Cree Corpus: A Collection of nêhiyawêwin Resources

Daniela Teodorescu, Josie Matalski, Delaney Lothian,
Denilson Barbosa, Carrie Demmans Epp
Dept. of Computing Science
University of Alberta
{dteodore,jmatalsk,dlothian,denilson,cdemmansepp}@ualberta.ca

Abstract

Plains Cree (nêhiyawêwin) is an Indigenous language that is spoken in Canada and the USA. It is the most widely spoken dialect of Cree and a morphologically complex language that is polysynthetic, highly inflective, and agglutinative. It is an extremely low resource language, with no existing corpus that is both available and prepared for supporting the development of language technologies. To support nêhiyawêwin revitalization and preservation, we developed a corpus covering diverse genres, time periods, and texts for a variety of intended audiences. The data has been verified and cleaned; it is ready for use in developing language technologies for nêhiyawêwin. The corpus includes the corresponding English phrases or audio files where available. We demonstrate the utility of the corpus through its community use and its use to build language technologies that can provide the types of support that community members have expressed are desirable. The corpus is available for public use[1].

1 Introduction

Recent work with Indigenous persons has shown that some want advanced technologies to support the learning and use of their languages. The Cree and Métis persons involved in this study stated a desire for technologies such as an app to help with learning the structure of the language for conversation, translation, and AI agents that resemble a speaker (Lothian et al., 2019). Participants wanted these tools to support interaction in nêhiyawêwin (Plains Cree) or the learning of this language. All of these larger ideas are dependent on core language technologies such as language models, speech recognition, speech synthesis, or machine translation. However, a lack of publicly available corpora hinders the development of such technologies for low-resource languages like nêhiyawêwin.

Government policies have contributed towards supporting the preservation and revitalization of some Indigenous languages, e.g., Inuktitut (Joanis et al., 2020). However, many have not benefited from this level of support for developing resources and technologies. Recently, some government informational material such as voter guides or COVID-19 pamphlets have been translated into nêhiyawêwin. Nevertheless, the availability of resources is still limited and short texts or other resources are distributed across libraries and the Internet. To understand why this is the case, we need to reflect on the colonial practices that have attempted to eradicate a language and people. Previous and on-going government policies and practices, such as the implementation of residential schools (Bombay et al., 2011), have left a small number of fluent speakers and language resources for nêhiyawêwin-speaking communities.

These practices prevented and continue to prevent the development of language technologies because state-of-the-art statistical and neural models require large amounts of text. To work towards addressing this issue, we created a nêhiyawêwin corpus from various sources. Our corpus is composed of 49,038 words and 3,727 lines of text in Standard Roman Orthography (SRO), 10 texts in syllabics, and 1,026 lines of English-nêhiyawêwin parallel data.

To the best of our knowledge, this is the first collection of processed nêhiyawêwin data ready for use to build language technologies. The most similar existing work includes a small collection of nêhiyawêwin text, lexical, and audio resources in their original formats (Open Language Archives Community). There is also a morphosyntactic tagged corpus (Arppe et al., 2020) which can be accessed by searching for words, lemmas, and mor-

[1] https://github.com/EdTeKLA/IndigenousLanguages_Corpora

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 6354 - 6364
May 22-27, 2022

phosyntactic information through a web interface. A targeted corpus of child-directed speech in Cree has been shared through the ACQDIV Database (Moran, 2016). However, this corpus contains materials in the northern dialect of East Cree (iyiyiu-Ayamiwin) rather than Plains Cree (nêhiyawêwin).

In response to the limited availability of resources and tools, this work contributes a collection of ready to use resources to enable the development of language technologies that can support the preservation and revitalization of nêhiyawêwin. We demonstrate the practicality of the corpus through its use by community-based teachers of nêhiyawêwin. Using these materials has informed their lesson plans. Further, we describe the ongoing development of predictive language models using the contributed corpus. These models enable predictive text that is expected to provide some of the language support needs that have been expressed by nêhiyawêwin speakers. With this work, we aim to inspire future data collection and sharing of nêhiyawêwin resources that are aligned with community interests.

2 nêhiyawêwin and Technology

Plains Cree is called nêhiyawêwin by its speakers, and it is not capitalized. nêhiyawêwin is a widely-spoken dialect of the Indigenous language that English-speakers call Cree: nêhiyawêwin is the mother tongue for approximately 3,655 speakers, and it is the language spoken most at home for approximately 2,165 persons (Statistics Canada, 2018). nêhiyawêwin is an extremely low resource language, with the official designation of being a "developing" language; it is at stage 5 on the Expanded Graded Intergenerational Disruption Scale (EGIDS) (Lewis and Simons, 2010) so it "is in vigorous use, with literature in a standardized form being used by some though this is not yet widespread or sustainable". Therefore, the ability to create language technologies for nêhiyawêwin is limited due to the minimal amount of monolingual and parallel data available.

Current language technologies for nêhiyawêwin include Finite State Transducers (FSTs) that have been used for tasks such generating word forms and conjugating verbs in online dictionaries (Arppe et al., 2016), representing nominal morphology (Snoek et al., 2014), and spell checking (Arppe et al., 2016).

It is not surprising that FSTs are one of the few technologies that exist given that nêhiyawêwin is a polysynthetic, agglutinative, and highly inflective language, which complicates the task of creating language technologies. These characteristics allow the meaning of a single token or word to map to that of a full phrase or sentence in English. For example, 'kimîciso' maps to 'you all eat' in English.

nêhiyawêwin has two writing systems: SRO and syllabics. A single character in syllabics represents one or more SRO characters (e.g., σ is ni in SRO and △ is i). Complicating this, is the variability in how these writing systems have been used and continue to be used across regions and time. This variability means that choices must be made with respect to the writing systems and 'standards' that are followed when developing language technologies. These are difficult choices and each community may have different preferences, which means that tools for converting across varied writing systems would help to maintain community norms. An example of such a tool is the SRO-syllabics converter (Antonio Santos, 2021). While any one project cannot address all considerations, these considerations are an important part of developing language technologies to support the revitalization and use of this language.

3 Corpus

Our corpus contains text from several domains making it a diverse collection of nêhiyawêwin resources (see Table 1). We collected materials from different genres such as Bible hymns, educational resources, and children's stories as well as content from social media such as Twitter and Facebook. As such, our corpus spans several time periods. For example, Bible translations are based on a bible from 1908, whereas social media content and educational documents are from the 2000's, with some being from the last couple of years. The category 'Other' contain texts such as election pamphlets, voter guides, speaker stories, and a first year university nêhiyawêwin workbook.

The material is organized into folders by category or source along with its copyright information for how the public can use them. Where nêhiyawêwin-English parallel texts exist, the folder contains a cleaned and aligned version of these texts; a given line in one language file corresponds to the same line in the other language file. Syllabics versions of texts are provided where available. Some texts also have an accompanying audio file.

Source	Source Count *SRO*	 *Syllabics*	Parallel Lines	Number of Tokens *nêhiyawêwin*	 *English*
Bible Songs	4	0	0	32,569	36,427
K- 12 Educational	5	0	494	5,912	8,764
Social Media and Blogs	13	4	203	604	1,064
Scholarly Articles	3	0	130	285	550
Children's Stories	3	3	56	284	263
Other	14	3	143	10,197	11,043
Total	*42*	*10*	*1,026*	*49,851*	*58,111*

Table 1: Data source counts by category. Line counts do not include BoW materials. Syllabics texts are included in the nêhiyawêwin token count.

Before adding a text to our corpus, we checked the copyright and license or obtained permission from the content creator. We provide a bag-of-words (BoW) representation when text was under copyright or the content owners felt this was an acceptable alternative to sharing the original text. These BoW files contain a list of words from the original text and their usage counts. As these files only contain individual words, there is no nêhiyawêwin-English mapping because there is often no one-to-one translation between nêhiyawêwin and English words.

Table 2 provides descriptive statistics for all text sources in the corpus. We provide mean (*M*) and standard deviation (*SD*) of data throughout this paper.

4 Creating the Dataset

To build this corpus we first identified sources of nêhiyawêwin text. We then extracted the text. Following extraction, we aligned the texts across languages and performed additional processing.

4.1 Identifying Texts

We used Google search to find nêhiyawêwin text online and entered keywords such as 'nêhiyawêwin text' and 'plains cree text'. Please see Appendix A for a full list of keywords. Some websites continually updated their content with new material (e.g., Cree Literacy Network[2]) so we returned and checked those sites for additional content.

Data were identified as nêhiyawêwin by carefully inspecting the source and its description. The contents of the text were also checked by one of our team members who had been trained in how to differentiate between dialects of Cree. This step ensured the text was in the targeted dialect. If uncertainties arose, such as when facing unfamiliar accents, hyphens, or characters, a nêhiyawêwin speaker would verify whether the text was Plains Cree.

Language	Vocab.	Sentence Length	Token Length
nêhiyawêwin	15,202	4.6 (3.58)	7.6 (5.36)
English	3,972	6.8 (4.42)	3.6 (2.06)

Table 2: Number of unique vocabulary (Vocab.), sentence length, as *M (SD)* of tokens, and token length in characters, as *M (SD)*. BoW resources are excluded from sentence length.

4.2 Permission

Copyright information was verified to see if the text could be shared or perhaps if the copyright would allow BoW format. For texts that contain Elders' stories, described below, permission from speakers was obtained to share the stories. The resources in the corpus can be publicly used as allowed by the copyright information detailed on GitHub for a particular source.

4.3 Obtaining Texts

Text was extracted from the original sources (e.g., PDFs, webpages) and converted into plain text. Care was taken to ensure the text was properly copied and that it excluded irrelevant information (e.g., HTML markup or English annotations).

Some data was collected by scraping websites, where licensing allowed it. When licensing did not permit scraping, we contacted site owners to obtain permission. In some cases, they shared the raw materials with us for inclusion in the corpus. Parallel phrases in English and nêhiyawêwin were extracted when available. The retrieved nêhiyawêwin texts

[2] https://creeliteracy.org/

Language	Text
nêhiyawêwin	**Before:** êwîpîk’skwâtamân tân’si êkîpêhisikiskinohamâsoyân nêhiyawêwin. âskaw âyiman ôma ôta, ôtênâhk. tâpitaw mâna ayisiyiniwak êhâpacihtâcik âkayâsîmowin. **After:** êwîpîk’skwâtamân tân’si êkîpêhisikiskinohamâsoyân nêhiyawêwin. âskaw âyiman ôma ôta, ôtênâhk. tâpitaw mâna ayisiyiniwak êhâpacihtâcik âkayâsîmowin.
English	**Before:** I’m going to speak about how I came to teach myself Cree. Sometimes it’s hard here, in the city. People usually always use English. **After:**I’m going to speak about how I came to teach myself Cree. Sometimes it’s hard here, in the city. People usually always use English.

Table 3: Aligning text where the number of sentences in one language maps to the other. This example is from the oral stories that Neil, an Elder, shared with us.

nêhiyawêwin	**English**
êpêkakwêcim’kawiyân ôma, tanêhiyawîyân. êkwa anima âya, k’tisipîk’skwêwin’nân niyanân kayâs kâkîpêhohpikêyâhk.	I’ve been asked this, to speak Cree, and well, of our language a long time ago when we were growing up

Table 4: Aligning text where the number of corresponding sentences in one language does not map to the same number in the other. This example is from an Elder’s story (Theresa).

may have used SRO or syllabics. Some were accompanied by an audio file. The availability of formats varied from resource to resource.

Beyond these publicly available online resources, we collected resources from the field. These resources are recordings of Elders who chose a story to tell us. They gave us permission to use and share these stories for the purposes of supporting learning and developing language technologies that could do the same. Most of the shared stories relate to their personal lives or socio-political issues. These recordings were made over a summer by attending cultural events and interacting with community members. The recordings were transcribed and translated into English in some cases. Three speakers of nêhiyawêwin took part in the transcription, translation, and verification process.

4.4 Aligning Texts Across Languages

Where parallel texts were available, alignment was performed before other preprocessing or data cleaning. Most parallel texts contained some spacing markers, such as line breaks for paragraphs or spaces for phrases. In these scenarios, single sentences or phrases were easily aligned to each other. Challenges arose when a paragraph contained a different number of sentences across languages. Since we aimed to provide sentence or phrase alignments in the corpus, we needed to distinguish how a sentence in one language is expressed in the other.

In longer texts, when multiple sentences in nêhiyawêwin mapped to one sentence in English, or vice versa, this mapping was used as the alignment to maintain the original meaning of the text. This situation was prominent in Biblical texts. In shorter texts, a nêhiyawêwin speaker reviewed the text and decided on the appropriate alignment. We note that this process of aligning paragraphs, then text within paragraphs is demonstrated to outperform alignment that does not account for paragraph boundaries (Joanis et al., 2020). We provide examples of aligning sentences in the simple case and more challenging case in Table 3 and Table 4.

4.5 Preprocessing

Preprocessing was only performed on texts that used the SRO writing system. Texts in syllabics did not undergo the below-described preprocessing.

We focused on preprocessing SRO texts for several reasons. It was relatively easy to obtain texts in SRO, which meant that there were more of them. SRO representations of the language vary in their use of diacritics and other conventions, which means that combining sources requires some element of normalization so that the texts can be jointly used. Moreover, one of the intended uses for our corpus is to support instructional activities for local courses, and SRO is the first writing

Before		After	
nêhiyawêwin	**English**	**nêhiyawêwin**	**English**
ê-wâpamikot	S/he was seen by him/her	ê wâpamikot	she was seen by him
		ê wâpamikot	she was seen by her
		ê wâpamikot	he was seen by him
		ê wâpamikot	he was seen by her
kimâmitonêyimitinân	We are thinking of you (one)	kimâmitonêyimitinân	we are thinking of you
Piko tanima Kânata Pimipahtâwin atoskêwêkamik akâmi Kânata (pônipayiwin ihtakon)	At any Elections Canada office across Canada (deadlines apply)	piko tanima kânata pimipahtâwin atoskêwêkamik akâmi kânata pônipayiwin ihtakon	at any elections canada office across canada deadlines apply

Table 5: Manual preprocessing examples. From (Muehlbauer, 2011; Ogg, 2020; Elections Canada, 2019b,a)

system that students learning nêhiyawêwin at the University of Alberta are taught.

The writing system used by speakers differs by community, where some use SRO and others use syllabics. This is also the case for the communities with which we have worked. The choice of writing systems and the considerations surrounding that choice are further discussed in Sections 5 and 7.

4.5.1 Manual Preprocessing

Before running the processing script[1], we manually identified the use of slashes or parentheses. When slashes were used, usually in English text to denote gender or possible alternative phrasings, we ensured that the nêhiyawêwin data would represent all possibilities (see Table 5). For example, we would remove the slash from the English sentence, generate a new English sentence with the alternative gender or phrase, and duplicate the nêhiyawêwin sentence to represent that the nêhiyawêwin text could have this alternative meaning in English. As the aim of this corpus is to develop language technologies, we wanted to ensure that all alternative genders or meanings from the text were included so that it would support the development of models that were as robust as possible given the data. See Table 5 for an example.

Parentheses were mainly used in English sentences to provide additional context. If the text in parentheses provided alternative phrasing, the alternative sentence in English would be constructed with the same nêhiyawêwin meaning mapped to it. This follows a similar pattern to that used with slashes for options like he or she. If the parenthetical expression did not provide an alternative or additional context, it was removed. Parentheses were removed manually in this process and not considered as punctuation to be kept in the preprocessing script, which we describe below. This initial manual process addressed the varying nature of each case and our desire to extract as much information as possible from the text.

4.5.2 Automated Preprocessing

Following the manual preprocessing, a Python script was run on the data files. The script follows a similar pattern for both nêhiyawêwin and English, with slight modifications for each.

Since nêhiyawêwin can be written with different types of diacritics used to represent the same information in SRO (e.g., ā, á, â), we converted all accents to circumflex to maintain consistency within the corpus. A different choice could have easily been made. Because each community may have a different preference, we have included a script that can be modified so that the corpus can be re-standardized according to a specific community's preferences.

All text was converted to lowercase. The only punctuation the script does not remove is periods, exclamation marks, question marks, colons, commas, apostrophes, and single quotes. Each of these punctuation markers are represented as a single token by inserting a space before them.

Hyphens are preprocessed differently from other punctuation. Because nêhiyawêwin and English use hyphens differently, we applied rules specific to each language. In English, hyphens were removed and replaced with a space because the words surrounding the hyphen could often stand alone

Language	Text
nêhiyawêwin	**Before:** ātiht kinosēwak misikitiwak māka ātiht apisīsisiwak **After:** âtiht kinosêwak misikitiwak mâka âtiht apisîsisiwak **Before:** kâ-pimwêwêhahk okakêskîhkêmowina **After:** kâ pimwêwêhahk okakêskîhkêmowina
English	**Before:** Some fish are big, but some are small. **After:** some fish are big , but some are small . **Before:** he-drums-people-into-the-afterlife's counselling speeches **After:** he drums people into the afterlife 's counselling speeches

Table 6: Automated preprocessing examples. The first excerpt is from Twitter and the second is from (Muehlbauer, 2011)

and maintain their meaning. There are no vowel combinations in nêhiyawêwin; however combining morphemes can cause two vowels to border each other. To address this, some authors insert a hyphen and some insert an "h". The justification for the latter is that the transition in speaking these vowels is not harsh, and an "h" indicates a softer transition. We chose to follow the "h" joiner standard. Consequently, the hyphen was replaced with the letter "h" when there was a vowel (i.e., a, i, o, â, ê, î, ô) on both sides of the hyphen. In all other cases, the hyphen was removed. Any other remaining markers were removed, including ellipsis, double quotes, and numbers. See Table 6 for a text-cleaning example.

5 Ethical Considerations

Now that we have described how the corpus was created, we need to discuss ethical considerations around the creation and use of such resources. The process of creating language technologies for any community of speakers should be guided by the goals and interests of the respective community. Natural language processing (NLP) research should directly involve the language communities for which the technologies are being designed, as it will directly impact the speakers of the language. Further, the process of constructing these technologies should be clear to the community so there is an understanding of the data required for the model and how it will be used. For example, communities may wish to see language technologies such as text-to-speech to honor an oral tradition. However, these systems require an underlying model trained on corresponding audio and text for the language, which may or may not be in accordance with a community's wishes.

In direct terms, the existence of this corpus in itself is not an invitation to make Indigenous language models and technologies independently and without consultation. As discussed by Pine and Turin (2017), successful Indigenous language revitalization projects must be "grounded in local understandings of impact and success, rooted in the lived experiences and aspirations of Indigenous communities."

An important consideration when developing language technologies using corpora and language models is the nature of the language used to train those models. For example, language models trained on Internet texts (e.g., GPT-3) have been subject to scrutiny following the revelation of racist and generally offensive outputs (Floridi and Chiriatti, 2020). Those who use the developed nêhiyawêwin corpus should note the potential for problematic outcomes when the data is used to support certain types of language technologies. This potential comes from the inclusion of biblical texts. While biblical texts are widely used for tasks such as machine translation (Mohler and Mihalcea, 2008), they could advance the harmful legacies of Christianity-related efforts and government policies that used religion to control and harm Indigenous groups (Bradford and Horton, 2016). The translation of bibles into local Indigenous languages was a means of furthering colonization (Pine and Turin, 2017). In addition, there are certain bible passages within our corpus that may be considered violent or aggressive in nature, e.g., "May sinners be destroyed from the earth… may the wicked be no more" (Psalm 1). This kind of text, paired with the history between the church and Indigenous peoples, should be used with caution, especially when designing language technologies that produce language (e.g., machine translation).

An additional important note is that, between aligning texts and automatically extracting text from varied sources, it is possible for there to be mistakes or inconsistencies. This should be taken into consideration when using the corpus. We also welcome edits and contributions.

Beyond the above considerations, each of the choices that we made during data cleaning has the potential to have normative effects on the language. Some may view norming and standardization as a benefit (Mager et al., 2018; United Nations, 2019). However, it also risks the loss of language variety that is often valued by community members. Consequently, we include our data cleaning scripts within the repository so that others may adapt them and transform the data into the version of SRO or syllabics that meets their needs.

6 Corpus Use

The corpus is already being used to support community needs as part of a broader project for developing language learning technologies and technologies to support language use. Within this context, corpus materials are being used to help people learn nêhiyawêwin. Materials are also being used to develop language models that support tasks that community members who are learning nêhiyawêwin would like supported. We briefly discuss these ongoing activities to demonstrate the utility of the corpus.

6.1 Supporting Instruction in Communities

As part of developing language-learning technologies, several teachers of nêhiyawêwin who work in and come from different nêhiyawêwin-speaking communities have joined our group. These teachers provide guidance on how to teach the language and help us to develop curricula and teaching materials.

Upon listening to the recordings in the corpus, one of the teachers was struck by the richness of the language and thematic content of the personal stories that Elders told. As a result of this experience with the corpus materials, she decided to work with those recordings to develop learning materials. She started by identifying the relevant cultural themes and values that were conveyed through the recorded stories. She then developed lesson plans around those recordings, the thematic and cultural content, and the grammatical structures used within the stories. This resulted in up to four lessons per recording.

She developed accompanying worksheets to allow students to practice the grammatical concepts she decided to add to her course. She also developed read-along activities. To do this, she had to convert the recordings from .m4a to .mp3 so that they could be played using technologies that are provided in her classroom, which demonstrates the potential barriers that file formats can introduce.

Building on her work, we have developed interactive online learning activities using her newly created worksheets. These interactive learning activities provide students with feedback and have been integrated into a computer assisted language learning (CALL) system.

In addition to the interactive worksheet activities, we have been developing a read-along activity as part of this CALL system. This read-along activity specifically uses the shadowing approach (Kadota, 2019), where a learner must read along while keeping pace with the audio. This approach helps to develop oral fluency among learners, which is a goal that many learners of nêhiyawêwin and their teachers have set. Since we are using the same karaoke-like approach that this teacher added to her classroom, we need to align the text with the audio. So, we are currently testing methods for supporting the automation of this alignment.

As the above case illustrates, the corpus materials can be used to develop and expand teaching materials. As reported by collaborating teachers, these materials have also influenced how teachers approach their students and courses. One teacher decided to start teaching certain aspects of the language, such as the transitive animate verb paradigm, sooner. Before listening to the stories from Elders, she would only teach the transitive animate paradigm to more advanced students. She thinks it is not taught in many settings because of its inherent complexity. Listening to the stories helped her realize what a central part it was of fluent speakers' speech. This realization came after analyzing the recorded stories. Upon reflection, she recognized that the adults in her life would use it when speaking to her as a child. Consequently, she now teaches it to young children with the expectation that they will gain knowledge and familiarity with this paradigm even though they are unlikely to produce language using verbs in the transitive animate form soon after they learn it. She expects that they will start using the transitive animate paradigm once they are older and more fluent.

Beyond supporting the development of learning materials and activities for use in person or online, the corpus has helped to identify gaps in existing materials. As part of preparing accompanying learning materials for students, language teachers often decompose new vocabulary items into their constituent morphemes because this helps students to learn the language and build upon their existing knowledge when they encounter new words (Wagner et al., 2007). One of the words that helped this teacher identify a gap in existing language support resources was ‘intopakwanikamik’. As part of preparing instructional materials for her students, she wanted to provide a formal definition of the ‘into’ prefix. However, ‘into’ was not present in any of the dictionaries she had access to. As a result, she plans to take this word and others like it to a meeting with Elders so that she can formally document the deeper cultural and semantic connotations of the words and prefixes that are in our corpus and not documented elsewhere.

6.2 Text Prediction

Text prediction is a language technology that many people use daily without noticing it. For many, they rely on it when typing on their phones to compose an email or text. They also use it to help them fill in forms. This language technology may be taken for granted in high-resource languages. The absence of support tools like these for nêhiyawêwin speakers has been noted, and learners of nêhiyawêwin have expressed a desire for similar types of support (Lothian et al., 2019). The nêhiyawêwin language has a rich morphology, where words are often composed of several morphemes. Therefore, we chose to support text prediction at the morpheme level for nêhiyawêwin rather than at the word level, which is how predictions are usually made for English and French.

Text prediction is a subtask of one of the projects that is being run out of the National Research Council Canada. This project aims to create “software to assist Indigenous communities in preserving their languages and extending their use” (Kuhn et al., 2020). The tasks they are working on have been derived from community needs and performed in collaboration with communities via the empowerment paradigm. A predictive text feature was enabled in the Keyman[3] keyboard software for those who wish to implement the model in a desired language when using the keyboard. However, Kuhn et al. (2020) note the predictive model is based on unigrams since there is often not enough language data available to create more complex models based on longer sequences of text.

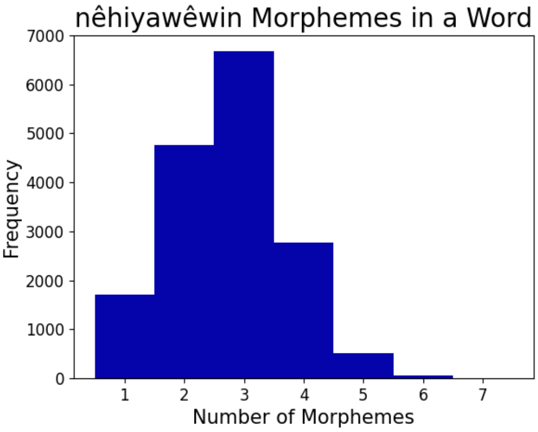

Figure 1: The distribution for the number of morphemes per word in the corpus.

To extend the work by Kuhn et al. (2020), we built n-gram models using the present corpus. These models consider what was typed previously when predicting text. To allow the model to learn sequences of morphemes, we first had to prepare the data so that it could be used to train such a model. We used an FST (Arppe et al., 2014–2019) to divide words in our corpus into their constituent morphemes. The corpus contains 3,650 unique morphemes and 45,220 morphemes in total. Figure 1 shows the distribution of the number of morphemes found in a single word.

The corpus was divided into 90% for the training set and 10% for the development set. We used KenLM (Heafield, 2011) to train n-gram models on the sequences of morphemes within a word. Hyperparameter tuning was then performed by training several models with different values of n, in the range of 2 to 7. We considered the model with the lowest average perplexity on the development set, as the best model. Although the models with different values of n performed similarly, the best performing model was the 5-gram model, with an average perplexity on the development set of 133.12 (SD = 242.05).

The COVID-19 pandemic has brought about informational materials translated into several languages in an attempt to reach as many members of the public as possible with general health guidance around this issue. Usually these pamphlets contain

[3] https://keyman.com/

Data Set	No.	Length
Train	14822	2.75 (0.98)
Development	1647	2.73 (0.98)
Test	478	2.13 (1.04)
Provincial Health	297	2.21 (1.07)
Health Canada	181	2.01 (0.97)

Table 7: Number (No.) of words and morphemes in a word, as *M (SD)* in the train, development, and test set.

a small amount of text and are shared as PDF files. We selected 2 of the longer nêhiyawêwin texts from the provincial health ministry, Alberta Health Services, and Health Canada as testing material. The 5-gram model achieved an average perplexity of 181.75 (*SD* = 325.08) on the test set.

The training, development, and test set characteristics are shown in Table 7.

7 Future Work

This corpus can be used to support several lines of future work. An immediate next direction would be further supporting the development of nêhiyawêwin learning materials using the corpus. For example, creating additional read-along activities and other game-based learning activities. SoundHunters is one such game that aims to improve learner phonological awareness (Lothian et al., 2020). The frequency statistics of different sounds, syllables, and words could be used to select learning materials for use in this and other games. The corpus could also be used to provide additional content.

Another avenue, would be applying the corpus to support the further creation of NLP technologies for nêhiyawêwin. As mentioned, predictive text models were created for nêhiyawêwin because this type of language technology is both desired and can be supported through the corpus. To determine if these models are helpful for nêhiyawêwin speakers when typing, we will perform user studies. From these studies, we aim to learn if the predictive models support text entry in a timely way and whether people perceive them to be useful. We will collect perceptual data and feedback from potential users after they have completed several text-entry tasks through the developed predictive-text system. We will use the same measures that are commonly employed to determine the performance of new text-entry techniques. These measures include response time, error rates, and key strokes per character (Soukoreff and MacKenzie, 2003). We will also analyze how often predictions are used and the ranking of the prediction selected. With this information, we can determine if the predictive text model meets a community's needs and preferences. It is simply not enough to rely on model performance metrics without obtaining feedback from potential users.

We recognize that by preprocessing SRO text, we have enabled easier use of this writing system for developing language technologies compared to syllabics. Future work should create a similar pipeline for syllabics that aligns with language rules used by communities, so that it can receive the same status and attention in the development of language technologies.

8 Conclusion

This work contributes a collection of nêhiyawêwin resources that have been cleaned, processed, and shared for creating language technologies. Care was taken to collect, align, and preprocess the material so it could be used by others. It is hoped that sharing these resources along with the documentation of how they have been prepared will support language preservation and revitalization efforts.

The utility of this corpus was shown via its community use in teaching nêhiyawêwin and by building language models to enable the creation of language technologies desired by speakers. This preliminary and on-going work demonstrates the value of the developed corpus for this low-resource language. Through these efforts in developing the corpus we hope to pave the way for the future creation of language technologies for and by nêhiyawêwin speakers.

Acknowledgements

We thank the many high school and summer students who helped to collect corpus materials. Anaka Sparrow, Kelly Shih, Divya Prasad, Sabrina Lou, and Adya Dutt helped to collect and align Cree text materials. Ronan Sandoval helped create the web scraper which allowed for more text collection.

This project was supported, in part, by funding from the the National Research Council Canada (NRCC), Natural Sciences and Engineering Research Council of Canada (NSERC), and Social Sciences and Humanities Research Council (SSHRC).

References

Eddie Antonio Santos. 2021. Cree SRO syllabics converter.

Antti Arppe, Atticus Harrigan, Katherine Schmirler, Lene Antonsen, Trond Trosterud, Sjur Nørstebø Moshagen, Miikka Silfverberg, Arok Wolvengrey, Conor Snoek, Jordan Lachler, Eddie Antonio Santos, Jean Okimāsis, and Dorothy Thunder. 2014–2019. Finite-state transducer-based computational model of Plains Cree morphology. Accessed on 06.2021.

Antti Arppe, Jordan Lachler, Trond Trosterud, Lene Antonsen, and Sjur N Moshagen. 2016. Basic language resource kits for endangered languages: A case study of Plains Cree. In *Proceedings of the 2nd Workshop on Collaboration and Computing for Under-Resourced Languages Workshop (CCURL 2016), Portorož, Slovenia*, pages 1–8, Portorož, Slovenia.

Antti Arppe, Katherine Schmirler, Atticus G Harrigan, and Arok Wolvengrey. 2020. A Morphosyntactically Tagged Corpus for Plains Cree. In *Papers of the Forty-Ninth Algonquian Conference*, volume 49, pages 1–16, Montreal, Quebec, Canada.

Amy Bombay, Kimberly Matheson, and Hymie Anisman. 2011. The impact of stressors on second generation Indian Residential School survivors. *Transcultural psychiatry*, 48(4):367–391.

Tolly Bradford and Chelsea Horton. 2016. *Mixed Blessings: Indigenous Encounters with Christianity in Canada*. UBC Press.

Elections Canada. 2019a. Guide to the federal election.

Elections Canada. 2019b. kiskinohtahiwēwin isi okimānāhk pimipahtāwin.

Luciano Floridi and Massimo Chiriatti. 2020. GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines*, 30(4):681–694.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

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

Shuhei Kadota. 2019. *Shadowing as a Practice in Second Language Acquisition: Connecting Inputs and Outputs*. Routledge, London.

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

M Paul Lewis and Gary F Simons. 2010. Assessing endangerment: expanding Fishman's GIDS.

Delaney Lothian, Gökçe Akçayir, and Carrie Demmans Epp. 2019. Accommodating Indigenous People When Using Technology to Learn Their Ancestral Language. In *Workshop on Supporting Lifelong Learning (SLL) at the International Conference on Artificial Intelligence in Education (AIED)*, volume 2395, pages 16–22, Chicago, Illinois, USA. CEUR-WS.

Delaney Lothian, Gökçe Akçayir, Anaka Sparrow, Owen Mcleod, and Carrie Demmans Epp. 2020. Soundhunters: Increasing Learner Phonological Awareness in Plains Cree. *Artificial Intelligence in Education*, 12163:346 – 359.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018. Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Michael Mohler and Rada Mihalcea. 2008. Babylon parallel text builder: Gathering parallel texts for low-density languages. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Steven Moran. 2016. The ACQDIV database: Min(d)ing the ambient language. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4423–4429, Portorož, Slovenia. European Language Resources Association (ELRA).

Jeffrey Muehlbauer. 2011. nêhiyawêwin katawasisin: THE PLAINS CREE LANGUAGE IS BEAUTIFUL. *The Canadian Journal of Native Studies*, 31(1):73–95.

Arden Ogg. 2020. Stay home: Learn Cree 29. Simon Bird – Thinking of you. *Cree Literacy Network*.

Open Language Archives Community. OLAC resources in and about the Plains Cree language.

Aidan Pine and Mark Turin. 2017. Language Revitalization. *Oxford Research Encyclopedia of Linguistics*.

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

R. William Soukoreff and I. Scott MacKenzie. 2003. Metrics for text entry research: An evaluation of msd and kspc, and a new unified error metric. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, page 113–120, New York, NY, USA. Association for Computing Machinery.

Statistics Canada. 2018. Language Highlight Tables, 2016 Census - Aboriginal mother tongue, Aboriginal language spoken most often at home and Other Aboriginal language(s) spoken regularly at home for the population excluding institutional residents of Canada, provinces and territories.

United Nations. 2019. Indigenous Languages Face Extinction Without Concrete Action to Protect Them, Speakers Warn General Assembly, as International Year Concludes | meetings coverage and press releases.

Richard K Wagner, Andrea E Muse, and Kendra R Tannenbaum, editors. 2007. *Vocabulary Acquisition: Implications for Reading Comprehension*. Guilford Press, New York.

A Keywords used to search for nêhiyawêwin resources

Keywords used to search for nêhiyawêwin resources include: 'nêhiyawêwin text', 'nêhiyawêwin text online', 'nêhiyawêwin text pdf', 'nêhiyawêwin children stories', 'nêhiyawêwin workbook', 'learn nêhiyawêwin'. Additionally, these searches were replicated by replacing 'nêhiyawêwin' with 'plains cree' and 'plains cree y dialect'.

University library websites were also searched and the language was specified as 'Cree'.