Computational Tools in Support of Niitsi'powahsin (Blackfoot Language)

Dominik Kadlec, Inge Genee, Antti Arppe

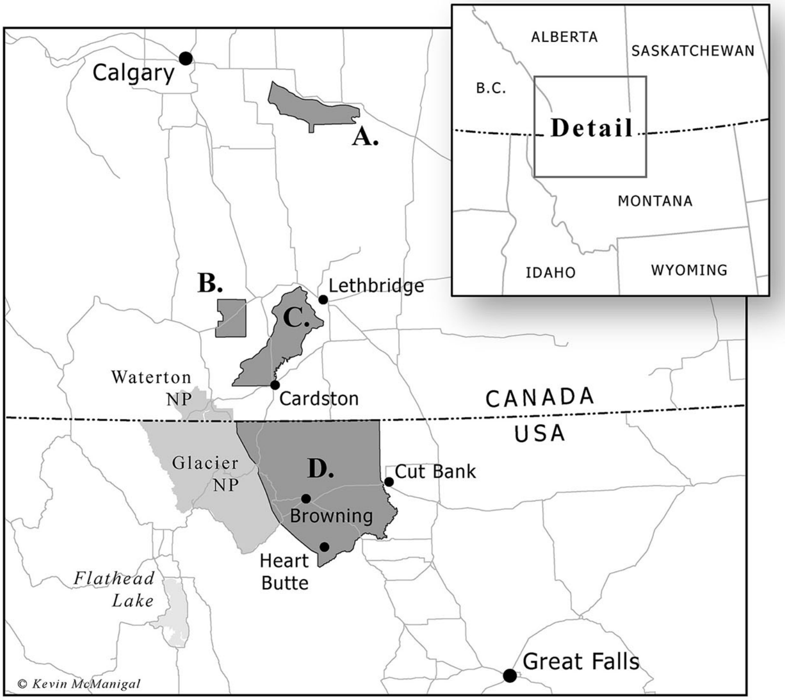
Computing technologies are increasingly being used to develop resources for endangered and under-resourced languages, including Indigenous languages of Canada. This paper presents some of the work being done at Iniskim University of Lethbridge to develop tools and resources in support of Niitsi'powahsin (Blackfoot language). We first discuss work completed over the past few years, and then present the Blackfoot computational modeling project, which is intended to take the resources developed until now to a higher level of technological sophistication, making them more useful to speakers and learners of the language. We show how the model of Blackfoot noun and verb morphology was developed on the basis of an existing model of Plains Cree, a related language, in combination with available documentation of the Blackfoot language. We explain what this model allows us to do to improve the resources previously created and make both revitalization and documentation efforts easier by producing automated corpus parsers to support documentation efforts and morphologically intelligent tools such as dictionaries and lesson platforms that can be used in classrooms and by individual teachers and learners. We hope to demonstrate that technology can meaningfully contribute to language stabilization efforts, provide a greater potential for democratizing documentation work and bring accessible pedagogical tools into the language revitalization classroom. These tools can be developed by experts as long as enough primary documentation work has been undertaken, and then used and applied by activists, Indigenous community members, educators, and researchers.

Where We Started: Background

The Algonquian Dictionaries and Language Resources project (https://www.algonquianlanguages.ca/), directed by Marie-Odile Junker at Carleton University in Ottawa since 2005, hosts resources for about a dozen Algonquian languages, co-created with language communities and scholars. In addition to dictionaries this includes, for some but not all languages, verb conjugations, grammar pages, spelling and orthography converters, lessons, oral stories databases, downloadable conversation apps (https://www.atlas-ling.ca/apps) and teaching manuals. The Algonquian Linguistic Atlas (https://www.atlas-ling.ca/) presents about 400 conversational phrases divided into 21 topics of conversation, recorded by 68 speakers from 58 communities, representing 20 languages and 47 dialects from Alberta and Wyoming in the West to Nova Scotia, Quebec and Labrador in the East.

Cite as from J. Reyhner, J. Martin, & L. Lockard (Eds.). (2024). *Honoring Our Indigenous Languages and Cultures* (pp. 75-84). Flagstaff, AZ: Northern Arizona University.

Among the Algonquian languages included in the project is Niitsi'powahsin (the Blackfoot language). Niitsi'powahsin, also called Siksikai'powahsin, is an Indigenous language spoken mainly in Alberta, Canada, and Montana, in the USA. The historic Blackfoot territory covered a vast area in the northern Great Plains region of North America (Van Beek, 2019; Grinnell, 1892). The Blackfoot Confederacy nations now control four reservations in Canada and the United States, as shown on Map 1. Siksika is the northernmost nation, and sits just east of Mohkinsstsis (Calgary). The Kainai (Blood) Nation is situated southwest of Sikoohkotoki (Lethbridge). The Piikani ((Northern) Peigan; a.k.a. Aapatohsipikani) Nation is located west of Fort Macleod. The Aamsskaapipikani (Southern Piegan) are located on the Blackfeet reservation in Montana. Like many North American Indigenous languages, Niitsi'powahsin has been declining in use over the past century. Nonetheless, attempts at revitalizing the language are coming from a variety of sources (Mills, 2011, Varadi, 2020, Waugh, 2018).



Map 1: Modern locations of Blackfoot reserves: A. = Siksiká, B. = Piikani/ Aapátohsipikani (Northern Peigan), C. = Kainai (Blood), D. = Aamsskáápipikani (Southern Piegan, Blackfeet) (Map by Kevin McManigal, University of Montana)

The Blackfoot Language Resources project (https://blackfoot.algonquianlanguages.ca/), directed by Inge Genee at Iniskim University of Lethbridge since 2016, hosts resources in support of Niitsi'powahsin (Blackfoot). The project was recently given the Blackfoot name *Kiistónnoon aiitsí'poyio'pa* 'We are speaking Blackfoot' by our Elder advisor Ninnaisipistoo Francis First Charger. The central element of the website is an on-line dictionary based on the database used for the most complete and up-to-date lexical resource available for Blackfoot, the third edition of the *Blackfoot Dictionary of Stems, Roots and Affixes* (Frantz & Russell, 2017). After joining the Algonquian Dictionaries and Language Resources project in 2016, the database for this dictionary was digitized and imported into the Algonquian Dictionaries infrastructure and is now available at https://dictionary.blackfoot.algonquianlanguages.ca/. Additional functionalities include audio and images, help videos, enhanced search capabilities, and a sophisticated relaxed search feature (Genee & Junker, 2018, Genee, 2020). The website also hosts a story archive, grammar and resource pages, and an oral conversations database. Dozens of community members have contributed to the project in a variety of ways (https://blackfoot.algonquianlanguages.ca/contributors/).

The second major project is called 21st Century Tools for Indigenous Languages (https://21c.tools/), directed by Antti Arppe at the University of Alberta in Edmonton since 2019. This project develops technological tools in support of revitalization of Indigenous languages, such as web-based dictionaries, searchable databases of spoken and written texts, spell-checkers, and computer-aided language-learning applications. Software development and computational language modeling form the basis of this tool development.

This work would also allow the development of other tools such as spell-checkers. This paper discusses a first round of computational work, which models the noun and verb morphology of the Blackfoot language (Kadlec, 2023). We describe the model as developed to date, including results of testing the model, and then discuss potential tools resulting from the project and some possible future developments.

Where We Are Now: Computational Modelling

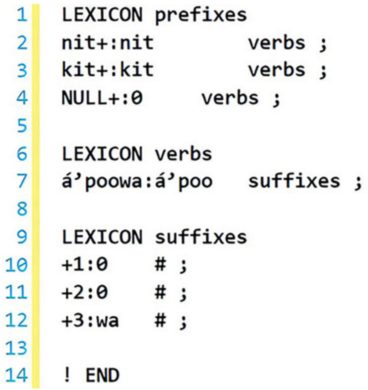
Blackfoot, like all Algonquian languages, is a polysynthetic language. This means that utterances often contain very few separate words, but that each individual word can contain many meaningful parts (morphemes), which together translate into a whole phrase or sentence in English. Example (1) illustrates this:

(1) *Aohkanaikstsitsahpohtominiki, (ni)taakiitomahkaasokapiiksii'pa.*
'When I finish putting it all together, I will mix it well.'

This sentence comes from a recipe for making bannock or fry bread, told by Beverly Hungry Wolf (https://stories.blackfoot.atlas-ling.ca/). In English it contains 12 words, but in Blackfoot it contains only two words. The first word has six morphemes, and the second word has eight morphemes. Currently, neither of these words will return a usable result when typed into the dictionary's Blackfoot-to-English search box. For our dictionary to be able to interpret such long words, it needs to know how to split the word into morphemes so it can tell which part is the stem. This can then be looked up in the dictionary. A computational model of the structure of Blackfoot words will allow the dictionary to “understand” the internal grammar of these long words.

A major recent step towards this goal is Kadlec (2023), who used previously published Blackfoot descriptive materials (in particular Frantz, 2017 and Frantz & Russell, 2017) to develop a computational model of Blackfoot nouns and verbs. The model itself is a descriptive model of Blackfoot noun and verb morphology, so is a specialized model that can be used to parse and to construct nouns and verbs.

The model was created using LEXC, a computational formalism that allows the user to define a set of lexica and constrain how they can combine (Kadlec 2023:12-25). The lexica basically work like closed dictionaries that can be added to within the formalism itself, but cannot accept entries which are not defined as part of the lexicon. These lexica can contain morphemes, in which case the model can be used to build words by combining elements in the lexicon in the order in which they are set to occur. Figure 1 below shows what this can look like in a small model of Blackfoot.



```
 1  LEXICON prefixes
 2  nit+:nit        verbs ;
 3  kit+:kit        verbs ;
 4  NULL+:0     verbs ;
 5
 6  LEXICON verbs
 7  á’poowa:á’poo   suffixes ;
 8
 9  LEXICON suffixes
10  +1:0    # ;
11  +2:0    # ;
12  +3:wa   # ;
13
14  ! END
```

Figure 1: Small Model of Blackfoot Verb Inflection

The model in Figure 1 is written in LEXC. Lines 1-4 represent the first lexicon. This lexicon has the name *prefixes* because the lexical items in this lexicon are all verb prefixes. The lexicon contains the prefixes *nit-* (first person, I), *kit-* (second person, you) and *NULL-*. The *NULL-* prefix is used when verbs do not take a prefix, such as when they are inflected for third person agreement (he/she/it). For each lexical item, the underlying analysis is found to the left of the colon, and the item that will appear on the surface of the word form is to the right of the colon.

The next lexicon is defined by the term *verbs*; this signals that the next lexicon for the model is the verb lexicon. For purposes of illustration there is only one lexical item shown in the verb lexicon in Figure 1. This item has the verb stem *á’poo* ‘travel/move about’, which is represented underlyingly (to the right of the colon) as a lemma (a basic inflected form of the verb), inflected to agree with the third person with the suffix -*wa*. As with the prefixes, the surface form appears to the right of the colon; it is followed by the term *suffixes*, which signals that the next lexicon is the *suffixes* lexicon. It includes three items, two of which are null in the surface form, but underlyingly indicate agreement with the first and second person prefixes. The other lexical item is the third person suffix, -*wa* (he/she/it).

These lexica are organized and can be constrained in the ways in which they combine in order to reflect the grammar of a given language. For example, the Blackfoot noun model would include a possessor prefix lexicon, a prenoun lexicon, some noun stem lexica (reflecting the Blackfoot noun subclasses animate and inanimate), a possessor suffix lexicon, and an animacy/number/obviation suffix lexicon. The connection between the possessor prefixes and suffixes is controlled through flags, which allow the model to restrict long distance relationships between morphemes. For instance, a noun with a first person plural possessor ("our") would have a prefix *nit*- and a suffix -*(i)nnaan*. The flags would specify that *nit*- can combine with -(*i)nnaan* but not with other suffixes, and conversely that the suffix -*(i)nnaan* requires the presence of the prefix *nit*-.This model was tested in two ways. One was by creating so-called YAML files containing manually created inflectional paradigms for Blackfoot stems alongside their underlying analysis as it would be produced by the model. The other was to explore quantitative results using a corpus of publicly available online Blackfoot texts.

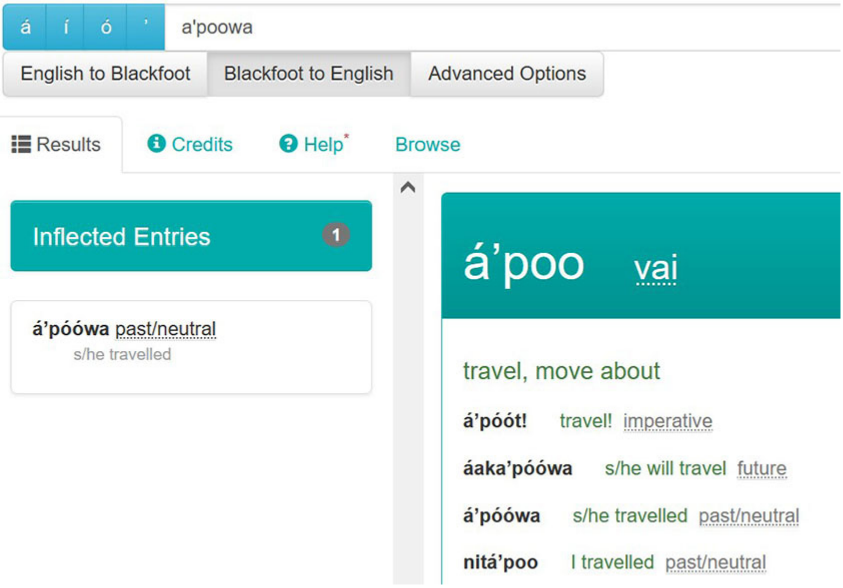
In the YAML files, 29 stems were tested altogether; 19 of them were nouns, and 10 of them were verbs, making for a total of 889 inflected word forms paired with underlying analyses. Of these 889 pairs, 885 were replicated by the model, thus missing the results for only 4 word forms, or 0.6% of the total word forms. This means the model performed very well on this test.

The corpus tests were conducted by determining how many words were parsed by the model, and then sampling parsed words to determine whether the words were accurately parsed. The corpus contained a total of 13,784 unique word-forms (types). Of these, 9,882 received at least one parse by the model, making for a total of 71.69%. We took a random sample of 300 parsed word forms, and manually evaluated whether there was a plausible morphological parse within the top ten results, giving a pass to any word that had a reasonable parse, and a fail to any that did not. 280 of the 300 words received a pass, which is 93.33% of the sample.

While the performance of the model is not 100% perfect, the results of both tests are encouraging. They show that the model performs well enough to be implemented for various uses.

What We Can Do: Using the Computational Model

While the resources created to this date are useful, several desirable features are not yet included. In particular, the dictionary is not "intelligent," meaning it is unable to interpret inflected forms that do not occur in its database. For a highly inflected language like Blackfoot this is a problem: words can have hundreds of different forms depending on inflection for person, tense/aspect, mode, and the incorporation of a range of other elements into the word. In practice this means that if you type a full word into the Blackfoot-to-English search bar, your chances of getting a result are not that high. For example, the entry for the verb stem *á’poo* ‘travel, move about’ has several examples that will produce a result when searched in the Blackfoot-to-English search bar. Typing *a’poowa* will find a hit in the Inflected Entries database, which will link to the main entry stem *á’poo*. This is shown in Figure 2. Similar results will be returned when typing any of the other three examples.



á’póówa past/neutral
s/he travelled

á’poo vai

travel, move about

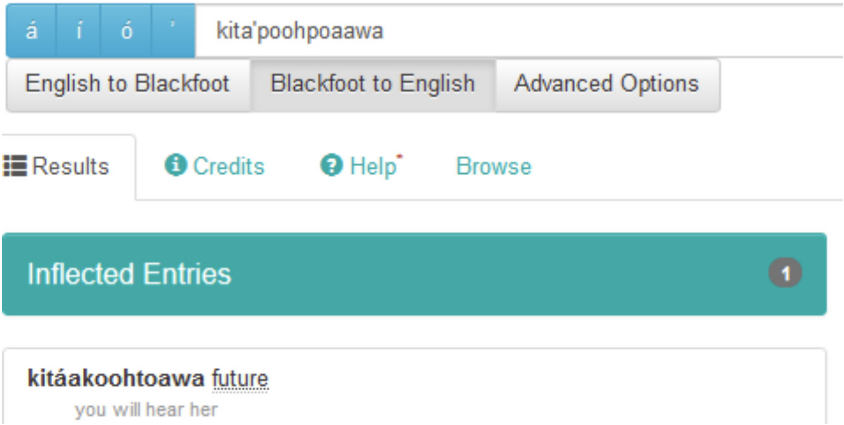
á’póót!	travel!	imperative
áaka’póówa	s/he will travel	future
á’póówa	s/he travelled	past/neutral
nitá’poo	I travelled	past/neutral

Figure 2: Example of a Successful Search of an Inflected Form of the Verb *a’poo*

But typing an inflected form that does not already occur in the database will not return a useful result, as shown in Figure 3. The search engine does return a result here, but it is not correct.

The computational work described in this paper is needed to develop an intelligent dictionary that can interpret forms such as given in the example above, by actually analyzing them rather than guessing at superficially similar but unrelated forms. This would make the dictionary more usable for those who do not have much explicit training in Blackfoot grammar and linguistics by allowing them to search for inflected word-forms. A search for an inflected word would, in the immediate future, return at least a dictionary entry for the stem contained in the inflected form. Going further, searches will also be able to break down and return items for which there is not yet a stem included in the dictionary. This would mean that at least some novel words could be searched by users, where that was not previously possible.

The computational model can also be used to develop parsed corpora, potentially much faster than was possible previously. Without computational tools, corpora need to be parsed and labeled by hand by people with highly

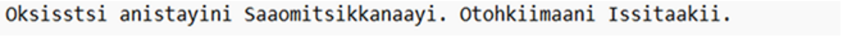


kitáakoohtoawa future

you will hear her

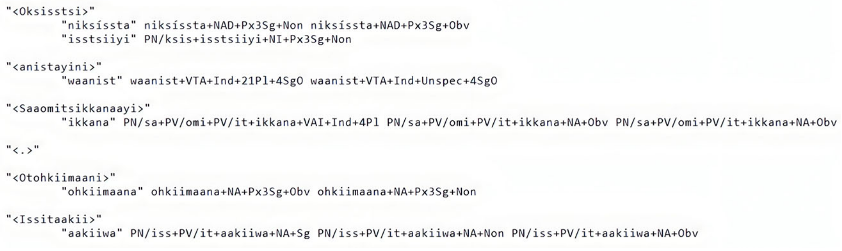
Figure 3: Example of an Unsuccessful Search of an Inflected Form of *a'poo*

specialized linguistic training. Even when such people are available, the work is very time consuming, which slows down progress and limits the size of any corpus. With the Blackfoot morphological model, large amounts of Blackfoot text can ultimately be parsed automatically, either with the help of researchers, or with the help of other computational tools that are designed to disambiguate word-forms. An example of three sentences taken from the test corpus can be found below in Figure 4 followed by the parses made by the model in Figure 5.



```
Oksisstsi anistayini Saaomitsikkanaayi. Otohkiimaani Issitaakii.
```

Figure 4: Example of Original Corpus Text (from Russell & Genee, 2014)



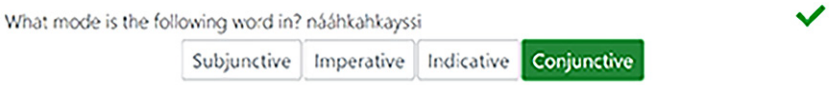
```
"<Oksisstsi>"
        "niksíssta" niksíssta+NAD+Px3Sg+Non niksíssta+NAD+Px3Sg+Obv
        "isstsiiyi" PN/ksis+isstsiiyi+NI+Px3Sg+Non

"<anistayini>"
        "waanist" waanist+VTA+Ind+21Pl+4SgO waanist+VTA+Ind+Unspec+4SgO

"<Saaomitsikkanaayi>"
        "ikkana" PN/sa+PV/omi+PV/it+ikkana+VAI+Ind+4Pl PN/sa+PV/omi+PV/it+ikkana+NA+Obv PN/sa+PV/omi+PV/it+ikkana+NA+Obv

"<.>"

"<Otohkiimaani>"
        "ohkiimaana" ohkiimaana+NA+Px3Sg+Obv ohkiimaana+NA+Px3Sg+Non

"<Issitaakii>"
        "aakiiwa" PN/iss+PV/it+aakiiwa+NA+Sg PN/iss+PV/it+aakiiwa+NA+Non PN/iss+PV/it+aakiiwa+NA+Obv
```

Figure 5: Example of Parsed Text from the Corpus

In Figure 5, each word is found between the double quotation marks and the less-than and greater-than signs ("< and >"). Below them, in double quotes, we find the potential stem that the model identified to be the stem of the word, followed by the possible parses of that stem. If the output is simply "+?", then this means that the word was not parsed by the model.

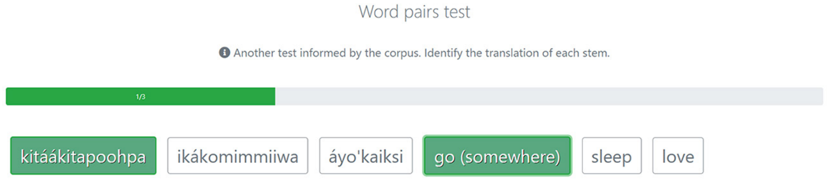
We can see that each word may have multiple possible stems, and multiple possible parses for each stem. In order to find the correct parse, there are two possible choices. One is to go through the whole corpus by hand and choose the most correct parses according to the context in which the word occurs. This option is not ideal because it is very labor-intensive, which will slow down progress. The second option would be to automate this process with disambiguation software. This was done for a corpus of Plains Cree (Schmirler, 2022) using a constraint grammar model, which evaluated the syntactic context of words, and chose most likely parses based on context. Such work remains to be done for Blackfoot, but it is certainly feasible.

Another feature of the Blackfoot Online Resource site is a Lessons page where users can complete a variety of online activities that can be used to help them learn Blackfoot. Those involved with online lesson development can create many different lessons of different types. This process can be slow going without real examples of Blackfoot words and sentences. This could be remedied with the help of the model and corpus together. Figures 6 and 7 show examples of an activity created for a lesson that was made using the parsed corpus.



What mode is the following word in? nááhkahkayssi

Figure 6: Example of Digital Multiple-choice Question



Word pairs test

Another test informed by the corpus. Identify the translation of each stem.

Figure 7: Example of Digital Matching Question.

We created the multiple-choice question in Figure 6 by searching for +CNJ in the corpus to find a word parsed in the conjunctive mode paradigm and used that as the basis for the question. For the questions in Figure 7, we found some parsed verb stems in the corpus by searching for +V and used some of the common word forms according to their counts from the corpus. In the activity, the user must select the stem translation that matches the word form. I suggest that the morphological model can be used by developers to find examples in corpora, and create digital lessons based around morphology faster, and in greater quantities than was previously possible.

Where Else We Can Go: Improving the Model and More Applications

The model of Blackfoot noun and verb morphology can be continually improved by adding more lexical items to the lexicons and by refining it so that it more accurately reflects the grammar of Blackfoot nouns and verbs. In order to create a full model of Blackfoot grammar, other word classes will need to be included as well. Currently, the model does not handle demonstratives, pronouns, and particles. A model for demonstratives is close to completion (Schmirler et al., subm). Once these remaining models are completed, peripheral technologies can be used to improve the functionality of the model. In this section, we provide some examples of technologies that can be used to improve the model based on previous research and developments.

The first such technology is a machine learning model that could learn from the production of the morphological model. Experiments with this machine learning model were done using Arapaho, which is also an Algonquian language and thus has a very similar grammar (Moeller et al., 2018). In the experiment, the Arapaho morphological model produced surface forms paired with underlying analyses which the machine learning model learned from. The result was a model that could recognize and produce novel words which the morphological model could not produce. The addition of machine learning technology such as this could be used to parse novel word forms which cannot be parsed by the model because it contains a noun or verb stem which is not present in the noun or verb stem lexica.

Technology could also be used to enhance the current capabilities of the model. For example, Harrigan and Arppe (2022) used an English machine learning model that used corpora to identify semantically related words in a Plains Cree dictionary. This allowed users to search for terms in either Plains Cree or English, and even if the term was not available through the dictionary, the search would return semantically related results. For instance, if a user was to search for the term *dinosaur* the dictionary would return entries for related words such as *animal*.

Another enhancement would allow users to enter Blackfoot words and the dictionary would return English phrases, or vice versa. This capability has been implemented for the Plains Cree dictionary (Arppe et al., 2021). This is done by mapping underlying analyses from the morphological model with English words and phrases. As a result, someone could type a phrase such as 'I saw you', and the dictionary would return the Cree word *kikî-wâpamitin* 'I saw you'.

Other advancements would include the development of more Blackfoot corpora. Kadlec (2023) curated a corpus of Blackfoot that was sufficient to test the model, but could be improved by adding more texts. The corpus contains many isolated word forms, and relatively less text from various other genres. This corpus will soon be made available to the public via the Blackfoot Language Resources website. Work is currently underway at the University of Lethbridge to create more Blackfoot corpora which can be used for various purposes.

Conclusion

The model of Blackfoot noun and verb morphology represents a new step in digital language documentation for Blackfoot. There are many advantages to developing this type of technology for Indigenous languages. Models like this open up new avenues for research and documentation which are not possible otherwise, or make research and documentation much easier.

References

Arppe, A., Santos, E., Neitsch, A., Poulin, J., Harrigan, A., Schmirler, K., Hieber, D., & Wolvengrey, A. (2021). Towards a morphologically intelligent and user-friendly on-line dictionary of Plains Cree – next next round [Presentation]. 53rd Algonquian Conference. Carleton University.

Bliss, H., & Witschko, M. (2023). *Stsíkiistsi ki stsíkiistsi*: The ubiquity of Blackfoot demonstratives in discourse. In Å. Næss, A. Margetts, & Y. Treis (Eds.), *Demonstratives in discourse* (pp. 123–147). Language Science Press.

Frantz, D. (2017). *Blackfoot grammar* (3rd Edition). University of Toronto.

Frantz, D., & Russell, N. J. (2017). *Blackfoot dictionary of stems, roots and affixes* (3rd Edition). University of Toronto.

Genee, I. (2020). "It's written *niisto* but it sounds like *knee stew*." Handling multiple orthographies in Blackfoot language web resources. *Written Language & Literacy, 23*(1), 1-27.

Genee, I., & Junker, M.-O. (2018). The Blackfoot Language Resources and Digital Dictionary project: Creating integrated web resources for language documentation and revitalization. *Language Documentation and Conservation, 12,* 274–314.

Grinnell, G. (1892). Early Blackfoot history. *American Anthropologist, 5*(2), 153-164.

Harrigan, A., & Arppe, A. (2022). Leveraging majority language resources for Plains Cree semantic classification. In M. Macaulay & M. Noodin (Eds.), *Papers of the Fifty-Second Algonquian Conference (PAC52*) (pp. 1-20). Michigan State University.

Heavy Shields Russell, L., & Genee, I. (2014). *Ákaitsinikssiistsi: Blackfoot stories of old*. University of Regina.

Kadlec, D. (2023). *A computational model of Blackfoot noun and verb morphology*. Master's Thesis, University of Lethbridge.

Mills, D. (2011). Niitis'po'siin (Revitalizing & preserving the Blackfoot language program). University of Saskatchewan I-Portal: Indigenous Studies Portal.

Moeller, S., Kazeminejad, G., Cowell, A., & Hulden, M. (2018). A neural morphological analyzer for Arapaho verbs learned from a finite state transducer. In J. L. Klavans (Ed), *Proceedings of the workshop on computational modeling of polysynthetic languages* (pp. 12–20). Association for Computational Linguistics. https://aclanthology.org/W18-4802

Schmirler, K. (2022). *Syntactic features and text types in 20th century Plains Cree: A constraint grammar approach*. PhD Dissertation, University of Alberta.

Schmirler, K., Kadlec, D., Genee, I., & Arppe, A. (submitted). A computational model for Blackfoot demonstratives. Submitted to I. Genee, M. Macaulay & N. Weber (Eds.), *Papers of the fifty-fifth Algonquian Conference*. Michigan State University.

Van Beek, S. (2019). *Beyond the border: Buffalo and Blackfoot tenure on traditional territories.* Master's Thesis, York University.

Varadi, Z. (2020, January 30). The Blackfoot language revival. Glenbow Museum Blog. https://www.glenbow.org/blog/the-blackfoot-language-revival/

Waugh, N. (2018, Feb. 7). Race against time: Scholars and teachers follow complex paths to revitalize Indigenous languages in Treaty 7 region. *UCalgary News*.