que cela aura aussi favorisé plus largement la collaboration lors de séjours subséquents.

3.3.2 Entente de collaboration finale et processus d'élaboration

L'entente conclue avec l'école prévoit deux activités de collaboration visant à mettre en oeuvre des ateliers pédagogiques ou à créer du matériel pédagogique pour l'apprentissage de l'innu-aimun, tout en permettant une collecte et validation de données de recherche à mêmes ces activités. Seule la première activité de collaboration a pu être mise en oeuvre dans le cadre de ce mémoire, c'est-à-dire la préparation de matériel pédagogique avec le personnel enseignant (voir section 3.4). Les activités décrites dans l'entente correspondent, du point de vue de la recherche, à l'étape de collecte et de validation de données en vue de mener de premières expérimentations en traitement automatique du langage. Cela correspond à la première étape du processus illustré dans le schéma de l'approche présentée à la figure 3.1.

Il est prévu dans l'entente que l'école reste propriétaire de toute donnée produite dans le cadre de cette collaboration. Si elle le désire, elle peut demander à l'équipe de recherche de lui rendre ces données et de ne plus les utiliser. Elle pourra en tout temps de son côté réutiliser ces données pour ses besoins, tel que la création de matériel pédagogique ou la mise en oeuvre de projets scolaires. De plus, si l'usage de ces données dans la recherche résulte un jour en un outil informatisé utilisable (tel une application), il est prévu que l'école en soit le propriétaire. Les aspects de l'entente qui concernent la propriété ont été élaborés pour respecter les principes de PCAP décrits à la section 3.1, mais aussi parce qu'ils sont une des demandes explicites de l'école. Ces demandes ont été faites durant l'élaboration, au cours d'échanges concernant les besoins de chaque partie. De tels échanges ont eu lieu avec la direction de l'école, et avec l'agent culturel et l'enseignante d'innu-aimun de l'école. Ces échanges sont aussi ce qui a permis la co-conception avec l'école

[4] Radio-Canada - L'intelligence artificielle pour la traduction français-innu

Kanatamat de l'approche de développement collaboratif, par et pour la communauté, des outils d'assistance à l'apprentissage et à la traduction.

3.4 Validation d'alignements et préparation d'exercices de niveau primaire

La principale référence en matière d'innu-aimun à l'école Kanatamat, ainsi que l'une des principales références dans la communauté de Matimekush - Lac John en général, est l'enseignante d'innu-aimun. Hormis un cours de niveau secondaire (pour les niveaux 4 et 5), le principal espace d'enseignement de la langue à l'école par cette enseignante sont les cours d'innu-aimun que recoivent tous les niveaux du primaire, à partir de la 3e année (les 1e et 2e années du primaire étant enseignées intégralement en innu-aimun).

Puisque la charge de l'enseignante d'innu-aimun est importante, et en vue de l'importance qu'elle a dans l'enseignement de la langue à l'école Kanatamat, il a été décidé que tout travail de validation de données effectué conjointement avec elle aurait comme objectif d'alléger sa charge de travail, plutôt que le contraire. Nous lui avons donc offert, dans le cadre de la collaboration, de l'aider à préparer des exercices avec corrigés pour ses classes de niveau primaire.

Plus spécifiquement, il a été décidé que le mode de fonctionnement pour la collecte de données validées serait la préparation conjointe d'un exercice d'alignement de textes bilingues adapté au niveau primaire. La synergie avec la collecte et la validation de données s'effectue donc au moment de préparer le corrigé pour l'exercice. Ce ne sont pas les élèves qui créeront ou valideront les données, mais bien l'étudiant-chercheur, avec validation par l'enseignant d'innu-aimun. Le résultat final de la collaboration est un corrigé d'exercices pour l'école, et une série de textes bilingues alignés et validés pour l'équipe de recherche.

La sous-section 3.4.1 présente la description de la forme d'exercice d'alignement retenue et la sous-section 3.4.2 présente la méthode de travail conjoint mise en oeuvre par l'étudiant-chercheur et l'enseignante d'innu-aimun pour la préparation de l'exercice et du corrigé. La sous-section 3.4.3 fait un retour global sur le processus de collaboration pour cette activité.

Enfin, il est à noter que l'enseignante a aussi collaboré à l'analyse et la validation des résultats obtenus grâce aux tests de traduction basés entre autres sur les alignements ici obtenus. Cette analyse collaborative est présentée au chapitre 5.

3.4.1 Description de l'exercice d'alignement

L'exercice d'alignement se base sur le recueil de poésie *Nin Auass* (Bacon et Morali, 2021), projet dont l'objectif était de permettre aux jeunes élèves dans les écoles des communautés innues de rédiger des poèmes en innu-aimun et en français avec l'aide de poétesses innues [5]. Ce recueil a été ensuite distribué dans les classes des écoles primaires et secondaires à travers les communautés et l'école Kanatamat en a des exemplaires.

L'exercice d'alignement a pour but de présenter aux élèves du cours d'innu-aimun un court poème jeunesse, d'un côté écrit en innu-aimun et de l'autre en français. Les poèmes sont à la fois distribués aux élèves et présentés au tableau. L'élève doit numéroter sur sa feuille chaque vers, dans les deux versions du poème, puis inscrire les associations entre chaque vers (par exemple : « 1-1 », « 1-2 », « 2-3 », etc.). La difficulté de l'exercice tient du fait que les vers peuvent ne pas être exactement les mêmes d'une langue à l'autre. Par exemple, il pourrait y avoir plus de vers d'un côté que de l'autre. Ou alors, le nombre de vers peut être le même, mais certains vers pourraient être inversés, fusionnés, etc. À la fin, à l'aide de son corrigé, l'enseignant présente au tableau les bonnes équivalences entre les vers.

L'exercice permet aux élèves d'établir des comparaisons entre le français et l'innu-aimun, voir de poser un certain regard critique sur les équivalences faites entre les deux langues. Si certains mots dans les poèmes sont inconnus des élèves, cela donne une occasion à l'enseignant de les leur apprendre. Enfin, cela permet aussi à l'enseignant d'introduire les élèves à ce qu'est la poésie, et de montrer en exemple des poèmes ayant été écrits par d'autres jeunes de leur âge.

3.4.2 Méthode de travail pour la préparation d'exercices

La méthode de travail suivante a été mise en oeuvre pour l'élaboration conjoint d'alignements et d'exercices avec l'enseignante d'innu-aimun.

1. **Sélection par l'étudiant-chercheur de poèmes selon certains critères établis avec l'enseignante d'innu-aimun.**

   Ces critères sont les suivants : les poèmes choisis pour les exercices doivent être plutôt courts (autour d'une dizaine de lignes maximum, selon la complexité), avoir un niveau langage approprié pour des élèves de niveau primaire et avoir un alignement innu-aimun/français qui est simple. Ce dernier

[5] Nin Auass - Moi l'enfant (Mémoire d'encrier)

critère vise à ce que l'exercice d'alignement ne soit pas être trop complexe, pour ne pas que la logique d'alignement prenne le dessus sur l'objectif pédagogique principal, qui est celui d'apprendre l'innu-aimun et de faire des liens avec le français.

2. **L'étudiant-chercheur fais une première tentative d'alignement.** Cette tentative est effectuée manuellement, sur la base d'indices tels que le nombre de lignes dans les deux langues et la ponctuation, ainsi qu'à l'aide d'une consultation du dictionnaire innu (Ambroise *et al.*, 2023) et de certaines connaissances de base sur la langue. La tentative d'alignement pour chacun des poèmes sélectionnés est formatée en tant que corrigé d'exercice pour le poème en question. Le corrigé indique le titre du poème, si les lignes sont les mêmes dans les deux langues ou non, et l'alignement proposé.
3. **L'enseignante valide l'alignement.** Les corrigés des poèmes sont imprimés et accompagnés des poèmes en question, qui sont présentés dans les deux langues, l'une à côté de l'autre. La validation peut s'effectuer conjointement, auquel cas l'enseignante identifie les erreurs et l'étudiant-chercheur peut les corriger sur le champs. Alternativement, l'enseignante peut valider par elle-même, auquel cas elle indique seulement la présence d'erreur ou non dans la tentative d'alignement. La première option permet de corriger plus facilement les alignements erronés (qui sont d'habitude plus présents dans les poèmes asymétriques), alors que la seconde permet de valider davantage d'alignements dans un temps donné.
4. **L'enseignante confirme la pertinence des poèmes sélectionnés et identifie le niveau scolaire.** Pour chaque poème, l'enseignante confirme s'il s'agit d'un exercice d'alignement approprié pour le cours. Un poème pourrait être jugé inapproprié par l'enseignante si, par exemple, les versions innues et françaises du poème ne sont pas assez concordantes (une traduction inexacte ou trop libre selon elle). Elle pourrait aussi être inconfortable d'enseigner le langage utilisé (s'il vient du dialecte d'une autre communauté, par exemple) ou juger qu'il est trop complexe pour des élèves de niveau primaire. Les poèmes jugés inappropriés sont retirés de la sélection d'exercices et du corrigé, mais leur alignement est conservé pour les fins de la recherche. Alternativement, il peut arriver qu'une traduction soit plutôt libre et ne concorde pas exactement, mais que l'enseignante souhaite tout de même conserver ce poème comme exercice ; une note est alors faite de cette inexactitude de la traduction. L'enseignante indique aussi le ou les niveaux scolaires pour lesquels le niveau de langage serait le plus approprié (par exemple, 5e et 6e année du primaire).
5. **L'étudiant-chercheur finalise le corrigé.** La rétroaction de l'enseignante est intégrée au corrigé préalablement préparé. Les corrigés et poèmes finaux sont imprimés et fournis à l'enseignante comme banque d'exercices dont elle pourra se servir en classe.

CORRIGÉ

Poème : L'ÎLE / MINISHTIK^{u}

Niveau : 4e-5e année

Mêmes lignes dans les deux langues: non

Alignement :

(1,1)
(2,2)
(3,3)
(4,4)
(4,5)
(5,6)

FIGURE 3.4 – Exemple d'un corrigé de poème à aligner

Le tableau **??** et la figure 3.4 montrent respectivement un exemple de poème bilingue choisi comme exercice d'alignement, avec le corrigé qui l'accompagne.

TABLE 3.1 – Exemple d'un poème à aligner, tiré de *Nin Auass* (Bacon et Morali, 2021))

Langue	**français**	**innu-aimun**
Titre	L'île	Minishtik^{u}
Vers 1	Sur l'île	Minishtik^{u}
Vers 2	dans la forêt	minashkuat
Vers 3	les moyaks se posent	missipat tueuat
Vers 4	ils rendent visite à leurs petits	natshi-uapameuat
Vers 5	et les protègent	umissipissuaua
Vers 6		nakatuenimeuat

3.4.3 Retour sur le processus de collaboration

Puisque la disponibilité de l'enseignante à même les heures de travail était limitée (et puisqu'il était hors de question d'imposer des heures supplémentaires à celle-ci, même avec compensation financière), le processus a pris un certain temps à aboutir. Ce long processus a toutefois permis une certaine immersion dans le milieu scolaire, avec au final une meilleure adapation à la réalité du terrain.

Notamment, les échanges avec l'enseignante, suivis d'essais en classe, ont été cruciaux pour aboutir à une forme d'exercice qui était réellement utile à l'enseignante. Après discussion, la première forme d'exercice jugée comme intéressante pour l'enseignement était une forme où les mots des vers dans chaque langue étaient numérotés, afin de pouvoir être alignés avec les mots correspondants. Cela permettait d'avoir un alignement direct et assez précis des mots dans les poèmes, bien que certains mots ou parties des vers étaient parfois absents ou différents dans l'autre langue (ils étaient alors considérés comme supprimés).

Après de premiers essais en classe avec les élèves du primaire, il est apparu que cette version de l'exercice n'était pas au point. Puisque la numérotation pouvait prendre différentes formes, plusieurs élèves sont arrivés à des réponses différentes qui pouvaient être valides, rendant la correction plus longue. De plus, cette forme de numérotation étendue et complexe apportait une certaine confusion auprès des élèves et compliquait la tâche de l'enseignante. Globalement, l'exercice était assez complexe et très long à mettre en oeuvre, avec beaucoup d'hésitations.

Suite à cet essai, l'étudiant-chercheur et l'enseignante ont conjointement décidé de conserver une forme de l'exercice où la numérotation était effectuée uniquement vers par vers. Cela a à la fois l'avantage de simplifier l'exercice pour tout le monde, en plus d'aboutir à un format d'alignement qui est directment en ligne avec celui recherché (alignement des phrases/vers) pour les buts de la validation de données.

3.5 Participation des étudiants en traduction innue au Cégep

Suite à l'annonce de la création d'un nouveau programme d'Attestation d'études collégiales (AEC) en traduction et d'interprétation d'innu-aimun au Cégep de Sept-îles [6], et puisque la toute première cohorte d'étudiants était à l'oeuvre, il a été décidé de proposer des activités de recherche participative avec ces derniers.

Les étudiants de ce programme sont des acteurs très pertinents pour le projet, pour plusieurs raisons. D'abord, parler l'innu-aimun est une condition d'admission au programme [7]. Ensuite, les étudiants, de futurs traducteurs professionnels, sont de futurs usagers potentiels des outils d'assistance à la traduction dont nous tentons ici de jeter les bases.

Vu le nombre d'heures d'enseignement limité dont dispose le programme, la quantité importante de sujets

[6] Le Cégep de Sept-Îles lance un programme de traduction en langue innue - Le Manic

[7] Description du programme: Traducteur / Interprète en langue innue

que ce dernier doit couvrir et la charge de travail déjà exigeante qu'ont les étudiants, il a été décidé qu'il était préférable que la participation ne se fasse pas à même le programme, mais plutôt en dehors, comme activité supplémentaire à la discrétion des étudiants. Cependant, il a aussi été déterminé qu'il serait primordial que l'activité soit enlignée avec les objectifs du programme (l'apprentissage de la traduction et interprétation de l'innu-aimun) et avec les intérêts des participants.

3.5.1 Définition de l'activité

L'activité proposée est un exercice d'alignement de phrases bilingues en innu-aimun et en français, sous la forme d'une activité parascolaire. La participation à cette dernière se fait sur une base volontaire, avec une compensation financière. On peut voir cet exercice comme une pratique supplémentaire optionnelle, en marge du programme en traduction et interprétation d'innu-aimun. Le formulaire de consentement signé par les étudiants participants est présenté a été approuvé par le Comité d'éthique de la recherche pour les projets étudiants impliquant des êtres humains (CERPÉ plurifacultaire) de l'UQAM, présenté l'Annexe B.

En plus de l'occasion de validation de données qu'offre l'exercice d'alignement, la tenue d'une séance avec les étudiants et la présentation du projet en amont offre aussi l'opportunité d'échanger de manière informelle à propos du projet lui-même. Quel est l'avis des étudiants par rapport à la proposition de créer un outil d'assistance à la traduction ? Est-ce que ce genre d'outil serait utile ou souhaitable ? Etc.

Les phrases à aligner sont des phrases tirées de deux livres de l'autrice innue An Antane Kapesh, et des vers de poèmes tirés du recueil de poésie jeunesse innue *Nin Auass*. Une description plus complète de ces textes est donnée au chapitre 4.

3.5.2 Déroulement de l'activité

La méthode d'alignement manuel utilisée est un document en ligne de type *Google Doc*, dans lequel les particpiants peuvent inscrire directement la numérotation d'alignement des phrases pour chaque paragraphe. Tel que nous avons pu le rapporter dans un article présenté à une conférence du domaine (Cadotte *et al.*, 2023), ce mode de fonctionnement permet d'éviter une période de formation pour que les participants sachent comment utiliser l'outil, le document étant proche de ce que ces derniers ont l'habitude d'utiliser dans leurs études ou dans leur travail. Cet aspect est non-négligable, vu le peu de temps qu'ont les participants et le peu de participants potentiels pouvant être recrutés.

Français	Innu-Aimun	Alignement	Note
[0] À mes huit enfants. [1] Préambule	[0] Nitauassimat umenu nishuaush etashiht. [1] Tshitshipanu aimun	0.0 1.1	
[0] Je remercie chacun de ceux qui m'ont aidée à faire ce livre que j'ai fait. Et je serais heureuse de voir d'autres Indiens écrire, en langue indienne.	[0] Ume mashinaikan ka tutaman ka itashiht ka uitshiht tshetshi tutaman ninashkumauat kassinu. [1] Kie nipa minueniten tshetshi uapataman kutak innu tshetshi mashinaitshet e innushtenit.	0.0 0.1	

FIGURE 3.5 – Échantillon d'un document d'alignement utilisé par les participants ; les phrases présentées ici proviennent de l'oeuvre de Kapesh (2019).

La figure 3.5 présente un exemple de formulaire d'alignement utilisé par les participants. On propose aux participants d'écrire les alignements dans un format de type Gale & Church (voir figure), mais ceux-ci conservent la liberté de représenter les alignements de la façon qu'ils trouvent la plus pertinente. Les participants peuvent aussi écrire des notes pour chaque alignement, voir surligner le texte ou y apporter des corrections. Ces corrections peuvent être reliées à une traduction que le participants trouve inexacte, ou alors pour indiquer une erreur dans le formattage ou la présentation des phrases à aligner, qui se serait par exemple produite lors de l'extraction des textes [8]. Les participants étaient aussi invités, s'ils le souhaitaient, à partager leurs avis généraux à l'oral, en tout temps durant l'activité.

3.5.3 Avis, commentaires et notes des participants

Dès la fin des activités d'alignement, les participants ont confirmé la pertinence de l'exercice dans le cadre de leur programme scolaire et ont dit apprécier pouvoir s'exercer de cette manière.

La principale note relevée par les participants est le fait que les poèmes jeunesse étaient des traductions très libres, souvent inexactes. Aux dires des étudiants, les aspects poétiques (tels les rimes) étaient visiblement priorisés au détriment de la rigueur de la traduction.

Les phrases des livres d'An Antane Kapesh sont, de l'avis unanime des participants, traduites de manière beaucoup plus rigoureuse et exacte. Les principales notes ou commentaires des participants portaient sur quelques rares choix de traduction qu'ils n'auraient pas fait de la même manière, ou alors sur des erreurs qui

[8] Une description complète du processus d'extraction et de pré-alignement des phrases est fournie au chapitre 4

avaient été commises lors de l'extraction du texte depuis les documents PDF. Ce processus d'extraction est présenté plus en détail au chapitre 4, dans le cadre de notre étude comparative sur l'alignement.

Les étudiants connaissaient non seulement les textes d'An Antane Kapesh, mais les avaient lus et s'étaient déjà exercés avec eux dans le cadre du programme. Ces oeuvres étaient visiblement appréciées et en accord avec la vision des participants, bien que certains ont fait part d'une certaine colère qu'ils ont ressentie à la relecture, puisque les oeuvres relatent de nombreuses injustices ayant été vécues par les Innus. D'ailleurs, un participant a relevé que la traduction française semblait plus aggressive et revendicatrice, alors que le texte en Innu était écrit dans un style plus factuel et droit au but.

3.6 Discussion

3.6.1 Analyse comparative des deux activités collaborative et participative

On pourrait être tentés de comparer les deux activités, l'une collaborative et l'autre participative, d'un point de vue purement informatique, c'est-à-dire celui de la recherche en traitement automatique du langage. Ce point de vue en est un qui s'intéresse à la capacité de produire le plus de données valides possibles pour des fins d'expérimentations. Si l'on se prête à ce jeu on peut noter que, alors que le travail des étudiants en traduction a permis de produire l'alignement de près de 1500 phrases (voir chapitre 4 pour un sommaire complet des corpus alignés) et d'une douzaine de poèmes, la collaboration avec le personnel enseignant de l'école Kanatamat a permis l'alignement de 33 poèmes. Même en considérant que les étudiants en traduction étaient plus nombreux (au nombre de 4 en tout) et ont pu investir plus de temps que l'unique enseignante d'innu-aimun avec qui nous avons collaboré, on pourrait dire que l'activité parascolaire avec les étudiants a été beaucoup plus productive que la préparation d'exercices de niveau primaire.

Or, est-il vraiment pertinent d'analyser les deux activités en ces termes productivistes pour notre projet, tel qu'on le ferait avec un projet dont le but final est l'annotation à large échelle ? Oui et non. Oui, car il faut bien, pour développer des outils basés sur le TAL, réussir à obtenir une certaine quantité de données. Toutefois, cette analyse ne saurait répondre à tous les aspects importants de notre projet qui, comme nous l'avons expliqué, repose en grande partie sur la collaboration et la participation. L'approche de collaboration est ici fondamentale, surtout dans l'optique d'un développement par et pour la communauté. Il faut donc impérativement analyser ce que la communauté tire dès maintenant de ces activités de collaboration et de participation, en plus de ce que le projet de développement collaboratif à plus long terme tire de ces activités,

au-delà de la simple quantité de données.

Avec une vision plus large, on pourrait dresser le constat que la communauté a pu tirer davantage pour le nombre d'heures données dans le cas de la préparation d'exercices de niveau primaire. En effet, la préparation de ces exercices n'a pas demandé beaucoup de temps à l'enseignante, mais il a été possible d'en produire une quantité appréciable du point de vue didactique, et avec des exercices adaptés à plusieurs niveaux du primaire. Cette collection d'exercices va perdurer dans le temps, après la fin de l'activité, et pourra même être possiblement utilisée par d'autres enseignants ou d'autres écoles (cela restsera à la discretion de l'enseignante). Du côté de l'activité avec les étudiants en traduction, cette participation aura été bénéfique en tant qu'exercice sur le moment, en plus de la compensation financière, mais l'impact direct pour ces derniers s'arrête là.

Bien sûr, le fait de produire plus de données peut contribuer au potentiel de développement d'outils qui aideront peut-être les étudiants en traduction dans le futur. Et pour le projet, les échanges avec de futurs usagers auront permis de valider en quelque sorte la pertinence de développer ces outils. Cela reste toutefois une perspective à plus long terme. Aussi du point de vue du projet, le temps de collaboration investi dans le milieu scolaire de niveau primaire peut être considéré comme bénéfique pour la compréhension de la réalité sur le terrain. Cette compréhension est primordiale si l'on veut un jour déployer des outils d'assistance qui seront réellement utiles dans une classe de niveau primaire.

Somme toute, on peut considérer les deux activités, collaborative et participative, comme complémentaires : l'un ou l'autre ne suffit pas à elle-même pour développer sur le long terme des outils TAL dont le but est de favoriser l'usage et l'apprentissage de l'innu-aimun. Même que l'une peut contribuer à l'autre, dans le sens qu'une participation de davantage d'acteurs de différents milieux et de différentes communautés permet une meilleure compréhension globale de la réalité de la langue et de la culture innue, et facilite toutes les formes de collaboration ou de participation. On peut aussi ajouter que, les communautés innues étant des milieux très petits et le nombre de personnes maîtrisant l'innu-aimun et pouvant potentiellement participer au projet étant limité, il faut absolument prioriser une approche de collaboration et de participation respectueuse des personnes et qui prend en compte tous les aspects, et non simplement la productivité immédiate en nombre de données.

3.6.2 Constats et implications plus larges pour la recherche

Que peut-on tirer de ces activités collaborative et participative, surtout face aux prétentions de l'approche collaborative présentée à la section 3.2 ? Les expériences vécues dans le cadre de ce mémoire permettent-elles de répondre à la question posée en début de chapitre, à savoir si les exigences de la recherche en informatique et en TAL peuvent s'arrimer aux besoins d'une communauté, d'un milieu scolaire et de ses acteurs, non seulement au regard des objectifs ultimes de la recherche mais aussi tout au long du processus de cette dernière ?

D'abord, suite aux résultats obtenus grâce aux activités décrites précédemment, il nous apparaît possible d'arrimer les besoin de la recherche en TAL à ceux de la communauté, dès les premières étapes. Jusqu'ici, d'un point de vue TAL, et tel qu'il sera décrit dans les chapitres 4 et 5, les premières collectes et validations de données, ainsi que les premières expérimentations ont été mises en oeuvre. Ces dernières constituent une étude de faisabilité et, le projet étant à ses débuts, un outil fonctionnel et déployable dans la communauté n'est pas immédiatement à portée de main. Toutefois, on peut constater que même sans cette disponibilité de l'outil, il a été possible de collaborer d'une manière immédiatement bénéfique pour la communauté. L'approche suivie est donc concluante dès les premières étapes du développement collaboratif.

La question des limites de cette forme de collaboration a aussi été posée en début de chapitre. On peut répondre que la principale limite est le temps nécessaire pour bâtir une quantité significative de données. À ce rythme, bâtir des outils TAL avancés pourra potentiellement prendre beaucoup de temps. Toutefois, cette limite n'est-elle pas une limite due en bonne partie aux réalités de la communauté, notamment la faible disponibilité de personnes maîtrisant à la fois l'innu-aimun et le français ? Qui plus est, étant donné que le projet s'inscrit dans des objectifs à plus long terme, qu'il doit être développé par et pour la communauté, cette disponibilité ne serait-elle pas au contraire maximisée sur le temps plus long avec une approche plus participative ? Et avec une approche qui rendraient les personnes plus intéressées et disposées à s'investir dans le projet ? Dans tous les cas, si l'on veut réellement prétendre développér des outils arrimés aux besoins et à la réalité de la communauté, une telle approche apparaît tout à fait nécessaire, même si elle nécessite une plus grande patience et une plus long engagement.

Ces constats pourraient potentiellement s'appliquer à d'autres projets de recherche en TAL ou en informatique qui ont des visées communautaires. Toutes les collaborations avec les communautés ne se font pas en milieu autochtone, avec les considérations que cela impose. Néanmoins, on pourrait trouver des milieux

communautaires où les conditions sont similaires : faible disponibilité et temps précieux des acteurs, besoin d'établir une relation de confiance pour pouvoir collaborer, culture différente, nécessité d'une sensibilité face aux réalités historiques, etc. Dans ces cas, nous sommes d'avis qu'une approche comme celle que nous avons présentée et mise en oeuvre pourrait être utile, considérant les résultats obtenus. Nous soulignons donc le potentiel d'application de cette méthode au-delà du sujet qui nous occupe, avec les limites précédemment identifiées.

3.7 Conclusion

Au cours de ce chapitre portant sur notre méthode de collaboration, nous avons décrit les principes qui ont guidé notre approche, décrit l'approche générale que nous proposons pour la collaboration, co-concue avec l'école Kanatamat, puis décrit la conclusion d'une entente de collaboration, ainsi que le déroulement de deux activités ayant permis de collecter et valider des données de recherche. L'entente et les deux activités respectant les principes et l'approche générales énoncés en début de chapitre, nous avons conclu que la proposition de départ, soit la réponse aux besoins d'une communauté, d'un milieu scolaire, à chaque étape d'un développement collaboratif par et pour cette dernière, était possible.

Les prochaines étapes du projet, au-delà du cadre de ce mémoire, qu'elles soient de futures activités de collaboration telles que définies par l'entente, ou de futures expérimentations et évaluations dans le cadre du développement de l'outil visé, permettront de confirmer cette conclusion ou d'exposer davantages les limites de notre proposition.

CHAPITRE 4
ÉTUDE COMPARATIVE DES MÉTHODES D'ALIGNEMENT

Au cours de ce chapitre, nous étudierons de manière comparative la performance de méthodes d'alignement existantes, en évaluant les alignements qu'elles produisent face aux jeux de référence créés de manière collaborative et participative (voir chapitre précédent).

L'alignement est une première étape nécessaire lorsque l'on veut effectuer un étude de traduction automatique pour une paire de langues peu dotée dont les textes bilingues disponibles ne sont pas tous déjà alignés. Ici, puisqu'on a un alignement de référence pour une partie de ces textes, notre étude comparative aura comme objectif de nous éclairer sur les meilleures méthodes d'alignement automatique à utiliser pour aligner les textes qui ne le sont pas.

La formation, pour les fins de la présente étude, de corpus bilingues innu-aimun/français alignés en collaboration et avec la participation de personnes maîtrisant l'innu-aimun offre par ailleurs une rare opportunité d'évaluer sur un nouveau jeu de référence peu commun des méthodes d'alignement déjà éprouvées. Ce jeu est peu commun, car il implique une langue sur laquelle ce genre d'évaluation n'a jamais été fait (l'innu-aimun), mais aussi parce que le contexte offre des difficultés supplémentaires. Il s'agit d'un contexte où une langue polysynthétique fait face à une langue qui ne l'est pas, mais aussi d'un contexte où une très faible quantité de données est disponible pour cette paire de langues. Ce contexte permet une évaluation en quelque sorte plus exigeante des méthodes connues. À cela s'ajoute le fait que les personnes ayant participé à l'alignement de référence ont produit certains alignements qui sortent du cadre habituel des méthodes existantes et ont laissé des réflexion qualitatives sous forme de notes concernant les textes à aligner.

La section 4.1 présente le cadre expérimental (méthodes d'alignement et méthodes d'évaluation utilisées), la section 4.2 présente les textes de référence utilisés et offre une analyse des corpus alignés, la section 4.3 présente les résultats de l'étude comparative, la section 4.4 en fait l'analyse détaillée avec une discussion et la section 4.5 dresse les conclusion de l'étude en proposant quelques pistes d'amélioration.

4.1 Cadre expérimental

Pour l'étude comparative, trois méthodes ont été choisies : Vecalign (Thompson et Koehn, 2019), Moore (Moore, 2002) et Gale & Church (Gale et Church, 1993). À cela s'ajoute une comparaison à une méthode que l'on nommera « naïve », qui produit un alignement n'étant pas sensé avoir de valeur ajoutée par rapport à l'alignement déjà présent dans le texte. Cette dernière comparaison est effectuée afin de voir quelle est l'efficacité réelle des méthodes, vis-à-vis de l'alignement déjà présent dans le texte d'origine.

Les trois méthodes d'alignement ont été choisies entre autres parce qu'elle fonctionnent chacune sur un mode différent : Gale & Church se base uniquement sur la longueur des phrases, Moore se base à la fois sur la longueur des phrases et sur un apprentissage statistique de correspondances lexicales, et Vecalign se base sur la distance entre les plongements des mots dans les phrases des deux langues. La méthode Vecalign a aussi été choisie en particulier parce qu'elle représente l'état de l'art dans le domaine de l'alignement des phrases. Moore et Gale & Church ont été choisies, malgré leur ancienneté, car elles nécessitent moins de données que les méthodes neuronales (telle Vecalign), ce qui est très pertinent dans notre cas. C'est Moore qui a obtenu les meilleurs résultats d'alignement dans l'étude de Joanis *et al.* (2020), sur l'anglais et l'inuktitut (autre langue autochtone polysynthétique). Dans le cas de Gale & Church, la quantité de données n'importe pas, ce qui peut être intéressant dans un cas comme le nôtre, où une très faible quantité de texte est disponible.

Parmi les méthodes présentées dans l'état de l'art (chapitre 2), on peut identifier trois catégories de méthodes d'alignements définies selon les types d'opérations d'alignement permises. La première catégorie ne permet que les alignements de type 1-1 et 1-2 (la méthode Gale & Church en fait partie), la seconde permet les alignements de type 1-1, 1-2 et les suppressions (la méthode de Moore en fait partie) et la troisième permet les alignements de type 1-1, 1-N, N-N et les suppressions (la méthode Vecalign en fait partie).

Ainsi les trois méthodes choisies, en plus de représenter des modes de fonctionnement différents, représentent chacune une étendue différente de types d'alignements possibles.

4.1.1 Méthode Vecalign

Tel que rapporté à la section **??**, la méthode Vecalign se base sur un modèle de plongements multilingues. La publication d'origine utilise le modèle multilingue LASER (Artetxe et Schwenk, 2019). Ce modèle n'inclut

toutefois pas l'innu-aimun. Il a donc été nécessaire de former un modèle de plongements bilingues sur la base des phrases innues des textes ici étudiés. La méthode choisie pour ce faire est le modèle Bi-Sent2Vec (Sabet *et al.*, 2020), de façon similaire à Joanis *et al.* (2020), qui ont utilisé cette méthode pour entraînement un modèle de plongements bilingues inuktitut-anglais afin de tester la méthode Vecalign sur le corpus du Nunavut Hansard.

La méthode Vecalign permet la configuration de plusieurs paramètres, susceptibles d'influencer les résultats d'alignement. Il est possible de configurer le type d'opération de fusion permise (le nombre de phrases maximum qu'il est possible de fusionner dans une seule opération) et la pénalité que donne l'algorithme à la suppression de phrases. À ces paramètres s'ajoute la taille des plongements entraînés par le modèle Bi-Sent2Vec.

4.1.2 Méthode de Moore

Pour rappel, la méthode de Moore recherche les paires de phrases pour lesquelles la probabilité de correspondance est la plus élevée, en se basant à la fois sur une comparaison de la longueur des phrases et sur des correspondances lexicales dérivées à partir du texte bilingue qu'on fournit à l'algorithme. Les alignements de phrases sont conservés seulement si la probabilité en question est au-delà d'un certain seuil. Ce seuil est configurable.

4.1.3 Méthode de Gale & Church

La méthode de Gale and Church, qui se base uniquement sur une distribution probabilistique interne, n'a quant à elle aucun paramètre qui puisse être ajusté. Toutefois, son fonctionnement nécessite de spécifier des séparateurs au sein du document à aligner. Ces séparateurs, qui sont des frontières au-delà desquelles les phrases ne peuvent être alignées, représentent typiquement des paragraphes. L'utilisation de Gale & Church dépend donc soit d'un certain traitement du texte en amont, que celui-ci manuel, automatisé ou semi-automatisé, afin d'identifier des portions de textes dont on est confiants de la correspondance entre les langues. Il doit y avoir le même nombre de séparateurs pour les deux langues.

4.1.4 Comparaison à une méthode naïve

Pour pouvoir réellement mesurer l'efficacité des méthodes d'alignement sur les corpus, nous proposons de comparer les résultats de ces dernières à une méthode que nous appelons naïve. Cette méthode effectue un alignement basé uniquement sur l'ordre des phrases tel qu'il l'est dans le texte d'origine. Si un texte comporte plus de phrases que l'autre, les phrases en trop à la fin seront tout simplement supprimées. Si les phrases d'un corpus sont déjà bien alignées à 100%, alors l'efficacité de la méthode naïve sera de 100%. Toutefois, dès qu'une phrase ne correspond pas à une autre, un décalage s'opérera pour le reste du corpus.

Formant en quelque sorte un groupe contrôle, les résultats de cette méthode permettent de distinguer l'efficacité due à un réel apport d'une méthode d'alignement de celle qui n'est due qu'à la chance, ou qui seraient basée sur la qualité d'alignement qui existe déjà à même le corpus.

4.1.5 Évaluation

Deux types d'évaluation des méthodes sont effectuées : l'une pour l'alignement du texte en entier et l'autre pour l'alignement du texte paragraphe par paragraphe. Le second type d'évaluation nécessite une séparation et un alignement préalable des paragraphes.

Trois mesures sont utilisées pour évaluer les différentes méthodes d'alignement : la précision et le rappel, mesures classiques en apprentissage statistique (James *et al.*, 2014), ainsi que la mesure F1, qui constitue une synthèse de ces deux dernières (Geron, 2019). Nous présentons pour rappel le calcul de ces mesures aux équations 4.1 à 4.3. TP (*True Positives*) symbolise le nombre de vrais positifs, FP (*False Positives*) symbolise le nombre de faux positifs et FN (*False Negatives*) symbolise le nombre de faux négatifs.

$$Precision = TP/(TP + FP) \tag{4.1}$$

$$Rappel = TP/(TP + FN) \tag{4.2}$$

$$F1 = 2 * (Precision * Rappel)/(Precision + Rappel) \tag{4.3}$$

Dans notre cas, le nombre de vrais positifs constitue le nombre de phrases alignées du jeu de référence

qu'on retrouve bel et bien dans les phrases alignées hypothétisées (celles que l'on évalue). Le nombre de faux positifs constitue le nombre de phrases alignées hypothétisées que l'on ne retrouve pas dans le jeux de référence. Et le nombre de faux négatifs constitue le nombre de phrases alignées du jeu de référence que l'on ne retrouve pas parmi les phrases alignées hypothétisées.

4.2 Textes bilingues et jeux de référence utilisés pour l'étude

Nous présentons dans la section qui suit les oeuvres bilingues choisies pour construire les jeux de référence et le processus d'extraction et de formattage qui a été utilisé pour en extraire les phrases. Nous analysons ensuite les corpus alignés de référence qui ont résulté de la collaboration avec le personnel enseignant de l'école Kanatamat et de la participation des étudiants en traduction et interprétation du Cégep de Sept-Îles (voir chapitre 3).

4.2.1 Présentation des textes utilisés pour l'étude d'alignement

Trois oeuvres publiées en éditions bilingues ont été choisies pour la présente étude. Les deux premiers, *Eukuan nin matshi-manitu innu-ishkueu* (Kapesh, 2019) et *Tanite nene etutamin nitassi ?* (Kapesh, 2020), sont de l'auteure Innue An Antane Kapesh. La troisième oeuvre choisie, *Nin auass* (Bacon et Morali, 2021), est un recueil de poésie participatif, ayant mis à contribution des jeunes de niveau primaire et secondaire provenant de différentes école innues.

Les deux textes d'An Atane Kapesh, parus pour la première fois avec leur traduction française dans les années 1970, ont été réédités plusieurs fois, notamment pour utiliser l'orthographe innue uniformisée. La version utilisée dans le cadre de la présente étude est celle publiée par la maison d'édition Mémoire d'encrier [1]. Le premier texte, *Eukuan nin matshi-manitu innu-ishkueu*, ou *Je suis une maudite sauvagesse*, est un essai de 217 pages (incluant les deux langues) portant sur les injustices vécues par la communauté de l'auteure. Il relate des mémoires à la fois autobiographiques et familiales, et son le style se veut réaliste et revendicateur. Le second, *Eukuan nin matshi-manitu innu-ishkueu* ou *Qu'as-tu fait de mon pays ?*, fait 90 pages et est écrit plutôt comme un conte ayant comme thème le colonialisme et la dépossession subis par les autochtones de manière générale.

Nin auass, ou *Moi l'enfant*, est un recueil de 363 pages co-édité par Mémoire d'encrier et l'Institut Tsha-

[1] An Antane Kapesh - Mémoire d'encrier

kapesh [2], contenant 176 courts poèmes présentés en Innu-Aimun et en Français. Portant sur des thèmes variés, mais pour la plupart reliés à l'identité innue, ils ont été rédigés par des jeunes sous la supervision de poétesses Innues.

Ce sont des textes provenant de deux domaines, soit la littérature et la poésie, dont il est intéressant d'étudier l'utilité pédagogique dans le cadre à la fois d'activités collaboratives et participatives dans des milieux éducatifs innus. Par ailleurs, il s'agit de textes rédigés par des Innus, portant sur des thématiques centrées sur l'identité et l'histoire de ce peuple. Dans l'esprit de développer des outils de langage représentatifs de la culture et de l'identité innues, il donc est très pertinent de se baser sur ces textes pour notre étude.

4.2.2 Processus de traitement des textes

Les étapes suivantes, automatisées par scripts, ont été mises en oeuvre pour traiter les textes choisis :

1. Retrait des éléments non nécessaires dans les textes (en-têtes, numéros de pages, etc.) et retrait des caractères spéciaux (par exemple, puces) via expressions régulières
2. Formattage des paragraphes et phrases via expressions régulières : recollage des mots scindés par trait d'union en fin de ligne, recollage des phrases scindées par un retour chariot, suppression des retours multiples, etc.
3. Standardisation de certains caractères qui ne le sont pas (par exemple, points de suspension).

En plus du formattage du texte, une étape supplémentaire a dû être mise en oeuvre pour distinguer les « u exposant » (caractère spécial innu) du « u » conventionnel. En effet, dans les textes extraits, les « u exposant » étaient représentés par des « u » avec un formattage en exposant, plutôt qu'avec le caractère dédié (unicode U+1D58 [3]). La logique suivante a donc été mise en oeuvre, pour remplacer les « u conventionnels » qui étaient sensés être des « u exposant » dans le texte publié :

1. En se basant sur les règles grammaticales de l'innu-aimun (Drapeau, 2014), identifier les mots qui contienne un « u » à un endroit où un « u exposant » serait grammaticalement correct. Cet endroit est à la toute fin du mot, après l'une ou l'autre des consonnes suivantes : « k », « sh », « m » ou « t ».

[2] Nin Auass • Moi l'enfant, Mémoire d'encrier

[3] Comment trouver le u exposant unicode dans mon ordinateur? - Innu-aimun.ca

2. Rechercher les mots précédemment identifiés parmi ceux du dictionnaire en ligne (Ambroise *et al.*, 2023), afin de vérifier s'ils existent tels qu'ils sont écrits (c'est-à-dire avec une terminaison en « u » conventionnel). S'il n'existent pas tels quels, vérifier s'ils existent plutôt avec une terminaison en « u exposant » et, si c'est le cas, en remplacer le « u » conventionnel par un « u exposant ».

Il est possible qu'après avoir mis en oeuvre la logique qui précède, certains caractères « u exposant » soient encore représentés dans le texte comme des « u » conventionnels. Sans une relecture et un remplacement manuels exhaustifs, il est impossible d'en garantir totalement l'absence.

Pour effectuer l'évaluation des alignements paragraphe par paragraphe (voir section 4.1.5), les paragraphes ont d'abord été séparés automatiquement via les règles décrites ci-haut basées sur le formattage du texte original, puis cette séparation a été corrigée manuellement sur la base d'une inspection visuelle et d'une consultation du dictionnaire innu (Ambroise *et al.*, 2023). La séparation a ensuite été validée durant la création de l'alignement manuel de référence (voir chapitre 3).

Lorsqu'on traite des texte bilingues, la quantité de phrases est parfois beaucoup plus élevée que celle que nous avions à notre portée. De plus, l'expertise de personnes maîtrisant la langue n'est pas toujours à portée de main. Ce genre de travail de séparation et validation des paragraphes n'est pas toujours possible, ou du moins pas avec la qualité qui a pu être obtenue ici, et il faut en prendre compte dans l'analyse des résultats sur le corpus séparé par paragraphe. Toutefois, cette manière de procéder devrait être considérée comme pertinente dans le cas de langues pour lesquelles très peu de phrases sont disponibles, comme l'innu-aimun, puisque la quantité de phrases à traiter est petite et que le temps exigé pour faire cette séparation semi-automatisée reste réaliste dans le cadre d'un projet comme le nôtre. Cette possibilité pourrait être considérée comme l'envers de la médaille du manque de phrases : bien que les méthodes basées sur l'apprentissage profond puissent être moins performantes que sur des langues mieux dotées, au moins la faible quantité de phrases permet l'utilisation de méthodes qui ne sont que partiellement automatisées.

4.2.3 Corpus alignés d'An Antane Kapesh

Le tableau 4.1 présente une analyse comparative des deux corpus alignés basés sur les livres d'An Antane Kapesh, soit *Eukuan nin matshi-manitu innu-ishkueu* Kapesh (2019) (identifié comme **kapesh-1** dans le tableau et dans le texte qui suit et pour la suite du mémoire) et *Tanite nene etutamin nitassi ?* Kapesh (2020)

(identifié comme **kapesh-2**). Ceux-ci sont le résultat d'un alignement manuel, fruit de la participation des étudiants en traduction du Cégep de Sept-Îles (voir chapitre 3).

TABLE 4.1 – Analyse comparative des deux corpus alignés basés sur les livres d'An Antane Kapesh

	kapesh-1	**kapesh-2**	**Combinés**
nb de paragraphes	163	149	312
nb de phrases	796	484	1280
% alignments non-standards	14.7%	11.4%	13.1%
nb de mots moyen par phrase (ratio français/innu-aimun)	1.46	1.41	1.44
nb de characatères moyen par mot (ratio innu-aimun/français)	1.63	1.58	1.62

Il est intéressant de noter que les tailles de vocabulaire en innu-aimun et en français reflètent bien la nature polysynthétique de la langue innue. Il y a en moyenne un plus grand nombre de mots par phrase en français qu'en innu-aimun, et un plus grand nombre de caractères par mot en innu-aimun qu'en français. Par ailleurs, en combinant les deux corpus de Kapesh, le nombre total de mots différents réduit en français comme en innu-aimun (puisque les deux corpus ont certains mots en commun), mais cette réduction est beaucoup plus importante en français qu'en innu-aimun. Le pourcentage de vocabulaire commun entre les deux corpus est aussi plus grand en français qu'en innu-aimun, tel montré dans le tableau 4.2

TABLE 4.2 – Vocabulaire en commun entre kapesh-1 et kapesh-2

Langue	**% de vocabulaire en commun**
Innu-Aimun	10,8 %
Français	23,2 %

Ceci correspond bien au fait que les mots innus ont tendance à être plus long et à s'infléchir de nombreuses façons, pour finalement correspondre à l'équivalent de plusieurs mots en français (Drapeau, 2014).

Le tableau 4.3 présente la compatibilité des alignements de référence avec les différentes méthodes d'alignement existantes. Les alignements considérés comme « non-standards » sont des alignements qui ne peuvent être effectués par les méthodes d'alignement de phrases existantes conventionnelles. On considère ici que les

méthodes d'alignement conventionnelles sont celles qui permettent uniquement des opérations de fusions ou de suppressions de phrases voisines. Les alignements qui sortent de ce cadre ne peuvent donc pas être effectués par ces méthodes. Par exemple, certains participants ont identifié la nécessité de scinder des phrases en portions afin qu'elles puissent être alignées correctement à leur contre-parties dans l'autre langue. Le tableau 4.3 présente une exemple de cas considéré non-standard où les portions de textes marquées rouge et noir coincident respectivement dans les deux langue. Dans cet exemple, il a été considéré comme nécessaire par la personne alignant le texte de scinder la première phrase d'innu-aimun en deux, puisque la seule manière correcte d'aligner les deux paires est de joindre la deuxième phrase d'innu-aimun avec la seconde moitié (en rouge) de la première.

TABLE 4.3 – Exemple d'alignement « non-standard »

	Innu-aimun	**Français**
#1	« Ne Kauitenitakusht katshi minikut nenua auassa tat^{u} shuniau-aueshisha, ekue mishta-papit ekue itenitak : « Apu nita tshika ut tshiueian ute katshi takushinian. » (Kapesh, 2020)	« Après avoir reçu de l'enfant les fourrures, le Polichinelle éclate de rire. » (Kapesh, 2020)
#2	« Ekute ute tshe ut uenutishian ! » itenitam^{u}. » (Kapesh, 2020)	« « Maintenant que je suis venu ici, jamais je ne m'en irai, c'est ici que je vais faire fortune ! » se dit-il. » (Kapesh, 2020)

Dans d'autres cas, des participants ont identifié la nécessité d'aligner des phrases distantes dans le texte, parfois en réordonnant l'ordre des phrases. Au-delà de ces alignements non-standards, le tableau 4.4 identifie le pourcentage d'alignements qui ne pourraient pas être effectué par la méthode de Moore (par exemple, une correspondance 1-N où N est supérieur à 2) ou pas la méthode de Gale & Church (par exemple, une suppression de phrases).

On peut voir que, dans le cas des corpus kapesh-1 et kapesh-2, la plupart des alignements proposés par les participants pourraient être faits par les méthodes de Moore ou de Gale & Church (ou méthodes analogues), et que la proportion d'alignements non-standards est très mince (moins de 3%).

TABLE 4.4 – Alignements non-standards ou non-compatibles avec certaines méthodes dans les corpus alignés

Type de compatibilité	**Kapesh-1**	**Kapesh-2**	**Combinés**
% alignments non-standards	2,45%	2,01%	2,24%
% alignments non-compatibles avec Moore	9,82%	10,1%	9,94%
% alignments non-compatibles avec Gale & Church	14,7%	11,4%	13,1%

4.2.4 Échantillon aligné du recueil de poésie jeunesse

Le tableau 4.5 présente les caractéristiques sommaires de l'alignement de référence produit pour un échantillon de 44 poèmes jeunesse, tirés du recueil *Nin Auass*. Cet échantillon, qui représente 25% des poèmes complets du recueil, a été aligné manuellement en collaboration avec le personnel enseignant de l'école Kanatamat de la communauté de Matimeksuh et avec la participation des étudiants en traduction d'innu-aimun du Cégep de Sept-Îles (voir chapitre 3).

TABLE 4.5 – Caractéristiques de l'échantillon de poésie jeunesse

Caractéristiques	
nb de paragraphes (poèmes)	44
nb de phrases (vers)	339
nb de mots moyen par phrase (ratio Français/Innu-Aimun)	1.7
nb de characatères moyen par mot (ratio Innu-Aimun/Français)	1.87

Les ratios du tableau 4.5 sont similaire à ceux observés pour kapesh-1 et kapesh-2, mais sont légèrement plus élevés, ce qui signale une utilisation d'un plus grand nombre de courts mots français pour équivaloir les mots d'innu-aimun. De l'avis des participants à l'alignement de référence, la traduction des poèmes n'est pas toujours exacte entre l'innu-aimun et le français. Souvent, la qualité artistique du poème est priorisée, au détriment de la précision ou de la rigueur. Par exemple, dans certains poèmes, on préfère obtenir des rimes plutôt que de garder une correspondance exacte entre chaque vers, quitte à utiliser des mots un peu différents d'une langue à l'autre ou à inverser l'ordre de certains vers.

Le tableau 4.6 présente les pourcentages d'alignement non-standards, ainsi que non-compatibles avec les méthodes de Moore et de Gale & Church, parmi ceux de l'échantillon de poèmes jeunesse. On remarque que ces pourcentages sont significativement plus élevés que pour kapesh-1 et kapesh-2, ce qui fait écho à liberté de traduction qu'on remarquée les participants à l'alignement manuel.

TABLE 4.6 – Alignements non-standards ou non-compatibles dans les corpus alignés

Type de compatibilité	**Proportion d'alignements**
% alignments non-standards	9.1%
% alignments non-compatibles avec Moore	20.5%
% alignments non-compatibles avec Gale & Church	25%

4.3 Résultats d'alignement

La section qui suit présente l'évaluation des résultats d'alignement obtenus avec les trois méthodes d'alignement choisies, en comparaison aux corpus de référence (kapesh-1, kapesh-2 et l'échantillon de poésie jeunesse), ainsi que l'exploration des paramètres ayant mené à ces résultats.

4.3.1 Résultats

Les tableau 4.8 et 4.7 présentent une comparaison des résultats d'alignement obtenus pour le corpus kapesh-2 à partir des méthodes de Gale and Church et de Moore, de la méthode Vecalign et de la méthode dite naïve. Le premier tableau présente les résultats d'alignements obtenus sur le texte en un seul bloc, alors que le second présente les résultats obtenus après une séparation préalable des paragraphes. Il est à noter que la méthode de Gale and Church ne s'applique qu'au second cas, puisqu'elle nécessite qu'on identifie des frontières entre paragraphes au sein du texte.

Tel qu'indiqué en gras dans les tableaux, c'est la méthode de Gale and Church qui performe le mieux lorsqu'on génère des alignements par paragraphe, alors que c'est la méthode de Moore qui performe le mieux pour le texte non-séparé. Toute catégorie confondue, la technique la plus performante est celle de Gale and Church avec un F1 de 0.9172. Toutefois, la technique de Moore a une efficacité qui s'y approche, avec un F1 de 0.7965, moins de 0.15 d'écart. Du côté de la précision, les résultats montrent Moore et Gale & Church encore plus près l'une de l'autre. Ainsi, s'il n'est pas possible de procéder à une séparation semi-

automatisée des paragraphes et de la valider par un expert de la langue, la méthode de Moore offre une performance qui est acceptable, si on peut accepter d'obtenir un rappel qui est autour de 75%.

TABLE 4.7 – Comparaison des résultats de référence sur le corpus kapesh-2 (avec séparation préalable des paragraphes)

Méthode	**Précision**	**Rappel**	**F1**
Moore	0.7586	0.6925	0.7232
Vecalign	0.6547	0.7072	0.6799
Gale and Church	0.8734	0.9656	0.9172
Naïve	0.6005	0.6805	0.6380

TABLE 4.8 – Comparaison des résultats de référence sur le corpus kapesh-2 (sans séparation préalable des paragraphes)

Méthode	**Précision**	**Rappel**	**F1**
Moore	0.8571	0.7438	0.7965
Vecalign	0.6199	0.6570	0.6379
Gale and Church	N/A	N/A	N/A
Naïve	0.0204	0.0227	0.0215

Les tableaux 4.9 et 4.10 présentent les résultats d'alignement obtenus avec les différentes méthodes, paragraphe par paragraphe et pour le texte complet, respectivement.

Si l'on compare les résultats de kapesh-1 à ceux de kapesh-2, on peut voir que l'ordre de performance des méthodes est le même autant pour l'évaluation paragraphe par paragraphe que pour celle sur le texte complet. Les résultats paragraphe par paragraphe montrent une performance légèrement meilleure des méthode Vecalign et Moore sur kapesh-1 vis-à-vis. Sur les résultats du texte complet, il y a une performance significativement supérieure de Moore sur kapesh-1 comparé à kapesh-2 (une différence de près de 0.10). Cela peut probablement s'expliquer par le fait que le corpus kapesh-1 est deux fois plus volumineux que le corpus kapesh-2, ce qui est favorable à la méthode statistique utilisée par Moore. Autre observation : l'écart entre la méthode naïve et les méthodes de Moore et Vecalign est plus important pour l'alignement paragraphe par pagraphe de kapesh-1 que pour kapesh-2, et l'écart entre les performances de Vecalign et de Moore et celle de Gale & Church est plus petit sur kapesh-1. Encore une fois, ces différences sont probablement dues

au plus important volume de phrase dans kapesh-1 que dans kapesh-2, qui est plus favorable aux méthodes Vecalign et Moore qu'à celle de Gale & Church.

TABLE 4.9 – Comparaison des résultats de référence sur le corpus kapesh-1 (avec séparation préalable des paragraphes)

Méthode	**Précision**	**Rappel**	**F1**
Moore	0.9249	0.6356	0.7534
Vecalign	0.7411	0.6597	0.6980
Gale and Church	0.8856	0.8836	0.8846
Naïve	0.7032	0.6059	0.6509

TABLE 4.10 – Comparaison des résultats de référence sur le corpus kapesh-1 (sans séparation préalable des paragraphes)

Méthode	**Précision**	**Rappel**	**F1**
Moore	0.9392	0.8485	0.8916
Vecalign	0.1280	0.6570	0.6379
Gale and Church	N/A	N/A	N/A
Naïve	0.0300	0.0318	0.0309

Le tableau 4.11 présente les résultats d'alignement obtenus paragraphe par paragraphe sur les poèmes jeunesses (autrement dit poème par poème), évalués en comparaison à l'échantillon de référence précédemment présenté. À noter : seul les résultats par paragraphe sont présentés pour les poèmes jeunesse, puisque ces poèmes sont déjà séparés et alignés à même le recueil. Il serait artificiel et futile de générer et évaluer un alignement pour le texte complet, sans séparation.

On peut noter ici que ce sont à nouveau les méthodes de Gale & Church de Moore qui obtiennent les meilleurs résultats, dans cet ordre. Toutefois, probablement à cause de la très petite quantité de paires de phrases dans l'échantillon, la performance de Vecalign est moins bonne que celle de la méthode naïve.

4.3.2 Exploration des paramètres

Les résultats présentés dans les tableaux de la section précédente représentent les meilleurs scores F1 obtenus après une exploration des paramètres pour chacune des méthodes, sur le corpus kapesh-2. Ce sont ensuite

TABLE 4.11 – Comparaison des résultats de référence sur l'échantillon du corpus jeunesse (alignement séparé pour chaque poème)

Méthode	**Précision**	**Rappel**	**F1**
Moore	0.7836	0.7050	0.7422
Vecalign	0.6977	0.4425	0.5415
Gale and Church	0.7486	0.8083	0.7773
Naïve	0.6677	0.6342	0.6505

ces paramètres qui ont été utilisés pour obtenir les résultats présentés au tableau 4.8. Cette exploration des paramètres est présentée ci-après.

Le tableau 4.12 présente les résultats de référence obtenus avec la méthode de Moore, pour différents seuils de probabilités. Le résultat considéré comme le meilleur et conservé pour la comparaison est celui avec un seuil de 0.1 (en gras).

TABLE 4.12 – Résultats avec la méthode de Moore pour différents seuils sur le corpus Kapesh-2 (séparation manuelle des paragraphes)

Seuil	**Précision**	**Rappel**	**F1**
0.1	**0.7586**	**0.6925**	**0.7232**
0.2	0.7637	0.6495	0.7020
0.3	0.7637	0.6495	0.7020
0.4	0.7637	0.6495	0.7020
0.5	0.7606	0.6092	0.6765
0.6	0.7554	0.6010	0.6694
0.7	0.7633	0.6293	0.6898
0.8	0.7590	0.6146	0.6792
0.9	0.7485	0.5980	0.6649

Le tableau 4.13 présente les résultats de référence obtenus avec la méthode Vecalign, pour différentes dimensions de plongements bilingues français/innu-aimun. Le résultat considéré comme le meilleur et conservé pour la comparaison est celui avec une dimension de 50 (en gras).

TABLE 4.13 – Résultats avec la méthode Vecalign pour différentes dimensions de plongements (avec la méthode Bi-Sent2Vec) sur le corpus Kapesh-2 (séparation manuelle des paragraphes)

Dimension des plongements	**Précision**	**Rappel**	**F1**
25	0.6256	0.7278	0.6728
50	**0.6547**	**0.7072**	**0.6799**
100	0.6848	0.6224	0.6521
200	0.6829	0.5744	0.6240
400	0.6779	0.5696	0.6190
800	0.6890	0.5765	0.6278
1600	0.6829	0.5744	0.6240

Le tableau 4.14 résume l'exploration des différentes configurations (nombre de fusions permises par alignement et pénalité à la suppression d'une phrase) avec les performances correspondantes, obtenues sur kapesh-2 paragraphe par paragraphe. On peut y voir que ce sont les configurations les moins permissives qui sont les plus performantes dans notre contexte. Ceci handicape de manière importante la méthode Vecalign, par rapport à la variété d'alignements qu'elle peut normalement effectuer.

TABLE 4.14 – Résultats avec la méthode de Vecalign pour différentes configurations sur le corpus Kapesh-2 (séparation manuelle des paragraphes)

Nombre de fusions permises	**Pénalité à la suppression**	**Précision**	**Rappel**	**F1**
1	**1**	**0.6073**	**0.5628**	**0.6595**
2	1	0.518	0.456	0.4861
2	0.2	0.2616	0.2006	0.2271
3	1	0.514	0.4183	0.4613
N	0.2	0.1595	0.09202	0.1167

4.4 Analyse et discussion

La section qui suit présente une analyse par point de discussion, selon les principales observations qui peuvent être faites suivant la présentation des résultats. D'abord, la sous-section 4.4.1 discute du fait que les

méthodes d'alignement dites classiques (Moore, Gale & Church) surperforment par rapport à la méthode considérée comme l'état de l'art (Vecalign). Ensuite, la sous-section 4.4.2 discute de ce qu'on peut tirer comme réflexion des bonnes performances de la méthode de Moore. Enfin, la sous-section 4.4.3 discute l'identification de ce qui doit être considéré comme la meilleure méthode d'alignement, selon le contexte.

4.4.1 La surperformance des méthodes classiques

La première et principale conclusion que l'on peut tirer de notre étude comparative est le fait les méthodes plus anciennes de Moore et de Gale & Church offre une bien meilleure performance que celle, plus récente et plus avancée, de Vecalign.

Pour les deux corpus et dans les deux types d'alignement (avec ou sans séparation préalable des paragraphes) la performance de Vecalign est bien en-deçà des meilleurs résultats. D'ailleurs, ni la précision, ni le rappel ne surpassent les 70%, dans tous les cas. Les résultats montrent même que Vecalign est presque inutile lorsque utilisée après séparation des paragraphes : par exemple, les scores F1 obtenus par Vecalign sur kapesh-1 et kapesh-2 s'approchent de de ceux obtenus par la méthode naïve. Si les résultats sans séparation préalable démontrent tout de même une grande démarcation par rapport par rapport à la méthode naïve (par exemple, F1 de 0.6379 plutôt que 0.0215 sur kapesh-2), l'écart reste significatif face à aux autres méthodes.

Pourquoi une telle sous-performance de la méthode qui est considérée comme l'état de l'art ? D'abord il s'agit d'une méthode basée sur les plongements de mots, qui normalement nécessitent une grande quantité de donnée. Puisque ce n'est pas notre cas, il n'y a pas de surprise ici. C'est plutôt l'écart avec Moore qui est intéressant, car cette dernière nécessite aussi une certaine quantité de données. C'est donc dire qu'une méthode statistique comme Moore peut bien performer dans un contexte avec une faible quantité de données, quantité qui serait insuffisante pour profiter d'une méthode neuronale comme Vecalign.

4.4.2 Ce qui ressort de la bonne performance de Moore

En plus des analyses présentées précédemment, deux considérations supplémentaires font en sorte que les résultats de Moore sont d'autant plus intéressants.

La première considération est que les résultats de la méthode Moore démontrent qu'elle conserve une efficacité intéressante même dans le cas de très petites collections de phrases (un texte complet d'environ 500

phrases, dans le cas de kapesh-2 par exemple) et ce, même si dans les notes d'usage de cette méthode on précise qu'elle n'a été testée que sur des textes de plus de 10000 phrases. Nous avons donc pu tester en quelque sorte un cas extrême de l'utilisation de Moore, cas qui confirme que la méthode reste efficace même lorsqu'on a affaire à de très petits textes dans une paire de langues peu dotée.

La seconde considération est que les résultats de Moore sur le texte complet et paragraphe par paragraphe permettent de comparer l'efficacité de l'alignement interne des paragraphes (effectué automatiquement par la méthode) à celle de alignement manuel des paragraphes. En effet, pour aligner les phrases, la méthode de Moore identifie d'abord des points d'ancrages, au-delà desquels la méthode ne se permet pas d'aligner des phrases. Ces points d'ancrages sont en quelque sorte l'équivalent des *hard-delimiters* de la méthode de Gale & Church, qui sépareraient des paragraphes. Or, on peut voir en comparant les résultats paragraphe par paragraphe à ceux pour le texte complet que séparer manuellement les paragraphes n'aide pas Moore à mieux performer. En effet, la méthode obtient de meilleurs résultats sans séparation préalable des paragraphes. C'est donc dire que les séparations effectuées à l'interne par la méthode, même si ces séparations n'ont peut-être pas rapport avec les paragraphes d'origines à proprement parler, sont plus utiles à l'algorithme que celles forcées par l'utilisateur.

4.4.3 La meilleure méthode selon le contexte et l'objectif

La comparaison des meilleurs résultats d'alignement obtenus toutes catégories confondues, c'est-à-dire ceux obtenus par la méthode de Gale & Church sur les paragraphes préalablement séparés, aux meilleurs résultats obtenus sur un texte complet est pertinente pour guider le choix d'une méthode selon le contexte et selon l'objectif. En effet, dans un contexte où l'on a peu de phrases disponibles, mais où l'on a un certain accès à des experts pouvant préalablement valider un alignement des paragraphes dans les textes disponibles, il est utile de savoir que c'est la méthode de Gale and Church qui pourra offrir la plus grande qualité d'alignement. Cela pourrait aussi s'appliquer à un contexte où l'on a confiance en notre capacité d'obtenir un alignement des paragraphes de grande qualité, de manière automatisée ou semi-automatisée. Toutefois, lorsque cette expertise n'est pas du tout disponible, les résultats présentés ici nous montrent que la meilleure méthode à utiliser est celle de Moore.

Par ailleurs, cette comparaison entre alignement par paragraphe et sur texte complet met en lumière un autre dilemme qu'on peut avoir dans le développerment de la traduction automatique pour langues peu dotées. Est-

il mieux de favoriser l'obtention d'un alignement de grand qualité ou alors l'obtention d'une plus grande quantité de phrases alignées ? La question se pose surtout sachant que tous les textes disponibles ne pourront pas nécessairement se prêter à un travail de division préalable des paragraphes. On a aussi pu constater que la précision des alignements peut varier d'un texte à l'autre, de par leur nature.

La réponse dépendra de l'usage recherché pour ces alignements. Si l'usage recherché est le développement d'un outil d'assistance comparable à des mémoires de traduction [4], où la précision de la traduction fournie par l'outil est importante pour que celui-ci soit utile, alors il est clair qu'il est préférable de favoriser la qualité d'alignement. Toutefois, dans le cas du développement d'un système de traduction automatique, qui nécessite de grande quantités de phrases et qui ne fournira pas directement à l'usager les traductions sur lesquelles il est entraîné, la réponse est moins claire. Le chapitre 5 tentera d'y répondre, notamment en évaluant l'impact de différentes qualités d'alignement sur les résultats de traduction.

4.5 Conclusion et perspectives

On peut conclure de cette étude comparative qu'elle a surtout su démontrer les limites des méthodes neuronales et la surprenante efficacité de méthodes classiques dans des cas où la quantité de phrases disponibles est très faible. Cette étude a par ailleurs permis d'identifier quelles méthodes d'alignement utiliser selon le cas, pour les futurs textes qui devront être alignés dans la recherche de davantage de paires de phrases, pour la traduction automatique notamment.

Pour pousser davantage notre étude ou pour tenter d'obtenir de meilleurs résultats d'alignement, plusieurs perspectives s'offrent à nous. D'abord, plutôt qu'utiliser le modèle statistique d'origine utilisé par Moore dans leur outil de base (l'IBM Translation Model 1), il pourrait être intéressant de tester d'autres types de modèles statistiques, plus récents. Ensuite, dans le cas de Vecalign, pour tenter d'obtenir de meilleures performances, il serait intéressant de tester des méthodes de plongements multilingues qui permettent de profiter des données disponibles pour d'autres paires de langues mieux dotées. Par exemple, la méthode *teacher-student* de Heffernan *et al.* (2022) pourrait possiblement permettre à l'innu-aimun de profiter d'un certain transfert depuis un modèle multilingue. Peut-être que cela permettrait à Vecalign de mieux profiter des bénéfices d'une représentation vectorielle multilingue comme celle du modèle LASER, utilisé à l'origine par les auteurs.

[4] Note : le site web Linguee est un exemple d'un tel outil

CHAPITRE 5
ÉTUDE DE FAISABILITÉ POUR LA TRADUCTION AUTOMATIQUE

Le présent chapitre a pour objectif d'étudier la faisabilité de développer la traduction automatique pour l'innu-aimun, avec les textes accessibles et les techniques existantes.

Cette question de faisabilité va bien au-delà d'une simple réponse de type oui ou non. D'abord, elle est centrale à la recherche d'une réponse aux questions plus larges posées à l'introduction de ce mémoire. Pour rappel, ces questions sont les suivantes. Est-il possible, avec les textes disponibles et les techniques existantes, de développer des outils d'aide à l'apprentissage et à la traduction de l'innu-aimun basés sur le traitement automatique du langage (TAL) ? Ou, plus précisémment, quelle est l'efficacité des techniques TAL existantes sur les textes aujourd'hui disponibles pour l'innu-aimun et comment pourrait-on améliorer cette efficacité ?

Entraîner un modèle de traduction automatique avec les textes bilingues disponibles est une première étape qui peut nous fournir une indication sur la possibilité qu'un traducteur automatique voit le jour pour l'innu-aimun ou sur l'effort qui est nécessaire pour y arriver. Or, bien que ce type d'outil puisse potentiellement aider les traducteurs, il ne s'agit pas de l'unique outil TAL que l'on pourrait développer grâce aux données bilingues disponibles. Par exemple, ayant à disposition un modèle de traduction automatique neuronale (TAN), mais sans nécessairement mettre ce dernier à disposition direct d'usagers (si l'on n'est pas assez confiant dans la qualité de ses traductions, par exemple), il pourrait toujours être possible d'utiliser ce dernier dans le cadre d'un système de recherche d'information interlingue. Entre autres exemples, Bi *et al.* (2020) ont mis à contribution un modèle de TAN pour traduire des requêtes de l'anglais vers le chinois, langue dans laquelle se trouve l'information recherchée. Dans le cas de l'innu-aimun, on pourrait imaginer des applications potentielles dans une recherche plus avancée des définitions au sein de la base de donnée du dictionnaire innu (Ambroise *et al.*, 2023).

En plus d'aborder la question de faisabilité, cette étude nous offre l'occasion d'obtenir de premiers résultats de base pour la traduction automatique de l'innu-aimun, qui serviront de référence que ce soit pour de futurs développement dans ce sens ou pour le développement d'autres outils de TAL pouvant potentiellement aussi bénéficier d'un apprentissage sur des corpus alignés. Plus que la seule évaluation quantitative des modèles de traduction, nous souhaitons profiter de cette occasion pour examiner en détail ce que permettent les données

à disposition et quel genre de traduction les modèles produisent.

La section 5.1 présente le cadre expérimental pour cette étude. La section 5.2 présente les résultats de référence obtenus par une approches de modèles de traduction neuronale. Cette approche utilise un modèle de type Transformer simple entraîné sur des textes bilingues innu-aimun/français. La section 5.3 présente les résultats d'essais utilisant l'approche de traduction statistique, entraînés sur les mêmes textes que pour les essais de traduction neuronale. À la lueur des résultats neuronaux et statistiques, la section 5.3 propose aussi une analyse plus détaillée et qualitative, en collaboration avec le personnel enseignant ayant participé à l'alignement des texte (voir chapitre 3). La section 5.4 présente une comparaison quantitative des deux méthodes de traduction automatique utilisées dans ce chapitre (neuronale et statistique) et évalue la significativité statistique des écarts observés entre les deux méthodes. La section 5.5 propose une synthèse des différents résultats obtenus, ainsi qu'une discussion sur la question de faisabilité posée au départ. La section 5.6 conclut avec une réflexion sur les limites de la présente étude et sur les différentes perspectives pour la suite des choses.

5.1 Cadre expérimental

5.1.1 Approche et modèle de base choisi

L'approche choisie pour l'étude de faisabilité est d'examiner s'il est possible de construire un corpus commun et général d'innu-aimun et de français aligné qui puisse servir à développer un modèle de traduction neuronal. Puisqu'il n'y a pas suffisamment de textes pour la paire innu-aimun et français pour effectuer une étude de généralisation, l'entraînement et l'évaluation du modèle sont tous deux effectués à même ce corpus général. Cette approche est analogue à celle utilisée par Joanis *et al.* (2020) pour obtenir les résultats de traduction basés sur le corpus du Nunavut Hansard (paire inuktitut-anglais). Elle permettra d'évaluer où en est la paire innu-aimun et français dans la construction d'un corpus qui permette le développement d'un traducteur automatique, outil qui est désormais disponible pour l'inuktitut (voir notre état de l'art, à la section 2.1).

Pour obtenir les résultats de traduction neuronale, c'est le modèle Transformer classique (Vaswani *et al.*, 2017) qui est choisi. Ce choix est fait pour deux raisons principales. La première est que cette méthode reste celle de référence lorsqu'on est dans un contexte d'une paire de langue individuelle, ce qui est notre cas puisque les langues les plus proches de l'innu-aimun ne comptent pas de données en parallèle de la

langue française. La seconde est qu'utiliser cette méthode permet d'établir des résulats de référence pour cette paire, sans apport de langues autres ou de méthode plus spécialisée. Ceci permet de les comparer aux résultats obtenus par d'autres paires de langues ayant des quantités de données plus ou moins importantes. Cela permet aussi de comparer à d'autres résultats éventuellement obtenus pour la paire grâce à d'autres techniques (la traduction statistique, présentée à la section **??** en est un exemple).

5.1.2 Pré-traitement et entraînement

Pour entraîner et évaluer le modèle, une portion des données représentant 15% du tout est mise de côté puis séparée en parts égales pour constituer des jeux de validation et d'évaluation. Le reste de données sert à l'entraînement. Ce schéma de division a été choisi en raison de la très faible quantité de données disponible, qui rend difficile la constitution de plus grands jeux d'évaluation ou de validation, sans trop nuire à l'entraînement du modèle.

Après division et avant de procéder à l'entraînement et l'évaluation, le prétraitement suivant est mis en oeuvre sur les paires de phrases des trois jeux de données :

1. Normalisation de la ponctuation
2. Suppression des caractères non-standards
3. Tokenisation
4. Segmentation
5. Binarisation

Pour les trois premières étapes (avant segmentation), ce sont les scripts de la librairie Moses qui sont utilisés [1].

Pour la segmentation, c'est la technique Byte-Pair-Encoding (BPE) de Sennrich *et al.* (2016b) qui est utilisée. Cette technique se base sur les combinaisons de caractères les plus fréquentes dans le texte pour constituer une liste de sous-mots, qui formera le vocabulaire d'entraînement du modèle. Pour ce faire, un modèle BPE commun doit d'abord être entraîné sur l'entièreté des données de la paire de langues, excepté celles du jeu

[1] Documentation de la librairie Moses (en anglais)

d'évaluation. L'entraînement et l'évaluation, tout comme l'étape de binarisation, sont effectués par le biais de la plateforme fairseq [2].

5.1.3 Évaluation

Les résultats sont évalués via le score BLEU (Papineni *et al.*, 2002), en utilisant l'implémentation SacreBLEU (Post, 2018), ainsi que via le score ChrF++ (Popović, 2015). Alors que le score BLEU compare les *n-grams* des mots entiers provenant des traductions de références à ceux des traductions générées par le modèle (voir chapitre 1 pour une définition plus complète du score BLEU), le score ChrF++ prend en compte dans sa comparaison les *n-grams* au niveau des caractères. La première méthode d'évaluation est la plus courante dans le domaine, alors que la seconde peut s'avérer particulièrement pertinente dans le cas où une des langues de la paire est à morphologie riche (comme l'innu-aimun). En effet, comme les mots sont typiquement constitués par aggrégation de différents morphèmes, un résultat où seulement une partie des morphèmes d'un mot sont corrects peut s'avérer tout de même intéressant. D'autant plus que, souvent, les mots en question peuvent être longs et équivaloir à plusieurs mots dans l'autre langue de la paire.

Les deux scores utilisés pour l'évaluation (BLEU et Chrf++) fournissent une mesure entre 0 et 100. Dans le cas du BLEU, 100 est le score pour une phrase correspondant parfaitement à la traduction de référence. Sachant que typiquement plusieurs traductions peuvent être considérées valides pour une seule et même phrase, l'atteinte d'un score parfait n'est pas l'objectif recherché.

5.2 Essais de traduction neuronale

La section qui suit présente les résultats d'essais de traduction automatique neuronale pour la paire innu-aimun/français. Ces essais sont effectués en utilisant des corpus basés sur des textes bilingues, les mêmes ayant été utilisés dans l'étude comparative des méthodes d'alignement, au chapitre 4.

La capacité à entraîner un modèle de traduction automatique neuronale avec chacun de ces corpus est d'abord évaluée individuellement. Puisque le domaine et la qualité d'alignement differt d'un corpus à l'autre, nous pourrons examiner les résultats vis-à-vis ces aspects. Puis, l'entraînement et l'évaluation sont effectuées pour ces corpus combinés ensemble, dans la logique d'une construction d'un corpus commun général.

[2] Documentation de la librarie Fairseq (en anglais)

Les corpus utilisés pour les essais qui suivent proviennent de textes culturels innus, respectivement des textes littéraires et poétiques. Les essais invididuels et combinés permettent donc de répondre à la question suivante : peut-on constituer un corpus dédié à la traduction automatique grâce à de textes culturels, et ce malgré leur style particulier et leur traduction libre ? Cette question est pertinente dans le sens où ce type de textes peut parfois être l'une des seules sources de données bilingues pour une langue autochtone qui est peu dotée, comme l'innu-aimun.

5.2.1 Corpus utilisés

Deux différents corpus alignés sont utilisés dans de le cadre de l'étude, de façon individuelle ou combinée [3]. Le tableau 5.1 présente tous ces corpus, avec leur taille et domaine respectifs.

TABLE 5.1 – Taille et domaines des corpus pour la traduction automatique

Corpus	**Domaine**	**Nombre de phrases (avant division)**
kapesh	Roman/Essai	1280
jeunesse	Poèmes jeunesse	1907

Le premier corpus est constitué de livres d'An Antane Kapesh, dont les phrases ont été alignées manuellement (voir chapitre 4) par les étudiants en traduction (voir chapitre 3). Ces livres ont été combinés en un seul corpus, qui sera dénommé **kapesh** pour la suite du chapitre.

Le second corpus est constitué de vers provenant du receuil de poésie jeunesse *Nin Auass*, qui ont été eux aussi présentés dans le cadre de l'étude d'alignement au chapitre 4. Les vers tirés du corpus de ce poétique ont été alignés avec la méthode de Gale & Church, par soucis de cohérence puisque l'échantillon aligné manuellement ne représentait qu'une partie du des vers. Pour rappel, une des conclusions du chapitre 4 était que la méthode de Gale & Church constituait la meilleur méthode pour aligner un texte paragraphe par paragraphe, comme pour le recueil de poèmes jeunesse. Ce corpus sera dénommé **jeunesse** pour la suite du chapitre.

[3] Les corpus alignés utilisés dans le cadre de cette étude, basés sur les oeuvres présentées au chapitre 4, sont des corpus formés uniquement pour la présente étude académique sur l'alignement et la traduction automatique pour l'innu-aimun et ne sont pas destinés à la diffusion, contrairement aux corpus publiques et libres d'utilisation (tel le Nunavut Hansard de Joanis *et al.* (2020)) qui sont monnaie courante en traduction automatique.

Le tableau 5.2 présente pour chaque corpus la taille du vocabulaire et la longueur moyenne des phrases, en nombre de mots. Ces caractéristiques sont présentées indépendemment pour l'innu-aimun et le français.

TABLE 5.2 – Caractéristiques des corpus pour la traduction automatique (portion Innu-Aimun)

Corpus	**Taille du vocabulaire**	**Longueur moyenne des phrases (mots)**
Innu-Aimun		
kapesh	3934	16.6
jeunesse	2168	2.77
Français		
kapesh	3058	24.1
jeunesse	1811	4.89

Sachant comment chacun de ces corpus a été constitué, on peut d'entrée de jeu établir que c'est le corpus kapesh qui devrait avoir la meilleure qualité d'alignement, puisque l'alignement a été fait par des personnes maîtrisant la langue. Ensuite, si on se fit aux résultats du chapitre 4, on s'attend à ce que le corpus jeunesse ait une qualité d'alignement acceptable, bien qu'inférieure à celle du corpus kapesh, puisqu'elle se base entre autres sur des paragraphes (poèmes) préalablement séparés et alignés.

Hormis la taille des corpus, leur domaine et la qualité de leur alignement, la longueur typique des phrases varie beaucoup d'un corpus texte à l'autre. Sans surprise, le corpus jeunesse est constitués de très petites phrases (des vers de poésie), alors que le corpus kapesh, de style littéraire, présente des phrases plus longue longues.

5.2.2 Résultats

Le tableau 5.3 présente les résulats de traduction des corpus individuels et combinés. La combinaison des deux corpus est indiquée par un « + » entre leurs dénominations.

Globalement, on peut noter que les résultats de la traduction automatique neuronale sont insatisfaisants, puisque les évaluations montrent toutes un score BLEU inférieur à 1, ce qui peut être considéré comme néligeable. Ceci est probablement dû à la taille insuffisante des textes utilisés (environ 1000 à 2000 paires de phrases par corpus, et environ 3000 lorsque combinés), puisque la traduction neuronale exige typiquement une importante quantité de phrases pour donner des résultats intéressants.

TABLE 5.3 – Scores obtenus sur les corpus individuels et combinés

Corpus	Source	Cible	Score BLEU	ChrF++
kapesh	Innu-Aimun	Français	0.28	**13.4**
kapesh	Français	Innu-Aimun	**0.62**	13.2
jeunesse	Innu-Aimun	Français	0.05	7.0
jeunesse	Français	Innu-Aimun	0.11	5.2
kapesh+jeunesse	Innu-Aimun	Français	0.19	12.3
kapesh+jeunesse	Français	Innu-Aimun	0.25	13.1

On peut faire une lecture de ces résultats au regard de la question posée au début de cette section : les textes culturels tels que les romans ou les poèmes peuvent-ils être utilisés dans la construction d'un corpus pour la traduction automatique ? Pour l'instant, les essais de traduction neuronale donnent une réponse négative avec l'échelle des corpus utilisés. Les essais d'autre méthodes (tel que ceux de traduction statistique présentés à la section suivante) pourraient toutefois nous permettre de déterminer si cette incapacité à obtenir des modèles de traduction automatique intéressants est due en partie à la nature de ces textes ou à leur taille.

5.2.3 Analyse sur les différences entre corpus

Au-delà des résultats insatisfaisant, une question se pose : si la taille des corpus utilisés est insuffisante, pourquoi observe-t-on aussi peu d'amélioration lorsqu'on double, voire triple pratiquement, la taille du corpus en passant de kapesh à kapesh+jeunesse ? La présente sous-section, qui examine les différences entre les deux corpus, propose des pistes de réflexions et d'hypothèses.

Le tableau 5.4 présente la proportion, en pourcentage, de vocabulaire commun et distinct entre les corpus kapesh et jeunesse. Les figures 5.1 et 5.2 présentent visuellement le nombre de mots communs et distincts, afin d'en visualiser les proportions. Il apparaît clairement que, bien qu'il y ait une part de vocabulaire commun entre les deux corpus, celle-ci reste plutôt petite. Les gains de fréquences des association de mots ou d'expressions seront donc probablement limitées lors de la combinaison des deux corpus. Il est important d'ailleurs de rappeler que ces deux corpus proviennent de domaines différents (littérature et poésie). En créant une corpus plus important dans un des deux domaines de spécialisation, on aurait probablement de meilleures performances, mais cela n'est visiblement pas le cas lorsqu'on cherche à combiner des domaines spécifiques différents.

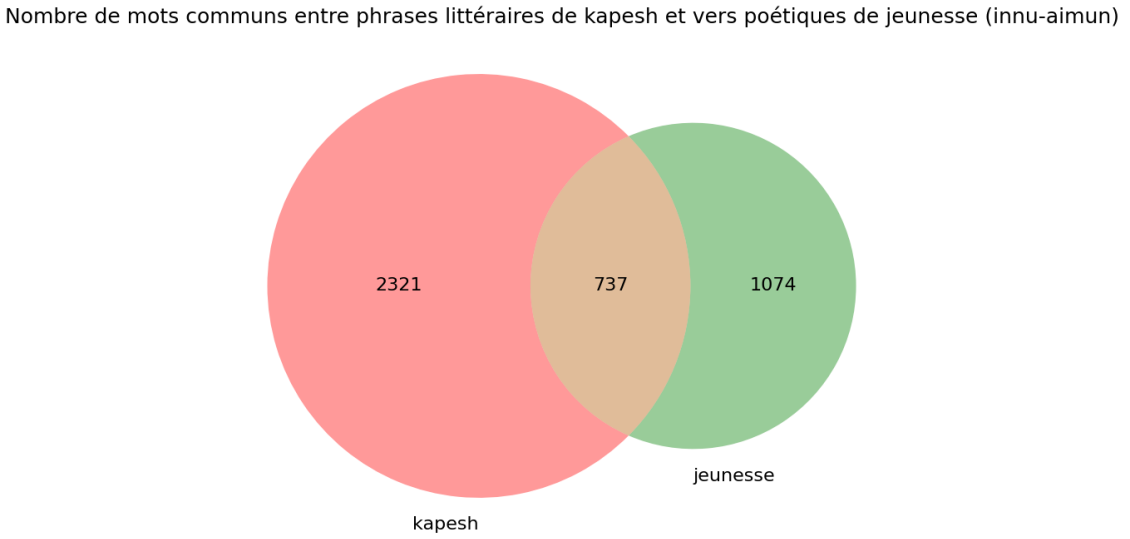

FIGURE 5.1 – Proportions de communalité du vocabulaire entre les différents corpus (Innu-Aimun)

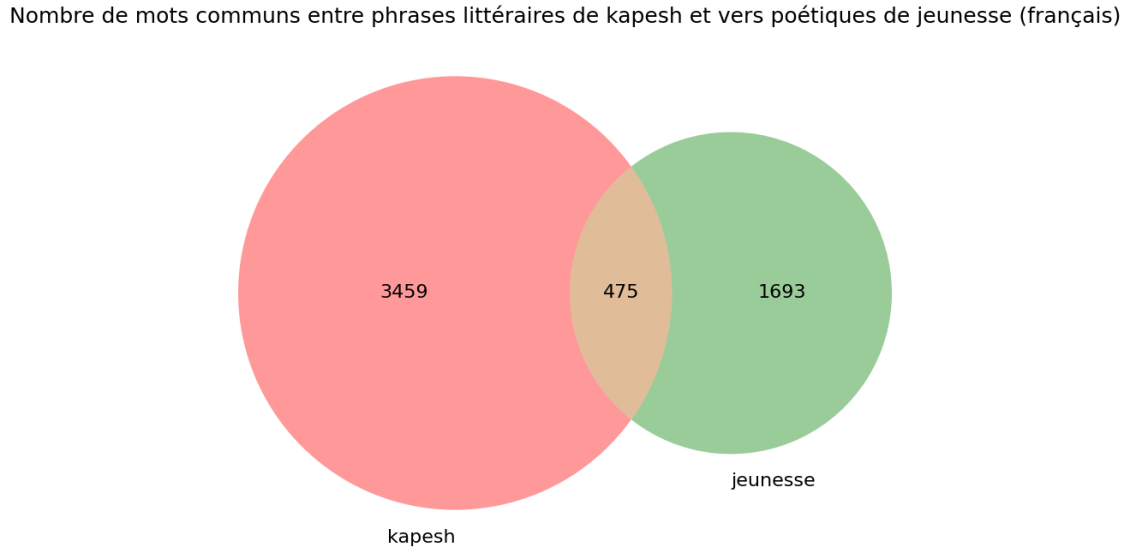

FIGURE 5.2 – Proportions de communalité du vocabulaire entre les différents corpus (Français)

TABLE 5.4 – Pourcentage du vocabulaire commun et distinct des différents corpus

Corpus	**Langue**	**% commun**	**% distinct**
kapesh	Innu-Aimun	24.1 %	75.9 %
kapesh	Français	12.1 %	87.9 %
jeunesse	Innu-Aimun	40.7 %	59.3 %
jeunesse	Français	21.9 %	78.1 %

Un autre aspect des corpus qui peut potentiellement avoir un impact sur l'apprentissage des modèles est la variété du vocabulaire par rapport à sa taille. Le tableau 5.5 présente le ratio de la taille du vocabulaire sur le nombre de phrases, pour les différents corpus. Ce ratio peut être vu comme en quelque sorte comme la densité de vocabulaire, ou de mots différents, par phrase. Par exemple, dans la portion en Innu-Aimun du corpus kapesh, on retrouve en moyenne 3.07 mots différents par phrase. Ce ratio permet d'évaluer à quel point les phrases d'un corpus contribuent à créer ce qu'on appelle le «contexte» d'une phrase, notion qui est importante pour le mécanisme d'attention au coeur du modèle Transformer de Vaswani *et al.* (2017).

TABLE 5.5 – Ratios de taille du vocabulaire sur nombre de phrases des différents corpus

Corpus	**Langue**	**Ratio taille vocab vs. nb de phrases**
kapesh	Innu-Aimun	3.07
kapesh	Français	2.39
jeunesse	Innu-Aimun	1.14
jeunesse	Français	0.950

On peut voir que le corpus kapesh, par rapport au corpus jeunesse, offre une plus grande variété de vocabulaire par rapport au nombre de phrases qui s'y trouve. On peut émettre l'hypothèse qu'il s'agit peut-être d'une explication pour la surperformance de kapesh par rapport à jeunesse : le corpus kapesh serait en mesures de fournir de plus de variétés de contextes à chaque phrase, ce qui favoriserait l'apprentissage du modèle. Cette possibilité est toutefois incertaine, et serait à vérifier plus rigoureusement.

D'autre matériel d'analyse des vocabulaires est fourni à l'annexe C, portant sur les mots les plus fréquents dans les corpus kapesh et jeunesse, pour les deux langues.

5.3 Essais de traduction statistique

La section qui suit présente les résultats des essais de traduction statistique.

Le système de traduction utilisé est le système *Moses*, qui utilise le modèle de traduction statistique par phrases de Koehn *et al.* (2003).

Le processus d'obtention des résultats est constitué de quatre étapes : le prétraitement des données, l'entraînement du modèle statistique, l'ajustement du modèle statistique, puis l'évaluation. Les étapes de prétraitement sont les mêmes que pour la traduction neuronale. Pour l'entraînement, l'ajustement et l'évaluation, les jeux de données sont séparés dans les mêmes proportions que pour les essais de traduction neuronale : 85% pour l'entraînement, 7.5% pour l'ajustement (plutôt que la validation) et les 7.5% restants pour l'évaluation.

Les phrases bilingues utilisées pour les essais statistique effectués ici sont les mêmes que celles des sections précédentes. Dans un premier temps le modèle est entraîné, ajusté et évalué sur les corpus individuels, puis combinés (tel qu'effectué à la section 5.2), dans les deux directions (innu-aimun vers français et l'inverse). Dans un second temps, le modèle résultant des corpus combinés est évalué sur les jeux d'évaluations des corpus individuels (kapesh et jeunesse), afin de mesurer l'impact des textes de l'un sur les résultats de l'autre.

Les types scores utilisés sont les mêmes que pour la traduction neuronale, soit le score BLEU et le score ChrF++, tous deux implémentés par sacrebleu (Post, 2018).

5.3.1 Résultats individuels et combinés

Le tableau 5.6 présente les résultats de traduction statistique innu-aimun/français pour les corpus kapesh et jeunesse, individuels et combinés.

5.3.2 Analyse générale

Dans un premier temps, nous pouvons effectuer l'observation que les résultats de traduction statistiques semblent être meilleurs que ceux de traduction neuronale simple (la significativité statistique de cette comparaison est fournie à la section 5.4). Cet écart semble se produire non seulement pour les corpus individuels,

TABLE 5.6 – Résultats de traduction statistique obtenus sur ajout/combinaison de différents corpus

Corpus	**Source**	**Cible**	**Score BLEU**	**ChrF++**
kapesh	Innu-Aimun	Français	**4.41**	22.7
kapesh	Français	Innu-Aimun	4.16	**33.6**
jeunesse	Innu-Aimun	Français	0.652	11.4
jeunesse	Français	Innu-Aimun	0.249	20.9
kapesh+jeunesse	Innu-Aimun	Français	4.22	20.9
kapesh+jeunesse	Français	Innu-Aimun	3.27	30.7

mais aussi pour le corpus constitué par combinaison. Cette différence semble toutefois beaucoup plus marquée pour le corpus kapesh et le corpus combiné que pour le corpus jeunesse.

Dans un second temps, on peut noter que l'ordre de meilleure performance des corpus est le même ordre que celui de meilleure qualité d'alignement, établit en début de section (kapesh étant aligné à 100% par des locuteurs, alors que jeunesse l'est à 25%, le reste étant aligné automatiquement). Il semble donc que la qualité d'alignement soit un facteur important pour la construction d'un corpus avec de meilleures performances de traduction. Et ce possiblement plus que le domaine ou que la taille (les deux corpus contiennent entre 1000 et 2000 phrases, jeunesse étant celui qui en a le plus). La longueur des phrases est peut-être un autre facteur (les vers de jeunesse sont beaucoup plus courts que les phrases de kapesh).

Dans un troisième temps, on peut noter les principales différentes entre le score BLEU et le score ChrF++. Selon le score BLEU, la direction innu-aimun vers français obtient toujours de meilleurs scores que l'inverse. C'est toutefois le contraire qui se produit, si l'on se fie au score ChrF++ seulement. Autrement dit, lorsqu'on analyse les mots dans leur ensemble uniquement (score BLEU), la traduction vers le français semble être de meilleur qualité, alors que si on se permet d'analyser les mots en partie (score ChrF++), alors c'est la traduction vers l'innu-aimun qui semble être de meilleure qualité. Cette différence est très probablement due au fait que l'innu-aimun est une langue à morphologie riche, alors que le français ne l'est pas. Dans l'ensemble, ces deux mesures doivent être vues comme complémentaires, puisqu'elles donnent chacunes des informations différentes.

Par ailleurs, dans le cas précis du corpus jeunesse, on peut aussi noter que l'hypothèse qu'une paire de vers d'un poème bilingue peut agir à titre de paire de phrases pour l'entraînement d'un modèle de traduction se

confirme quelque peu : en combinant les phrases de kapesh aux vers de la poésie jeunesse pour constituer un seul corpus, on obtient une amélioration des résultats de traduction par rapport à ceux de kapesh seul. Il est aussi intéressant de noter que, malgré les imprécisions constatées par les traducteurs innus dans le corpus de poésie jeunesse (voir chapitre 3), les vers tirés des poèmes permettent dans leur totalité une amélioration des résultats de traduction.

On peut aussi noter que les meilleurs scores de traduction statistiques dans les deux directions sont obtenus sur le corpus kapesh. Ainsi, si le corpus combiné kapesh+jeunesse semble offrir, selon ces scores quantitatifs, une meilleure performance que sur le corpus jeunesse, le score quantitatif, lui est plus faible sur le jeu combiné kapesh+jeunesse que sur le jeu kapesh simple.

Enfin, il est toutefois important de noter que l'évaluation quantitative à travers des a ses limites et qu'elle ne permet pas de juger à elle seule la qualité des traductions, surtout à une échelle de score aussi basse.

5.3.3 Analyse qualitative des phrases

L'évaluation qualitative complète, par des traducteurs professionnels, des modèles obtenus est hors du cadre de ce mémoire. Toutefois, afin de mieux comprendre les différences entre modèles, nous proposons ici un échantillon de phrases et traductions à comparer qualitativement. La comparaison et l'analyse qualitative des phrases est faite grâce à la participation de l'enseignante d'innu-aimun qui a contribué à l'alignement de vers poétique, dans une activité collaborative ayant aussi permis la création d'exercices pédagogiques (voir chapitre 3). Cet apport est important pour obtenir, dans notre approche collaborative, un dernier retour qualitatif sur les résultats qui découle des activités collaborative avec le personnel de l'école.

Dans les exemples présentés, la direction de traduction est de l'innu-aimun vers le français, afin d'en faciliter la lecture à ceux qui ne maîtrisent pas l'innu-aimun, incluant l'auteur de ce mémoire. Dans les tableaux, les éléments des phrases générées qui concordent avec des portions de la phrase de référence sont identifiées en **gras**.

5.3.3.1 Modèles TAS et TAN

Les tableaux 5.7 et 5.8 présentent deux exemples échantillonnés aléatoirement, pour qualitativement comparer la performance des modèles TAN et TAS entraînés sur le corpus kapesh (qui a obtenu les meilleurs

résultats avec les deux méthodes). Un phénomène qui intéressant à noter dans les deux exemples, est qu'on voit que le modèle TAS tend à fournir des traductions comportant des mots d'innu-aimun, même si la langue cible est le français. Ce n'est pas le cas du modèle TAN, qui fournit seulement des mots de français dans ses traductions. En examinant les mots d'innu-aimun suggérés par le modèle TAS (mis en *italique* dans les tableaux d'exemples), on se rend compte que ce sont des mots qui ont été tirés de la phrase d'origine en innu-aimun.

Dans le premier exemple, que le modèle TAS arrive à générer deux expressions différents correctes par rapport à la traduction de référence, c'est-à-dire l'ensemble de mots « **ce n'est pas** » et le mot « **toujours** » (Kapesh, 2020). Le modèle TAN, quant à lui n'en génère qu'une des deux et ne le fait qu'imprécisément et avec répétition (**c'est pas** et **ce n'est**). Il y a là un exemple qualitatif qui pourrait à expliquer la surperformance quantitative du modèle TAS observée globalement, si cet exemple est représentatif des autres. D'ailleurs, pour cette paire de phrase précise, le score BLEU obtenu par le modèle TAS est de 11.7 vs 2.01 pour le modèle TAN.

TABLE 5.7 – Comparaison qualitative entre modèles littéraires TAS et TAN : exemple #1

Phrase source (Innu-Aimun)	« Tshitshisseniteti ka ishpish tshi takushiniek^{u} ute, namaieu nene inniun nin ka ishinniuian. Nanitam nikushtatshiti. » (Kapesh, 2020)
Traduction (référence)	« Savais-tu que depuis votre arrivée ici, ce n'est pas une vie que j'ai vécue ? J'avais tou-jours peur. » (Kapesh, 2020)
Traduction (TAN)	C'est ce que c'est là que **c'est pas** notre territoire, c'est **ce n'est** là que c'est pas notre territoire.
Traduction (TAS)	il ne nous a *takushiniek^{u} Tshitshisseniteti* à moi, **ce n'est pas** ici, il n'y a *ishinniuian* nene *inniun nikushtatshiti*. **toujours**.

Dans le second exemple ici présenté, on voit que le modèle TAN n'arrive qu'à obtenir l'expression « **il n'** » qui soit correcte et qu'il la duplique. Le modèle TAS, lui, arrive à obtenir l'expression « **il n'en** », en plus du mot « **trouvé** » et de l'expression « **un seul** » (Kapesh, 2020). Par ailleurs, l'enseignante d'innu-aimun note que dans cet exemple, l'expression « notre territoire » que nous fournit le modèle TAN est complètement absent de la phrase d'origine en innu-aimun. C'est donc que cette notion a complètement été hallucinée par le modèle neuronal, peut-être à cause de sa prépondérance dans le corpus kapesh. Ce phénomène n'apparaît

toutefois pas dans la traduction du modèle statistique. Il est aussi intéressant de noter ici que le mot innu « *shutshenimat* » de la traduction du modèle TAS correspond, selon l'enseignante d'innu-aimun, à la notion de confiance, qui serait autrement manquante dans cette traduction.

Au final, la différence de qualité de traduction entre les deux modèles se reflète dans le score BLEU : le modèle TAN obtient 0.1130 alors que le modèle TAS obtient 9.8647 pour cette paire de phrases.

TABLE 5.8 – Comparaison qualitative entre modèles littéraires TAS et TAN : exemple #2

Phrase source (Innu-Aimun)	« Shash apu mishkuat peik^{u} tshetshi shutshenimat. » (Kapesh, 2020)
Traduction (référence)	« Il n'en trouve plus un seul en qui avoir confiance. » (Kapesh, 2020)
Traduction (TAN)	**Il n'**a pas à l'intérieur des terres, **il n'**a pas de l'intérieur des terres.
Traduction (TAS)	**il n'en** avoir **trouvé** un à *shutshenimat* déjà qu'**un seul**.

5.3.3.2 Modèles TAS littéraires avec et sans vers poétiques

Afin de mieux comprendre, les différence de performance quantitative entre les modèles TAS entraînés sur le corpus kapesh individuel et celui entraîné sur le corpus combiné avec jeunesse, quatre exemples de paires de phrases sont présentés aux tableaux 5.9, 5.10, 5.11 et 5.12. Ces exemples sont examinés en collaboration avec l'enseignante d'innu-aimun.

Dans le premier exemple, on voit que les deux modèles proposent correctement les expressions « **cette histoire** » et « **jamais** » (Kapesh, 2020). Les deux modèles arrivent aussi à cerner la notion contenue dans l'expression «pense-il», qui est traduite par **se dit-il**. Les deux modèles n'arrivent pas à traduire le mot « *petakuak* » (Kapesh, 2020). Toutefois, alors que le modèle kapesh garde le mot «*Tshima*» (Kapesh, 2020) dans sa langue d'origine, le modèle kapesh+jeunesse parvient à le remplacer par **Si seulement**, ce pourrait être une traduction valide de ce mot, selon notre collaboratrice. Notre collaboratrice est aussi de l'avis que la traduction du modèle kapesh+jeunesse est meilleure (ou moins pire) que celle du modèle jeunesse.

Dans le second exemple, si les deux modèles proposent correctement dans leur traduction l'expression « **l'enfant** » (Kapesh, 2020), les deux traductions semblent assez éloignées celle de référence lorsqu'on

TABLE 5.9 – Comparaison qualitative entre modèles TAS littéraire, avec et sans vers poétiques : exemple #1

Phrase source (Innu-Aimun)	« « Tshima eka nita petakuak ume tipatshimun.. itenitam^{u}. » (Kapesh, 2020)
Traduction (référence)	« « Pourvu qu'on n'entende jamais cette histoire..» pense-il. » (Kapesh, 2020)
Traduction (kapesh)	« Tshima **cette histoire** de ne **jamais** *petakuak*.. se **dit-il**.
Traduction (kapesh+jeunesse)	« Si seulement lui n'était pas *petakuak* **jamais** de **cette histoire** de.. **se dit-il**.

en compare les mots. Toutefois, selon notre collaboratrice, on arrive à comprendre, dans la traduction du modèle kapesh+jeunesse, le sens général de la phrase. Dans cette traduction, c'est plutôt la syntaxe qui n'est pas bonne. Selon l'enseignante d'innu-aimun, la phrase aurait pu être considérée comme une meilleure traduction sans les cinq derniers mots (« à ce qu'il dit »). Mieux encore, selon elle la traduction aurait été bonne si elle avait été « l'enfant est d'accord avec ce qu'il lui dit » plutôt que « l'enfant est d'accord avec ce que lui dit à ce qu'il dit ».

On peut noter par ailleurs, que la traduction du modèle entraîné grâce à l'apport des poèmes de jeunesse est meilleure notamment parce qu'elle traduit le mot *tapuetueu* (Kapesh, 2020), plutôt que de le laisser tel quel. Ce mot, selon notre collaboratrice, représente la notion d'accord, qui est sinon manquante parmi les mots français de la traduction du modèle kapesh.

TABLE 5.10 – Comparaison qualitative entre modèles TAS littéraire, avec et sans vers poétiques : exemple #2

Phrase source (Innu-Aimun)	« Ne auass tapuetueu nenu etikut. » (Kapesh, 2020)
Traduction (référence)	« L'enfant se laisse convaincre. » (Kapesh, 2020)
Traduction (kapesh)	**l'enfant** est qu'il *tapuetueu* ce que lui dit.
Traduction (kapesh+jeunesse)	**l'enfant** est d'accord avec ce que lui dit à ce qu'il.

Dans le troisième exemple, on voit que la traduction du modèle kapesh est plus incomplète : elle est composée partiellement de mots d'innu-aimun. On peut voir dans la traduction du modèle kapesh+jeunesse que

les mots français « chose » et « je vais » (Kapesh, 2020) sont aussi utilisés dans la traduction de référence. L'enseignante d'innu-aimun nous confirme que aussi que le mot « et » peut être considéré comme une traduction correcte du mot innu *mak*, présent dans la phrase source. Toutefois, encore selon l'enseignante d'innu-aimun, pour la traduction du modèle kapesh+jeunesse soit une bonne représentation des concepts de la phrase source, il faudrait retirer le concept d'« autre » et qu'on ajoute le concept « dire ».

TABLE 5.11 – Comparaison qualitative entre modèles TAS littéraire, avec et sans vers poétiques : exemple #3

Phrase source (Innu-Aimun)	« Mak kutak tshekuan tshe uitamatan. » (Kapesh, 2020)
Traduction (référence)	« Je vais te dire encore une chose. » (Kapesh, 2020)
Traduction (kapesh)	la *uitamatan* et les autres.
Traduction (kapesh+jeunesse)	et les autres **choses** que **je vais**.

Dans le quatrième exemple, on constate que la traduduction du modèle kapesh est syntaxiquement incorrecte en français. Si elle ne représente pas une phrases complète, la traduction du modèle combiné kapesh+jeunesse pourrait elle être correcte, selon notre collaboratrice, si ce n'était du fait qu'elle accorde ses mots au pluriel alors que la phrase source serait plutôt au singulier. Pour que la traduction de kapesh+jeunesse soit correcte, il aurait fallu que la phrase source soit plutôt *Mak kutaka(t) tshekuana(t).*. Notons qu'ici le pluriel serait avec un *t* ou non dépendemment du genre, animé ou inanimé, des mots en coréférence, dans les phrases avoisinantes.

TABLE 5.12 – Comparaison qualitative entre modèles TAS littéraire, avec et sans vers poétiques : exemple #4

Phrase source (Innu-Aimun)	« Mak kutak tshekuan. » (Kapesh, 2020)
Traduction (référence)	« Et il y a autre chose. » (Kapesh, 2020)
Traduction (kapesh)	**et** les **autres** y.
Traduction (kapesh+jeunesse)	**et** les **autres choses**.

Il ressort de cette comparaison détaillée et appréciation qualitative des traductions des deux modèles que les phrases du jeu d'évaluation kapesh peuvent bénéficier, dans leur traduction, de la présence à l'entraînement des vers poétiques du jeu jeunesse. Cela semble être en contradiction avec le score plus faible obtenu par le modèle commun.

Pourquoi certaines phrases semblent être clairement mieux traduites par le modèle commun alors que le score de ce dernier est inférieur ? Une partie de la réponse pourrait résider dans le fait que nous n'observons ici que les phrases d'origine du jeu d'évaluation kapesh, alors que le jeu d'évaluation commun contient aussi les phrases du jeu d'évaluation jeunesse. Ces dernières pourraient être en général moins bien traduites, ce qui ferait baisser le score global du jeu commun, même si les phrases du jeu kapesh sont mieux traduites lorsqu'entraînées de manière commune.

Pour mieux comprendre le phénomène, la sous-section 5.3.4 proposera une évaluation sur les jeux individuels kapesh et jeunesse du modèle entraîné conjointement.

5.3.3.3 Modèles TAS littéraires TAS poétique avec et sans phrases littéraires

De façon analogue aux modèles kapesh et kapesh+jeunesse, les différences de performance qualitative entre les modèles TAS entraînés sur le corpus jeunesse individuel et celui entraîné sur le corpus combiné avec kapesh sont ici examinées. Quatre exemples de paires de phrases sont présentés à cet effet, aux tableaux 5.13, 5.14, 5.15 et 5.16. Les quatre exemples sont des vers assez court, ce qui est représentatif de ce corpus (voir analyse plus détaillée des différences entre corpus à la section précédente).

Dans le premier exemple on voit que le modèle combiné atteint une exactitude parfaite sur sa traduction, par rapport à la traduction de référence. Le modèle jeunesse individuel, lui, utilise le mot innu « *kie* » plutôt que le mot « et » (Bacon et Morali, 2021). Autrement dit, la présence des phrases de kapesh à l'entraînement a permis au modèle de parfaire sa traduction, peut-être en profitant de la fréquence de la paire de mot kie/et dans cet autre corpus.

TABLE 5.13 – Comparaison qualitative entre modèles TAS poétique, avec et sans phrases littéraires : exemple #1

Phrase source (Innu-Aimun)	« kie nutin » (Bacon et Morali, 2021)
Traduction (référence)	« et le vent » (Bacon et Morali, 2021)
Traduction (jeunesse)	*kie* **le vent**
Traduction (kapesh+jeunesse)	**et le vent**

Dans le second exemple, le modèle jeunesse individuel traduit le vers source mot-à-mot par «je loup». Les deux notions de ce court vers de deux mots sont bel et bien présentes dans la traduction du modèle, mais

cette dernière est syntaxique incorrecte. La traduction produite par le modèle combiné est une séquence plus longue qui elle est syntaxiquement correcte. On peut aussi dire qu'elle représente bien le sens du vers d'origine, bien qu'elle ne soit peut-être pas exactement ce qui est recherché par ce dernier, si l'on se fit à la traduction de référence. Ici, on pourrait dire que l'entraînement conjoint avec le corpus kapesh a permis d'atteindre un meilleur niveau syntaxique, qui va au-delà du simple remplacement mot-à-mot. Ceci pourrait être dû au fait que le corpus kapesh comprend en général beaucoup plus de phrases longues et complètes, alors que les vers poétiques ne sont souvent longs que de quelques mots et peuvent ne pas être des phrases complètes (voir analyse à la section précédente).

Une validation collaborative avec l'enseignante d'innu-aimun nous permet aussi de confirmer que «je suis un loup» est une traduction correcte de « nin maikan » (Bacon et Morali, 2021), au même titre que « moi le loup » (Bacon et Morali, 2021).

TABLE 5.14 – Comparaison qualitative entre modèles TAS poétique, avec et sans phrases littéraires : exemple #2

Phrase source (Innu-Aimun)	« NIN MAIKAN » (Bacon et Morali, 2021)
Traduction (référence)	« MOI LE LOUP » (Bacon et Morali, 2021)
Traduction (jeunesse)	je **loup**
Traduction (kapesh+jeunesse)	**moi** je suis un **loup**

Le troisième exemple montre un autre cas où la traduction d'un vers de jeunesse est améliorée par l'entraînement conjoint avec les phrases de kapesh. On passe de la phrase incomplète et syntaxiquement incorrecte « la longtemps » à une traduction beaucoup plus complète : « il y a longtemps déjà ». La validation collaborative nous permet par ailleurs de conclure non seulement que « il y a longtemps déjà » est une traduction correcte de l'expression innue « shashish shash » (Bacon et Morali, 2021), mais qu'en plus il s'agit d'une traduction plus littérale et respectant plus le sens propre de cette dernière. La traduction de référence, « les jours s'éternisent » (Bacon et Morali, 2021) serait une traduction représentant davantage un sens figuré, qui privilégie la valeur artistique.

Dans le quatrième exemple, on peut voir une autre figure de style dans la traduction de référence : dans cette dernière, « Utshashumeku, toi le saumon »,(Bacon et Morali, 2021) on voit un usage à la fois de mots français et d'un mot d'innu-aimun. Il est à noter aussi que le mot *Utshashumeku* aurait dû être *Utshashumekᵘ*,

TABLE 5.15 – Comparaison qualitative entre modèles TAS poétique, avec et sans phrases littéraires : exemple #3

Phrase source (Innu-Aimun)	« shashish shash »(Bacon et Morali, 2021)
Traduction (référence)	« les jours s'éternisent »(Bacon et Morali, 2021)
Traduction (jeunesse)	la longtemps
Traduction (kapesh+jeunesse)	il Y A LONGTEMPS déjà

une erreur s'étant probablement introduite lors du pré-traitement de ce vers (voir chapitre 4). À travers une validation des résultats de traduction avec l'enseignante d'innu-aimun, on peut confirmer que l'usage du mot innu *Utshashumek*^{u} constitue une forme de répétition, un peu comme si l'on disait de manière bilingue *le saumon, toi le saumon*. Dans ce cas précis, les traductions des deux modèles (jeunesse et jeunesse+kapesh) sont les mêmes ; l'ajout de kapesh n'a donc pas d'impact. Toutefois, la validation collaborative nous permet de confirmer qu'il s'agit d'un autre cas où la traduction fournie par modèle statistique est plus proche de la forme d'origine, du sens propre de la phrase source. Ainsi, le système favorise une traduction plus proche du format de la phrase source et ne fait pas de figure de style, contrairement à la traduction de référence.

TABLE 5.16 – Comparaison qualitative entre modèles TAS poétique, avec et sans phrases littéraires : exemple #4

Phrase source (Innu-Aimun)	« tshin, Utshashumek^{u} »(Bacon et Morali, 2021)
Traduction (référence)	« Utshashumeku, toi le saumon »(Bacon et Morali, 2021)
Traduction (jeunesse)	toi, le saumon
Traduction (kapesh+jeunesse)	toi, le saumon

Globalement, l'analyse collaborative de ces quatre exemples poétiques de traduction statistique montrent un phénomène très intéressant : le meilleur système de traduction (kapesh+jeunesse) peut permettre d'obtenir des traductions françaises plus proches du sens propre des vers d'origine en innu-aimun. On peut par ailleurs noter que ce phénomène a pour effet de diminuer le score BLEU obtenu pour le corpus jeunesse : puisque les traductions hypothétisées par le système s'éloignent de celles de référence, la comparaison à ces dernières, qui est la base du score BLEU, devient désavantageuse. Toutefois, on examinant de manière qualitative, on voit bien que ce sont des traductions non seulement valides mais aussi parfois plus exactes. On pourrait d'ailleurs conclure que les scores d'évaluations quantitatives tels le score BLEU ne sont peut-être pas adaptés

à l'évaluation de traductions de type poétique ou littéraire, qui sont parfois plus libres et artistiques.

5.3.4 Résultats d'évaluation par corpus

La présente sous-section fait état des résultats d'évaluation des modèles TAS combinés sur les jeux d'évaluations individuels. Ces résultats font suite à l'analyse présentée à la sous-section précédente, analyse selon laquelle les phrases du corpus kapesh pouvaient bénéficier de l'entraînement combiné avec les phrases du corpus jeunesse, bien que le score global, lui, soit plus faible pour la combinaison. Les résultats ici présentés visent donc à déterminer quel est l'impact d'un entraînement commun sur les phrases corpus individuels. Le tableau 5.17 propose donc les résultats de l'entraînement conjoint évalué sur chacun des jeux d'évaluation individuels. Il présente aussi à nouveau les résultats d'entraînement individuels avec évaluation sur les jeux individuels, pour en permettre la comparaison.

TABLE 5.17 – Comparaison des résultats d'évaluation par corpus

Corpus	**Jeu d'évaluation**	**Source**	**Cible**	**Score BLEU**	**ChrF++**
kapesh	kapesh	Innu-Aimun	Français	4.41	22.7
kapesh	kapesh	Français	Innu-Aimun	4.16	**33.6**
kapesh+jeunesse	kapesh	Innu-Aimun	Français	**4.97**	23.7
kapesh+jeunesse	kapesh	Français	Innu-Aimun	3.27	30.7
jeunesse	jeunesse	Innu-Aimun	Français	0.652	11.4
jeunesse	jeunesse	Français	Innu-Aimun	0.249	20.9
kapesh+jeunesse	jeunesse	Innu-Aimun	Français	**1.08**	13.8
kapesh+jeunesse	jeunesse	Français	Innu-Aimun	0.879	**22.1**

La comparaison des évaluations montrent confirme que le modèle kapesh+jeunesse atteint une meilleure performance globale que le modèle kapesh, lorsqu'évalué uniquement sur les phrases du corpus kapesh, dans la direction innu-aimun vers français. Toutefois, ce n'est pas le cas pour la direction français vers innu-aimun. Ainsi, ces résultats d'évaluation sur les phrases françaises du corpus kapesh confirment nos observations qualitatives (effectuées sur les traductions françaises), bien qu'elles ne puissent permettre de les être étendre à la direction inverse.

Ces résultats sur kapesh permettent ainsi de confirmer que la baisse de performance vers le français sur le jeu d'évaluation combiné kapesh+jeunesse, est due à la présence des vers de jeunesse dans ce dernier, sur

lesquels le modèle kapesh+jeunesse obtient de moins bonnes performances que les phrases de kapesh, et non à la dégradation de performance. Toutefois, pour la performance vers l'innu-aimun, il y a bel et bien une dégradation de performance vers l'innu-aimun pour les phrases de kapesh lors de l'entraînement combiné.

Dans le cas des évaluations sur des vers poétiques du corpus jeunesse, les deux directions de traduction montrent des améliorations quantitatives lorsque l'entraînement du corpus jeunesse est effectué conjointement avec celui de kapesh. Cela confirme à la fois les chiffres quantitatifs observés sur les jeux combinés et les obsevations effectuées les deux exemples aléatoires.

5.4 Comparaisons et significativité statistique

La présente section compare entre eux les scores de traduction automatique des deux différents méthodes évaluées, soit la traduction automatique neuronale (basée sur Transformer) et la traduction automatique statistique, et évalue la significativité statistique de cette comparaison. Les modèles et résultats utilisés pour cette comparaison sont les mêmes que ceux utilisés pour présenter les scores des sections précédentes.

5.4.1 Méthode de test

La méthodes choisie pour évaluer la significativité statistique des comparaisons est le rééchantillonnage *bootstrap* (ou *bootstrap resampling*), tel que proposé par Koehn (2004) pour la traduction automatique. Il s'agit d'ailleurs du test de significativité statistique recommandé lorsque les mesures qui nous servent à cette évaluation sont des score BLEU (Dror *et al.*, 2018).

Ce test se base sur l'hypothèse que l'échantillon de mesures que nous avons en notre possession est représentatif de la population que nous évaluons. L'algorithme de *bootstrap*, afin de vérifier si les écarts observés entre les mesures des deux méthodes comparées sont dûs au hasard, multiplie les mesures à disposition en effectuant un rééchantillonnage aléatoire basé sur les mesures des échantillons fournis, puis vérifie si les écarts observés sont toujours avérés (Koehn, 2004). Si la distribution des écarts (valeur p) se situe à l'intérieur d'un intervale de variation autour de la moyenne, défini par le seuil α , alors l'hypothèse de comparaison entre les deux méthodes est considérée comme concluante (statistiquement significative).

Dans le cas qui nous occupe, l'échantillon testé est composé des scores (BLEU ou ChrF++) individuels des phrases du jeu d'évaluation (*sentence-level*). La comparaison est donc faite pour chaque phrase qui a

générée et comparée à celle de référence. Les différences ou variations mesuréws sont celles entre les scores des phrases. Le seuil choisi est le seuil de significativité typique pour ces tests, soit $\alpha = 0.5$. Nous utilisons l'implémentation faite par Dror *et al.* (2018) de l'algorithme *bootstrap*, qui utilise la logique décrite par Berg-Kirkpatrick *et al.* (2012).

5.4.1.1 Limites de ce test

Il est crucial de mentionner que, dans le cadre de ce test, une valeur de $p = 0.0$ sur la comparaison entre une méthode A et une méthode B ne signifie pas que nous avons une certitude absolue que la méthode A est meilleure que la méthode B. Plutôt, dans le test de rééchantillonnage aléatoire comparatif, une valeur de $p = 0.0$ signifie simplement qu'avec l'échantillon de scores dont nous disposons (et suite à une rééchantillonnage de cet échantillon) 100% des comparaison entre scores individuels sont favorables à la méthode A. De la même manière, une valeur de $p = 1.0$ ne signifie pas qu'il est absolument certain qu'il n'y ait aucune différence entre la méthode A et la méthode B ; elle signifie seulement que sur tous les scores (rééchantillonnés aléatoirement) de notre échantillon, il n'y a aucune différence favorable à A qui ait pu être mesurée.

Cela nous rapporte à l'hypothèse de base du test de rééchantillonnage statistique, celle qui suppose que l'échantillon que nous avons à disposition est parfaitement représentatif de la population générale. Il est évident qu'un corpus donné n'est pas représentatif de la langue utilisée de manière générale, ou même du domaine ou sous-domaine duquel il est issu. Il est parfaitement représentatif uniquement de lui-même. Ayant cela en tête, et supposant que notre échantillon de test (jeu d'évaluation) pour ce corpus est représentatif de tout le corpus, une valeur $p < 0.0$ nous confirmerait avec certitude que la méthode A est meilleure que la méthode B pour ce corpus, mais pas nécessairement pour n'importe quel corpus de même taille dans ce domaine ou pour cette langue en général.

5.4.2 Résultats

Résultats de significativité statistique pour la comparaison des trois méthodes entre elles sont présentés ci-bas, avec le code de couleur suivant : un résultat en vert signifie que le test de significativité est concluant ($p < 0.05$), alors que qu'un résultats en rouge signifie que le test de significativité n'est pas concluant ($p < 0.05$).

5.4.2.1 Comparaison entre traduction statistique et traduction neuronale

Le tableau 5.18 présente la significativité statistique sur l'hypothèse que la traduction automatique statistique (TAS) est plus performante que la traduction automatique neuronale (TAN), sur les résultats des corpus kapesh et jeunesse, individuels et combinés (voir section **??**).

TABLE 5.18 – Significativité statistique de la comparaison TAS > TAN pour les scores des ajouts/combinaisons de différents corpus

Corpus	**Source**	**Cible**	**Score BLEU**	**ChrF++**
kapesh	Innu-Aimun	Français	$p < 0.05$ avec $p = 0.0$	$p < 0.05$ avec $p = 0.0$
kapesh	Français	Innu-Aimun	$p < 0.05$ avec $p = 0.0$	$p < 0.05$ avec $p = 0.0$
jeunesse	Innu-Aimun	Français	$p > 0.05$ avec $p = 0.0009$	$p > 0.05$ avec $p = 0.0$
jeunesse	Français	Innu-Aimun	$p > 0.05$ avec $p = 0.0014$	$p < 0.05$ avec $p = 0.0$
kapesh+jeunesse	Innu-Aimun	Français	$p < 0.05$ avec $p = 0.0$	$p < 0.05$ avec $p = 0.0$
kapesh+jeunesse	Français	Innu-Aimun	$p > 0.05$ avec $p = 0.0$	$p < 0.05$ avec $p = 0.0$

5.4.3 Conclusions générales

Les résultats de significativité statistique sont tous concluants. Ils nous permettent de confirmer les observations effectuées aux sections précédentes, soit la supériorité des résultats de traduction automatique statistique face à ceux obtenus par la traduction automatique neuronale. Comme cette méthode est celles qui a obtenu tous les meilleurs résultats et ce de manière statistique significative, on peut affirmer que c'est celle qui a le plus de potentiel pour les développement futurs de traduction automatique pour l'innu-aimun. Du moins c'est ce qu'on peut établir pour le court terme et à des échelles de données similaires à celles ici examinée. À d'autres échelles de données, ou avec d'autres types de données et de modèles, la traduction neuronale qu'elle mériterait d'être testée davantage.

5.5 Synthèse et discussion sur la faisabilité

La section qui suit effectue une synthèse des résultats de traduction en regard à la question fondamentale de faisabilité posée en début de chapitre : est-il possible de développer un outil de traduction automatique pour l'innu-aimun avec les techniques actuelles et les textes disponibles ? Étant donné qu'il s'agit de la toute première étude de ce type pour l'innu-aimun, et sachant pertinnement qu'il est difficile d'obtenir des

résultats appréciable avec des quantités de données tel qu'il en existe pour l'innu-aimun, l'étude vise à savoir quelles sont les principales pistes d'amélioration.

L'ensemble des résultats rapportés au présent chapitre permet de dresser quatre principaux constats. Premièrement, les résultats montrent que la quantité de données disponibles pour l'innu-aimun est trop faible pour que des modèles de traduction puissent être immédiatement utilisés en tant que traducteurs automatiques. Deuxièmement, il semble que les données bilingues culturelles, telles que la littérature et la poésie, puissent être mises à contribution dans la construction d'un modèle de traduction automatique, à des degrés d'efficacité divers. Troisièmement, la qualité d'alignement apparait comme primordial pour l'amélioration des résultats de traduction par ajout de données. Quatrièmement et finalement, la traduction statistique offre les résultats les plus prometteurs.

Le premier constat est dressé surtout en comparant les performances obtenus dans la présente étude à ceux atteints par Joanis *et al.* (2020) pour l'inuktitut, la seule langue autochtone du Canada qui possède son traducteur, et pour laquelle une évaluation de corpus similaire à la nôtre a été effectuée. Il peut être considéré hasardeux en traduction automatique de comparer les métriques d'évaluation tels que le BLEU entre différentes paires de langues [4]. Toutefois, ne serait-ce que par l'écart qu'on voit entre les scores BLEU obtenus par les précédents tests et ceux ayant menés au premier traducteur automatique pour une langue autochtone au Canada [5], il est clair que les performances de traduction automatique actuelles pour l'innu-aimun sont loin du stade d'utilisabilité immédiate. Les meilleurs scores BLEU, toutes catégories confondues, de traduction automatique obtenus ici sont de **4.97** vers le français (entraînement combiné kapesh+jeunesse et évaluation sur kapesh seulement) et de **4.16** vers l'innu-aimun (entraînement sur kapesh seulement et évaluation sur kapesh seulement). Les scores d'évaluation obtenus à partir du Nunavut Hansard à sa publication, quant à eux, étaient de 35.0 vers l'anglais et 20.3 vers l'inuktitut. Et il n'est pas à exclure que les auteurs de l'outils Microsoft aient eu à apporter des améliorations significatives avant de pouvoir rendre disponible leur modèle. Par contre, étant donné la très faible quantité de données ayant été utilisées pour obtenir le meilleur score (total combiné de 3187 phrases et 1280 phrases seulement pour kapesh) et sachant qu'il s'agit des tous premiers essais de ce type, les résultats sont tout de même encourageants et la tâche n'apparaît pas du tout comme impossible.

[4] Google, dans sa documentation portant sur l'évaluation de la traduction automatique, l'explique bien : Google - Evaluating models

[5] Microsoft ajoute l'inuktitut au Traducteur Microsoft (en anglais)

Le second constat montre qu'il est possible de mettre à contribution des textes culturels, poétiques et littéraires, pour développer des outils de traduction automatisée pour l'innu-aimun. Si les textes disponibles pour une langue très peu dotée sont donc surtout des textes culturels, ceux-ci peuvent être utiles. Nous a pu d'ailleurs constater, à travers une analyse qualitative en collaboration avec une enseignante d'innu-aimun, que certaines des phrases traduites par nos modèles peuvent être considérées comme des traductions adéquates bien que différentes de celles de référence. Certaines traductions de référence favorisent un sens figuré ou une interprétation libre et artistique (vu les domaines littéraires et poétiques des textes étudiés), alors que les modèles de traduction automatique fournissent une traduction plus proche du sens propre. Il semblerait donc que les scores BLEU faibles obtenus ne signifient pas nécessairement des traduction incorrectes, dans le contexte qui nous occupe.

Le troisième constat montre que le plus important dans la construction d'un corpus pour la traduction automatique serait de porter une attention particulière à leur alignement. Et ce que les textes soient des textes littéraires, poétiques ou autres, du moins dans un contexte de faible quantité de données disponibles tel que le nôtre. Notre étude a en effet permis de conclure que les textes culturels peuvent bien être mis à contribution pour la construction d'un corpus servant à la traduction automatique, mais qu'il est possible que la qualité d'alignement qui ait un grand impact.

Le quatrième constat, que la traduction statistiques offre les résultats les plus prometteurs, témoigne entre autres de la très petite quantité de textes bilingues innu-aimun/français disponibles actuellement. À court terme, il semblerait ainsi avantageux de prioriser l'ajout de textes bilingues en innu-aimun/français, comme en témoignent notamment nos résultats de traduction statistiques obtenus par combinaisons de corpus. On peut entre autres penser à l'apport potentiel des phrases trilingues (innu-aimun, français et anglais) présentées en exemple dans le dictionnaire innu en ligne (Ambroise *et al.*, 2023)).

5.6 Conclusion, limitations et perspectives

Au cours de ce chapitre, nous avons pu établir de premiers résultats de traduction automatique pour l'innu-aimun. Nous avons pu étudier l'impact de l'ajout de deux principaux jeux de phrases bilingues aux caractéristiques différentiées. Nous avons pu tester deux approches différentes, soit l'approche neuronale et l'approche statistique. Face aux résultats obtenus et aux analyses que nous en avons fait, ils nous a été possible de dresser quatre constats au regard de la question de faisabilité posée en début de chapitre. Nous

avons pu ainsi constater que (1) la quantité insuffisante de données disponibles pour l'innu-aimun empêche l'utilisation immédiate de modèles de traduction automatique, bien que les premiers résultats obtenus soient tout de même encourageants. Que (2) l'utilisation de données bilingues culturelles, comme la littérature et la poésie, peut être envisagée en contexte de faible quantité de données. Que (3) dans l'ajout de données supplémentaires, peu importe leur nature, la qualité d'alignement est d'une importance cruciale pour améliorer les résultats de traduction. Que (4) la traduction statistique présente les résultats les plus prometteurs pour le moment.

Comme le développement de la traduction automatique en est à ses débuts pour l'innu-aimun, ces pistes d'améliorations pourront renseigner sur les aspects où il sera plus pertinent d'investir des ressources (temps des personnes-ressources, fonds disponibles, etc.) pour obtenir les plus grandes avancées.

Les résultats obtenus au présent chapitre et les constats rapportés ci-haut doivent être interprétés à la lueur des limitations du cadre expérimental proposé. La principale limitation est que les modèles de traduction ont été entraînés et évalués sur les mêmes corpus. Les évaluations sont donc nécessairement biaisées et il est très probable que les modèles aient des performances inférieures si on évalue leur capacité à généraliser sur d'autres jeux d'évaluation. Par ailleurs une limitation additionnelle peut être notée dans le cas des combinaisons de textes : en ajoutant les textes à la fois à l'entraînement et à l'évaluation (puisqu'on traite désormais les divers textes comme un seul corpus), il est certain qu'on ajoute ainsi un certain biais à l'évaluation. Enfin, une limitation qu'on peut noter pour toutes les expérimentations effectuées est la taille généralement petite des jeux d'évaluation, qui s'impose par la faible quantité de données disponible.

Par ailleurs, bien que le présent mémoire s'intéresse principalement à la paire innu-aimun et français (suivant le contexte sociolinguistique du Québec, où l'on retrouve la majorité des Innus), il est important de mentionner que la paire innu-aimun et anglais correspond elle aussi à une réalité, pour les communautés innues du Labrador. Tel que mentionné au chapitre 1, certains des ouvrages publiés en édition bilingue innue et française sont aussi proposées en édition bilingues innue et anglaise, bien que leur nombre soit moins important. Ainsi, dans une optique de développement d'outils TAL bilingues pour les communautés innues du Labrador, une étude des méthodes de traduction neuronale multilingue mériterait d'être explorée. On pourrait considérer notamment des modèles permettant de mettre à l'usage d'autres paires de langues mieux dotées dans lesquelles l'anglais est associé (on peut penser notamment à l'inuktitut, dont le cas a été discuté précédemment). Moyennant un effort plus considérable, on pourrait penser qu'il y a aussi un grand potentiel

dans la traduction multilingue pour l'innu-aimun, si des corpus multilingues en venaient à être construits avec d'autres langues de la famille algonquienne qui sont en contact avec le français, tel que le cri de l'est et l'atikamekw. Des efforts en ce sens pourraient d'ailleurs profiter à toutes ces paires de langues, du fait de leurs caractéristiques communes.

Pour la suite des choses, quelques perspectives expérimentales à court terme s'offrent à nous. D'abord, tel que mentionné à la section **??** dans notre état de l'art, il pourrait être intéressant d'amener la contribution de traducteurs professionnels, ou étudiants, sous la forme de post-éditions. Un tel processus, dont le bénéfice pour les réslutats de traduction a déjà été démontré (Sanayai Meetei *et al.*, 2020), permettrait d'amener une évaluation qualitative de nos modèles de traduction par des experts en la matière. En créant de nouvelles traduction de référence, cela augmenterait par le fait même la collection de données textuelles bilingues et pourrait améliorer les résultats des modèles, dans un cercle vertueux. Cette post-édition pourrait d'ailleurs contribuer à un processus de *back-translation*. Tel que proposé par Sennrich *et al.* (2016a), les post-éditions pourraient permettre de synthétiser, grâce à nos modèles, de nouvelles paires de phrases bilingues, qui pourraient être mise à contribution dans un nouvel entraînement de nos modèles.

Pour améliorer les performances sur les vers poétiques et les phrases littéraires, deux autres avenues pourraient aussi être explorées : celle de la traduction de document entier (ou du moins d'un plus grand nombre de phrases ou de vers en même temps) et celle de la résolution de coréférences. Dans les deux cas, l'objectif serait d'éviter soit des phrases trop courtes et manquant de contexte (ce qui peut être le cas avec les vers poétiques) ou dont les sujets ou objets font référence au contexte provenant des phrases avoisinantes.

Enfin, sur le plan de la mise en application des modèles obtenus, une prochaine étape envisageable serait le développement d'un outil de recherche d'information interlingue innu-aimun/français pourrait dès maintenant mettre à application les modèles que nous avons développé. L'objectif pourrait être, par exemple, d'étendre la recherche du dictionnaire innu pour que celle-ci puisse prendre en entrée plusieurs mots français à la fois, et considérer le contexte en entier dans ses suggestions de mots innus. Dans un tel cas, les modèles de TAN/TAS pourraient servir à générer une phrase en innu-aimun basée sur celle en français, ce qui pourrait faciliter la recherche, par similarité, des mots innus les plus pertinents pour l'expression complète. Un tel développement aurait l'avantage d'amener un impact bénéfique immédiat à la communauté, plutôt que d'attendre d'éventuels meilleurs résultats de TAN/TAS à plus long terme. Une telle retombée à court terme devrait, tel que nous l'avons amplement souligné dans le chapitre 3, être priorisée dans l'esprit

d’une approche collaborative comme celle que suit le développement dans lequel nous nous inscrivons.

CONCLUSION

Avec les travaux effectués dans le cadre de ce mémoire, nous avons pu jeter les bases de futurs outils d'assistance à la traduction et à l'apprentissage, basés sur le traitement automatique du langage, pour l'innu-aimun. Nous avons apporté les premiers résultats de traduction automatique neuronale et statistique pour la langue. Ces résultats de référence nous ont permis d'évaluer où en est l'innu-aimun en terme de quantité de textes bilingues, et ce qu'il est possible d'obtenir comme performance en traduction automatique à l'heure actuelle, avec ces textes et en utilisant les méthodes existantes.

À travers plusieurs types d'expérimentations, nous avons pu identifier les pistes d'amélioration les plus prometteuses, afin de progresser vers un traducteur ou un assistant traducteur offrant des performances acceptables pour l'utilisation par la communauté. Nous avons de prime abord constaté que la qualité d'alignement était primordial lors de l'ajout de données, dans un contexte comme le nôtre où la quantité total de données disponibles est faible. Nous avons par ailleurs pu identifier, à travers une étude comparative, les méthodes d'alignement les plus pertinentes pour procéder à l'alignement automatique de nouveaux textes bilingues d'innu-aimun, à savoir les méthodes plus anciennes de Moore et de Gale & Church, puisqu'elles nécessitent moins de données pour fournir de bons résultats.

En comparant différentes méthodes de traduction automatique, nous avons constaté que la traduction statistique était très pertinente pour la quantité de données à disposition. En ce sens, il nous semble pertinent de favoriser à court terme la traduction statistique pour la paire innu-aimun et français. De plus, nous avons pu confirmer qu'il est possible de mettre à contribution les textes bilingues innus existants provenant des domaines littéraire et poétique. Malgré les traduction parfois libres et artistiques de tels textes, nous avons observé que le nombre de phrases contenues dans ces derniers permettent tout de même d'obtenir des modèles pouvant fournir des traductions exactes.

Ces travaux, nous avons pu les faire à l'aide de textes parallèles innu-aimun/français alignés de manière collaborative et participative avec des acteurs de milieux éducatifs innus. Nous avons mis en place cette collaboration et cette participation en tâchant d'éviter l'accaparement trop important du temps des experts de la langue, temps qui est précieux pour les communautés. Nous avons aussi tâché d'obtenir dès maintenant, à travers cette collaboration, des bénéfices pour les membres participants et pour la communauté. Ces bénéfices ont pris la forme de banques d'exercices du niveau primaire, et d'activités de pratique pour

les étudiants en traduction d'innu-aimun. La recherche de ces bénéfices immédiats vise à pallier au fait que nous ne pouvons pas atteindre d'outil fonctionnel dès maintenant, et qu'un tel développement collaboratif s'échelonnera plutôt sur le long terme.

En mettant en oeuvre une telle collaboration avec le personnel enseignant l'innu-aimun, ainsi que la participation d'étudiants en traduction, nous avons pu mettre à l'épreuve l'approche de collaboration que nous avions co-conçue avec l'école primaire-secondaire Kanatamat. L'élément central de cette approche est l'arrimage des besoins de recherche avec ceux de la communauté, ainsi que l'apport de bénéfices immédiats pour cette dernière, et ce pour chaque étape du processus de recherche. À la lueur des résultats obtenus, nous avons pu constater un certain succès pour cette approche, ce qui nous pousse à la recommander pour d'autres travaux de recherche en informatique ayant des visées communautaires ou s'effectuant dans un contexte similaire.

Ce mémoire est la première étape dans un projet collaboratif de développement d'outils à plus long terme. Nous pouvons identifier dès lors, les prochaines étapes qui permettront de s'approcher d'outils fonctionnels utiles pour la traduction et l'apprentissage de l'innu-aimun. En terme de traduction automatique, une prochaine étape que nous pouvons envisager est une évaluation plus substantielle et qualitative, par des traducteurs, des résultats de traduction obtenus automatiquement. Combinée à une post-édition de ces traductions automatiques, ces évaluations permettront à la fois d'augmenter la quantité de données bilingues disponibles pour l'entraînement, de mieux identifier les lacunes des modèles entraînés et de mieux cibler les usages qu'on pourrait en faire dès maintenant. Parmis les usages immédiats que l'on pourrait tester avec les modèles obtenus, nous suggérons la recherche d'information interlingue innu-aimun/français dans le dictionnaire. En traduisant (imparfaitement) des phrases du français vers l'innu-aimun, on pourrait possiblement arriver à mettre en oeuvre une recherche contextuelle pour les expressions multi-mots. Ceci pourrait constituer un premier outil d'assistance rudimentaire pour les traducteurs.

Si elles sont évidemment motivées par le but de dépasser les résultats de notre recherche, les deux prochaines étapes que nous proposons sont de premier chef motivées par notre approche collaborative qui, rappelons-le, recherche l'atteinte d'une utilité concrète après chaque étape, dans un développement par et pour la communauté.

Même si un outil basé sur la traduction automatique n'est pas dès aujourd'hui à portée de main, nous souhai-

tons être témoins dans le futur d'une amélioration de la qualité des traductions suggérées. Une amélioration qui, à moyen ou long terme, rendrait possible d'offrir des outils d'assistance plus complets, via une application mobile par exemple. De tels outils, nous l'espérons, pourraient fournir une assistance aux traducteurs d'innu-aimun dans leur travail, ainsi qu'aux jeunes et aux adultes qui apprennent ou se réapproprient leur langue. Si la quantité de données bilingues disponibles actuellement paraît insuffisante, cela pourait être amené à changer, et rapidement, si l'innu-aimun en venait à gagner un statut plus officiel (comme l'inuktitut au Nunavut) ou si plus de fonds étaient rendus disponibles. Par ailleurs, avec la formation récente de nouveux traducteurs pour l'innu-aimun, nous croyons qu'un cercle vertueux est possible.

ANNEXE A
ARTICLES PUBLIÉS DANS LE CADRE DE CE MÉMOIRE

L'annexe qui suit présente les trois articles qui ont été produits dans le cadre de ce mémoire. Le premier article a été soumis et est sous évaluation au moment du dépôt. Les trois autres articles ont été publiés.

Cadotte, A. et Sadat, F. (Soumis en 2023). Developing Innu-Aimun Neural Machine Translation Through Cultural Data : Can Verses and Prose Help Low-Resource Indigenous Models ? Soumis à Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024).

Cadotte, A., Thernish, A.-C. et Sadat, F. (2023). Lost in Innu-Aimun Translation - Re-defining Neural Machine Translation for Indigenous Interpreters and Translators Needs. Dans Proceedings of the International Conference on Human-informed Translation and Interpreting Technology, 342–353., Naples, Italy. Récupéré de https ://hit-it-conference.org/wp-content/uploads/2023/ 07/HiT-IT-2023-proceedings.pdf

Cadotte, A., Le Ngoc, T., Boivin, M. et Sadat, F. (2022). Challenges and Perspectives for Innu-Aimun within Indigenous Language Technologies. Dans Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages, 99–108., Dublin, Ireland. Association for Computational Linguistics. http ://dx.doi.org/10.18653/v1/2022.computel-1.13. Récupéré de https ://aclanthology.org/2022.computel-1.13

Tan Le, N., **Cadotte, A.**, Boivin, M. et Sadat, F.(2022). Deep Learning-Based Morphological Segmentation for Indigenous Languages : A Study Case on Innu-Aimun. Dans Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing, 146–151., Hybrid. Association for Computational Linguistics. http ://dx.doi.org/10.18653/v1/2022.deeplo-1.16. Récupéré de https ://aclanthology.org/2022.deeplo-1.16

ANNEXE B

CERTIFICAT D'ÉTHIQUE EN APPUI AU PROCESSUS DE COLLABORATION

Cette annexe présente le certificat d'éthique qui a été délivré par le Comité d'éthique de la recherche pour les projets étudiants impliquant des êtres humains (CERPÉ plurifacultaire) de l'UQAM pour l'approbation du protocole de recherche suivi dans le cadre de ce mémoire, suivi de l'avis final de conformité.

UQÀM | Comités d'éthique de la recherche avec des êtres humains

No. de certificat : 2023-5307
Date : 2023-01-17

CERTIFICAT D'APPROBATION ÉTHIQUE

Le Comité d'éthique de la recherche pour les projets étudiants impliquant des êtres humains (CERPE plurifacultaire) a examiné le projet de recherche suivant et le juge conforme aux pratiques habituelles ainsi qu'aux normes établies par la *Politique No 54 sur l'éthique de la recherche avec des êtres humains*(2020) de l'UQAM.

- Titre du projet : **Traitement automatique du langage pour la revitalisation de l'innu-aimun**
- Nom de l'étudiant :**Antoine Cadotte**
- Programme d'études : **Maîtrise en informatique (intelligence artificielle)**
- Direction(s) de recherche : **Fatiha Sadat**

Modalités d'application

Toute modification au protocole de recherche en cours de même que tout événement ou renseignement pouvant affecter l'intégrité de la recherche doivent être communiqués rapidement au comité.

La suspension ou la cessation du protocole, temporaire ou définitive, doit être communiquée au comité dans les meilleurs délais.

Le présent certificat est valide pour une durée d'un an à partir de la date d'émission. Au terme de ce délai, un rapport d'avancement de projet doit être soumis au comité, en guise de rapport final si le projet est réalisé en moins d'un an, et en guise de rapport annuel pour le projet se poursuivant sur plus d'une année au plus tard un mois avant la date d'échéance (**2024-01-17**) de votre certificat. Dans ce dernier cas, le rapport annuel permettra au comité de se prononcer sur le renouvellement du certificat d'approbation éthique.

Caroline Coulombe
Professeure, Département de management
Présidente du CERPÉ plurifacultaire

FIGURE B.1 – Certificat d'éthique

UQÀM | Comités d'éthique de la recherche avec des êtres humains

No. de certificat : 2023-5307
Date : 2023-10-27

AVIS FINAL DE CONFORMITÉ

Le Comité d'éthique de la recherche pour les projets étudiants impliquant des êtres humains (CERPÉ plurifacultaire) a examiné le projet de recherche suivant et le juge conforme aux pratiques habituelles ainsi qu'aux normes établies par la *Politique No 54 sur l'éthique de la recherche avec des êtres humains* (janvier 2016) de l'UQAM.

Titre du projet : Traitement automatique du langage pour la revitalisation de l'innu-aimun
Nom de l'étudiant : Antoine Cadotte
Programme d'études : Maîtrise en informatique (intelligence artificielle)
Direction(s) de recherche : Fatiha Sadat

Merci de bien vouloir inclure une copie du présent document et de votre certificat d'approbation éthique en annexe de votre travail de recherche.

Les membres du CERPÉ plurifacultaire vous félicitent pour la réalisation de votre recherche et vous offrent leurs meilleurs voeux pour la suite de vos activités.

Raoul Graf, M.A., Ph.D.
Professeur titulaire, département de marketing
Président du CERPÉ plurifacultaire

FIGURE B.2 – Avis final de conformité

ANNEXE C
TRADUCTION AUTOMATIQUE : MATÉRIEL D'ANALYSE SUPPLÉMENTAIRE

TABLE C.1 – Comparaison qualitative entre modèles des différents corpus : exemple additionnel #1

Phrase source (Innu-Aimun)	Nenua kassinu ka-ashteuani mashinaikana, tanenitsheni tshin ka tutamin ?
Traduction (référence)	Parmi tous les livres qui sont posés là, quels peuvent bien être ceux que tu as écrits ?
Traduction (dict)	**Qu'est-ce que** tu as vu ce que tu as pris ?
Traduction (dict+kapesh)	Qu'as-ce que tu as pris ?
Traduction (dict+jeunesse)	**Qu'est-ce qu**'on m'as appris ?

TABLE C.2 – Comparaison qualitative entre modèles des différents corpus : exemple additionnel #2

Phrase source (Innu-Aimun)	Nika paten utatshekata ek^{u} tshe pakashuian.
Traduction (référence)	Je vais passer à la flamme les pattes arrières de caribou puis je les fendrai pour retirer la moelle.
Traduction (dict)	**Je vais** aller chercher mon **filet** de l'endroit où je vais faire la rivière.
Traduction (dict+kapesh)	**Je vais faire sécher** mon **filet** de l'autre côté de l'autre côté et ensuite je vais faire sécher.
Traduction (dict+jeunesse)	**Je vais faire sécher** mon enfant à l'endroit où j'irai chercher mes collets.

TABLE C.3 – Comparaison qualitative entre modèles des différents corpus : exemple additionnel #3

Phrase source (Innu-Aimun)	Nui nakatuapatam^{u} ashinia mueshaue-kuakushuniti utauia.
Traduction (référence)	Louis surveille les roches alors que son père se dirige vers le large en poussant dans le fond.
Traduction (dict)	J'ai donné un castor à l'endroit où il est allé à l'église.
Traduction (dict+kapesh)	L'enfant va chercher son grand-père quand il va à l'intérieur des terres.
Traduction (dict+jeunesse)	À l'hiver dernier, il y a de l'autre côté de l'autre côté de l'autre côté, c'autre côté de **l'eau** est **très loin**.

[b]1.2

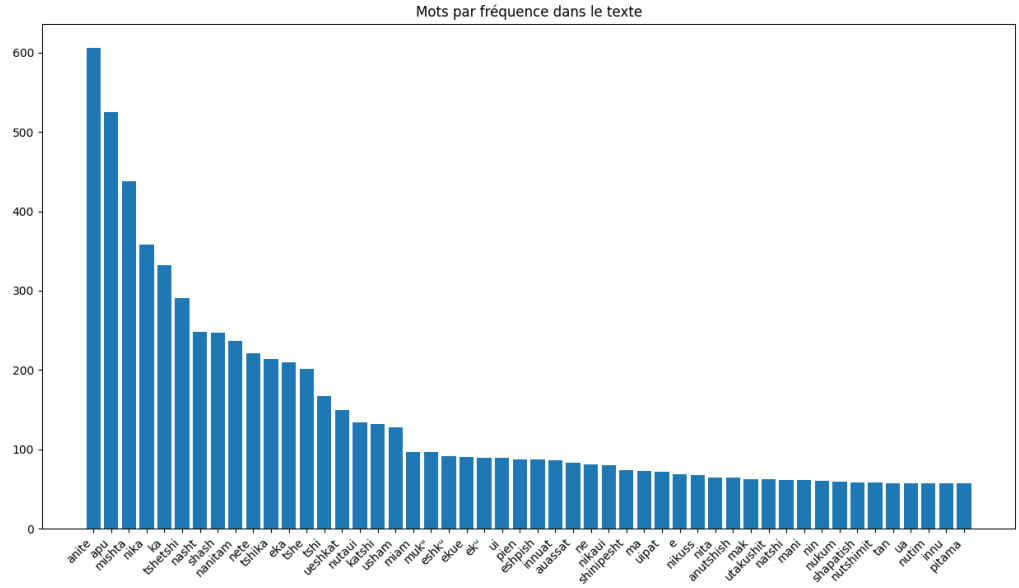

FIGURE C.1 – Dictionnaire : mots par fréquence dans les phrases en innu-aimun

[b]1.2

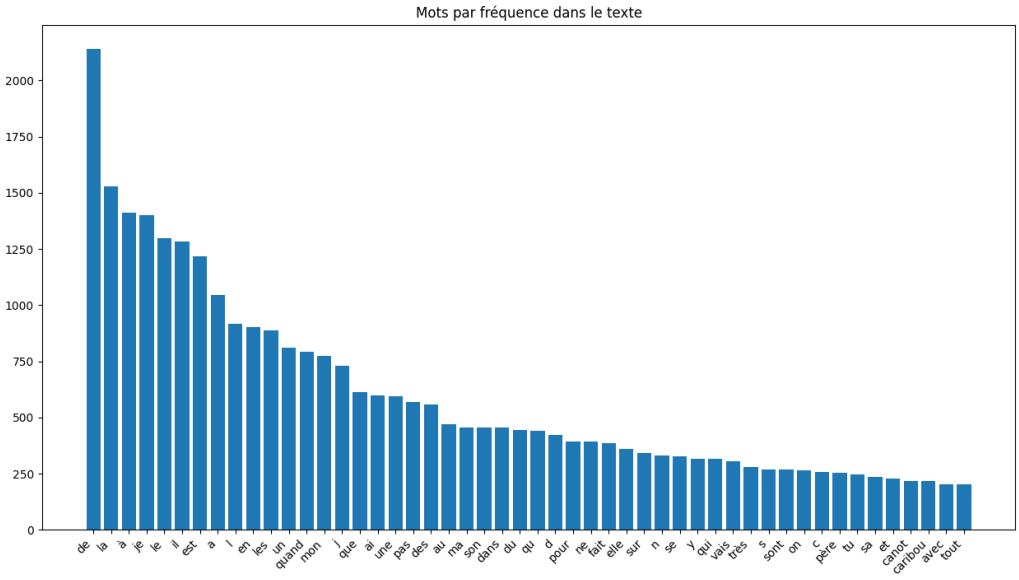

FIGURE C.2 – Dictionnaire : mots par fréquence dans les phrases en français

[b]1.2

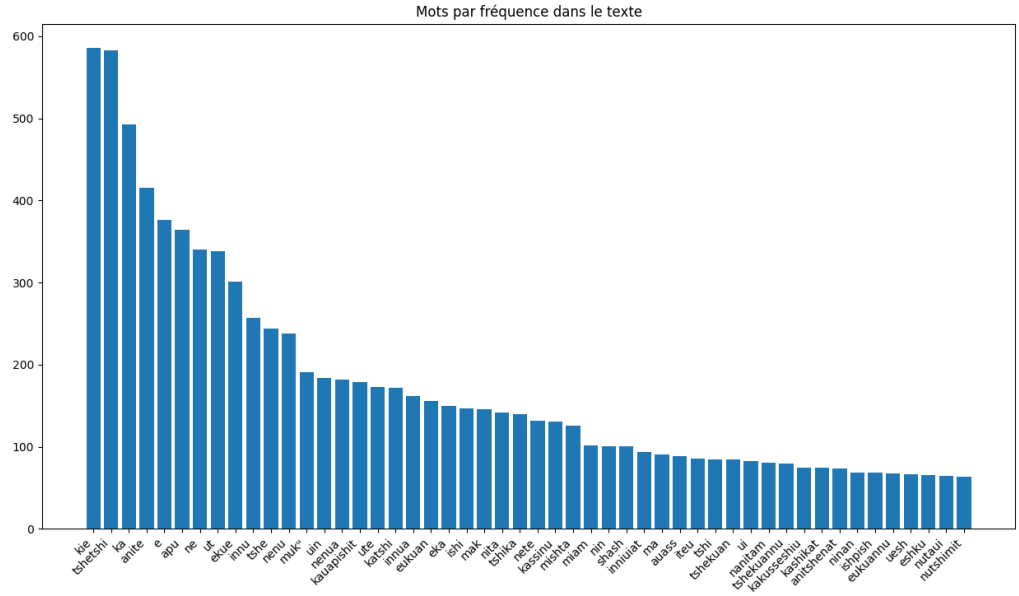

FIGURE C.3 – Kapesh-1+Kapesh-2 : mots par fréquence dans les phrases en innu-aimun

[b]1.2

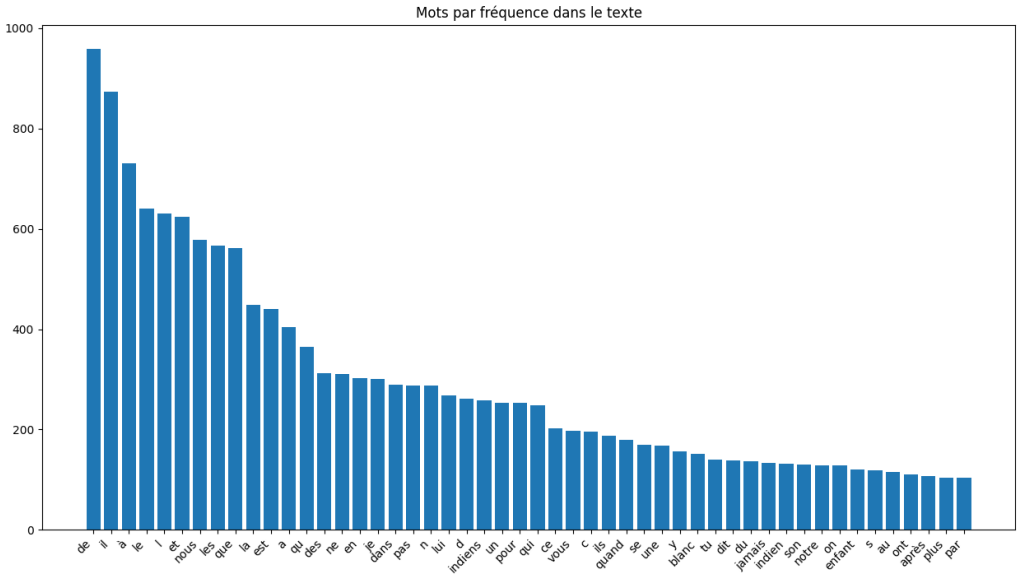

FIGURE C.4 – Kapesh-1+Kapesh-2 : mots par fréquence dans les phrases en français

[b]1.2

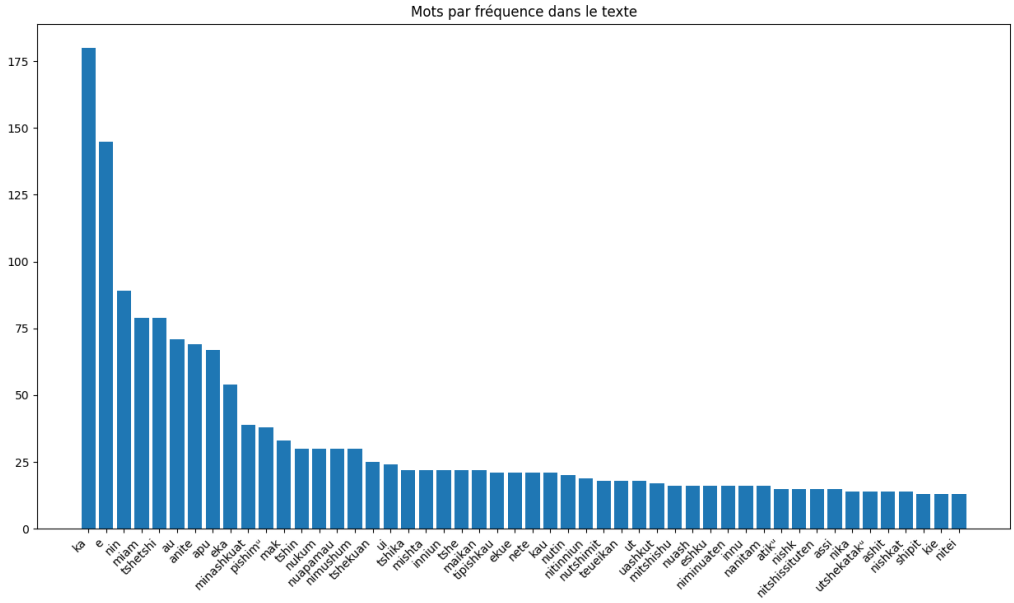

FIGURE C.5 – Jeunesse : mots par fréquence dans les phrases en innu-aimun

[b]1.2

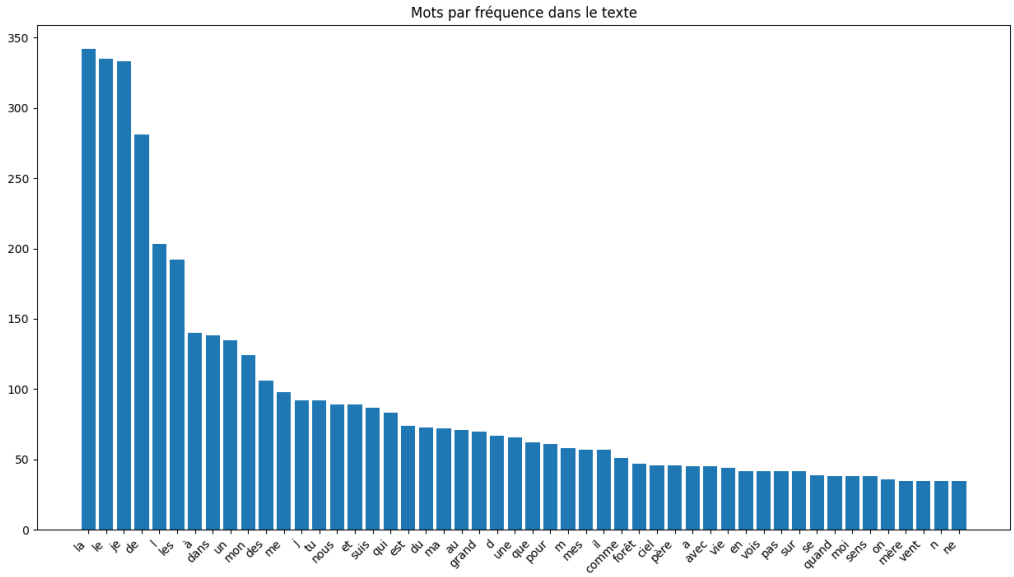

FIGURE C.6 – Jeunesse : mots par fréquence dans les phrases en français

BIBLIOGRAPHIE

Ambroise, J., Junker, M.-O., MacKenzie, M. et Mollen, Y. (2023). Dictionnaire innu en ligne. Récupéré de https://dictionary.innu-aimun.ca/

Arppe, A., Lachler, J., Trosterud, T., Antonsen, L. et N. Moshagen, S. (2016). Basic Language Resource Kits for Endangered Languages : A Case Study of Plains Cree. Dans *Proceedings of the LREC 2016 Workshop "CCURL 2016 – Towards an Alliance for Digital Language Diversity"*, 1–8., Portorož (Slovenia). Récupéré de http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-CCURL2016_Proceedings.pdf#page=8

Artetxe, M. et Schwenk, H. (2019). Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, *7*, 597–610. Place : Cambridge, MA Publisher : MIT Press, http://dx.doi.org/10.1162/tacl_a_00288. Récupéré de https://aclanthology.org/Q19-1038

Bacon, J. et Morali, L. (dir.) (2021). *Nin Auass. Moi l'enfant : Poèmes de la jeunesse innue*. Mémoire d'encrier.

Bahdanau, D., Cho, K. et Bengio, Y. (2016). Neural machine translation by jointly learning to align and translate.

Baraby, A.-M. (2000). The Process of Spelling Standardization of Innu-Aimun (Montagnais). Dans *Indigenous Languages across the Community. Proceedings of the 7th Annual Conference on Stabilizing Indigenous Languages*, p. 17., Toronto, Ontario, Canada. Récupéré de https://eric.ed.gov/?id=ED462244

Baraby, A.-M., Junker, M.-O. et Mollen, Y. (2017). A 45-year old language documentation program first aimed at speakers : the case of the Innu. Récupéré de http://hdl.handle.net/10125/41973

Berg-Kirkpatrick, T., Burkett, D. et Klein, D. (2012). An empirical investigation of statistical significance in NLP. Dans *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 995–1005., Jeju Island, Korea. Association for Computational Linguistics. Récupéré de https://aclanthology.org/D12-1091

Bi, T., Yao, L., Yang, B., Zhang, H., Luo, W. et Chen, B. (2020). Constraint translation candidates : A bridge between neural query translation and cross-lingual information retrieval.

Bird, S. (2020). Decolonising Speech and Language Technology. Dans *Proceedings of the 28th International Conference on Computational Linguistics*, 3504–3519., Barcelona, Spain (Online). International Committee on Computational Linguistics. http://dx.doi.org/10.18653/v1/2020.coling-main.313. Récupéré de https://aclanthology.org/2020.coling-main.313

Bontogon, M. A. (2016). *Evaluating nêhiyawêtân : A computer assisted language learning (CALL) application for Plains Cree*. (Thèse de doctorat). University of Alberta. Récupéré de https://doi.org/10.7939/R3VD6P81C

Cadotte, A., Le Ngoc, T., Boivin, M. et Sadat, F. (2022). Challenges and Perspectives for Innu-Aimun within Indigenous Language Technologies. Dans *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, 99–108., Dublin, Ireland. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2022.computel-1.13. Récupéré de https://aclanthology.org/2022.computel-1.13

Cadotte, A., Thernish, A.-C. et Sadat, F. (2023). Lost in Innu-Aimun Translation - Re-defining Neural Machine Translation for Indigenous Interpreters and Translators Needs. Dans *Proceedings of the International Conference on Human-informed Translation and Interpreting Technology*, 342–353., Naples, Italy. Récupéré de https://hit-it-conference.org/wp-content/uploads/2023/07/HiT-IT-2023-proceedings.pdf

Drapeau, L. (1991). *Dictionnaire montagnais-français*. Presses de l'Université du Québec.

Drapeau, L. (2014). *Grammaire de la langue innue*. PUQ.

Drapeau, L. et Lambert-Brétière, R. (2013). The Innu Language Documentation Project. Dans *Proceedings of the 17th Foundation for Endangered Languages Conference*. Récupéré de https://ir.library.carleton.ca/pub/13609

Dror, R., Baumer, G., Shlomov, S. et Reichart, R. (2018). The hitchhiker's guide to testing statistical significance in natural language processing. Dans *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, 1383–1392., Melbourne, Australia. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/P18-1128. Récupéré de https://aclanthology.org/P18-1128

Eisenstein, J. (2019). *Introduction to Natural Language Processing*. Adaptive Computation and Machine Learning series. MIT Press. Récupéré de https://books.google.ca/books?id=72yuDwAAQBAJ

Fontaine, N. (2017). *Manikanetish : Petite Marguerite*. Roman (Mémoire d'encrier (Firme))). Mémoire d'encrier. Récupéré de https://books.google.ca/books?id=zZqZtAEACAAJ

Gale, W. A. et Church, K. W. (1993). A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics*, *19*(1), 75–102. Place : Cambridge, MA Publisher : MIT Press. Récupéré de https://aclanthology.org/J93-1004

Geron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow : Concepts, Tools, and Techniques to Build Intelligent Systems* (2nd éd.). O'Reilly Media, Inc.

Haddow, B., Bawden, R., Barone, A. V. M., Helcl, J. et Birch, A. (2021). Survey of Low-Resource Machine Translation. _eprint : 2109.00486.

Heffernan, K., Çelebi, O. et Schwenk, H. (2022). Bitext mining using distilled sentence representations for low-resource languages. Dans *Findings of the Association for Computational Linguistics : EMNLP 2022*, 2101–2112., Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. Récupéré de https://aclanthology.org/2022.findings-emnlp.154

Hébert, L.-P. (1988). Le père jean-baptiste de la brosse, professeur, linguiste et ethnographe chez les montagnais du saguenay (1766-1782). *Sessions dx27 ;étude - Société canadienne dx27 ;histoire de*

lx27 ;Église catholique, *55*, 7–39. http://dx.doi.org/https://doi.org/10.7202/1006944ar

James, G., Witten, D., Hastie, T. et Tibshirani, R. (2014). *An Introduction to Statistical Learning : With Applications in R*. Springer Publishing Company, Incorporated.

Joanis, E., Knowles, R., Kuhn, R., Larkin, S., Littell, P., Lo, C.-k., Stewart, D. et Micher, J. (2020). The Nunavut Hansard Inuktitut–English Parallel Corpus 3.0 with Preliminary Machine Translation Results. Dans *Proceedings of the 12th Language Resources and Evaluation Conference*, 2562–2572., Marseille, France. European Language Resources Association. Récupéré de https://aclanthology.org/2020.lrec-1.312

Junker, M.-O., Mollen, Y., St-Onge, H. et Torkornoo, D. (2016). Integrated web tools for Innu language maintenance. Dans *Papers of the 44th Algonquian Conference*, 192–210.

Junker, M.-O. et Stewart, T. (2008). Building search engines for Algonquian languages. *Algonquian Papers-Archive*, *39*.

Jurafsky, D. et Martin, J. (2009). *Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall series in artificial intelligence. Pearson Prentice Hall. Récupéré de https://books.google.ca/books?id=fZmj5UNK8AQC

Kapesh, A. (2019). *Je suis une maudite Sauvagesse : Eukuan nin matshi-manitu innushkueu*. Mémoire d'encrier.

Kapesh, A. (2020). *Qu'as-tu fait de mon pays ? Tanite nene etutamin nitassi ?* Mémoire d'encrier.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. Dans *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 388–395., Barcelona, Spain. Association for Computational Linguistics. Récupéré de https://aclanthology.org/W04-3250

Koehn, P. (2005). Europarl : A parallel corpus for statistical machine translation. Dans *Proceedings of Machine Translation Summit X : Papers*, 79–86., Phuket, Thailand. Récupéré de https://aclanthology.org/2005.mtsummit-papers.11

Koehn, P., Och, F. J. et Marcu, D. (2003). Statistical phrase-based translation. Dans *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 127–133. Récupéré de https://aclanthology.org/N03-1017

LeCun, Y., Bengio, Y. et Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. http://dx.doi.org/10.1038/nature14539. Récupéré de https://doi.org/10.1038/nature14539

Leroux, S. (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. Publisher : Université du Québec à Montréal, http://dx.doi.org/https://doi.org/10.7202/1033619ar

Lewis, J. E. (2020). Indigenous Protocol and Artificial Intelligence. Dans *Position Paper*, Honolulu. Indigenous Protocol and Artificial Intelligence Working Group and the Canadian Institute for Advanced Research. Récupéré de https://spectrum.library.concordia.ca/id/eprint/986506/

Mager, M., Mager, E., Kann, K. et Vu, N. T. (2023). Ethical considerations for machine translation of indigenous languages : Giving a voice to the speakers. Dans A. Rogers, J. Boyd-Graber, et N. Okazaki (dir.). *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, 4871–4897., Toronto, Canada. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2023.acl-long.268. Récupéré de https://aclanthology.org/2023.acl-long.268

Mailhot, J., MacKenzie, M. et Oxford, W. (2012). *Dictionnaire Innu-Français*. Institut Tshakapesh.

Manning, C. et Schutze, H. (1999). *Foundations of Statistical Natural Language Processing*. Foundations of Statistical Natural Language Processing. MIT Press. Récupéré de https://books.google.ca/books?id=ZL34DwAAQBAJ

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

Moore, R. C. (2002). Fast and accurate sentence alignment of bilingual corpora. Dans *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas : Technical Papers*, 135–144., Tiburon, USA. Springer. Récupéré de https://link.springer.com/chapter/10.1007/3-540-45820-4_14

Møller, H. (2016). Culturally safe communication and the power of language in Arctic nursing. *Études/Inuit/Studies*, *40*(1), 85–104. Publisher : [Université Laval, Études/Inuit/Studies]. Récupéré de http://www.jstor.org/stable/44254675

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohungbe, T., Akinola, S. O., Muhammad, S., Kabongo Kabenamualu, S., Osei, S., Sackey, F., Niyongabo, R. A., Macharm, R., Ogayo, P., Ahia, O., Berhe, M. M., Adeyemi, M., Mokgesi-Selinga, M., Okegbemi, L., Martinus, L., Tajudeen, K., Degila, K., Ogueji, K., Siminyu, K., Kreutzer, J., Webster, J., Ali, J. T., Abbott, J., Orife, I., Ezeani, I., Dangana, I. A., Kamper, H., Elsahar, H., Duru, G., Kioko, G., Espoir, M., van Biljon, E., Whitenack, D., Onyefuluchi, C., Emezue, C. C., Dossou, B. F. P., Sibanda, B., Bassey, B., Olabiyi, A., Ramkilowan, A., Öktem, A., Akinfaderin, A. et Bashir, A. (2020). Participatory research for low-resourced machine translation : A case study in African languages. Dans *Findings of the Association for Computational Linguistics : EMNLP 2020*, 2144–2160., Online. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2020.findings-emnlp.195. Récupéré de https://aclanthology.org/2020.findings-emnlp.195

Papineni, K., Roukos, S., Ward, T. et Zhu, W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. Dans *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318., Philadelphia, Pennsylvania, USA. Association for Computational Linguistics. http://dx.doi.org/10.3115/1073083.1073135. Récupéré de https://aclanthology.org/P02-1040

Popović, M. (2015). chrF : character n-gram F-score for automatic MT evaluation. Dans *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 392–395., Lisbon, Portugal. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/W15-3049. Récupéré de https://aclanthology.org/W15-3049

Post, M. (2018). A call for clarity in reporting BLEU scores. Dans *Proceedings of the Third Conference on Machine Translation : Research Papers*, 186–191., Brussels, Belgium. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/W18-6319. Récupéré de https://aclanthology.org/W18-6319

Sabet, A., Gupta, P., Cordonnier, J.-B., West, R. et Jaggi, M. (2020). Robust cross-lingual embeddings from parallel sentences.

Sanayai Meetei, L., Singh, T. D., Bandyopadhyay, S., Vela, M. et van Genabith, J. (2020). English to Manipuri and Mizo Post-Editing Effort and its Impact on Low Resource Machine Translation. Dans *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, 50–59., Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI). Récupéré de https://aclanthology.org/2020.icon-main.7

Sennrich, R., Haddow, B. et Birch, A. (2016a). Improving Neural Machine Translation Models with Monolingual Data. Dans *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, 86–96., Berlin, Germany. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/P16-1009. Récupéré de https://aclanthology.org/P16-1009

Sennrich, R., Haddow, B. et Birch, A. (2016b). Neural Machine Translation of Rare Words with Subword Units. Dans *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, 1715–1725., Berlin, Germany. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/P16-1162. Récupéré de https://aclanthology.org/P16-1162

St-Gelais, M. (2022). *Une histoire de la littérature innue*. Montréal, Canada : PUQ, Institut Tshakapesh. Récupéré de https://www.puq.ca/catalogue/livres/une-histoire-litterature-innue-4253.html

Tan Le, N., Cadotte, A., Boivin, M. et Sadat, F. (2022). Deep Learning-Based Morphological Segmentation for Indigenous Languages : A Study Case on Innu-Aimun. Dans *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, 146–151., Hybrid. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2022.deeplo-1.16. Récupéré de https://aclanthology.org/2022.deeplo-1.16

Thompson, B. et Koehn, P. (2019). Vecalign : Improved Sentence Alignment in Linear Time and Space. Dans *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 1342–1348., Hong Kong, China. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/D19-1136. Récupéré de https://aclanthology.org/D19-1136

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u. et Polosukhin, I. (2017). Attention is all you need. Dans I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus,

S. Vishwanathan, et R. Garnett (dir.). *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc. Récupéré de https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Ziemski, M., Junczys-Dowmunt, M. et Pouliquen, B. (2016). The United Nations parallel corpus v1.0. Dans *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 3530–3534., Portorož, Slovenia. European Language Resources Association (ELRA). Récupéré de https://aclanthology.org/L16-1561