A COMPUTATIONAL MODEL OF BLACKFOOT NOUN AND VERB MORPHOLOGY

DOMINIK KADLEC
Bachelor of Arts, University of Calgary, 2021

A thesis submitted
in partial fulfillment of the requirements for the degree of

MASTER OF ARTS

in

INDIGENOUS STUDIES

Department of Indigenous Studies
University of Lethbridge
LETHBRIDGE, ALBERTA, CANADA

© Dominik Miroslav Kadlec, 2022

A COMPUTATIONAL MODEL OF BLACKFOOT NOUN AND VERB MORPHOLOGY

DOMINIK KADLEC

Date of Defense: Thursday, July 13th 2023

Elder Francis First Charger Ninnaisipistoo (‘Owl Chief’)	Eminent Scholar	Kainai Ph.D.
Thesis examination committee member		
Dr. Inge Genee	Professor	Ph.D.
Dr. Antti Arppe	Associate Professor	Ph.D.
Co-supervisors		
Dr. Conor Snoek	Associate Professor	Ph.D.
Thesis examination committee member		
Dr. Paul Mackenzie-Jones	Associate Professor	Ph.D.
Thesis examination chair		

Territorial Acknowledgment

Oki, and thank you for reading my thesis. The University of Lethbridge's Blackfoot name is Iniskim, meaning Sacred Buffalo Stone. I acknowledge and deeply appreciate the Siksikaitsitapii peoples' connection to their traditional territory. I, as a person living and benefiting from Blackfoot Confederacy traditional territory, honour the traditions of people who have cared for this land since time immemorial. I recognize the diverse population of Aboriginal peoples who attend the University of Lethbridge and the contributions these Aboriginal peoples have made in shaping and strengthening the University community in the past, present, and in the future.

Abstract

This thesis describes the construction of a computational model of Blackfoot word structure. This model was developed so that it could provide a foundation for Blackfoot language technologies such as spelling and grammar checkers, search suggestion generators, paradigm generators for pedagogical purposes, intelligent dictionaries, automated corpus parsers for linguistic research and more. Many Indigenous languages in Canada have been declining in use. In response, many Indigenous communities and activists have implemented revitalization strategies which vary in effectiveness. One way to help language efforts to be more effective is to ensure that tools for research and revitalization are freely available to community members. In the 21st century this can be achieved in part through technology, particularly with the help of the internet, which offers information freely (in most cases) to those who wish to access it. In this thesis I describe the early developments of a project that will be used to augment currently available digital resources and provide a basis for future technology for the Blackfoot language. I use Finite State Transducer technology to develop a computational model of Blackfoot noun and verb morphology and test the model using a corpus of modern Blackfoot text that was constructed from a curated collection of available texts.

Acknowledgments

First, I would like to thank the three Blackfoot nations, as this research was on their language, which belongs to them. This project was funded by the 21st Century Tools for Indigenous Languages project administered by the Alberta Technology Lab (ALTLab) at the University of Alberta. I would like to thank the Social Sciences and Humanities Research Council (SSHRC) for funding the 21st Century Tools project through its Partnership Grant program, and the Office of Research Services of the University of Lethbridge for its matching contribution to this project.

I want to thank my academic advisors, Francis First Charger, Inge Genee, Antti Arppe, and Conor Snoek. Their expertise and guidance made this project possible. Francis provided valuable guidance and knowledge that could only be given by a Blackfoot elder and fluent speaker. Inge provided her wealth of knowledge on Blackfoot linguistics, Antti provided expertise in engineering technologies for modelling the morphologies of Algonquian languages, and Conor provided guidance that helped me to greatly improve my academic writing.

I would also like to thank Natalie Weber and Katie Schmirler for their intellectual contributions. Natalie's work in developing Blackfoot corpora and systematically studying Blackfoot morphology and phonology proved very useful and important in developing a corpus for testing the technologies. Katie completed the bulk of the phonological modelling needed for the model, and expertise from her work in Plains Cree modelling.

Contents

List of Tables

List of Figures

List of Abbreviations

A. Linguistic glosses

0- Inanimate (inanimate 3rd person)
21- Inclusive first person
4- Obviative (4th person)
3/0- Third person animate, or third person inanimate 3
5- Further obviative (5th person)
AI- Animate Intransitive Verb
BNF- Benefactive
CN- Conjunctive nominalization (see Frantz 2017:132-133)
CNJ- Conjunctive mode (in Blackfoot examples), Conjunct mode (in Cree examples)
COND- Conditional
CONJ- Conjunction
DEM- Demonstrative
DIR- Direct
DUR- Durative
FUT- Future
IRR- Irrealis mode (unreal (Frantz 2017))
IMP- Imperative mode
INSTR- Associated instrument nominalization (see Frantz 2017:131)
INTERR- Interrogative
INV- Inverse
NAR- Narrative indicator
NONREF- Non referencing
OBV- Obviative (marked on animate nouns)
PAST- Past
PERF- Perfect
PL- Plural
POSS.TH- Possessive theme suffix (-im in both Blackfoot and Plains Cree)
PRO- Pronoun
PROX- Proximate (marked on animate nouns)
RECIPR- Reciprocal
REFL- Reflexive
SBJ- Subjunctive mode
SG- Singular
TA- Transitive Animate Verb
TH- Theme suffix
TI- Transitive Inanimate Verb

B. Computational modelling abbreviations

0PlO- Fourth person plural object
0SgO- Fourth person singular object
1Pl- First person plural
1PlO- First person plural object
1Sg- First person singular
2PlO- Second person plural object
2Sg- Second person singular
2Pl- Second person plural
3Pl- Third person plural
3PlO- Third person plural object
3Sg- Third person singular
DEM- Demonstrative
Der/Refl- Reflexive verb derivation
Dist- Distant (demonstrative indicator)
Dur- Durative suffix
Fut- Future
Imp- Imperative
INTERR- Interrogative suffix
Ind- Indicative mode
NA- Animate noun
Neg- Negation
NI- Inanimate noun
Obv- Obviative
Past- Past
PN/- Prenoun
Pl- Plural
PV/- Preverb
Px1Sg- First person singular possessor
Px21Pl- Firs person plural inclusive possessor
Px3Sg- Third person singular possessor
Sbj- Subjunctive
SG- Singular
VAI- Animate intransitive verb
VTA- Transitive animate verb
VTI- Transitive inanimate verb

C. In text abbreviations

CBLR- Community-Based Language Research
FST- Finite State Transducer
FSM- Finite State Machine
HFST- Helsinki Finite State Compiler
LEXC- Finite state lexicon compiler
OLD- Online Linguistic Database
PAR- Participatory Action Research
REGEX- Regular expressions
SRO- Standard Roman Orthography (for Cree)
VAI- Intransitive animate verb
VTA- Transitive animate verb
VTI- Transitive inanimate verb
VII- Intransitive inanimate verb
VAI+O- Animate intransitive verb plus an object
XFST- Xerox Style Finite State Compiler
YAML- Yet Another Markup Language

1. Background

This chapter of the thesis provides the background for subsequent chapters which describe the research and development of the foundational Blackfoot morphological technology. The chapter provides background in four major areas which are required to understand the work in this thesis.

This thesis draws from four subfields of linguistics, including Blackfoot language and Algonquian linguistics, documentation and revitalization, and computational linguistics. The first two sections of this chapter provide background relating to language documentation and revitalization, and to Blackfoot and Algonquian anthropology, history, and linguistics and the two latter sections provide background relating to technology and computational linguistics. Section 1.1 contains a background for the wider national context of Indigenous languages in Canada and the role of linguists in language revitalization. Section 1.2 provides a brief description of the Blackfoot-speaking peoples (Niitsi'powahsiniiksi) and the current state of their language. Section 1.3 is a summary of the context of this project, the grants which funded it, the larger projects that it falls within and the previous work that has been done relating to Blackfoot language technology. The chapter ends with section 1.4, a description of the computational technologies that made the project possible and provides a description of the development process using examples.

1.1. Indigenous languages in Canada and the role of the linguist

Indigenous languages in Canada have seen a decline in use over the past century. The Blackfoot language, spoken in Montana and southern Alberta is no exception. Very few young people are able to speak the language, and the population of speakers is aging. Table 1 shows that for Canadian Indigenous languages generally, there is a declining proportion of speakers among the younger generations, with the highest percentage of speakers being 55 or older. Younger people usually do not acquire it as a mother tongue. Without robust communities of younger speakers, Indigenous languages will not be able to recover.

Table 1: Indigenous identity population, by age and selected language characteristics, Canada, 2022. [From Statistics Canada, Census of Population 2022.][1]

Age Group	Total Indigenous identity population	Has knowledge of an Aboriginal language		Has an Aboriginal mother tongue	
	Number	Number	Percent	number	percent
Total	1,807,250	243,155	13.5%	185,510	10.5%
0 to 14 years	459,210	53,620	11.7%	39,145	8.5%
15 to 24 years	284,890	31,750	11.1%	23,325	8.2%
24 to 54 years	688,005	96,655	14%	72,015	10.5%
55 years and older	375,140	61,135	16.3%	51,020	13.6%

With so many Indigenous languages in Canada declining in use, and a general lack of resources supporting their revitalization and growth, linguists at institutions across Canada have begun working towards solutions. Linguists can help to support language stabilization efforts for the language communities they work with through Participatory Action Research (PAR) (Genee and Junker 2018). This form of research centers members of the language community as experts and ensures that research is done with their interests as the basis for the research. This is considered by some to be a core pillar of ethical linguistic research (Rice 2011; see also, Gerdts 1998; Yamada 2007; Greenwood and Levin 1998). Czaykowska-Higgins (2009) takes research centering Indigenous people a step further by suggesting that data be produced and owned by Indigenous people, making linguists into active partners. She calls this Community-Based Language Research (CBLR).

Evidence has shown that language can be connected to health and well-being for Indigenous people. Knowledge of their traditional languages can help to connect Indigenous people to their elders, to traditional knowledge and spirituality, and to their identity (van Beek 2016:6–11; see also, McIvor 2005; Puchala et al. 2013). This is one of the most compelling reasons why it is so important to support language reclamation and revitalization efforts.

One effort made by academics to support Indigenous language revitalization in Canada is the Alberta Language Technology Lab, which was founded by its director Dr. Antti Arppe of the University of Alberta. The lab aims to produce language technologies which support both Indigenous communities and researchers in documenting and revitalizing Indigenous languages,

[1] According to the Canadian census, Indigenous people include the First Nations, Metis and Inuit people of Canada.

as well as providing higher levels of technological equality to low resource languages. Through their work and their partnership programs with other institutions, the lab has been able to reach experts in academic spaces and in Indigenous communities across Canada (see their websites, https://21c.tools/ and https://altlab.ualberta.ca/ for more information). The research for this thesis was conducted in the context of the 21st Century Tools for Indigenous Languages Project directed by Dr. Antti Arppe at the University of Alberta. The partnership found here at the University of Lethbridge was led by Dr. Inge Genee with the goal of developing language technologies for and in collaboration with the Blackfoot community.

1.2 Blackfoot language, history, and culture

This section provides a very brief cultural, historical, and linguistic background of the Blackfoot people. This section also provides an overview of the Blackfoot language and general Algonquian grammatical features. A more in-depth discussion of Blackfoot noun and verb morphology can be found in chapters 2 and 3, sections 2.1 and 3.1 respectively.

The Blackfoot language is part of the Algonquian language family. It is spoken by the three nations of the Blackfoot confederacy, the Kainai (also known as the Blood nation), the Piikani (now subdivided into the Aamsskapipiikani[2] in the United States, and the Aapatohsipiikani in Canada), and the Siksika. Members of the Blackfoot nations refer to themselves as the Niitsitapi 'real/true people' or Siksikaitsitapi, literally, 'Blackfoot people'.

The three Blackfoot speaking nations often refer to themselves as the Blackfoot-speaking people (Niitsi'powahsini, 'real/true language speaking people' or Siksikai'powahsini, 'Blackfoot speaking people'). The Blackfoot people inhabited a large territory in the north-western part of the Great Plains region of North America. Figure 1 shows the extent of this traditional territory.

[2] They often refer to themselves as the Blackfeet nation in English.

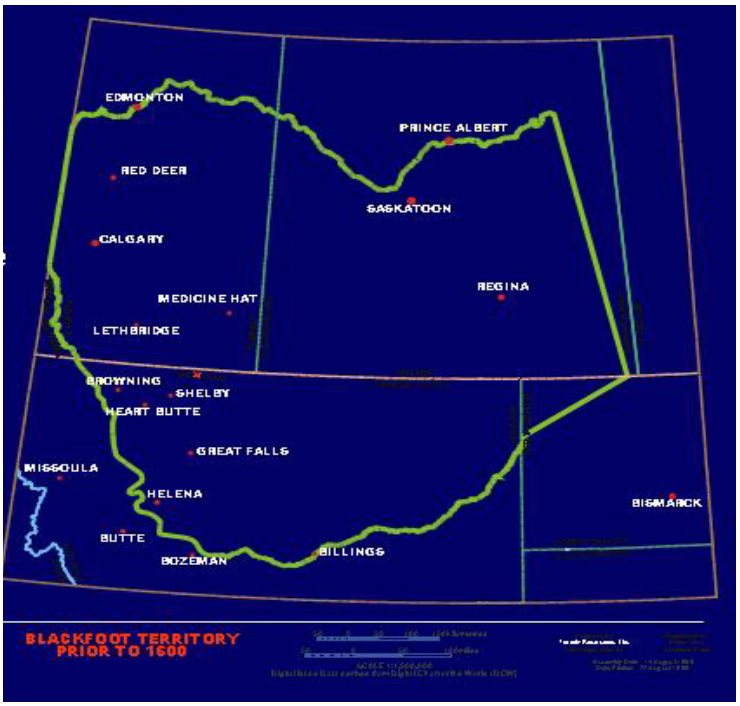

Figure 1: The traditional territories of the Blackfoot people prior to 1600.[3]

The map shows that the Blackfoot people inhabited a territory that stretched as far west as the Rocky Mountains, and as far east as what is now Saskatchewan. Their territories changed after 1600 due to the influences of colonization.

Daschuk, Hackett, and MacNeil (2012), drawing on primary historical sources report on the history of the Northern Plains region during the late 1800s, when some of the most radical changes occurred among the Plains Nations. After the Dominion of Canada was formed, the Canadian government began looking to secure Western territories that were held by sovereign Indigenous Nations. In 1876, they passed the Indian Act, which was designed to make Indigenous peoples in Canada subject to the authority of the government. Tuberculosis, smallpox, and influenza epidemics periodically swept through the region, leading to population changes. In 1874, the Dominion sent the North-West Police and physicians to Blackfoot territory to tend to the epidemics and to secure greater control over the region. Physicians were ordered to attend to any Indigenous people who asked for help with the intent to foster trust among the

[3] This map was provided by my advisor and elder of the Kainai nation, Francis First-Charger. It is used by members of the Blackfoot nation to educate non-Blackfoot people about the extent of their traditional territories before colonization began in earnest.

Blackfoot, particularly the Piikani who were gaining affluence in the area (Daschuk, Hackett, and MacNeil 2012:73).

The Blackfoot, like many Plains Nations, traditionally relied on the bison herds that were abundant on the North American plains as a natural resource. The decline in the bison populations caused by colonial activity created a scarcity that drove many plains nations including the Blackfoot to change their way of life. Over the course of the decade from 1870 to 1880, the bison herds rapidly disappeared from the Plains and caused widespread famine. By the mid-1870s, the famine had spread to the Western Plains, and by 1878, Colonel James Macleod reported that the Blackfoot experienced their first failed Bison hunt,[4] and were suffering from starvation and an especially cold winter. For the Blackfoot, this calamity was closely followed by a tuberculosis epidemic. These conditions led to the signing of Treaty 7 in 1877 by the Blackfoot alongside the Stoney Nakoda[5] and Tsuut'ina nations. As seen in the report from Colonel Macleod, the signing of the Treaty did not immediately remedy the situation, and in some cases, exacerbated the problem (Daschuk, Hackett, and MacNeil 2012:75–76).

According to the Report of Acting Superintendent MG Dickieson (July 1879), the Plains Nations including the Blackfoot Nations were driven to shift to a more European-style labor-based economy as wood cutters or farmers.[6] They no longer could rely on their traditional lifeways, but now needed to conform to colonial economics, supplemented by government assistance. The Blackfoot people, although generally confined to their reserve lands, retained their cultural traditions whenever it was possible. C.C. Uhlenbeck was a Dutch linguist and anthropologist who was among the first to describe the Blackfoot language academically. During his documentation expedition to the Blackfeet reserve of Montana in the summer of 1911, his wife reported that many of the Blackfoot continued to live in tipis, and to travel about their territories (Genee 2009:108–112).

The Blackfoot retained their traditions and language where possible, but the next change that occurred caused a further shift in Blackfoot culture. The Government of Canada mandated that Indigenous children, subject to the Indian Act, attend residential schools. These schools were used in both Canada and the United States, and the expressed intent of these schools was to

[4] An annual activity that was essential to the Blackfoot economy and involved all the Blackfoot nations.
[5] A Siouan speaking nation that live along the edge of the Rocky Mountains from Southern to Central Alberta.
[6] I say European-style to distinguish the type labour economy that the Blackfoot engaged in traditionally from the type of economy that they engaged in later on.

change the culture, language, and religion of the Indigenous people of Canada, often referred to by its early proponents as a civilization process (Davin 1879, Ulysses Grant circa. 1870, quoted in Haig-Brown (2012:222–23). The residential system is no longer in effect, but its effects still influence Indigenous people today. It is this system that led to dramatic decreases in Indigenous language use over the course of the 20th century.

Many Blackfoot people now live on reserved land set aside for them by the Canadian and United States governments. There are three Blackfoot reserves in Alberta, one each for the Siksika, the Kainai, and the Aapatohsipiikani, and one in Montana for the Aamsskapipiikani. Figure 2 shows a map of the locations of the current Blackfoot reserves in both Canada and the United States and Figure 3 provides a map of the modern treaty territories in Alberta.

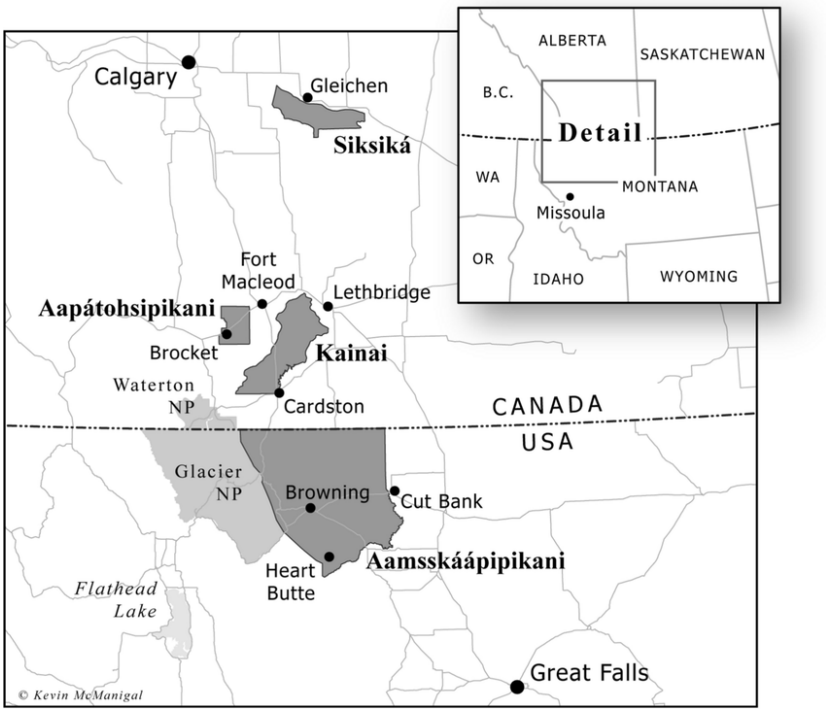

Figure 2: Modern locations of Blackfoot reserves [Map by Kevin McManigal, from Weber 2022]

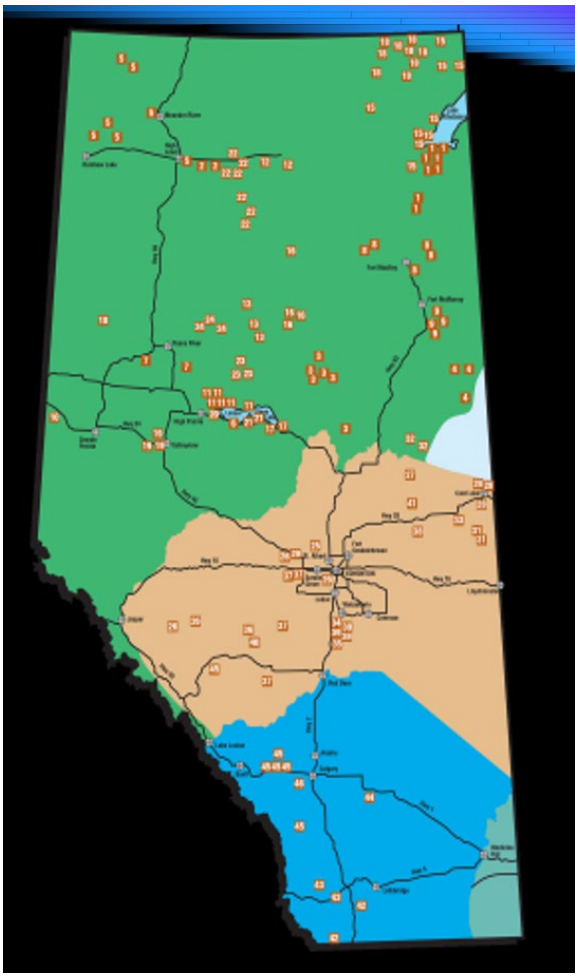

Figure 3: Modern Treaty territories in Alberta (Treaty 7 is in blue, Treaty 6 is in pink, and Treaty 8 is in green).

The Blackfoot people in Canada, along with the Tsuut'ina and the Stoney Nakoda peoples now fall under the numbered Treaty 7 (in blue), which was signed in 1887. Thus far, I have provided background on some of the relevant history of the Blackfoot people, but their history and culture is too rich and diverse to adequately cover in a single section. I now move on to a brief discussion of their language and its linguistic context.

The Blackfoot language is part of the Algonquian language family, which is a large family stretching across a vast territory in North America, from the Plains Cree in the North-West of what is now Alberta Canada, to Mohican and Delaware on the Mid-Eastern coast of the United States in what is now North Carolina. The Algonquian language family traditionally covered a territory nearly covering the width of northern North America, spreading as far north

as the North-West Territories of what is now Canada, and as far south as Colorado in what is now the United States.

The map in Figure 4 shows our current scientific understanding of the pre-colonial distribution of the Algic languages, of which the Algonquian languages are a large subgroup. The Algonquian languages are shown outlined on the map in orange, not including Yurok and Wiyot (on the Pacific coast), which are outlined in green, and are related to the Algonquian languages, but more distantly. An approximation of the traditional territory of the Blackfoot people is highlighted on this map. The relationship between Algonquian, Wiyot, and Yurok within the larger Algic family is shown in Figure 5.

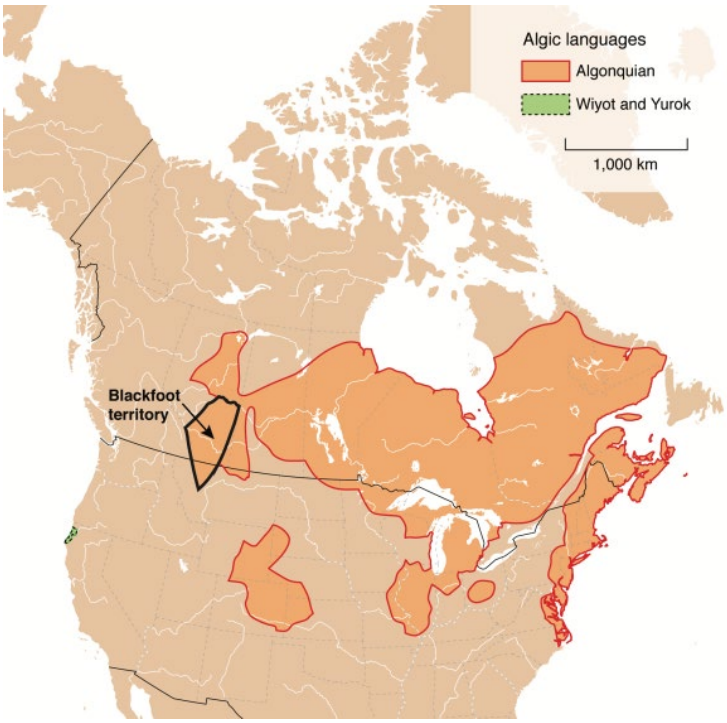

Figure 4: Historical distribution of major Algic languages. [Image by Erin Leinberger, from Weber 2022]

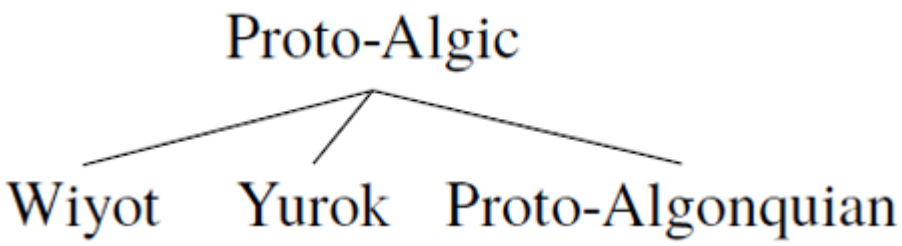

Figure 5: Algic family tree. [Image from Oxford (2019)]

Algonquian languages are described as polysynthetic and agglutinating because most grammatical information is conveyed by stringing morphemes together. They share many grammatical features, including an animacy and obviation system, a direct-inverse person hierarchy, and a very verb-centric morpho-syntactic marking system requiring up to four classes of verbs based on transitivity and animacy parameters (Oxford 2019:506–19).

Blackfoot has a very small phoneme inventory, as do many Algonquian languages, effectively represented orthographically in an alphabet of eleven consonants and three vowels, making for 14 total characters (Frantz 2017:1-7). Vowels and consonants can both be distinguished by their length, (represented by doubling the long vowel or consonant), and most words have a pitch accent in a distinct position represented by an acute accent over the vowel of the high-pitched syllable. A syllable may have a falling or rising pitch if the syllable's nucleic vowel is long or is a diphthong. The grammatical features of Blackfoot are typical of an Algonquian language with high degrees of agglutination. It is similar to Cree in that the majority of affixes for both nouns and verbs are prefixed to the stem. Blackfoot words can take prefixes that mark tense, aspect, negation, person-agreement, and clause type (see sections 2.1 and 3.1 for a detailed description and breakdown of sources).

The history of Blackfoot language documentation begins with missionization efforts made by Catholic and Anglican missionaries (Genee 2020:4-7). Their sources represent the earliest Blackfoot documentation work that we have to date. The first were Catholic missionaries, but their written records are few and far between and were used only by other missionaries. Methodist John Maclean worked on the Kainai reservation from 1880 to translate religious resources and he developed a dictionary. The first published grammar and dictionary of Blackfoot were created by John William Tims (Tims 1889), an Anglican missionary who worked

on the Siksika reservation. This was followed by a translation of the gospel of Matthew (Tims 1890).

Academic linguists have been conducting research on the Blackfoot language for close to 100 years. Some of the earliest and most comprehensive academic work was undertaken by the Dutch linguist C. C. Uhlenbeck. His field studies of Blackfoot resulted in a Blackfoot grammar (Uhlenbeck 1938) and dictionary (Uhlenbeck and Van Gulik 1930, 1934). His work was largely theoretically motivated so, as was the academic custom of the time, the linguistic resources which he gathered from the community were not made available to them for their benefit (Genee 2009).

The next stage in documenting the language came from Allan Taylor in the 1960s. Allan Taylor's doctoral dissertation (Taylor 1969) was titled *A Grammar of Blackfoot* which was the first detailed description of Blackfoot phonology, morphology, and syntax. Donald Frantz published his doctoral dissertation in 1971, titled *Toward a Generative Grammar of Blackfoot: With Particular Attention to Selected Stem Formation Processes* which formed the basis for his later works, *Blackfoot Grammar* (Frantz 1991) and the *Blackfoot Dictionary of Stems, Roots, and Affixes* (Frantz and Russell 1989) drawing on Taylor's work, along with his own extensive documentation. These volumes contain grammatical and lexical descriptions of the Blackfoot language. It is for this reason that the groundwork of this thesis was largely based upon the information contained in the most recent editions of these two volumes (Frantz 2017; Frantz and Russell 2017).

Today, each of the three Blackfoot nations represent a variety of the language. Differences between regional and tribal varieties are small enough that speakers have no trouble communicating with Blackfoot speakers from other areas, but they are significant enough to be noticed by speakers. A significant variation of Blackfoot can be found between Old (ages 80 and up) and New Blackfoot speakers (ages 60-80). The Old and New varieties are mutually intelligible, but New Blackfoot speakers cannot produce Old Blackfoot (Miyashita and Chatsis 2015:109-115). The specifics of how each variety differs from the others is not yet well documented, but I cover variation in more detail as it applies to this research in chapter 4, section 4.1 of this thesis when describing the challenges with creating and using a Blackfoot corpus to test the Blackfoot computational model.

Over the past 100 years, the Blackfoot language has been declining in use. This is mainly due to pressures from the colonial majority language, English which is privileged in Canadian law and society, and to the active efforts made by the Canadian government to carry out a cultural genocide against Indigenous peoples[7] (Kingston 2015:64-89; Truth and Reconciliation Commission of Canada 2015). As a result, most first-language users are elderly, and the language is rarely passed on to the next generation. According to a 2022 census undertaken by the government of Canada, the total population in Canada who reported Blackfoot identity was 23,200. Of them, 4,945 people reported that they spoke Blackfoot as their mother tongue, and only 4,970 claimed to speak Blackfoot in the home. The census reported an estimate of 6,680 people who had knowledge of the Blackfoot language. Table 2 breaks these statistics into age groups, following Table 1 in section 1.1.

Table 2: Blackfoot identity population, by age and selected language characteristics, Canada, 2022. [From Statistics Canada, Census of Population 2022.]

Age Group	Total Blackfoot identity population[8]	Has knowledge of Blackfoot		Speaks Blackfoot as a mother tongue	
	number	number	percent	number	percent
Total	23,200	6,680	28.8%	4,945	21.3%
0 to 14 years	6,500	1,280	19.7%	895	13.8%
15 to 24 years	3,615	815	22.5%	555	15.3%
25 to 54 years	8,895	2,420	27.2%	1,570	17.6%
55 years and older	4,185	2,175	51.9%	1,925	45.9%

When we compare the national statistics on Indigenous languages shown in Table 1 (section 1.1) with Table 2, the Blackfoot language appears to have many speakers. However,

[7] This term is used to describe a process where one group attempts to eradicate another group, not necessarily violently (although violence is usually involved), by eradicating cultural and linguistics distinctness. The Canadian government intentionally tried to assimilate Indigenous people into settler society. They sought to achieve this goal by creating residential school programs for children, withholding certain rights and privileges from Indigenous Canadians, outlawing traditional practices and lifeways, and a variety of other tactics. They are now working towards reconciliation with Indigenous peoples in Canada.

[8] In parallel to Table 1, I am only including the information from the 2022 Canadian census. This means that it only included people with Indigenous status in Canada, who identified as having full or partial Blackfoot ancestry, living on reserves or not. The world Blackfoot population is greater than what is reported in Table 2, as this does not include the population in the United States, and because the census parameters may be excluding some Canadians with Blackfoot ancestry.

when we compare the age statistics, the proportion of speakers and language users drops drastically as the age range decreases. Further, the parameters 'has knowledge of Blackfoot,' and 'speaks Blackfoot as a mother tongue' are problematic, because they are vague and do not define levels of skill or fluency. As a result, this has likely led to an inaccurate portrayal of how many individuals speak Blackfoot, and to what degree. Regardless, if this general trend continues, the Blackfoot language may no longer be spoken within a few generations. The survival of the language now lies with revitalization efforts at all levels, undertaken by individuals and Blackfoot communities, at home, in schools, and in a variety of other contexts. If the community desires it, linguists can help to support these efforts by using their knowledge to develop tools.

1.3 Technologies for Blackfoot and other Indigenous languages

Researchers in the field of computational linguistics study the use of computation in modelling and processing language, and the uses of language in computation. Linguists in the latter group may design or study programming languages which are used to communicate with computational machinery. Others might use computation as a framework to theorize about or to study human language capabilities. Linguists from the first group focus on the computational processing of natural languages. These linguists design computational technology that is capable of mimicking human abilities to process or produce natural languages. This thesis describes a Natural Language Processing project and was written with the goal of creating, evaluating, and describing a computational model of Blackfoot noun and verb morphology. This section provides a background for this model by describing the work that has been done for Blackfoot in digital language preservation and revitalization efforts, and in producing computational processing technologies for the Blackfoot language.

The foundational computational technology that preceded this modelling project, is Finite State Transducers (henceforth FSTs). Understanding this type of computational machine is integral to this project. FSTs are a type of computational machine that allows us to define which types of states the machine can have, and how the machine can get there. For example, a light switch can be considered a finite state machine with two states, an ON state and an OFF state. Networks between the states provide legal pathways between the two states, allowing a user to change the state of the machine from ON to OFF, or from OFF to ON. One cannot change the state

to ON when the machine is already ON, so that would be considered an illegal state change. FSTs can become more and more complicated by adding states and networks between those states and by defining initial and terminal states. The light switch example has two networks defined, going from OFF to ON and vice versa, but has no terminal state, so it can loop endlessly between the two states. If we define an initial state, and a terminal state at OFF, and a terminal state at ON, then the machine will only be able to go from the ON state to the OFF state. This example of the light switch as an FST was used in Beesley and Karttunen (2003:1-3).

In Natural Language Processing, the capabilities of FST technology are useful because they can be used to mimic human linguistic capabilities. For example, if we set the initial state of an FST to be the word LOVE, and then define three acceptable end states, D, S, and ∅ (represented in the diagram as having no morpheme on the surface), this FST would be capable of producing the words LOVED, LOVES, and LOVE. Taking this further, we could design the FST to have two "sides" to each state, which would map two pieces of information to each other. In linguistics, this could be a surface form and an underlying analysis. The resulting FST could then produce the surface forms LOVED, LOVES, and LOVE which could be directly mapped to the underlying forms, LOVE+PAST, LOVE+3SG, and LOVE+PRESENT. Xerox style FST technology is explained generally in Beesley and Karttunen (2003:1-37). In section 1.4, I describe how this is designed and implemented using formalisms which can be implemented to design FSTs. I illustrate the *love* example using an FST network diagram in Figure 6.

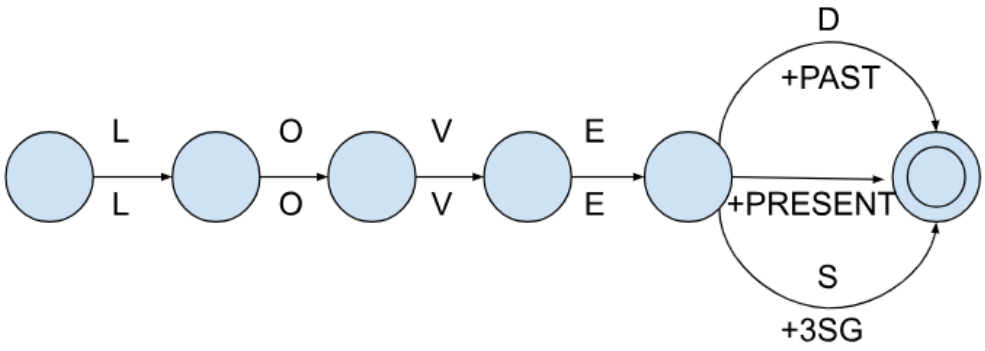

Figure 6: FST network diagram example

The diagram shows an FST network that begins at the left with the initial state and ends with the terminal state at the right, represented by a circle with a second concentric circle within it. Each

state is connected by an arrow, representing a network, which ties each state together and defines its directionality. The surface forms appear above each network arrow, and the underlying analyses tied to each surface form is found below the network arrow. The parts above and below each arrow are tied to each other through their corresponding states and networks.

I have briefly illustrated how FST networks can be constructed as a type of computational machine. These networks can become increasingly complex, from something as simple as a binary ON and OFF network, to representing the morphological relationships between thousands of morphemes that can have hundreds of thousands of states and networks. In sections 2.3 and 3.3, I diagram the FST technology that I designed, but I do so on a more abstract level as it would not be feasible to portray the hundreds of thousands of states and networks present in the model. Now that I have established the foundational technology behind this project, I can address the history of Blackfoot language technology and situate this project within its context.

The first attempt at creating a Blackfoot morphological parser was made by Dunham in his doctoral dissertation (2014), which reported on a parsing tool that he developed to automatically gloss text in the Online Linguistic Database (henceforth, OLD). In principle, OLD was designed to be an open-source database for linguists and community members to enter linguistic data from a range of languages that had been collected for documentation purposes. The positive side to this is that it was designed to help address some of the ethical issues which often plague language documentation work, such as ownership of data and who should be in control of documentation work. This would be done by allowing speakers of languages to collaborate in language data collection and analysis if they chose to do so. It was designed to allow both language community members and linguists to have access to data which is normally extraordinarily difficult to find otherwise (Dunham 2014:101-158). Unfortunately, this has not turned out to be the case, as the OLD project is no longer actively maintained, and the Blackfoot portion of the corpus is no longer easily accessible.

The downside to such a collectively assembled corpus is that spelling and glossing conventions are difficult to regulate, especially these are undertaken by untrained users. Thus, Dunham designed a software that could parse text from any language represented in the OLD automatically, so that users would only need to check how well the automated parsing was done and make corrections where needed. The software he designed was a bidirectional Finite State Transducer with two components, the first of which was a phonology script handmade by

Dunham himself, and the other was a morphological parser in a LEXC script which was generated by a machine learning model. Together, they were designed to be used to analyze any text from any language in the OLD, but Dunham tested the technology on Blackfoot text along with an N-Gram Language Model, which served the purpose of disambiguating ambiguous word forms (Dunham 2014:159-205).

Although Dunham's work was specialized for the OLD and, because of the factors described above, is not suited to the needs of the 21st century tools for Indigenous languages project, it was nonetheless valuable, and was investigated for this project, particularly the phonological model which he created, as it could be used in early testing. The reasons that Dunham's parsers could not be used in this project was mainly because they were designed to be flexible, so they could be used on many languages represented in the OLD corpus, but that meant they were not very precise, and relied partially on linguists, community members, or other language specialists to edit and check the results of the parser. In order to create intelligent dictionaries, spell checkers, and other technologies which can easily be used by the Blackfoot community, the technology must be more precise and specifically designed for Blackfoot. Even when used as corpus parsers, Dunham's designs still required a lot of work to fill in the blanks left by the parsers, and to correct the result of their work, which in turn, would require a certain level of user expertise. The Blackfoot modelling project was undertaken with a final product in mind; an inflectional model that could generate morphological paradigms, and tools that could parse Blackfoot texts. Dunham's parser was not designed to generate morphological paradigms, nor to parse only Blackfoot texts, thus, it was not suitable to use for this project.

The foremost goal of the Blackfoot model is to provide a free, intelligent, online dictionary. Free-to-use dictionary software is already available through the Algonquian Dictionaries and Language Resources project (https://www.algonquianlanguages.ca/) (Junker and Torkornoo 2017). The Algonquian Dictionaries project seeks to provide online dictionary resources for Algonquian language speaking communities. The Blackfoot dictionary project is part of this project and is headed by Inge Genee at the University of Lethbridge. The dictionary project essentially takes the best available language resources and makes them freely available in an online dictionary, and in some cases, corpus formats. In the case of the Blackfoot dictionary, it is based on Frantz and Russell's Blackfoot dictionary (2017) plus some additional content such

as stories and example sentences which have been collected from fluent Blackfoot speakers. It is also being continually developed through ongoing documentation work.

Currently, the website can only return searches for isolated Blackfoot stems, lexical prefixes, and examples of their use. Examples can be searched using the advanced search function, which is very useful, but not all lexical entries include examples, and there is no lexical entry which can provide full inflectional paradigms. Unfortunately, this still limits searches to items which have been manually entered into the dictionary. This is problematic because Blackfoot morphemes and stems do not occur in isolation in text or speech and will always be inflected when in context and each word could have upwards of thousands of different inflected forms, too many for even a team of researchers to painstakingly enter into the dictionary database by hand. This limits the capabilities of the Blackfoot dictionary in a way that excludes many members of the Blackfoot community, particularly those who do not have any explicit training in Blackfoot grammar, or who are unfamiliar with the language. Figure 7 illustrates this problem using real examples from the dictionary.

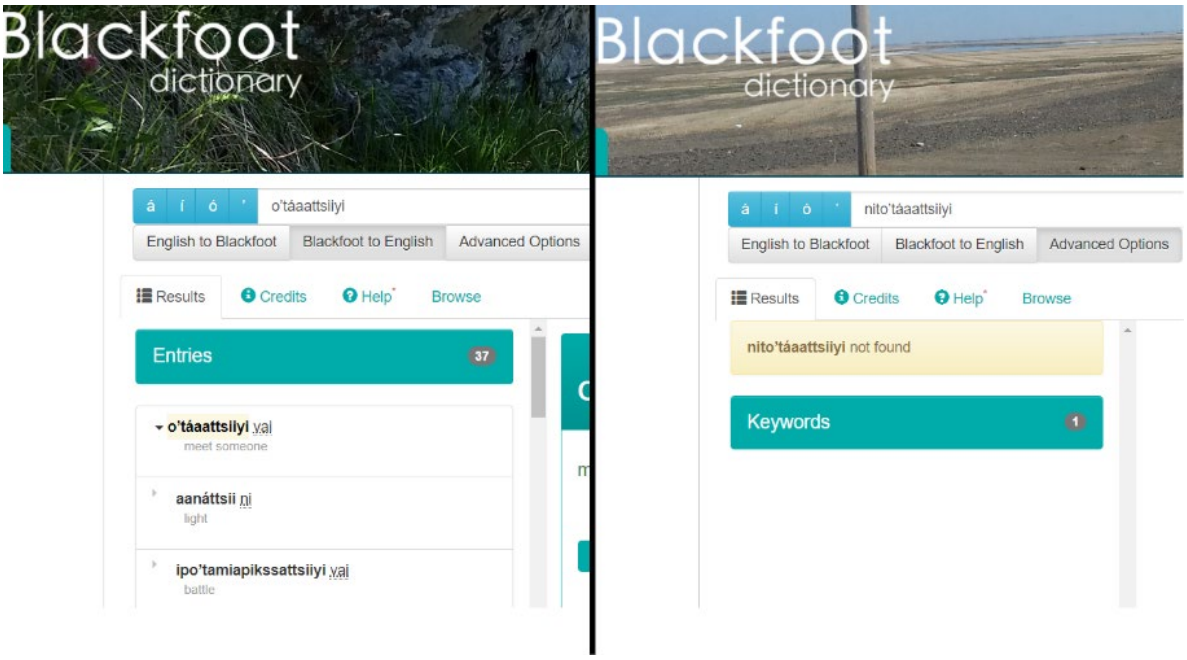

Figure 7: Current Blackfoot digital dictionary search results.

Notice that on the left side of Figure 7, the search result is successful because the user is searching for the isolated stem *o'taaattsiiyi,* which is the verb stem for 'eat.' As such, a list of

matching lexical entries appears under the *entries* tab, and the first entry that appears is the one the user is looking for. On the right side of the Figure, the stem *o'taaattsiiyi* is inflected with a basic verb prefix *nit-*, which agrees with a first person singular possessor. This form is not found anywhere else in the dictionary, thus searching for it does not return any results. The implementation of FST technology can change this situation, allowing community members to search for a full variety of complex word forms found in the context of spoken or written Blackfoot (Johnson, Antonsen, and Trosterud 2013).

The other immediate use of the Blackfoot model is in documentary linguistics. Researchers can use the model to accurately parse corpora containing only Blackfoot text. Such a tool would allow linguists and community members to document Blackfoot with minimal supervision required for disambiguating ambiguous word-forms, and correcting errors.

1.4 Technological background

This section provides an overview of the technological tools which supported this project. Such timely and advanced development of Blackfoot morphological models would not have been possible without the technology described in this section. In this section, I provide a computational explanation of the compilers and formalisms used in development. I also use examples from a simplified Blackfoot morphological FST to illustrate the uses of these technologies. This section provides the necessary background for chapters two and three of this thesis, which describe the structure of the full computational models of Blackfoot noun and verb morphology.

Before discussing the technology that I utilized in this project, I provide some definitions and explanations of terms used in this section. In computation, a *compiler* is a program that translates software, which can be written and read by humans as though they are natural languages, into a format that can be processed by computational hardware. *Compilers* are specialized for code in specific programming languages or formalisms.

In programming, the term *syntax* is used to refer to the allowable set of symbols that are recognizable by a given *compiler*, and how they combine, making it somewhat similar to the linguistic definition of *syntax*. Using the correct *syntax* for a given programming language will allow the *compiler* for that language to easily transfer the text in the program to machine code.

Incorrect *syntax* will cause the *compiler* to be unable to compile the program and will usually crash or will have unintended consequences for how the program functions (often called *bugs* by programmers).

Formalism is a term used by logicians and mathematicians to describe a set of allowable symbols, what they denote, and how they can be combined to create a string. For example, in linguistics, we have a formalized system of writing phonological rules in shorthand (exemplified later in this section). By this definition, we could also consider a given programming language's *syntax* to represent a type of *formalism,* or rather, the formalized logic behind the *syntax*. If we consider programming languages or *formalisms* to be analogous to natural languages, then, just like natural languages, all formalisms have a wide range of expressions, and can vary quite significantly. Knowledge of one programming language does not mean one can use any other programming language, although knowing one makes learning another easier. Unlike natural languages, not all programming languages or *formalisms* are designed to perform every possible function, and some, like the *formalisms* described in this section, have very specialized capabilities. LEXC for example, is a *formalism* that can be implemented to compile FSTs using a *formalism* that can be used for modelling the morphologies of natural languages.

This project made extensive use of the FOMA compiler. FOMA is a finite-state compiler which was created using the C programming language.[9] FOMA can compile programs written using either the LEXC or XFSCRIPT formalisms. LEXC is a formalism that can be used to define what are referred to as lexicons and to restrict how those lexicons combine with each other. In this thesis, I use the term *lexicon* to refer to a type of *lexicon* that is defined through the LEXC formalism, unless I specify otherwise.[10] These lexica are similar to the linguistic concept of the mental lexicon as a sort of dictionary of words or morphemes and their underlying meanings, roles and uses contained within the grammar. LEXC programs use FST technology (explained in section 1.3) to computationally formalize the concept of 'dictionary' and 'grammar,' allowing the designer to organize various lexica (the dictionary) and define how they

[9] C is a functional, multi-use programming language. It is a third-generation programming language that was written using an Assembly programming language, which itself is derived directly from the computer architecture using Binary. That is what allows written software to effectively communicate with the computational machine. C has also been used to develop fourth-generation compilers and libraries for general-use languages such as Python, Perl, Ruby and PHP.

[10] Not to be confused with the definitions of lexicon used by linguists, or the definition of lexicon used in computer science.

can combine generatively (the grammar) (Hulden 2009, Beesley and Karttunen 2003). The XFSCRIPT formalism is written using regular expressions, which are discussed in greater detail later in this section. The general type of computational technology described in this thesis is an FST, a type of computational technology that I describe in section 1.3 above.

The Helsinki Finite-State Toolkit library (henceforth HFST), is a library and compiler which was developed as an alternative to the XFST compiler, also using the LEXC, XFST, and REGEX formalisms. It implements all aspects of the FOMA compiler with some expanded capabilities, including the ability to compile TWOLC scripts, and to apply weighting. The HFST compilers and library were written using the C++ programming language (HFST is first described in Lindén et al. 2013). Most importantly for my purposes, HFST includes a library that allows users to assign weights to components in LEXC files, allowing for a simple intelligent computational model that can be informed by a corpus of texts. Weighting is discussed in detail in chapter 4, section 4.4 of this thesis, and is not relevant to the preceding sections.

The LEXC formalism allows linguists to use their ability to understand and describe grammatical structures to implement their knowledge into a computational model of the grammar. For the remainder of this section, I use examples to demonstrate how LEXC and regular expressions can be used to produce a model of Blackfoot morphology and phonology. Figure 8 is a simple model of Blackfoot verb morphology.

```
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 8: Simple FST for Blackfoot verb morphology.

Figure 8 provides a real example of how the LEXC formalism can be implemented to build a model of Blackfoot morphology using a Xerox-style compiler such as FOMA. This was the method I used to design the Blackfoot technology, so in this section, I describe the components in this small Blackfoot FST to demonstrate the uses of FOMA, LEXC, XFST and HFST.

Since the FOMA and HFST compilers read the LEXC program from top to bottom, the program starts on line one with a lexicon defined as *prefixes*. For the sake of brevity, only three regularly agglutinating person prefixes are included in this lexicon. For each person prefix, the prefix tag is found to the left of the colon, with a + symbol on their right edge to indicate a morpheme boundary for prefixes, and the surface form of the morpheme is to the right of the colon. In this case, the underlying analysis is the same as the surface form. Following the surface forms of each prefix, the command *verbs* defines the next lexicon in the path. The *verbs* lexicon starts on line 6, where the structure is essentially the same as in the *prefixes* lexicon, but contains a single, regularly inflecting Blackfoot verb stem *á'poo* 'go.' The surface form of the verb is the basic inflecting surface form, and the underlying form of the verb to the left of the colon represents the lemma of the verb (a basic inflected form of the stem, in this case, the third-person, proximate, singular suffix -*wa*). The *verb* lexicon then goes to the next defined lexicon labelled *suffixes,* containing person suffixes. This lexicon is similar to the *prefixes* lexicon, but the underlying analyses are the grammatical meanings which are added to the verb by each respective suffix in the lexicon. The hash mark following each suffix signals the termination of the process, or in this case, the end of a word.

The example shows how the LEXC formalism allows users to define the positions of morphemes that can concatenate together. However, in the example, the morphological components are able to combine freely, therefore producing ungrammatical forms. Table 3 below shows some possible outputs of this FST. Ungrammatical forms are marked with an asterisk.

Table 3: Outputs of a simple FST containing Blackfoot morphemes.

Surface form	**Underlying analysis**
á'poowa	NULL+á'poowa+3[11]
kitá'poo	kit+á'poowa+2
*nitá'poowa	*nit+á'poowa+3

[11] I would like to note that in the actual model, I do not employ an underlying representation of null morphemes as NULL+. This is only here to explicitly show how the model comes together. In the actual model, a null morpheme simply would not appear in the underlying analysis of the word.

Since the FST accepts both unacceptable input and output, the next step in development is to constrain the rules of the grammar to ensure that they do not accept or produce ungrammatical forms. This can be done through the use of flags.

Flags can be used to restrict the allowable combinations of morphemes from one lexicon to another. A LEXC script making use of flags is shown in Figure 9, using the same model as shown in in Figure 8.

```
LEXICON prefixes
@P.person.nit@:@P.person.nit@nit verbs ;
@P.person.kit@:@P.person.kit@kit verbs ;
@P.person.null@:@P.person.null@0 verbs ;

LEXICON verbs
a'poowa:a'poo   suffixes ;

LEXICON suffixes
@R.person.nit@+1:@R.person.nit@0    # ;
@R.person.kit@+2:@R.person.kit@0    # ;
@R.person.null@+3:@R.person.null@wa # ;
```

Figure 9: Simple Blackfoot FST utilizing flags.

In the LEXC program shown in Figure 6, the flags are first set in the *prefixes* lexicon. The flags are set on both sides of the colon, which sets them for both the underlying and surface forms. The flags appear between the @ symbols, which are used to signify a flag in the LEXC syntax. These are P flags used to pre-set a flag condition. Tags and flags need to be set as multicharacter symbols since each set of tags and flags can be different for each language being modelled in this way. I chose to label them as person flags because they control which person prefixes can cooccur with which person suffixes in the same word. The prefix lexicon then goes to the *verbs* lexicon as normal. Then, when it goes to the *suffixes* lexicon, the R flags restrict which prefix flags the suffixes can combine with. For example, the @R.person.nit@ flag can only cooccur with the @P.person.nit@ flag and excludes the possibility of it cooccurring with any other flags of the same type such as the @P.person.kit@ flag. Now that the morphology has been corrected, the next step is to implement phonological rules that will allow the user to include a wider

variety of verbs in our computational morphology of Blackfoot. These phonological rules are implemented using Xerox Finite State Technology (henceforth, XFST).

The XFST formalism makes use of regular expressions and is most useful for defining phonological rewrite rules. XFST programs apply phonological rewrite rules in the order in which they are compiled, allowing the program creator to define the rules and then order them, in a similar way to Optimality Theory in phonology. Figure 10 is an example using the same program in Figure 9, with the addition of two words which require phonological rewrite rules.

```
LEXICON prefixes
@P.person.nit@:@P.person.nit@nit verbs ;
@P.person.kit@:@P.person.kit@kit verbs ;
@P.person.null@:@P.person.null@0 verbs ;

LEXICON verbs
a'poowa:a'poo       suffixes ;
ikskimaawa:ikskimaa suffixes ;

LEXICON suffixes
@R.person.nit@+1:@R.person.nit@0    # ;
@R.person.kit@+2:@R.person.kit@0    # ;
@R.person.null@+3:@R.person.null@wa # ;
```

Figure 10: Simple Blackfoot verb FST in need of rewrite rules

The addition of the word *íkskimaa* 'hunt game' means that the program now requires rewrite rules, otherwise its surface form will appear as **nitíkskimaa*, which is not correct, as it does not follow the t-Affrication rule (Frantz 2017:178), which would make it *nitsíkskimaa*. To solve this, I can use a simple XFST file to define a rule which turns the /t/ in *nit-* and *kit-* into an affricate when it appears before any /i/. Figure 11 shows this simple rule.

```
define tAffrication [ 0 -> s || t _ [ i | í ] ] ;

```

Figure 11: Rewrite rule for simple Blackfoot verb FST

The rule is defined as tAffrication on the left side of the line. Then, in the brackets, the surface manifestation of the rule is described, as 0 -> s, which can be read as ∅ becomes s, or /s/ is

inserted. The double pipe (||) signals that everything to the right of it defines the environment where /s/ is inserted. In this case, /s/ is added in an environment wherever there is a /t/ followed by an /i/. This is formalized as t _ [i | í], meaning 'between /t/ and /i/…'. Computationally, both the /i/ with and without the acute accent marking must be specified because the software treats them as different characters. The pipe (|) that separates them can be considered as a logical operator for an 'or' statement. The whole thing can be read as "I define t-Affrication as /s/ insertion in an environment where /t/ appears to the left of /i/." This formalism is called a regular expression. Regular expressions can be used to easily write phonological rules because they are formatted in a similar way to the types of formal phonological rules used by linguists and taught in introductory phonology courses. When we combine this phonological[12] model with the morphological model, it will always check if /s/ will be inserted by reading the environment that /t/ is in. So, this rule does not affect *a'poo,* or any morpheme that doesn't begin with an /i/.

The result of the parser is a Finite State Transducer (or FST) with two sides, one which can analyze underlying analyses and return a surface form, and one which can analyze surface forms and return an underlying analysis. An example of the output of this simple FST is shown in Table 4.

Table 4: A simple Blackfoot FST output

Underlying analyses	**Surface forms**
ikskimaawa+3	ikskimaawa
ikskimaawa+1	nitsikskimaa
a'poowa+2	kita'poo

When the FST is in use, entering the surface forms will return a corresponding underlying analysis and entering an underlying analysis will return a corresponding surface form.

The technology described throughout this section exists within a larger technological architecture. The GiellaLT infrastructure was created by Moshagen, Pirinen, and Trosterud (2013) and was designed to support technological development for a variety of languages using the FOMA compiler. The GiellaLT infrastructure compiles the scripts that make up a computational grammar through GitHub. First, compiling the morphological components

[12] I use the term phonological because this is probably the most accurate description of what it is, even though it only deals with phonological processes that are represented in the orthography. Terms such as ortho-phonological, may be used in some circles, or alternatively, phono-graphical rules, but for simplicity, I am sticking with the term phonological.

together, then combining those with the phonological model, and finally, combining that with any other supporting technology such as weighted models or orthography relaxers. *Figure 12* represents how the Blackfoot morphological and phonological model is compiled. Components written in LEXC are in blue and the components written using XFST are in red.

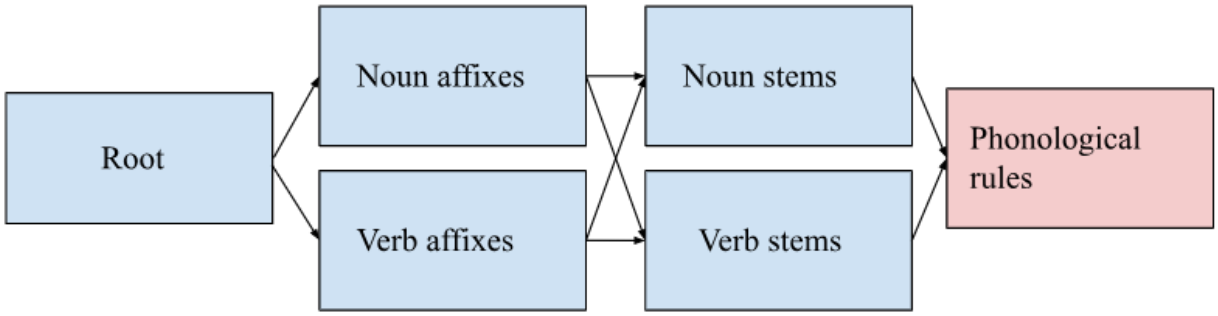

Figure 12: Blackfoot morphology and phonology architecture

The *root* contains a lexicon that defines the morphological paths through the model, representing the initial state of the FST. It also contains a list that define all of the multicharacter symbols including the underlying analyses of morphemes, flags and special characters that are used in the subsequent lexica. Setting these here allows them to be recognized by the FOMA compiler when compiling programs written in the LEXC formalism. This is important because the compiler does not automatically recognize these functions, as they are unique to every language that is being modelled.

Special characters include characters that are used in underlying forms, and any special character that is required by the phonology. For example, in Figures 8, and 9, the symbol used to signify the third-person singular in the underlying analyses, is +3Sg. It is essential to the performance of the program, to define this as a multi-character symbol. Any special symbols used in the phonology also need to be defined. For example, in Blackfoot, there are two alternating /i/ phonemes. One behaves regularly, only causing affrication when it occurs after /t/ as in the word *Niitsitapi* (*niit-* 'true/real' + *itapi* 'people'). This /i/ alternates with what is referred to by Frantz as breaking-i (Frantz 2017:28, 176), which also causes affrication when /i/ is preceded by /k/, or when /i/ precedes /t/. The alternation can be marked in the lexicon with regular /i/ being represented as /i/ and breaking-i being represented in the model as i2 /I/,

requiring /I/ to be defined as a special character in the computational model. Frantz uses /I/ to represent breaking-I, but I use i2 so that the model does not have to distinguish between upper- and lower-case characters as underlyingly different phonemes.[13]

The user defines the initial continuation lexica in the root lexicon, which allows the model to interact with the affix lexica which follow. This currently includes a verb and a noun affix set but can be expanded to include morphological models of the remaining parts of speech. These lexica contain all the prefixes and suffixes for their respective word categories, and the rules for how they can combine. Then, the user can set the verb and noun stem lexica. Notice that the verb affixes and noun affixes can go to either the noun or verb stems. This allows for nominalization and verbalization, which I explain in detail in chapters 2 and 3. The two stem lexica contain all the noun and verb stems that are used and organizes them according to their classes and subclasses. The script that generates acceptable sequences of morphemes constituting words for each part-of-speech is written in LEXC, making up the morphological model. The phonological changes taking place at morpheme boundaries and elsewhere, from the underlying morpheme sequences that the morphological model defines to the surface word forms, are written with Xerox-style rewrite rules, which make up the phonological model.

This section has provided a description of the technology that makes this project possible, how it works, and how it determines the architectural and engineering choices for the Blackfoot language technology described in this thesis. In chapters 2 and 3, I describe the structures of the morphological components of the Blackfoot LEXC files.

[13] This is important because the model would otherwise parse capitalized characters differently than lower-case characters, which may limit the usefulness of the model.

2. Modelling Blackfoot noun morphology

This chapter provides a description of the elements of Blackfoot nominal morphology which I modelled in this project. In section 2.1 of this chapter, I establish the basic qualities of the Blackfoot noun morphological system that I modelled. In section 2.2, I compare the Blackfoot noun system to the Plains Cree system and explain how I developed the Blackfoot model using the existing Plains Cree model (Snoek et al. 2014) as a base. The chapter ends with section 2.3, a summary of the Blackfoot noun morphology model's structure, and the elements from the Plains Cree LEXC file which I used to design the Blackfoot LEXC file.

2.1 Summary of Blackfoot noun morphology

All the grammatical information that informed this section comes from (Frantz 2017). To the best of my knowledge, this volume describes the entire morphological system for Blackfoot nouns, and so, it is what the Blackfoot model is based on. Section 2.1.1 discusses the basic morphology of noun stems and section 2.1.2 discusses the morphology of nouns derived from verb stems.

2.1.1. Noun stem morphology

Each Blackfoot noun is classified syntactically by its animacy,[14] like in the other Algonquian languages. Each noun is either animate or inanimate and this constrains which verbs can take them as arguments (more detailed information about their morpho-syntactic relationship to verbs is found in section 2.2, and chapter 3), and which types of affixes they can take. This animacy is realized in the surface forms of nouns through nominal suffixes, *-wa* for the singular animate, *-yi* for the inanimate singular and *-iksi* and *-istsi* for the animate and inanimate plural respectively. Nouns also undergo an obviation process where animate singular nouns take a *-yi*

[14] The literature frequently speaks of gender here, but many Algonquian scholars favor the term noun class or animacy over the term gender since this better represents the grammatical structure and, to some degree, the semantic classification of Blackfoot nouns (Little 2018:1; Wiltschko and Ritter 2015:1).

suffix when there is a more prominent[15] third person in a phrase (Frantz 2017:14-15). This happens in a few situations. For example, when an animate noun is possessed by a third person, the possessed noun becomes obviative. Another case is where there are two or more animate third-person arguments in one clause. This obviation process demotes animate nouns to a more peripheral category to disambiguate their roles in a sentence, creating a sort of discourse hierarchy. Animate nouns are either proximate or obviative and since obviative nouns are more peripheral than proximate nouns in the discourse hierarchy, Frantz refers to them as fourth-person nouns or minor third-person, while he refers to proximate nouns as third-person or major third-person (Frantz 2017:14-15).[16] This hierarchy is most clear in the verbal morphology, so I will describe it in more detail in section 3.2 of chapter 3. Inanimate nouns do not undergo obviation in Blackfoot.

Blackfoot also makes use of a non-referring suffix *-i*. Number is neutralized when nouns occur with this suffix, and the suffix can appear on animate or inanimate nouns (Frantz 2017:12-14).[17] The animacy, number, obviation, and non-referring suffixes are shown below in Table 5.

Table 5: Blackfoot animacy, number, obviation, and non-referring suffixes.

Animacy	Animate		Inanimate
Non-referring	*-i*		
Person	3 (proximate)	4 (obviative)	0
Singular	*-wa*	*-yi*	*-yi*
Plural	*-iksi*		*-istsi*

In Table 5, the animacy of the noun is indicated on the left, and the type of inflection is indicated across the top. The rest is straightforward, but for the obviative singular spot for inanimate nouns. This spot has only a dash in it, which I am using to indicate that inanimate nouns are not inflected for obviation.

In addition to these inflection animacy, plural, obviation, and non-referring suffixes, Blackfoot nouns can also be inflected for possessor. This is realized as a prefix indicating the

[15] Algonquian literature refers to proximate and obviative third persons as being part of a person hierarchy that interacts with verb morphology. In this hierarchy, the proximate is considered the true 3rd person, while the obviative is referred to as the 4th person, while Frantz (2017) refers to it as major third person and minor third person.

[16] Using this terminology is especially helpful when discussing the further obviative, which Frantz effectively refers to as the fifth-person (Frantz 2017). The further obviative is not relevant to nominal morphology as it is not marked on Blackfoot nouns, only on the verbs.

[17] Frantz uses the term non-particular rather than non-referring.

person who possesses the noun, and a suffix which indicates the number of possessors when the number of possessors is plural. This is shown in example (1) below.

(1) Kitápotskinaamoaawaiksi (áóhkomiiyaawa)
kit-apotskinaa-im-oaawa-iksi (a-ohkomii-yi=aawa)
2-cow.NA-POSS.TH-2PL-AN.PL (DUR-call.AI-3PL=PRO)
'Your (pl.) cows (are mooing).' (Frantz 2017:80)

In the first word in the example above, there is a second-person plural possessor indicated by the second person prefix *kit-* and the second person plural suffix *-oaawa*. There is also a possessive theme suffix *–im* (*-m* after a long vowel) that some nouns take when they have a possessor, but others do not. The semantic or syntactic reason why some nouns take this morpheme when possessed is not well understood either for Blackfoot or for other Algonquian languages, so nouns that require this suffix need to be lexically specified. We also have the actual noun stem, *apotskinaa*, and at the very end of the inflected word, the animate plural marker, *-iksi*.

The last element of Blackfoot nominal morphology is made up of what are referred to in the Algonquian literature as prenouns. These are essentially descriptive, and act as noun modifiers. It seems that these modifiers can combine freely and recursively as long as they make sense semantically. Example (2) below demonstrates the use of prenouns using the example prefixes *iik-* ('very') and *pok-* ('small').

(2) ííkohpokitapiwa
iik-pok-itapi-wa
very-small-person.NA-AN.SG
'a very small person' (Frantz 2017:85)

The morpheme *pok-* undergoes initial allomorphy which prefixes *oh* to the word whenever it is preceded by another prefix. This is one of the common allomorphic patterns in Blackfoot, which I discuss in more detail later in the thesis (see chapter 4, section 4.4.2.2).

Blackfoot also distinguishes between independent and dependent nouns. Independent nouns, like the ones shown in the examples above, can be inflected for a possessor optionally.

Dependent nouns must be inflected for possessor. They also take shortened person prefixes such as *n-* instead of *nit-* for first-person possession, *k-* instead of *kit-* for second person, and *w-* instead of *ot-* for third person. Dependent nouns are primarily kinship and body-part terms,[18] and Frantz refers to them as relational nouns (Frantz 2017:51).

2.1.2 Nominalization of verb stems

Both nouns and verbs can change their part-of-speech class in Blackfoot. In this section, I discuss the how nouns can be derived from verb stems. This is done in a few different ways depending on the type of noun being derived. The nominalization process is described in Frantz (128-145).

The simplest way to nominalize verb stems is to use noun inflection (Frantz 2017:128). For example a verb stem inflected with the suffix *-iksi* derives an animate plural noun from an animate intransitive verb stem, comparable to using *-er* to derive an agent noun in English. This is demonstrated in example (3) below.

(3) omiksi áyo'kaiksi

om-iksi a-yo'kaa-iksi

that-AN.PL DUR-sleep.AI-AN.PL

'*lit*. those sleeping ones / those sleepers' (Frantz 2017:128)

Notice that the derived noun in the example also takes a preverb marked as 'durative.' This indicates that the ones referred to by the noun are currently sleeping, but other preverbs could also be used such as *yáak-*, which would indicate that the referents will be sleeping, or *máát-* to indicate that the referents are not sleeping.

More complex derivations take nominalizing suffixes that indicate the type of noun being derived. These can indicate that the verb is being nominalized into an abstract idea such as *paaskaan*, 'the dancing,' which is derived from the VAI stem *(i)paaskaa* 'dance' using the suffix

[18] Kinship terms and body-part terms are different from each other in the way that they function. Kinship terms must always specify a person possessor while body-part terms can have an unspecified possessor. One can say 'someone's (unspecified person's) arm,' but one cannot say 'someone's elder.' I discuss my decision for modelling this difference in section 2.3.

-n. The suffix *-hsin* can also be used to derive abstract nouns from verbs. They can also derive a noun that is associated with a verb such as in *sinaakia'tsis* 'book', which is derived from *sinaaki* 'read' and the nominalizing suffix *-a'tsis*. What Frantz refers to as conjunctive nominals are nominalizations that are formed by fulfilling a role that is similar to the conjunctive verb mode but is inflected as the indicative. This is explained through example (4), which shows some different nominal patterns.

(4) a. o'kááni
yo'kaa-n-yi
sleep.AI-NOM-IN.SG
'sleep (n.)' (Frantz 2017:130)

b. kitsísttókimaa'tsinnooni
kit-isttokimaa-a'tsis-innoon-yi
2-drum.AI-INSTR-21-IN.SG
'our (incl.) drum (lit. what one drums with)' (Frantz 2017:132)

c. iitáóoyo'pi
it-á-ooyi-o'p-yi
there-DUR-eat.AI-21.CN-IN.SG
'where one eats/restaurant' (Frantz 2017:133)

Examples (4a) to (4c) show what Frantz refers to as abstract nominalization, associated instrument nominalization, and a locational conjunctive nominalization respectively. In (4a), we can see that suffix *-n* is added to the verb stem *yo'kaa*, which indicates that it is a derived noun. The suffix *-yi* is added indicating it is now animate and singular. The result is a noun expressing the idea of 'sleep,' which can be compared with example (3), which shows a simple derivation resulting in a concrete noun 'the ones who are sleeping,' or 'the sleepers.' In example (4b), the instrument 'drum' is derived from the verb 'drum' using the suffix *-a'tsis*. Notice that this noun now takes nominal inflection morphology expressing possessor and, like (4a), that it takes the inanimate singular suffix.

(4c) provides an example of a conjunctive nominalization, which can be used to derive nouns in a variety of ways, in this case, to create a location term. The location term is derived by using the locational preverb *it-* 'there,' which is normally used to link an inflected verb to a noun within the same sentence, making that verb locative. The verb is inflected with the first-person inclusive[19] plural suffix and the inanimate nominal suffix, coming together to mean 'where we eat,' or 'where one eats'[20] using the inclusive first person plural suffix *-o'p*. This derived form is treated as a noun from then on. There are several other patterns of conjunctive nominals that can express other types of meanings. A comprehensive breakdown of the various linking forms are described in Frantz (2017:133-140). Now that I have established the basic aspects of Blackfoot nominal morphology, I describe some the similarities and differences between Plains Cree and Blackfoot nominal morphology.

2.2 Comparison of Blackfoot and Plains Cree noun morphology

The development process and the techniques used to design the Blackfoot model were previously utilized to develop a model of Plains Cree morphology (Harrigan et al. 2017; Snoek et al. 2014; Arppe et al. 2014), and another Algonquian language, Odawa (Bowers et al. 2017). Due to similarities in their morphologies, the Plains Cree and Blackfoot noun models ended up being very similar to each other. So similar in fact, that the Blackfoot model was developed directly from the Plains Cree model, with a few structural differences. In this subsection, I discuss those similarities and differences.

The main morphological differences between the languages respecting nouns, include a lack of diminutive and locative suffixes which Plains Cree marks on nouns, while Blackfoot does not. The other structural difference that needed to be accounted for was the animacy specification typology. Plains Cree underspecifies the plurality of a noun when the noun is obviative, causing ambiguity as to whether a noun is plural or singular when it is obviative. Blackfoot underspecifies obviation causing ambiguity as to whether the noun is plural and

[19] In many Algonquian languages including Cree and Blackfoot, there is a difference between first person exclusive and first-person inclusive plural person agreement. Inclusive includes the person being addressed, and exclusive does not. This allows people to easily distinguish a phrase like 'we are going to the movies' as either meaning 'you and I are going to the movies' and 'me and some other people are going to the movies.'

[20] The first person inclusive suffix can be used as what Frantz refers to as an 'unspecified subject'. (Frantz 2017:57)

obviative, or plural and proximate, (Bliss and Oxford 2017). Examples (5a) and (5b) demonstrate this pattern of obviative specification by providing examples of Blackfoot sentences containing fourth-person plural and singular nouns. Example (5c) demonstrates the obviation specification pattern of Plains Cree.

(5) a. Isspómmoyiiwa aakííkoana póósi. (3>4)
isspommo-yii-wa aakiikoan-wa póós-yi
help.TA-DIR-3SG girl.NA-PROX cat.NA-OBV
'The girl$_3$ helped the cat$_4$.' (Frantz 2017:58)

b. Anna Anna iikáístsimmiiwa omiksi aakííks. (3>4s/4p)
ann-wa Anna iik-á-istsimm-yii-wa om-iksi aakii-iksi
DEM-PROX Anna very-DUR-respect.TA-DIR-3SG DEM-AN.PL woman-AN.PL
'Anna$_3$ respects those women$_4$.' (Bliss and Oxford 2017:7)

c. wāpamēw nāpēwa. (3:4/4p)
wāpam-ē-w nāpēw-a
see.TA-DIR.3SG man.NA-OBV
'He/she$_3$ sees the man/men$_4$.' (Okimāsis 2018:147)

Notice that in the Cree example, the singular and plural obviatives are ambiguous as they use the same suffix *-a*, while in the Blackfoot example, the ambiguity lies in whether the plural suffix *-iksi* is signifying a proximate or obviative noun. In both cases, the ambiguity is handled by specifying the number and obviative properties of the arguments in the verb agreement, so the nouns are not actually ambiguous in context, but on their own, they are ambiguous.[21] Tables 6 and 7 below follow Bliss and Oxford (2017:4) in summarizing the difference between the two specification patterns demonstrated in example (5).

[21] This is important because a computational model of Blackfoot noun morphology as described in this thesis would not be able to use context to disambiguate the proximity of a noun.

Table 6: Summary of Plains Cree obviation patterns

	Animate	
	PROX	OBV
SG	*-Ø*	*-a*
PL	*-ak*	

Table 7: Summary of Blackfoot obviation patterns (Bliss and Oxford 2017:6)

	Animate	
	PROX	OBV
SG	*-wa*	*-yi*
PL	*-iksi*	

The pattern of unique plural and obviative suffixes in tables 6 and 7 show how each language specifies plural and obviative differently. Table 6 shows how Plains Cree singular animate, proximate nouns have a null suffix and plural proximate nouns are marked with the suffix *-ak*. It also marks obviative on both singular and plural nouns using the same suffix, creating a three-way morphological distinction, while there is a four-way distinction in the grammar. Blackfoot also has a three-way morphological distinction, but with a different pattern of specification. Proximate and obviative singular are marked differently, but proximate and obviative plural are marked the same.

These were the most significant structural differences that needed to be accounted for between the languages for the purpose of developing a computational model of Blackfoot. Table 8 provides a summary of the morphological structures of Blackfoot and Plains Cree morphology.

Table 8: Summary of nominal morphological features in Blackfoot and Plains Cree

Features	**Blackfoot**	**Plains Cree**
Possessive person prefix	✓	✓
Descriptive prenouns (recursive)	✓	✓
Noun stem (animate, inanimate, dependent)	✓	✓
Possessive person suffix	✓	✓
Diminutive suffix	✗	✓
Plural, animacy, obviation, referring	✓	✓ (Except non-referring)
Locative suffix	✗	✓

The morphological features are shown in order from top to bottom, with the features at the top appearing at the left edge of the nouns, and the features at the bottom appearing at the right edge of the nouns. The grammatical features listed here are not non-existent in Blackfoot but are expressed differently than they are in Cree. For example, locatives are expressed as prefixes morphemes on Blackfoot verbs, and diminutives only occur on a few noun stems as unproductive archaisms, rather than productive morphemes as they are in Cree (Berman 2006:278). The consequence of these differences is that, for modelling purposes, Blackfoot noun morphology can be considered a subset of Plains Cree noun morphology. There are just a couple of significant structural differences, those being the differences in obviation specification, and the consequence of Cree not using the non-referring suffix.

Both Blackfoot and Cree are capable of nominalizing verb stems morphologically. Plains Cree is remarkably similar in the ways that it derives different types of nouns from verbs using morphology, just with a few differences. One difference is that Plains Cree uses a prefix *o-*, to mark nominalized agent nouns. For example, *mîcisow* is the animate intransitive verb root for 'eat.' *Omîcisow* is the derived agent noun for 'one who eats.' This is useful in Cree because the animate plural suffix *-ak* is the same as the animate plural verb agreement suffix, so the prefix distinguishes between them, while this is not the case in Blackfoot. To create an abstract noun, one can add the suffix *-win* to get *mîcisowin,* 'the eating, the meal, etc.' To form locational nouns from verbs, the suffix *–(i)kamikw* can be added to the verb stem, so *mîcisowikamik*, would mean 'the place where one eats,' or 'restaurant.' *-n* can be used to derive an instrument or product noun from the verb stem (the exact same suffix used in Blackfoot). For example, *kistikân,* 'grain' is derived from *kistik*, the verb stem for 'planting.' This shows that Plains Cree can derive nominals from verb stems in the same ways that Blackfoot can, just using slightly different morphology.[22]

[22] Example words taken from https://dictionary.plainscree.atlas-ling.ca/#/help. This is a site that is part of the larger Algonquian dictionaries project and includes links between words and their derivational affixes. Further explanations were verified from Okimāsis (2018:234-249).

2.3 Structure of the Blackfoot noun FST model

As stated in chapter 1, the goal of this project is to create Finite State Transducers for the Blackfoot language. Finite State Transducers are a Natural Language Processing tool that can process input to produce output. In this case, this is done by defining the morphological and phonological rules, and acceptable lexica of the Blackfoot language, as discussed in section 1.4. The goal of this section is to provide a description of the structure of the noun morphology component of the FST. This section also discusses the role of the Plains Cree model in the development of the Blackfoot model. A description of the phonological component can be found in Schmirler, Arppe, and Genee (Forthcoming), and the further developments that I implemented for this project are described in section 4.4.

The lexica for the model were derived from the Blackfoot Online Dictionary Database (https://dictionary.blackfoot.atlas-ling.ca/#!/help). This database contains all the lexical entries from the *Blackfoot Dictionary of Stems, Roots and Affixes* (Frantz and Russell 2017), as well as some additional entries that have been created from ongoing documentary research. The lexical items can be searched or browsed based on their semantic or grammatical categories. Browsing for grammatical category made it easy to gather the lexical items from the online dictionary and input them into the LEXC model. The only parts that needed to be accounted for after that were the inflectional person, animacy, obviation, number, specificity, and plural suffixes (all of which are described in the grammar).

This section is divided into two subsections each describing different morphological "paths". Henceforth, I use the term "paths" to describe different inflectional paradigms contained within a single morphological model. A diagram of the path system appears in *Figure 13*.

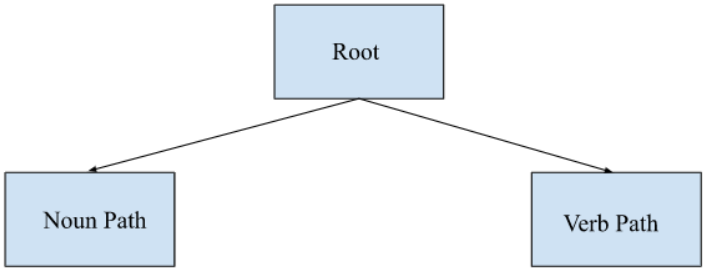

Figure 13: Root lexicon splitting into inflectional paths.

This diagram represents the overall structure of the whole morphological model. The root lexicon splits the model into two general morphological paths,[23] one for nominal morphology, and the other for verbal morphology. The result of each path is a word that functions syntactically as either a verb or a noun, depending on which inflectional path it takes. In the nominal path, noun stems can be inflected regularly, while verb stems are inflected by first forming a derivation structure that changes the part-of-speech class and inflects the verb as a regular noun. Section 2.3.1 describes the structure of the noun model as it can be used to inflect noun stems. Section 2.3.2 describes the structure of the noun model as it can be used to derive nouns from verb stems.

2.3.1 The noun stem morphological path

I began by setting the path in the root lexicon. This lexicon contains the multicharacter symbols used throughout the entire model, and then allows the noun model to split into two paths, one path for regular nominal morphology and one for deriving nouns from verb stems. I discuss each path separately, starting in this section with the noun morphology path.

I set the first lexica in the noun path as two separate person prefix lexica, one for independent nouns and one for dependent nouns. Flags (shown and described in greater detail in section 1.4 of chapter 1) are set in the person prefix lexicon, which tell the program to restrict which person suffixes they can combine with later in the model. The lexicon of shortened person prefixes skips the prenoun lexicon and goes directly to the dependent noun stems. The reason for this choice is that it is not well known how or whether dependent nouns can take descriptive prenouns.[24] The other difference between this lexicon and the regular noun prefix lexicon is that there is no null person prefix because these nouns are obligatorily possessed and cannot appear otherwise.

[23] I use the term paths to refer to ways of building inflectional paradigms. One path inflects stems using the nominal paradigm, and the other inflects stems using the verbal paradigm.
[24] Thank you to Francis First Charger for helping to clarify this during the thesis defence. According to First Charger, it is possible to say *nitsi'nakitana* 'my small/young daughter.' For now, I am leaving the model as it is because I believe evidence from a variety of Blackfoot speakers will be needed to work out the specifics of how this works.

After the independent noun person prefixes, I set a prenoun lexicon which contains all the Blackfoot prenouns. The next lexicon I defined was a gateway lexicon[25] which loops back to the prenoun lexicon allowing the prenouns to combine freely. The loop exits when the model selects for a null prenoun marked as 0, which exits the loop and collects the independent noun stems at the independent noun stem lexicon. The prenoun lexicon has an additional feature that uses flags to count the number of prefixes on a given word. This counter sets a numbered flag each time the recursive model loops through the lexicon and adds a new prenoun. Once the counter reaches 6, the counter forces the model to exit recursion and move on. This feature helps to streamline the model computationally so that it runs more smoothly and was informed by the Blackfoot corpus described in section 4.1. Otherwise, the model could theoretically loop through the prefix lexicon infinitely, which produces words forms which would not only never be produced by a native speaker but also slow the model down quite a lot. I found that 6 was the highest number of prefixes that words in the corpus could have while keeping the model running.

The next defined lexicon is the noun stem lexicon, which I split into two lexica, one for animate and one for inanimate nouns. I set these to go into their respective continuation lexica, which are used to set flags which are used to restrict the set of animacy-marking suffixes that they can take later. I decided not to set different continuation lexica depending on whether nouns take the possessive theme suffix *-im.* I found that, although there are many examples of words which must take the suffix when possessed, there are many examples where there is wide variation as to whether they can take the suffix. On the Blackfoot Online Resource site, the word *sinaakia'tsisi* has two different possessed forms documented in different places. One is a full paradigm where the word takes *-im* when possessed (for example, *nisínaakia'tsiimi*, 'my book'),[26] while in the example forms linked to the dictionary entry entry, it does not (nisínaakia'tsiistsi 'my books').[27] For this reason, I left it optional, as it is known to be obligatory for only a small handful of stems.

I used R flags in the person suffix lexicon to restrict which person prefixes they can combine with. For example, the prefix *kit-* can only combine with *-oaawa* (second-person

[25] I am using this term to describe lexica which do not contain lexical items in their surface forms, but which are used to elegantly define the next lexica and to assign tags and flags.
[26] https://blackfoot.algonquianlanguages.ca/grammar/nouns/noun-inflection/possession/paradigms/sinaakiatsis-book/
[27] https://dictionary.blackfoot.atlas-ling.ca/#!/results

plural), *-innoon* (inclusive first-person plural), or nothing (second-person singular). The dependent nouns take the same set of person suffixes as the independent nouns do and from this point on, I converged the paths for the dependent nouns and independent nouns to define the next defined lexicon as the noun suffixes. I split this into two lexica, one for animate, and one for inanimate nouns. These lexica provide us with the singular, plural, obviative and non-referring suffixes. At this point, the program has finished generating an inflected Blackfoot noun.

Figure 14 provides a visual representation of the structure of the nominal morphology path for noun stems in the Blackfoot model.

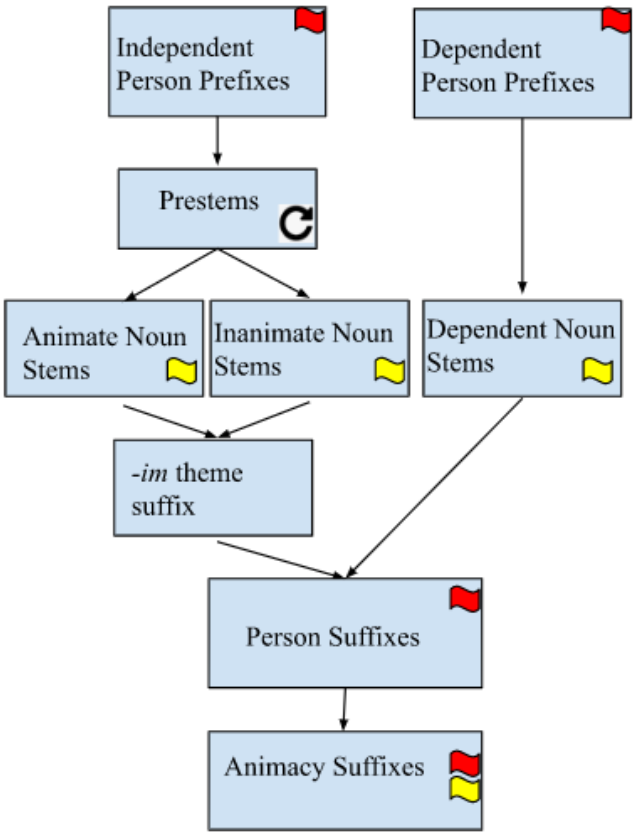

Figure 14: Structure of the Blackfoot noun model: noun path

Each box in the diagram represents a lexicon. For the sake of simplicity, the diagram does not include the multicharacter symbol set which is used primarily used to set flags, only the lexica containing stems and affixes. Arrows are used to show which lexicon is next in the concatenation process. Flags appear on the right of each lexicon box. Flags are colour coded, so flags with the

same colour indicate a long-distance restriction. For example, the red flags represent person restrictions so that first person possessive prefixes are restricted to combine with first person possessive suffixes, and so on. The "replay" symbol on the right of the *prestems* box represents recursion. They are labelled *prestems* because they are shared with verbs, so rather than calling them prenouns or preverbs, I use the term *prestems* to make it clear that it is essentially the same lexicon.

There are a few things relating to my design choices that are made explicit by the diagram and need to be addressed. One is that the dependent person prefixes go straight to the dependent noun stems, skipping the preverb lexicon. As previously discussed, this is because there is not a lot of direct evidence indicating whether relational noun stems can combine with prenouns, so the model is currently designed to skip straight from the dependent person prefix lexicon and go straight to the noun stem lexicon. This path can only take nouns from the dependent noun lexicon, and cannot take, for example, stems from any of the verb lexica because there are no dependent nominalized verb stems. The second thing to note is that there is a red flag on the animate suffixes, indicating a long-distance restriction connected to the person affixes. This relationship restricts the model so that third person possessive affixes do not co-occur with the proximate noun marker *-wa*, only obviative ones, following the obviation rule (Frantz 2017:14-15).

2.3.2 The nominalized verb stem path

The nominal path of the verb model is very similar to the general nominal path seen in section 2.3.1. The largest differences are that nominalized verbs can take several suffixes that provide additional semantic information about the noun being derived from the verb (see section 3.1 for a detailed description of these prefixes), and that nominalized verbs can take additional prefixes. In this section, I provide a descriptive overview of the verb stem nominalization path of the model.

Just like in the regular nominal path, in the nominalization path I first set the person prefixes. Unlike the person prefixes in the verbal path, these express possessive persons, and represent a separate person prefix lexicon from the person agreement prefixes in the verb path (described in section 3.3.1). Flags are set here which restrict the person prefixes to their corresponding suffixes. Then, unlike in the regular nominal path, I set the next lexicon for the

derived nouns to be the prefix set containing the negation, tense, and aspect prefixes. This is because, as explained in section 2.1, nominalized verb stems can take these prefixes, while noun stems in the regular nominal morphological paradigm cannot. Next, I set the stem lexicon, which is shared with the verb path, and is organized into the four lexica representing the animate intransitive, inanimate intransitive, transitive inanimate, and transitive animate verb stems. These four lexica and what they represent are irrelevant to the nominal morphology, so I do not discuss it in detail here, but simply acknowledge that this is how the verb stem lexica are organized.

I then set a suffix lexicon, where the nominalizing suffixes are concatenated to the verb. The nouns are marked with one of several suffixes, a null one for agent nominalization, *-n* and *-i2hsin* for abstract nominalizations and *-o'p* for conjunctive nominalizations. These suffixes define a definitive split from the verbal morphology path, taking the same set of nominal suffixes found in the noun model, where I set the lexicon for the possessive person suffixes that agree with the possessive person prefixes, and then the animacy, number, obviative and non-referring lexicon. At this point, the morphological process is terminated, ending the word form generation.

This entire nominalization path is shown in Figure 15.

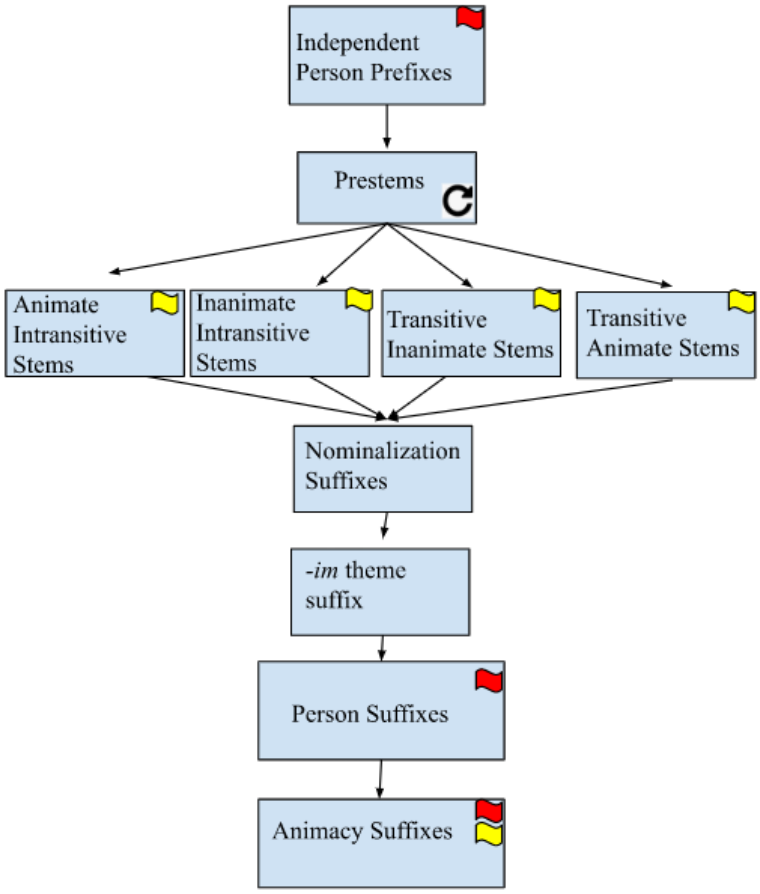

Figure 15: Structure of the nominalization path.

The red flags in the diagram signify the flags which mark long distance relationships between person agreement.

Derivational morphology such as nominalization was initially not modelled in the Plains Cree morphological model, at least not in the same way. Arppe et al. (2019) describes the implementation of a derivational model of Plains Cree, which included the modelling of nominalization morphology, and followed the design of the first nominal model which only modelled inflection. The result was two layers of analysis, one inflectional, and the other derivational. The paper describes how the derivational parser breaks up Cree stems into much smaller components. The parser was designed to create a detailed corpus that broke every word into its constituent parts. Implementing this feature into the Plains Cree model was not necessary to capture a variety of nouns as it was in Blackfoot for two reasons. One reason is because most derived stems were already available in the computational lexica for the Cree model. Having a

smaller number of Blackfoot noun stems meant that the number of nouns parsed by the model could be greatly increased by modelling nominalization morphology. The second reason is that nominalization morphology does not seem to be as productive in Plains Cree as it is in Blackfoot (Arppe et al. 2019:4-7). These points are evidenced by the corpus tests described in section 4.3, which showed that a significant number of words parsed by the model were nominalized verbs which were not already represented in the lexicon.

This work suggests that transferring models of Algonquian morphologies[28] can be an efficient way to develop novel technologies for different Algonquian languages. I was effectively able to save a lot of time by transferring the basic structure of the Plains Cree model, so that more time and effort could be invested in dealing with the specific aspects of Blackfoot grammar that differ from Plains Cree and in enhancing the model. The fast development of the project was its greatest strength since a lot of development could be done very quickly using available resources and when it comes to developing resources for minority languages, time is valuable. It has also helped to show the important similarities and differences between the two languages, revealing just how similar they are in terms of their morphology. Further work in developing Algonquian language technologies in this way should be able to make use of these methods.

[28] I note the Odawa model created by (Bowers et al. 2017). I do not compare the Blackfoot model with the Odawa model in this thesis despite many similarities between them. A direct comparison between the Plains Cree, Odawa, and Blackfoot models would make for the subject of an interesting paper in the future.

3. Modelling Blackfoot verb morphology

In this chapter, I discuss the structure of the Blackfoot verbal morphological path in the computational model. It is structured similarly to chapter 2. In section 3.1, I provide a basic description of Blackfoot verbal morphology. Then in section 3.2, I compare this with Plains Cree verbal morphology. In section 3.3, I describe how this I implemented this into the computational model, and how the model compares with the Plains Cree model.

3.1 A summary of Blackfoot verb morphology

This section explains Blackfoot verbal morphology, and it is split into two sections. Section 3.1.1 describes the regular verbal morphology as it is applied to Blackfoot verb stems, and section 3.1.2 explains how nouns undergo verbalization.

3.1.1 Verbal morphology of Blackfoot verb stems

Blackfoot verbs have four subclasses which specify combinations of their valency and the animacy of their arguments. There are transitive animate (VTA), intransitive animate (VAI), transitive inanimate (VTI) and intransitive inanimate verbs (VII). Table 9 illustrates the relationship between animacy, valency, and agreement with examples.

Table 9: Inflectional verb subclasses (Retrieved from https://dictionary.blackfoot.atlas-ling.ca/#!/help)

Features	**Animate**	**Inanimate**
Intransitive	Nitssskaa. nit-sskaa 1-break.AI 'I went bankrupt.'	Isskaawa.[29] sskaa-wa break.II-3/0SG 'It broke.'
Transitive	Nitsssinnoka. nit-ssinn-ok-wa 1-break.TA-INV-3/0SG 'He/she bankrupted me.'	Isskima. sski-m-wa break.TI-TH-3/0SG 'He/she broke it (inanimate object).'

[29] Note that there is an /i/ that appears on *isskaawa* and *isskima* when they do not have prefixes. This /i/ surfaces in some speakers when there is no prefix on the stem. I explain more about this in section 4.4.

Intransitive verbs can normally only take one argument that is, a subject (except in certain cases).[30] If it is a VII, that subject must be inanimate, and if the verb is VAI, the subject must be animate. Transitive verbs can be either transitive or ditransitive. In either case, they must take an animate noun as a subject, except in special grammatical circumstances.

Verb agreement in Blackfoot allows up to two arguments to be indicated on the verb. The subject is always marked, and in transitive verbs, the object is also indicated. On ditransitive verbs, the conjugation agrees with the indirect object. Blackfoot relies on a system of obviation to assign roles to participants (see section 2.1). This section goes into more detail about obviation as it applies to verbal morphology.

In an example where there are two different persons, for example, a third person and a first person, they are clearly marked on the verb. Example (6) illustrates how this appears in the morphology. The arguments are bolded.

(6) a. Nitsikákomimmawa nitána

nit-ikákomimm-a-**wa** **n-itan-wa**

1-love.TA-DIR-**3/0SG** **1-daughter.NA-AN.SG**

'I love my daughter.' (Frantz 2017:56)

b. Nitsikákomimmoka nitána

nit-ikákomimm-ok-**wa** **n-itan-wa**

1-love.TA-INV-**3/0SG** **1-daughter.NA-AN.SG**

'My daughter loves me.' (Frantz 2017:61)

In the example, both sentences have the same participants, a first person, and a third person (their daughter). In both sentences, the prefix *nit-* indicates a first person participant, and the prefix *-wa* indicates a third person participant. The suffix *-a* is a direct suffix found in example (6a), which indicates that the first person is the subject, and the third person is the object. The inverse suffix *-ok* in example (6b) indicates that the third person is the subject, and the first person is the object.

[30] Certain animate, intransitive verbs may take an object optionally. These are marked in the Blackfoot Online Dictionary as VAI+O. It is important to note that this differs from the VTA and VTI verbs, for which an object is obligatory.

This exemplifies how Blackfoot person agreement works when the arguments are distinct, but what if the arguments are both third person? Example (7) shows how to inflect VTA verbs with two third person arguments, and the arguments are bolded.

(7) a. Ikákommimmii**wa nohkówa kitáni.**
ikákomimm-yii-**wa** **n-ohko-wa** **k-itan-yi**
love.TA-DIR-**3/0SG** **1-son-AN.SG** **1-daughter.NA-OBV**
'My son loves your daughter.' (Frantz 2017:58)

b. **Ot**sikákommimmok**a nohkówa otáni.**
ot-ikakomimm-ok-**wa** **n-ohko-wa** **w-itan-yi**
3/4-love.TA-INV-**3/0SG** **1-son-AN.SG** **3-daughter.NA-OBV**
'Her daughter loves my son.' (Frantz 2017:62)

Example (7a) shows how verbs with two third person arguments are inflected. The subject remains third person, the object undergoes obviation, demoting it to fourth person, and the verb receives a direct third person suffix *-ii* and a fourth person agreement suffix, *-wa*. If the subject and object are explicitly stated, as in the sentence in (7a), the third person object is takes the obviative suffix, as I discuss in section 2.1. In example (7b), the subject *otani* is already fourth person because it is possessed by a third person. In this case, the fourth person can become the subject, and the fourth person the object by using the inverse suffix *-ok*, alongside the third person prefix *ot-*.

Another thing that needed to be accounted for was enclitic pronouns (Frantz 2017:51-55).[31] Blackfoot has unattached pronouns, but they are usually omitted and attached pronouns surface much more frequently. They appear as enclitics on the ends of verbs in two main environments. Examples of both these environments appear below in example (8), with example (8b) and (8c) showing cases where the third person argument is explicitly expressed and appears before the verb, and example (8a) showing a case where the third person argument appears as a pronoun.

[31] Frantz also notes that they are called enclitic pronouns. I will follow Frantz in calling them attached pronouns.

(8) a. Nitohpómmatoo'pi**aawa**.
nit-ohpommatoo-'p-yi=**aawa**
1-buy.TI-TH-3/0PL=**PRO**
'I bought them.' (Frantz 2017:51)

b. Saahkómaapiiksi áwaawahkaay**aawa**.
saahkomaapi-iksi á-waawahkaa-yi=**aawa**
boy-AN.PL DUR-play.AI-3/0PL=**PRO**
'Some boys are playing.' (Frantz 2017:51)

c. Nitákkawa itápsskonakiw**aiksi**.
n-itakka-wa itap-sskonaki-wa=**aiksi**
1-friend-AN.SG toward-shoot.AI-3/0SG=**PRO.PL**
'My friend shot at them (animate).' (Frantz 2017:53)

Through this example, we can see that an argument that is not overtly expressed can appear as an attached pronoun. When the syntactic construction is transitive, the attached pronoun must represent at least one of the arguments if the argument is not explicitly expressed, as seen in (8a). In (8b), the attached pronoun appears because the argument *saahkomapiiksi* does not directly follow the verb. Attached pronouns are also used when expressing Goal, Instrument, or Location, as seen in (8c).

Blackfoot verbs are also inflected for mode. The modes are called independent, subjunctive, conjunctive, irrealis,[32] and imperative. These are specified fusionally, through different sets of person prefixes and suffixes. Here, I provide a brief explanation of how each mode is used in Blackfoot clause construction. The independent mode is the most basic mode for verbs. It is primarily used to create simple, indicative phrases. Any verb that is not part of a conjunct phrase, is not being used to express conditionality, or is not being used as a command is independent.

[32]Frantz refers to the irrealis mode as the "unreal mode" (Frantz 2017: 126-127).

The conjunctive mode is used in subordinate clauses (adverbial clauses and complement clauses). Example (9) demonstrates these functions. The example demonstrates two types of subordinate clauses. Example (9a) demonstrates a purpose clause, and example (9b) demonstrates a subject complement clause.

(9) a. Nomohtó'too kááhksspommookssoaayi
n-omoht-o'too k-ááhk-sspommo-ok-i-hs-oaa-yi
1-source-arrive.AI 2-might-help.TA-INV-1-CNJ-2PL-CNJ
'I came for you (plural) to help me.' (Frantz 2017:122)

b. Iíksoka'piiwa otáísootaahsi
iik-soka'pii-wa ot-á-sootaa-hs-yi
very-good.II-3/0SG 3-DUR-rain.II-CNJ-CNJ
'It is good that it is raining.' (Frantz 2017:122)

The irrealis mode is used when speakers want to express something that they are unsure of. It can be used in hypothetical subordinate clauses or in counterfactual statements. Example (10) demonstrates both usages. Example (10a) shows how the irrealis is used in hypothetical subordinate clauses, and example (10b) shows how the irrealis is used in counterfactual statements.

(10) a. Nitsítssáyoyihtopi, nitáaksoyi ánnohka.
nit-it-say-Ioyi-htopi nit-aak-Ioyi annohka
1-then-NEG-eat.AI-IRR 1-FUT-eat.AI now
'If I hadn't eaten then, I'd eat now.' (Frantz 2017:127)

b. Nikkáminaanatao'topi!
n-ikkam-inaanat-a-o'topi
1-if-own.TA-DIR-IRR
'How I should like to own him!' (Frantz 2017:127)

Besides person and mode, verbs can be marked for negation, tense, degree, modals, and descriptors. These appear as prefixes or preverbs which surface between the person prefix and the verb stem. Common modals such as *ohkott-* 'can/able to' or *sstsina-* 'need to' appear as morphemes in this position too. Locatives such as *sap-* 'in,' *otats-* 'on,' or *itap-* 'towards' and comitatives such as *ohpok-* 'with,' also appear as preverbs, as do subjunctive and conjunctive particles like *ikkam-* 'if,' and *piitsyoohk-* 'when/as soon as.'

These preverbs provide an interesting challenge to the design of a Blackfoot computational model. The order in which they appear is not well described and has never been closely investigated. This would not be too much of a challenge if, like prenouns, there was only one type of preverb (ie. descriptive or modal), but as I have just described, they can convey many different types of grammatical and lexical information. To deal with this, I have identified the most common prefixes, and the more or less fixed orders in which they appear. The most common verb prefixes include the negation, tense, and aspect prefixes and I designed the model to allow them to appear in this order. I discuss how I implemented this in section 3.3.

Blackfoot verbs also make use of valency-changing derivational verb suffixes. They have the ability to increase the valency of a verb, such as in the benefactive[33] or decrease the valency of a verb, for example with the reflexive. Increasing valency in Blackfoot requires a change of the verb class, thus changing which person agreement prefixes and suffixes it takes. Example (11) shows how these suffixes work.

(11) a. Iihpómmoyiiwáyi ónnikii.

ii-ohpomm-o-yii-wa=ayi	onnikis-i
PST-buy.AI-BNF.TA-TH.DIR-3/0SG=PRO	milk-NONREF

'He bought some milk for her.' (Frantz 2017:113)

[33] The benefactive is used when an action was done for someone else's benefit. For example, in Blackfoot, *bake* is a VAI, but we can use the benefactive to say, 'I baked them for my daughter,' and the verb *bake* changes its subclass to VTA to agree with the indirect object (Frantz 2017:114).

b. Omiksi ponokáómitaiksi áísiksipotsiiyiyaawa.

om-iksi ponokaomitaa-iksi a-siksip-otsiiyi-yi=aawa

DEM-AN.PL horse-AN.PL DUR-bite.TA-RECIPR.AI-3/0PL=PRO

'Those horses are biting each other.' (Frantz 2017:117)

Example (11a) shows the use of the benefactive which turns the verb *ohpomm* from a VAI to a VTA verb. As a consequence, the verb is inflected for person using the VTA morphological paradigm, with the third-person singular being the subject, and the fourth-person singular being the indirect object. Example (11b) shows the use of the reciprocal. Here, the VTA *siksip* 'bite', becomes VAI, thus taking the third-person plural suffix *-yi.*

Blackfoot verbs can also be inflected with interrogative suffixes. These suffixes inflect verbs in the independent mode and can be used to form yes/no questions and open-ended questions (Frantz 2017:146-155). Example (12) shows how these suffixes can be used to form questions.

(12) a. Kikátai'nóókaiksaawa?

kit-Ikata'-ino-ok-waiksaawa

2-INTERR-see.TA-INV-3PL.INTERR

'Did they see you?' (Frantz 2017:147)

b. Kitsikákomimmokihpa?

kit-ikakomimm-ok-i-hpa

2-love.TA-INV-1-INTERR

'Do you love me?' (Frantz 2017:146)

c. Tsá anistápííwaatsiksi

tsa anist-apii-waatsiksi

what manner-be.II-SG.INTERR

'What is it (inanimate)?' (Frantz 2017:151)

Example (12a) shows the use of two different interrogative affixes. One is the interrogative prefix *ikata'*- which can be found among the preverbs. The second is the independent person suffix -*waiksaawa* which marks a third-person plural subject. The inverse theme suffix -*ok* is still used to indicate that the third person is acting on the first-person. Example (12b) gives an example of the interrogative suffix -*hpa* which is used when there is no third person-indicated on the verb. In this case, it is second person acting on first person, so again, we see the use of the inverse theme suffix -*ok.* Example (12c) shows the use of the interrogative in a simple, open-ended question, with the third person singular interrogative indicating a singular subject on the verb.

3.1.2 Verbalization morphology of Blackfoot noun stems

I also modelled verbalization morphology of noun stems. Blackfoot noun stems can change their grammatical category, and consequently, can take verbal morphology. The testing done in section 4.3 shows that this is a common phenomenon like nominalization of verb stems (see Table 16). An example of this happening in English is in the word 'vacuum,' which can be used as a noun, as shown in the sentence, 'There are many vacuums at the store,' or as a verb, as in 'He is vacuuming the house.'

In Blackfoot, verbs can be derived from nouns by utilizing verbalization morphology which, in most cases, involves simply inflecting the noun with the same prefix and suffix sets as are used to inflect intransitive verbs. I cover the morphological verbalization process as described in Franz (2017:26). Frantz gives the example in (13) to show how Blackfoot nouns can be made into animate intransitive verbs.

(13) (Frantz 2017:26)
Nítaakiiyihpinnaana
nit-aakíí-yi-hpinnaana
1-woman.NA-AI-1PL
'We are women.'

Example (13) shows a basic person inflection of a noun that has been converted to a VAI. The animate noun stem *aakii* 'woman' gets a verbalizing suffix *-yi*, and then simply adds the person prefix *nit-* and first person plural suffix *-ihpinnaan* which mark that the verb agrees with a first person plural subject. Notice that these are similar to the possessive person affixes marking a first person plural possessor but are distinct enough to recognize that the word is now an inflected verb, not an inflected noun.

3.2 Comparison of Blackfoot and Plains Cree verb morphology

The Plains Cree verb model was heavily consulted in the process of constructing the Blackfoot verb morphological model (Harrigan et al. 2017). I consulted with expert in Blackfoot grammar, Inge Genee, and the researcher who oversaw and directed the development of the Plains Cree model, Antti Arppe. We first mapped out the structure of the Cree model to see how it compared with Blackfoot verbal morphology, determining which elements of the Cree model would inform the development of the Blackfoot model, and which parts would need to be removed or restructured to accommodate Blackfoot grammar. With those meetings informing the development, I began designing the model of Blackfoot verb morphology. For aspects of Blackfoot grammar which were unique, such as the negative prefixes (Plains Cree uses free particles for negation), separate solutions were developed. Like with nominals, the structures of Blackfoot and Plains Cree verbal morphology are similar, but the differences were still more substantial than in the noun model. In this section, I describe those differences.

The most consequential structural difference was that Blackfoot verbs can be inflected with more paradigms to distinguish between more verb modes than Plains Cree. The functions of the subjunctive and conjunctive modes in Blackfoot are expressed using what many scholars refer to as the conjunct mode in Plains Cree (Okimâsis 2018:272–281). Thus, one of the most substantial differences between the models was that one more mode was required for the Blackfoot model than was required for the Cree model.

The person prefix sets required for each respective mode represented a consequential structural difference between the models. For example, in Plains Cree, person prefix sets for each mode were the same across the verb inflectional classes, but for Blackfoot, there were some differences between verb classes. This meant that for Plains Cree, a single continuation lexicon

could be used to set person pre-set flags, but for Blackfoot, the restrict flags needed to be specified for each lexical item in the person suffix lexica since the patterns varied somewhat.

The final significant differences regard the interrogative suffix set, which is not expressed morphologically in Cree, and the absence of attached pronouns in Cree. Attached pronouns represent a syntactic difference between Blackfoot and Cree, but for Blackfoot, this has morphological consequences because the pronouns attach like suffixes. Unlike Blackfoot, Cree does not require pronouns to follow verb phrases when a participant is not explicitly expressed. In Cree, there is a non-affirmative particle, but it is not expressed through a bound morpheme. This is shown in example (13a) later in the chapter.

There are additional, relatively insignificant differences between Plains Cree and Blackfoot which had little consequence for the model structures. For example, in Plains Cree, conditionals and prepositions are not bound morphemes as they are in Blackfoot. Blackfoot marks negation as a preverb, while Plains Cree marks negation with a free morpheme (Okimāsis 2018). Both Plains Cree and Blackfoot use different negatives depending on the mode. In Blackfoot, *máát-* is used only in independent verbs, *miin-/piin-* (interchangeable depending on dialect) are used in imperative verbs, and *saw-,* can appear in any verb mode. Because negatives were not part of the Plains Cree morphological model, the relationship between modes and preverbs was not accounted for in the Plains Cree verb morphological model. This is demonstrated in example (13b) and (13c).

The Blackfoot and Plains Cree verbal morphologies are structurally similar enough that the Blackfoot model could be engineered by first creating a comparison with the structure of the Plains Cree model, and deriving its structure from it, and then making the relevant changes to reflect the structures of Blackfoot grammar. Table 10 summarizes the similarities and differences of Blackfoot and Cree verb morphology.

Table 10: Summary of Blackfoot and Plains Cree verb morphology

	Blackfoot	**Plains Cree**
Person prefix	✓	✓
Preverbs (recursive) (negative, tense and aspect, modal, locative, descriptive)	✓	✓ (Except negative, linking, and mode preverbs.)
Verb stem (VAI, VII, VTI, VTA)	✓	✓
Derivational suffix	✓	✓
Person suffixes (directionality, person, number, mode)	✓	✓ (Except the conjunctive and subjunctive modes use the so-called conjunct paradigm.)
Non-affirmative suffixes	✓	✗
Attached pronouns	✓	✗

Table 10 illustrates that Plains Cree and Blackfoot make use of almost all of the same types of affixes but the Plains Cree model does not make use of the same types of functions contained within each morphological slot. Cree also lacks the non-affirmative affix used in questions and negations. Cree does not lack these grammatical features altogether but makes use of free words rather than bound morphemes, as mentioned earlier in this section. Example (14) illustrates this point.

(14) a. Okimāsis (2018:86)

kika-masinahēn cī ōta kiwīhowin?

kit-ka-masinahē-n cī ōta kit-wīhow-in?

2-could-write.AI-2 INTERR here 2-name-2

‘Can/could you write your name here?

b. Okimāsis (2018:89)

namōya, namōya ninōhtē-mētawān.

namōya, namōya nit-nōhtē-mētaw-ān.

NEG, NEG 1-want-play.AI-1

‘No, I don’t want to go play.’

c. Okimāsis (2018:91)

ēkā takohtēci wīpac, kika-ati-sipwēhtānaw.

ēkā	takohtēc-i	wīpac	kit-ka-ati-sipwēht-ānaw.
NEG.CNJ	arrive.AI-3	soon	2-FUT-begin-leave.AI-21.PL

'If he/she does not arrive soon, we will leave.'

Example (14a) shows an interrogative sentence marked by the word *cī*. In Blackfoot, this would be marked with the verbal suffix *-hpa*. Notice in example (14b) and (14c) that Plains Cree uses different negatives for different modes just as Blackfoot does, but the difference is that they appear as a prefix in Blackfoot and are free words in Plains Cree (see Schmirler and Arppe 2019 for a corpus-based account). In example (14b), the negative *namōya* is used to negate the indicative verb phrase *I do not want to play.* In example (14c), the negative *ēkā* is used to negate the conjunct verb phrase *if he/she does not arrive* and has an implied conditional meaning when combined with a verb in the conjunct mode (Okimāsis 2018:91). In Blackfoot, the negated conditional would be made using the subjunctive mode, and by using the negative preverb *saw-* alongside the conditional preverb *ikkam-*.

I do not discuss Cree verbalization morphology because it did not inform the creation of the Blackfoot verb model, and was not implemented in the Plains Cree model.

3.3 A description of the Blackfoot verb model

In this section, I describe the structure and design of the Blackfoot verbal morphology model. I also discuss the ways in which the Plains Cree model informed its development. The Blackfoot verb morphology model was designed as a Finite-State Transducer using the LEXC formalism, as was the noun model. This section describes the basic structure of the verb morphological model and some of the rationalizations for design decisions that were made. The lexical items for the verb morphology were also taken from the Blackfoot Online Dictionary (https://dictionary.blackfoot.atlas-ling.ca/) (Frantz and Genee 2016-2023). The section is divided into two subsections, with subsection 3.3.1 describing the regular, verbal inflection path, and subsection 3.3.2 describing how I modelled verbalization morphology. The overall design following two morphological paths follows the same overall structure as the noun model.

3.3.1 The regular verbal morphology path

I began with the root lexicon which is splits the verbal morphological model into two paths, one for the regular verbal morphology, and the other for the verbalized noun morphology. I then set the flags for the four verb classes in a general verb lexicon, and then set the flags for the five modes in the next lexicon, which subdivides the whole path into 20 different agreement patterns, which is significant for the person suffix lexica which appear later in the model. After that, I set the four-person verb prefixes, *nit-, kit-, ot-,* and 0 (null), with flags set to restrict which person suffixes they will combine with. Then I set the preverb lexica to appear next.

I made two preverb lexica, one containing only the tense, aspect,[34] and negation morphemes and the second contains all other preverbs. The second lexicon is entirely recursive, as was done with the prenouns in the noun model, allowing them to combine freely. Like with the prenouns, I used flags to limit the number of possible Blackfoot preverbs to a maximum of 6. This choice was also informed mainly by the corpus described in section 4.1 and my comparison with the Plains Cree model. There are almost no words produced by Blackfoot speakers which contain more than 6 preverbs. In the preverb lexicon, flags are used to restrict certain preverbs from combining totally freely. For example, the negation prefix *máát-* only occurs in independent verbs, while the negation *saw-* can occur in any mode. Flags are also set for prefixes which combine with mode suffixes later in the model.

I created a continuation lexicon, where I set flags restricting the four verb classes, with each one going to its respective lexicon containing the list of verb stems from each class. There are four verb lexica, one for each of the VAI, VII, VTI, and VTA verb classes. I then set the next lexicon as the secondary verb final lexicon.

The secondary verb finals are the derivational suffixes such as the causative and reflexive, which change the valency and therefore, the class of the verb (for example, VAI to VTA for causative and VTA to VAI for reflexive). This section is designed so that when verbs are analyzed by the model, the inflectional class of the verb stem is returned, and if the inflectional class changes due to the addition of a derivational suffix, then this is marked with a derivational suffix marker, its semantic role, and the appropriate verb class change. Table 11

[34] I am only referring to the perfect prefix *ákaa-* as aspect. See page 90 for an explanation of why I separated tense and aspect preverbs this way.

shows an example of the FST output making use of this process. Example (15) gives a linguistic analysis of the word found in the Table using a standard gloss.

Table 11: Example of a Blackfoot verb FST output

Surface Form	**Analysis**
Nitáínoohsspinnaan	Dur+ino+VTA+Der/Refl+VAI+1Pl

(15) Frantz (2017:117)
Nitáínoohsspinnaan
nit-a-Ino-ohsi-hpinnaan
1-DUR-see.TA-REFL.AI-1PL
'We see ourselves.'

When the inflectional class changes in an example such as this, the verb now also takes the suffix set from the new inflectional class. This is done by defining the next lexicon as the one which is determined by the derivational suffix. Otherwise, the next lexica are the person suffix lexica. The person suffix sets are divided into twenty separate lexica. Each one is a combination of the verb class and mode. For example, the first lexicon that appears in this list of lexica is the animate, intransitive verb suffixes for the independent mode.

The person suffixes could have been broken down further into constituent morphemes such as *-ok* for the inverse suffix, and *-hs* for the conjunctive suffix, but doing so causes two potential issues. First, it affects the readability of surface forms for the model, causing much greater levels of ambiguity when analyzing such short morphological segments. The second issue is that these morphemes can change shape, in some instances in ways which are unpredictable. Therefore, although the method of chunking them together may be more work in the short term, and may not be ideal linguistically, it is a computationally elegant and effective solution. This "chunking" approach was used for person suffixes in the Plains Cree model (Harrigan et al. 2017).

The third person interrogative suffixes are listed in the independent person suffix lexica because they were too complicated to simply suffix them after the inflectional person suffixes. So, like the other person suffixes, I enumerated them in the independent person suffix lexicon. The interrogative suffix that is used for first and second-person interrogatives *-hpa* is within its

own lexicon because its affixation is simple enough that it can be controlled by defining the next lexicon for first and second person suffixes in the independent mode as the interrogative lexicon. Otherwise, the next lexicon after the person suffix lexicon in the other modes is the enclitic lexicon where they can optionally take an attached pronoun.

I did not constrain how verbs can take attached pronouns for two reasons. First, doing so would require a massive number of flags to accounts for the large number of constraints that can be placed on attached pronouns. The second reason is that attached pronouns may appear in other environments. For example, an animate intransitive verb may be inflected for first or second person but may still take an attached pronoun if it is a VAI+O verb.[35] For those reasons, I leave the attached pronoun set unrestricted.

The diagram in Figure 16 visually represents the structure of the Blackfoot computational model of verbal morphology, as it inflects verb stems.

[35] An animate intransitive verb that can take a non-referring direct object. These are not distinguished from the VAIs in the model because they are generally inflected the same way.

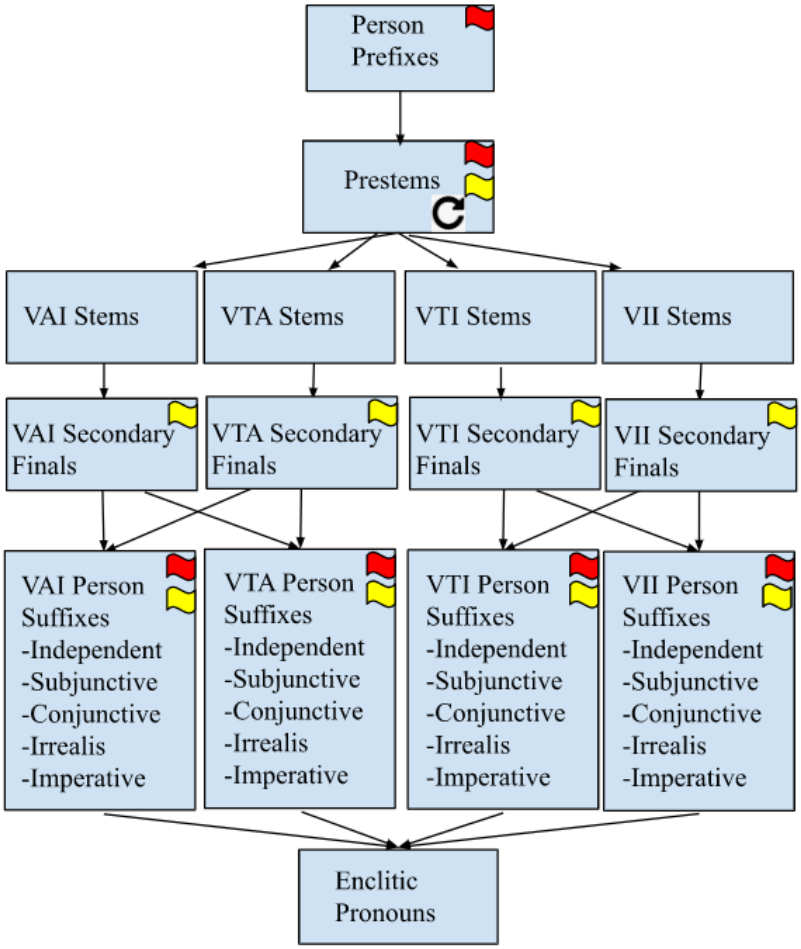

Figure 16: Structure of the Blackfoot verb model

The diagram of the verb model is organized the same as the noun model diagram, with arrows showing the lexicon order and coloured flags indicating long distance restrictions set by LEXC flags. It is important to note the following. The secondary verb suffixes change valency, but not animacy of their arguments, so the order of the suffixes is organized accordingly to indicate that the secondar suffixes can make VAIs into VTAs, and VIIs into VTIs, but cannot make VIIs into VTAs or VAIs into VTIs. The final set of boxes (the *person suffixes* boxes) do not each represent a single lexicon, but four sets of five lexica. There is one lexicon for each combination of animacy and transitivity suffix type (VAI, VII, VTI, and VTA), and mode, making for a total of twenty person-suffix lexica. The interrogative suffix set is not mentioned in here because it falls under the independent paradigm.

The other thing to note is the colour that is represented by each flag. In the case of the verb model, there are only two flag colours, red and yellow. The red flag represents person restrictions. The yellow flag represents mode restrictions. Both flags appear in the *preverbs* lexicon, restricting only a few prefixes which only appear in certain mode and person combinations.

3.3.2 Verbalization path for Blackfoot noun stems

I set the verbalization path in the root lexicon, which splits the model into a regular verbal morphology path and a verbalization path. The next lexica are the verb prefixes for the different modes, just as in the regular verb morphology path, and then the prefix lexicon, since verbalized noun stems can have negation, aspect, and tense prefixes. I then set the preverb set as the next lexicon, making this path the same as the regular verb path up to this point. It diverges here, as it reaches a continuation lexicon where the next defined lexica are either the animate or inanimate independent noun stems. The verbalization path did not include the dependent noun stems for the same reason that I did not allow the dependent noun stems to take prefixes because it is not fully documented how or whether Blackfoot dependent nouns can take other prefixes besides person prefixes.[36] I also did not allow for dependent nouns to become verbs because it is beyond the scope of this, although it may be implemented in the future if it is found to be useful.

The next step in the verbalized noun path is a continuation lexicon which subdivides the newly formed verbs into animate and inanimate intransitive verbs depending on the class of the noun stem. I defined the next lexica to be either the animate intransitive or transitive inanimate person suffixes, depending on the class of the verb, and these agree with the person prefixes, and the modes using flags. There are only two general verb classes resulting from verbalization, the animate intransitive, which are derived from animate nouns, and the inanimate intransitive, which are derived from inanimate nouns. The verb suffixes are also organized into five different mode lexica, each with a different inflectional paradigm, making for a total of 20 person suffix lexica.

[36] Unofficial investigation suggests that it is possible for dependent nouns to take prenouns at least in some varieties of Blackfoot, and even showed how this is done however, more investigation is required to provide conclusive evidence about this. Thank you to Francis First Charger for providing this information during the thesis defence.

Like in the regular verbal path, the verbalization path takes optional interrogative or attached pronouns, where I set the termination of the verbalization morphology path. Figure 17 provides a visual representation of the structure of the verb path in the Blackfoot noun model.

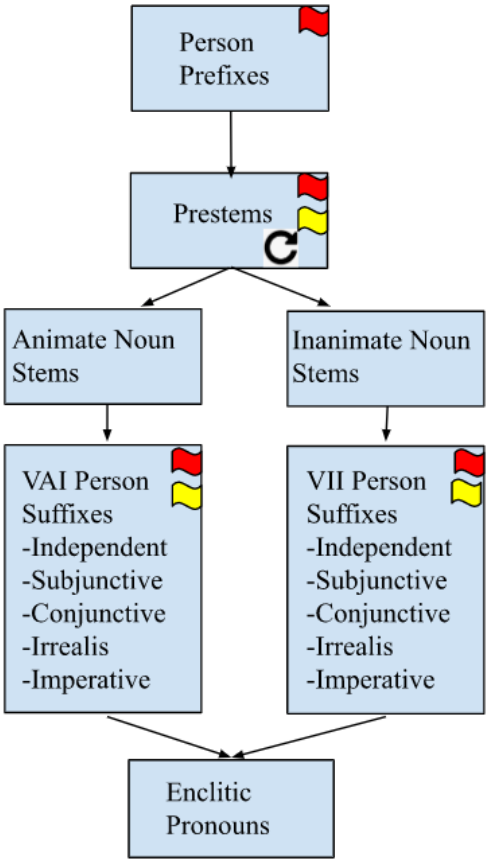

Figure 17: Structure of the verbalization path.

Comparing this diagram with Figure 16, the structures are similar, but in Figure 17, there are only two suffix lexica, and no secondary verb final lexica. I kept this part of the model simple, basing it on the derivation described in Frantz (2017:26). Future work on the model may implement more advanced morphology for verbalized nouns. The full model of Blackfoot morphology is publicly available on GitHub at https://github.com/giellalt/lang-bla/tree/main/src/fst.

4. Evaluating the Blackfoot FST Using a Corpus

The next step in developing the morphological model described in chapter two and three of this thesis is to test how well it models Blackfoot morphology. For this step, I test the model on an assembled corpus of Blackfoot text. In section 4.1, I describe the Blackfoot corpus, the variety of texts used in its assembly, and the requirements and reasons for choosing such texts. In section 4.2, I describe the methods and results of an initial test that guided model development using pairs of surface forms and underlying analyses within a set of test files. Section 4.3 explains the methods and results of a quantitative statistical analysis performed using the corpus described in section 4.1. In section 4.4, I describe the modelling design choices and some peripheral technologies that improved the model and explanations about how they improved it.

4.1 The corpus

Corpora play an integral role in developing language technology. They can be used for machine learning techniques (see section 4.4.1) or evaluating the technologies that they have designed. One of the challenges for developing Blackfoot technology is that there is no large, standard corpus of Blackfoot texts. To compensate for the lack of such material, I had to find the best solution possible within a specific timeline. The corpus is available at https://blackfoot.algonquianlanguages.ca/.

I would like to point out that there are in fact Blackfoot corpora, but I evaluated them as being unsuitable for this project. Two of the best were assembled or described by separate researchers. The first is the Blackfoot corpus of texts available in the OLD database, described by Dunham in his 2014 dissertation. Unfortunately, it was not possible to access this corpus of texts for this research. The second is a historical corpus assembled by Weber (2021). This corpus in its entirety was only unsuitable for testing the model because it lacked one important feature: orthographic standardization. The corpus was assembled by collecting various texts of Blackfoot, many of which were made well before any form of standard orthography was developed, so orthographic systems could vary wildly. There were parts of that corpus that were standardized, such as the Blackfoot dictionary corpus which could be utilized, and so I added those sections with Weber's permission for this study. Since I designed the model for contemporary uses, it

needed to adhere to a standard, widely used orthography system, and needed to be easily accessed and utilized by people who are not trained as linguists. As such, the corpus being tested on needed to utilize a relatively standardized orthography and needed to adhere to a conventional glossing standard (if any).

For this project I assembled a corpus of existing texts which utilized the standard orthographic system developed by Frantz (2017) and which had not yet been glossed. This resulted in the largest corpus of orthographically standardized Blackfoot text that exists to date. Due to time constraints, I was not able to provide linguistic glosses for the corpus, but the model provides a tool that will allow researchers to parse and gloss the corpus much faster than could be done without a computational parser (discussed in detail in section 5.2).

I chose texts for their orthographic homogeneity and their adherence to Frantz' orthographic system. The reasons for using this system are as follows. First, I used Frantz' descriptions of Blackfoot grammar and phonology extensively in this project, so it only makes sense to use the orthography system that he developed. Second, Frantz' orthographic system quite accurately captures the phonological and phonetic properties of the Blackfoot language in a straightforward way. Frantz' orthography is already becoming more widely adopted by Blackfoot speakers while other systems are not as widely used (Genee 2020:6–7). I am not using a glossed corpus. Although a glossed corpus would help with evaluation by allowing us to evaluate the machine-produced glosses against human made glosses, time constraints made it more desirable to create an unanalyzed corpus.[37] This allowed me to collect a larger corpus faster, which could be parsed using the technology described in this thesis.

Although Frantz' (2017) orthographic system is standardized, there are still many degrees of variation possible. This is due to several reasons. One is that the language is shifting from a primarily oral language, to one that is also written, so the standardization process is still in progress (see Genee 2020:2–3). Another is that Blackfoot has several dialects and sociolects (Miyashita and Chatsis 2015). These can affect both the grammar and the phonetic realization of a word, which often leads to orthographic variations. Table 11 shows some examples of how differences in spoken varieties can appear in written forms between normative orthographic conventions according to Frantz' orthography and colloquial variation.

[37] I did not use glossed corpora simply because glosses were not present in most of the texts that I used. Only a few had glosses available at all.

Table 12: Orthographic differences between Old and New Blackfoot varieties. (Miyashita and Chatsis 2015)

Normative form	**Colloquial variation**	**Gloss**
tsa anistapiiwa	tsistapii	'What is it?'
kitaakitamattsino	kiatamattsin	'goodbye (I will see you again)'
nitaakitapoo	taakitapoo	'I am going there'

In the table above, there are three examples of how surface forms can vary between Blackfoot varieties. In the first row, the two words have been contracted. In the second row, the shortened spelling conforms to a spoken variety of the language which pronounces a shortened version of the word. Not all dialects or individuals pronounce the word this way, but some do. In the final example, the first syllable has been dropped entirely. All variations are acceptable because they conform to the orthographic conventions based on how they are pronounced, but represent differences such as dropping syllables, contracting two words, or shortening words according to pronunciation.

Considering the information laid out in this section, I chose corpus texts based on the following principles. They had to be written in the Blackfoot language (regardless of dialect). They had to utilize Frantz' orthography, and they had to adhere fairly closely to that orthography, although I allowed variation to a limited degree. Additionally, the texts needed to be publicly available, so as not to inadvertently run into ethical issues by collecting private texts or using the intellectual property of groups or individuals without their permission to benefit this research.

The texts collected were from the following sources:

1. Examples and stories found in *Blackfoot Grammar* (Frantz 2017).
2. Examples found in *Blackfoot Dictionary of Stems, Roots, and Affixes* (Frantz and Russell 2017) and compiled by Weber (2021).
3. Web pages from the Siksika translations of the Jehovah's witness website (https://www.jw.org/en/library/?contentLanguageFilter=bla).
4. Blackfoot translations of web pages from the Glenbow museum website (https://www.glenbow.org/blackfoot/BL/html/index.htm).

5. Transcribed stories from the Blackfoot dictionary website (https://stories.blackfoot.atlas-ling.ca/#/stories).
6. Transcribed words and phrases from the *conversations* tab on the Blackfoot dictionary website (https://blackfoot.algonquianlanguages.ca/conversations/).
7. Stories from the book *Akaitsinikssistsi: Blackfoot Stories of Old* (Russell and Genee 2014).

The resulting corpus had an approximate[38] count of 13,784 unique word forms, and approximately 16,590 total word tokens. The quality of orthographic standardization varied a lot. For example, the web pages from the Glenbow website did not make use of the acute accent which is used to mark the placement of pitch accents, the placement of which can be lexically determined and can mark minimal pairs. Even some of the texts which were transcribed by Frantz himself could display orthographic variations. For example, in the story found at the end of *Blackfoot Grammar* (Franz 2017:187-197), the orthographic representation in example (16) appears.

(16) kiomá áattsistaaw itomátaniiyiihka

ki	om-wa	aattsistaa-wa	it-omat-aanii-yiihk-wa
CONJ	that-AN.SG	rabbit-AN.SG	then-start-say-NAR-3SG

'then the rabbit started to say…' (Frantz 2017:189)

Here, we can see that, although Frantz' orthographic rules would dictate that the final suffix *-wa* be represented in its fullness, it is represented as *-w*. The representation of animacy suffixes is a common variation across the texts, due to the variety of ways it they can be pronounced in spoken Blackfoot. The final vowels in Blackfoot words are usually devoiced, or dropped, and in some varieties of Blackfoot, the entire syllable *-wa* or *-yi* can be devoiced or dropped entirely. This is often reflected in Blackfoot orthography, and people sometimes write short versions of the suffixes, or simply omit them entirely. This is not limited to the singular animacy suffixes, but to most Blackfoot final syllables (Prins 2019:13–19). The example also shows another

[38] I say approximate because this count includes section labels, numeric characters, and punctuation found throughout the corpus. However, this count is very close to the number of unique Blackfoot word forms.

common variation, where the conjunction *ki* is analyzed as being cliticized to the demonstrative *oma* rather than as a free word.

Other common variations found across all the texts include the placement of long and short vowels and consonants, which are represented in the standard orthography by doubling the character. The placement of pitch accents can also vary, or pitch accents can be omitted in texts. The high pitch accent is also considered by some Blackfoot scholars to be mobile, meaning that orthographic representations of Blackfoot words can have pitch accents appearing in different syllables due to concatenation (Weber and Shaw 2022:10-12). This is shown in example (17).

(17) a. Issksinímayi
ssksini-ma-yi
know.TI-TH-3/0PL
'He knows it.' (Blackfoot Online Dictionary https://dictionary.blackfoot.atlas-ling.ca/#!/results)

b. Nítssksinihpa komohtsspíyihpi.
nit-ssksini-hp-wa k-omoht-ihpi-hp-yi
1-know.TI-TH-3/0SG 2-means-dance.AI-CNJ-IN
'I know (the reason) why you dance.' (Frantz 2017:158)

Example (17a) shows the stem *ssksini,* with the pitch accent appearing at the end of the stem before a suffix boundary. In example (17b), the pitch accent is represented to appear on the first syllable, in the first-person prefix. It may be possible to accurately model this as a morpho-phonological phenomenon, but for now, the best solution to this is an orthography relaxer that allows for variable placement of the pitch accent.

I have demonstrated how the articulation of final syllables can vary, but it is also common for initial syllables to be dropped. Example (18) contains an instance from the corpus where this variation is clearly displayed.

(18) tsinikíísínaakssini
atsinikiisinaakssin-yi
newspaper-IN.SG
'newspaper' (Frantz and Russell 2017 [from Blackfoot corpus])

This initial syllable dropping is also noted by Miyashita and Chatsis, and they provide an example which is represented in example (19).

(19) táakitapoo
nit-aak-itap-oo
1-FUT-toward-go.AI
'I am going there' (Miyashita and Chatsis 2015:112)

Although these variations present challenges, they also represent the most common ways in which the orthography can vary. This provided a good basis on which to build important tools such as orthographic relaxers and generally variable phonologies which allow a wide variety of surface forms to be captured. Since this technology is being designed both to assist in Blackfoot documentation work, and to provide tools for the wider Blackfoot community to use, taking these common variations into account represents a crucial part of the technological development process because they make the technology more useful to linguists and Blackfoot community members alike. In the sections that follow, I describe how I used the Blackfoot corpus to evaluate the computational model, and I provide the results of those evaluations.

4.2 YAML file evaluations

I performed the initial evaluations for the technology using Yet Another Markup Language files (henceforth YAML files). YAML files are encoded using a markup language with serialization capabilities. Markup languages can be decoded or serialized by most programming language compilers, meaning that they can be used to compare two sets of data, one data point at a time. Each data point can be entered into the file and treated as an object (such as a tuple, list, or

dictionary). In this case, the YAML files are being designed and treated like a dictionary,[39] with underlying analyses as the keys, and their corresponding surface forms as values. In morphological modelling, YAML files can be used to test whether the key:value pairs in the file can be matched exactly by the Finite State Model of a language's morphology, which is what I used them for. When using YAML files to determine the performance of a computational model of a language's morphology, the files ideally contain full paradigms of inflected stems that have been verified,[40] and their correct underlying analysis as it would be parsed by the machine.

This type of evaluation is largely qualitative. It is based on a few examples that have been identified by the researcher and verified using field or corpus methods. In this section, I describe the exemplary forms used in the YAML files, and how they were obtained and tested separately for verbs and nouns. The section ends with a report on the results of the YAML file tests.

4.2.1 Methods

When creating YAML files, ideally, we would create enough files to cover the most common inflectional patterns found across stem types. For example, when evaluating Blackfoot noun stems, it is important to identify the most common allomorphic variations, particularly regarding initial change (described by Frantz 2017:84-89; see section 2.5).

The YAML tests are performed to investigate how accurate the model is. In each file, full inflectional paradigms are provided, and each underlying analysis is paired with its correct generated word form, as should be provided by the model. This data is entered as a dictionary with the key being the analysis and the value being the surface form. See Table 11 below for an example of some inflected nouns and the corresponding expected analyses provided by the model. The colon at the end of each analysis separates the key from the value.

[39] In programming, a dictionary is a list of key: value pairs. Lists and dictionaries are objects that can be manipulated by programming languages.
[40] Verification can be done in a variety of ways. It can be done by collecting data from fluent speakers, running field experiments, or by taking forms from a corpus.

Table 11: YAML file pair examples

Analyses	**Surface forms**
imitááwa+NA+Px1Sg+Pl:	nitómitaamiksi
ikákomimm+VTA+Sbj+3Pl+2PlO:	ikákomimmotsiinoainiki
Neg+Fut+oowatoo+VTI+Ind+1Sg+0SgO+INTERR:	nimáátáaksowatoohpa

This testing process ensures that the model can accurately parse inflected word forms and can accurately generate inflected word forms from an underlying analysis. When complete inflectional paradigms are provided, and all known inflectional classes and patterns are represented in the YAML files, we assume that the computational model is accurate if it is completely successful.

The list of noun stems and the rationale for including each of them can be found in Table 12. I should first briefly explain some of the terms I use in Table 12. One is the term non-permanent. I use this term following Frantz to describe initial nasals which are deleted when they appear after a prefix (Frantz 2017:80-81). I also use an additional term based on my observations of Blackfoot inflection, which I made while adding the allomorphic triggers to the stems in the model. The disappearing /i/, is an initial /i/ which surfaces sometimes when there is no prefix on the stem. Sometimes, these disappearing /i/s were not consistently represented in the Blackfoot dictionary site where I got the stems for the model, thus, I needed to develop a solution, and test that solution in the YAML tests. I remain theoretically neutral as to the nature of this disappearing /i/. A more in-depth discussion about the allomorphic triggers I used can be found in section 4.4.

Table 13: Noun stems included in YAML file tests.

Stem	Grammatical class and gloss	Rationalization
aakíí	NA, ‘woman’	No allomorphy. Mobile pitch accent.
issk	NA, ‘pail	Irregular initial allomorphy.
imitáá	NA, ‘dog’	/i/ to /o/ initial allomorphy.
isttóan	NA, ‘knife’	Breaking /i/ to /o/ initial allomorphy.
miistsís	NA, ‘tree’	Permanent initial /m/.
moksís	NA, ‘awl’	Non-permanent initial /m/ and shortened person prefixes. Not a body part.
aaáhs	NAD, ‘elder’	Dependent noun beginning with /a/.
ităn	NAD, ‘daughter’	Dependent noun beginning with /i/.
ohkó	NAD, ‘son’	Dependent noun beginning with /o/.
ohkíímaan	NAD, ‘wife’	Dependent noun beginning with /o/.
póós	NA, ‘cat’	/oh/ initial allomorphy.
nínaa	NA, ‘man’	Non-permanent initial /n/.
aaapan	NI, ‘blood’	Disappearing initial breaking /i/.
aahkioohsa’tsis	NI, ‘boat’	No allomorphy.
kinakíni	NI, ‘liver’	/oh/ initial allomorphy, takes short person prefixes, body part term.
atsikín	NI, ‘shoe’	Takes short person prefixes, body part term.
míín	NI, ‘berry’	Non-permanent initial /m/ followed by breaking /i/.
mótokaan	NI, ‘head’	Non-permanent initial /m/, body part term.
sináákiatsis	NI, ‘book’	Begins with disappearing breaking /i/, takes short person prefixes.

Full paradigms for verbs are more difficult to come by than they are for nouns. A likely reason for this is that full inflectional paradigms for Blackfoot verbs are much larger than they are for Blackfoot nouns. The same principles that apply to the process of creating YAML files discussed in the previous section also applies to creating YAML files for verbs. That is, the YAML files ideally represent the most common inflection patterns, each one contains the full inflectional paradigm for its respective stem, and they all contain hand-parsed analyses of each surface form.

In order to obtain examples of inflected stems to be used in the YAML files, a test version of the verb model was used to parse the full Blackfoot corpus. The test version of the verb model, like the test version of the noun model contained only a handful of stems which I

considered to be exemplary based on Frantz (2017). I then isolated only the parsed forms and compiled a list of the unique parsed word forms and their parses. For each of these, I expanded the paradigms based on the patterns found in the corpus and the descriptions provided by Frantz and entered them into the YAML files for testing. Table 14 below displays the stems that were used, the rationales for including each, and additionally, the number of unique word forms in the corpus containing each stem.

Table 14: Verb stems included in YAML file tests.

Stem	**Grammatical class and gloss**	**Rationale**
á'po'taki	VAI, 'work'	Durative morpheme surfaces in second syllable.
itápoo	VAI, 'go there'	Begins with /i/.
okská'si	VAI, 'run'	Begins with /o/.
wáahkayi	VAI, 'go home'	Begins with a glide. Inflects regularly as a form beginning with /a/.
ináátsi	VII, 'appear to be'	Begins with breaking /i/.
soká'pii	VII, 'be good'	Begins with /s/.
ikákomimm	VAI, 'love'	Full paradigm. Begins with breaking /i/.
ino	VAI, 'see'	Short VAI. Begins with breaking /i/.
oowátoo	VTI, 'eat'	Full paradigm. Begins with disappearing breaking /i/.
ssksiní	VTI, 'know'	Begins with disappearing breaking /i/.

4.2.2 Results

I ran two different test versions of the Blackfoot model. One allowed for simple pitch accent flexibility in the phonology to account for pitch accent mobility, and the other did not. I cover the design of orthographic relaxers in more detail in section 4.4. The results are shown in tables 15 and 16. Table 15 provides the results for the noun stems that I tested and Table 16, for the verb stems. The tables include the stem, its inflectional class, and its performance when testing against a relaxed version of the model, which allowed for variability in the placement of pitch accents, and when tested without the relaxer. The results are shown with the number of successes over the total test words.

Table 15: Noun stem YAML test results

Noun Stem	Inflectional Class	Relaxed Model	Without Relaxer
aakíí	NA	30/30	19/30
issk	NA	30/30	26/30
imitáá	NA	30/30	30/30
isttóan	NA	30/30	26/30
miistsís	NA	30/30	29/30
moksís	NA	30/30	1/30
aaáhs	NAD	22/22	22/22
itán	NAD	30/30	29/30
ohkó	NAD	26/26	0/26
ohkíímaan	NA	24/24	23/24
póós	NA	30/30	0/30
nínaa	NA	30/30	30/30
aaapan	NI	24/24	3/24
aahkioohsa'tsis	NI	24/24	24/24
kinakíni	NI	24/24	1/24
atsikín	NI	24/24	24/24
míín	NI	24/24	24/24
mótokaan	NI	24/24	24/24
sináákiatsis	NI	24/24	24/24

On Table 15, we can see that the relaxed model produced perfect results for every stem. For both models, the results of the tests run with the phonology without the pitch accent relaxer are varied, with many tests producing perfect results, while others failed. This depended on the placement of the pitch accent. For example, the stem *aakii,* could surface as *áákii* or *aakíí,* due to the mobility of the pitch accent. The default form in the model was *aakíí,* so the 11 test forms with the pitch accent shifted to the first syllable failed. Another example is *nitán* which is represented in the model and the dictionary without a pitch accent, but which usually has a pitch accent on the second syllable. The solution of a relaxer was thus essential for the YAML tests until pitch accent mobility can be accurately modelled.

Table 16 provides the results for the noun stem YAML tests.

Table 16: Verb stem YAML test results

Verb Stem	Inflectional Class	Relaxed Model	Without Relaxer
á'po'taki	VTA	70/70	50/70
itápoo	VAI	25/25	8/25
okská'si	VAI	16/16	3/16
wáahkayi	VAI	13/13	9/13
ináátsi	VII	14/14	3/14
soká'pii	VII	18/18	18/18
ikákomimm	VTA	95/95	95/95
ino	VTA	26/26	11/26
oowátoo	VTI	89/93	71/93
ssksiní	VTI	9/9	2/9

The tests provided some slightly anomalous results when I tested them with the relaxed model. Most notably, *oowatoo,* which was the only verb which did not achieve a perfect score. It had a full paradigm which I took from the Blackfoot dictionary site, and which was elicited from a Blackfoot speaker. *Oowatoo* contained a few forms with attached pronouns which did not have sufficient documentation for me to add them to the model. It was unclear to me whether they were being used as interrogatives, or whether they were simply attached pronouns. One of the word forms in question is shown below in example (20).

(20) Maatáaksowatoomaatsayi

máát-áak-oowatoo-m-aatsayi

NEG-FUT-eat.TI-TI.TH-?

'He/she will not eat it.' (https://blackfoot.algonquianlanguages.ca/grammar/verbs/verb-classes/transitiveinanimate/oowatoo-eat/)

I designed the model to recognize up to two enclitic pronouns on each word, and the interrogative endings, but some of the forms in the *oowatoo* paradigm, were not recognized by the model. I tried different parses, but none seemed to fit. This needs to be explored more in the future.

In general, the results of the YAML file tests reveal that the model can accurately both parse and generate word forms given a phonological model with relaxed pitch accent, although

not quite perfectly (in the case of *oowatoo*). The model is also capable of accurately parsing a variety of stem and prefix types with different allomorphic patterns, and of parsing inflected word forms with stems from all the inflectional types. In the future, creating and testing more YAML files will help to increase the accuracy of the model, and assist with its continued development.

4.3 Evaluating the model with the Blackfoot corpus

The YAML files tests reported on in section 4.2 of this chapter provide insights into the accuracy of the Blackfoot computational model on a variety of controlled examples that were chosen by the researcher. In this section, I perform some simple quantitative tests that provide further insight into the accuracy of the model and additional insight into precision of the model using the Blackfoot text corpus that I describe in section 4.1. I provide the methodology used for testing in section 4.3.1 followed by a detailed report on the results in section 4.3.2.

4.3.1 Methods

I performed these quantitative tests because the qualitative YAML tests alone do not provide information about the precision of the model, only its accuracy. A well-designed computational model should be both accurate and precise. A model may perform perfectly on the YAML file tests, but still be too broad and imprecise to be considered an accurate general model of Blackfoot morphology. Thus, a quantitative statistical investigation performed on the entire Blackfoot corpus provides new insights into the precision of the model by recording the average and maximum number of parses provided by the model for each parsed word. It also provides us with further insight into the accuracy of the model by recording what percentage of the corpus is parsed.

I begin this section by explaining the technical difference between accuracy and precision as they apply to the tests that I performed. Table 17 provides examples of results parsed by an accurate, but imprecise model of Blackfoot. It gives the surface form, the correct analysis as parsed by the model, and then the number of analyses that the model generated. These examples come from some earlier version of the model.

Table 15: Input and output of an accurate but imprecise model

Surface form	**Generated correct analysis**	**Number of generated analyses**
nitsikákomimmayi	ikákomimm+VTA+Ind+1Sg+3PlO	268
ssááhsinaawa	Past+ssaahsinaa+VAI+Ind+3Sg	786
áakaakaapiksistsimaawa	Fut+yaakaapiksistsimaa+VAI+Ind+3Sg	6024

Table 17 is accurate but imprecise because it always generates at least one correct analysis, but that correct analysis is only one among many incorrect ones that are also generated by the model, as shown by the number of generated analyses.

Table 18 provides examples of results parsed by an inaccurate but precise model. It shows the surface form being analyzed, and the analysis as parsed by the model. Again, these examples were taken from an early version of the model.

Table 16: Input and output of an inaccurate but precise model

Surface form	**Analysis given**	**Number of generated analyses (all incorrect)**
Mi[41]	ini+VTI+Ind+3Pl+0PlO	2
Ma	ini+VTI+Ind+3Sg+0SgO	1
Niksísta	PN/ksis+sstaa+VAI+Ind+1Sg	1

The precision of the model in Table 18 lies in the fact that, although the model never provides the correct analysis, it always reliably provides one or two analyses for each form. It can be thought of as being inaccurate, yet precisely getting the wrong answer every time. The successful YAML tests performed in section 4.2 of this chapter largely rule out the possibility that the model is inaccurate, but they do not rule out the possibility that the model is imprecise.

There are also several other reasons to evaluate the model using the whole corpus. First, a test on the whole corpus evaluates the ability of the model to parse inflected Blackfoot words in context. Determining how many of the words are parsed, and what percentage of the unique wordforms these represent provides us with insights into how well the model is performing when

[41] For the first and second examples in Table 18 are demonstratives. Any time the model parses a demonstrative, there is no chance that the analysis is correct because demonstratives have not yet been modelled.

parsing a variety of words that are inflected in various ways. This type of evaluation is similar to the YAML file tests in principle but done in a much less controlled environment.[42]

Evaluating the model on the full corpus also provides the opportunity to evaluate a sample of the parsed wordforms manually. This adds more data to the accuracy tests and tells us two things that are not covered by the YAML file tests. First is that it allows us to evaluate the order in which the parses appear. For example, if there are fifty parses of a single word, and the only plausible parses appear at the end of the list of parses returned by the model, then evaluating those parses manually provides the opportunity to find a solution that privileges the most plausible parses and removes the implausible ones. Second, it allows us to evaluate a wider variety of parsed forms that are not present in the YAML files. This means that I could determine whether the model was effectively capable of parsing a wide variety of inflected words, not just common inflections. The normative YAML file tests mostly capture controlled inflectional paradigms, specifically those relating to person agreement, mode, and possession.

I also used these corpus evaluations in the development of the model to inform certain design choices, which are explained in section 4.4.1. This was especially important to inform grammatical choices which were not explicitly clear in the resources used to develop the Blackfoot morphological model. For example, determining which preverbs can appear where, and with which restrictions was difficult because these things have not yet been explicitly explored or discussed in the literature. If limiting some of the preverbs in the model would help to narrow down the number of possible parses, then doing so should be accompanied by a quantitative test to ensure that this modelling choice both decreases the number of possible parses, and either increases the number of unique wordforms parsed by the model, or at least keeps it the same.

4.3.2 Results

The results of the quantitative tests are shown in tables 19, 20 and 21 below. Table 19 shows a summary of the results without any other details. It simply shows the number of unique words in the corpus, the number of those words which were parsed by the model (regardless of whether or

[42] I acknowledge that it is not an entirely uncontrolled environment since I am the one who intentionally curated the corpus, thus controlling the content present in the corpus.

not they are correct), and what percentage of the total unique words were parsed. Table 19 also provides the total number of words parsed as verbs or as nouns, including words that were parsed as both verbs and nouns (which is why they add up to more than the total number of parsed words).

Table 17: Quantitative results of the morphological model

	Total word tokens	**Total unique words**	**Total unique words receiving at least one parse**	**Total unique words receiving no parse**	**Total unique words receiving at least one verb parse**	**Total unique words receiving at least one noun parse**
Full model	16590	13784	9882 (71.69%)	3902 (28.31%)	9158	7921

Table 20 shows more fine-grained details about word classes. It shows which words were parsed as belonging to which class (either nouns or verbs), and whether the class of the word changed, or remained the same (nouns parsed normally). Table 20 also includes the total number of words parsed as verbs or nouns. The total number of words parsed as verbs or nouns is greater than the total number of parsed words because there is some overlap between nouns and verbs which means some words are parsed as being both nouns and verbs. This table indicates that many of the words present in the corpus would not have been parsed at all if verbalization and nominalization morphology were not modelled.

Table 18: Quantitative results broken down into parts of speech[43]

	Total words with at least one verb parse	**Total words with at least one noun parse**	**Total words with at least one noun and one verb parse**	**Total words parsed**
Verbs	7071	5087	4082	8076
Nouns	1082	1829	77	2834

[43] The total number of noun and verb stems present in the corpus are not included in the corpus because collecting those numbers would need to be done by hand. This would take more time than is possible in this thesis project, but the table serves to communicate how many words were captured by the verbalization and nominalization features.

Table 21 provides a general overview of the precision of the model by providing the mean number of parses per each wordform in the corpus.

Table 19: Precision statistical evaluation

	Average parses per verb stem	**Max parses per verb stem**	**Average parses per noun stem**	**Max parses per noun stem**	**Full model average parses**	**Most frequently occurring value**
Precision stats	111.99	9332	134.37	10579	121.72	2 (72 times)

The results of these statistical tests are very encouraging. The model can parse 71.69% of the corpus, making for a total of 9,882 of the 13,784 unique words parsed (not taking into account whether the parses are correct). Considering how quickly the model was developed, these are quite satisfying results.

This test also reveals a problem with the initial model, that is, it is overproductive. Although the most commonly occurring number of parses per word form are between 2 and 8,[44] both the noun and verb models are commonly producing over 1000 parses per word and producing a maximum of 10,579 parses for a single word form. To investigate this, I hand-checked a sample of some of the words with high parse counts from the fully parsed corpus and found that the vast majority of the parses were entirely incorrect, but every parsed word in the sample had at least one correct analysis. This means that although the model is accurate, it is extremely imprecise. Investigating this sample allowed me to devise a simple solution that would greatly increase the precision.

To solve this issue, I designed a program using Python, that privileges the simplest parses, and cuts out the more complicated ones. The logic behind it works as follows. Table 22 provides an example of a word-form that returned thousands of parses. The majority of the parses break complex stems[45] themselves into all possible separate morphemes, rather than leaving the stem intact.

[44] The most common number of parses being: 2 (73 occurrences), 4 (67 occurrences), 6 (59 occurrences), 8 (59 occurrences), and 5 (49 occurrences). The most commonly occurring number of parses is shown in Table 21 under the caption 'most frequently occurring values.'

[45] Blackfoot stems can be made up of multiple lexicalized morphemes. For example, the VTA *ssksino*, 'know' can be evaluated as containing two parts, the directional preverb *issk-*, 'backwards, behind', and the stem *ino*, see. The unrestricted version of the model would parse these as two separate morphemes, and as a single stem.

Table 20: Example of an over-productive word-form

Word-form	**Correct parse**	**Total parses**
Áakaakaapiksistsimaawa	Fut+yaakaapiksistsimaa+VAI+Ind+3Sg	6024
Example of an incorrect parse		
PV/áka+PN/aká+PN/áápi+PV/iksist+PN/mii+imm+VTA+Ind+21Pl+3SgO		

Table 22 provides an example of two parses of the same word, one which is correct, and another which is among the 6023 incorrect or undesirable parses. The example in the Table is not just undesirable, but incorrect because the model has parsed the word using preverbs which cannot possibly be present in the original word form (a common issue when using the orthography relaxer to help parse corpora). This was a common pattern among the parsed words, where often a word would have hundreds or thousands of possible analyses because it was being segmented by the model with more prestems than were actually present. This suggested that the most desirable analyses would be the simplest ones, which in this case, would be the ones with the fewest morphemes. The reason why this works is that many of the Blackfoot noun and verb stems represented in the dictionary are already complex constructions made from simpler stems with several prefixes or other stems compounded to them. The most likely preverbs to be concatenated to the stem are the inflectional prefixes that express tense, aspect, and negation. Unlike the other preverbs, they are marked with their inflectional functions rather than their underlying forms as preverbs and prenouns are, so they are not counted by the Python program.

Table 23 below reports on how the precision of the model improved after implementing this program. The other stats remain the same as in tables 19 and 20 as the program does not do anything to restrict the number of words which are parsed by the model.

Table 21: Model test with preverb and prenoun restriction

	Average parses per verb	**Average parses per noun**	**Max parses (for all)**	**Average parses for whole model**
Precision stats	5.657	6.71	9	5.657

We can see in the Table that the precision of the parses is greatly increased by restricting the maximum number of parses to the eight parses with the lowest number of preverbs and prenouns.

Once the issue with precision was solved, I needed to report on the accuracy of those parses. For this, I hand-checked a sample of the corpus to ensure that the accuracy of the model was not negatively affected. My method was to use Python to collect 300 random parsed words

(3.03% of the parsed words) and to evaluate them. I gave each parsed word a pass or fail value, and they passed if at least one of the nine parses was correct, and they failed if none of the nine parses were correct. The result of this check is reported in Table 24.

Table 22: Hand checked accuracy statistics

	Number of words checked	**Number of words with one correct analysis**	**Percentage of correctly parsed words in sample**
Full model	300/9882	280/300	93.33%

The Table demonstrates that the accuracy of the model is not negatively affected, at least according to the sample. This means that the model, in conjunction with the extra software that I have developed is now both reasonably accurate and precise.

The largest identifiable weakness of the model is not its ability to inflect Blackfoot nouns and verbs accurately and precisely, but that it is too flexible. The data shows that the most common error of the model is in incorrectly identifying surface forms as nouns or verbs, as shown in Table 25, which provides a surface form, its incorrectly parsed analysis, and an example of a correct parse.[46]

Table 23: Incorrect identification of demonstratives as nouns and verbs

Surface form	**Example of incorrect parse**	**Correct parse**
Mi	ini+VTI+Ind+3Pl+0PlO	DEM+Dist+NI+SG
Omi	omii+VAI+Ind+3Sg	DEM+Dist+NI+SG
Anna	anááyi+NI	DEM+Prox+NA+SG

This over flexibility is problematic for the Blackfoot morphological model, especially since it cannot yet accurately[47] parse demonstratives or any word that does not contain a noun or verb stem.

[46] This correct parse is not a suggestion about how a demonstrative parse should look, simply an example of what demonstrative parses could look like.
[47] I say accurately because the model is parsing demonstratives, just incorrectly.

4.4 How the corpus informed the development of the model

The corpus played an essential part in testing the model, as demonstrated in sections 4.2 and 4.3. In addition to testing, the corpus also informed some development and decisions, which I describe in this section. The corpus informed the creation of a simple supervised machine learning model that uses weights to determine which parses are most likely to be correct. I describe how this works, and how the corpus was used to create this model in section 4.4.1. The testing process also informed development decisions and aided in maximizing the accuracy and precision of the model. The noun and verb morphology described in chapters 2 and 3 provide an overview of the final model structure as determined by the descriptions of Blackfoot grammar that have been completed thus far. This work was complete enough to create an effective computational model to parse Blackfoot, but said model also required some extra work in order to make it maximally effective. These design choices and other computational developments are described in section 4.4.2.

4.4.1 Weighting the model

Weighting is a simple, supervised machine learning technique that can be used to supplement morphological models to make them more useful for parsing. The software that I used in conjunction with the Blackfoot model (Silfverberg and Lindén 2009) is part of the Helsinki Finite State Technology (HFST) toolkit discussed in section 1.4 above. In this subsection, I describe how the weighted model operates in section 4.4.1.1. In section 4.4.1.2, I describe the process of determining and applying these weights, and why the HFST compiler is used to produce finite-state machines that make use of weights.

4.4.1.1 Weighted modelling

There are a few types of weighting techniques utilized in the field of natural language processing that vary in their complexity. Some are able to take relationships between words and morphemes into account while others, such as the one I utilized for this project, are simpler. In this section, I

explain how the weighted model works, and I justify my choice to use this specific weighting technique.

Before I discuss how the weighted model works, I provide some explanations about supervised machine learning. First, machine learning is a computational technique which relies on quantitative methods mixed with algorithmic mathematics to process data. Machine learning models are improved by adding to the data that the machine learning model learns from; thus, machine learning software allows computational machines to "learn" in a way that is analogous to the way that humans learn. Machine learning is a subfield in the field of artificial intelligence, which is more general term to describe computational techniques that are used to develop software that mimics the capabilities of human intelligence.

Machine learning models can be supervised, or unsupervised. Unsupervised models do not require human intervention in data development. An example of an unsupervised model in natural language processing might be one which compares two sets of data in order to learn from them. Such a model could use a set of texts in one language, and a set of those same texts translated into a different language, compare them as two sets of data, and produce a machine translator that can translate text inputted by a user. This model is unsupervised because the developer does not intervene in the data correlation process. In contrast, if the machine learning model used glossed Russian text to produce English translations, then the machine learning model would be considered supervised because the data processed by the model was produced specifically for the model to process with human intervention.

Both supervised and unsupervised techniques have advantages and disadvantages that guide the choices of developers. Unsupervised models require less control of the data and, if enough data exists, can be developed quickly using advanced statistical methods. The disadvantage to unsupervised techniques is that they require huge amounts of data to be effective. Supervised techniques do not require such massive amounts of data, but they do require manual supervision of the training data to make the data more usable. Since Blackfoot

does not have enough data to make a comprehensive[48] unsupervised model, I employ a simple supervised one.[49]

The weighted model works by simply assigning a value to each lexical item found under the lexica in a LEXC file. This can be done manually, perhaps based on linguistic descriptions, or it can be done more automatically. Figure 18 shows a simple model of Blackfoot morphology using HFST (see section 1.4 of this thesis for more details about the LEXC modelling process). This model is hypothetical, and is only being used to illustrate how weighting works, not to accurately reflect what weights would be for each morpheme.

```
LEXICON prefixes
@P.person.nit@:@P.person.nit@nit verbs "weight: 4" ;
@P.person.kit@:@P.person.kit@kit verbs "weight: 4" ;
@P.person.null@:@P.person.null@0 verbs "weight: 2" ;

LEXICON verbs
a'poowa:a'poo       suffixes "weight: 5" ;
ikskimaawa:ikskimaa suffixes "weight: 3" ;

LEXICON suffixes
@P.person.nit@+1:@P.person.nit@0    # "weight: 4" ;
@P.person.kit@+2:@P.person.kit@0    # "weight: 4" ;
@P.person.null@+3:@P.person.null@wa # "weight: 2" ;
```

Figure 18: Simple model of Blackfoot morphology utilizing weights.

The weights appear in double quotation marks between the next defined lexicon, and the semi colon used to signify the end of a function. This weight is assigned to the morpheme that it appears next to. Table 26 shows the output of the simple weighted morphological model in Figure 18.

[48] I use the term comprehensive because it may be possible to use an unsupervised model to learn the most common affixal morphemes or lemmas (see Lane, Harrigan, and Arppe 2022).

[49] The corpus I created and used for development and testing cannot be used for unsupervised learning for two reasons. First, it is not the right type of data. If the corpus had corresponding translated English text, then it might have been useful for an unsupervised model, which leads me to the second problem. Regardless of the type of data, the corpus is not large enough to utilize an unsupervised technique.

Table 24: Output of the simple weighted Blackfoot morphological model

Underlying analyses	**Surface forms**
íkskimaawa+3	íkskimaawa (weight: 7)
a'poowa+2	kita'poo (weight: 13)

In the Table, each weight is determined by adding the weight of each component morpheme. For example, in *kita'poo* the model will add the weight for the prefix *kit-* (weight: 4), the stem *a'poo* (weight: 5), and the null suffix (weight: 4). The result comes to 13, which is shown beside the surface form.

The weights are inconsequential when analyzing surface forms that only produce one analysis, but they are consequential when analyzing surface forms that are ambiguous such as *aaápani*, which is ambiguous as to whether the suffix is obviative *-yi* with the glide deletion, or non-referring *-i*. In Figure 18, the third-person suffix is given a lower weight than the null suffixes, so given the same stem, the null suffixes will be privileged and appear first.

The example model was compiled into an HFSCRIPT program using the HFST compiler. HFST is similar to the FOMA compiler that was used for most of the development process. FOMA was extensively used for testing because it could be used to quickly compile and test different versions of the model to determine the best design and development choices for the model. The HFST compiler is slower than FOMA but has additional capabilities such as support for weighted modelling, with the ability to output the weights for each parse. FOMA cannot be used to compile a weighted model because it ignores the weight indicators.

4.4.1.2 Weighted model integration

The model used as an example in section 4.4.1.1 has weights which were added manually. This makes it a closely supervised model. I used a different technique to add weights to the full Blackfoot nominal and verbal morphology models. I used a less supervised technique, or a semi-supervised technique to add the weights. I describe that technique in this section.

First, I used the full Blackfoot verb and noun model to parse the corpus. Alongside the raw model, I used the Python program (discussed in the section 4.3), that was designed to increase the precision and accuracy of the model to prune the analyses. Next, I used a GAWK script written by Antti Arppe to isolate each morpheme tag, including all inflectional,

derivational, phrasal, and stem morphemes and got individual counts for each of them. I then applied the regularization equation log2 to each count. This returns the natural logarithm of a number to base 2. An example of how this works is found in Table 27 and is followed by an explanation of why I used this regularization technique.

Table 25: Morpheme counts compared with their regularized forms.

Morpheme	**Original weight**	**Regularized weight**
nit	8	3
kit	8	3
wa	6	2.58
a'poo	5	2.32
ikskimaa	3	1.59

The Table shows how the log2 operation regularizes the weight values. No adjusted weight value in this example is more than twice as high as another, even if the original value was more than twice the other. This is useful when one count value that would otherwise be hundreds of times more than another value, becomes only 3 times higher. For example, a morpheme with an original weight of 500, is only 8.966 when adjusted, which is just less than three times greater than a morpheme with an original weight of 8, rather than 62.2 times greater. This ensures that there are no extremely asymmetrical weights assigned to any given morpheme. The model sorts parses according to their weights, so avoiding large quantitative asymmetries in the morpheme weights ensures that a range of parses are still made readily available by the model.

Although this model is simple, and there are some potential limitations to it, it still enhances the Blackfoot morphological model by making it more intelligent. There are two clear flaws to the weighted version of the model as described in this section. First, I collected the weights using the same corpus that the model is designed to parse and that was parsed using the same model for which the weighted model was developed. This is not an ideal situation, but not an uncommon solution.

Second, even with the use of a program to increase the accuracy and precision, the output of the model is not always correct or representative of what some of the weights should be. For example, when presented with the preverb *áák-*, the Python program will privilege any parse that analyzes it as *yáak-*, which marks the future tense. The reason is that the program does not count *yáak-* as a preverb because its analysis is represented as Fut+ in the model. This was

purposefully done to limit the number of parses, and to privilege negation, tense and aspect preverbs over other preverbs because they are far more likely to occur. However, if a word-form actually contains the preverb form of *áák-*, it may not be analyzed as such because the model currently has no way to distinguish between the two similar prefixes and will prioritize parses containing *yáak-*.

These issues can be solved in the future as researchers use the technology to help them produce a human-checked morphologically tagged corpus. The technology will speed up this process a lot, making it possible to create better parsing technology in the near future. More detailed discussion about the future of Blackfoot technological development and research is found in chapter 5 below.

4.4.2 Morphological and phonological modelling choices

Besides the weighting process discussed in section 4.4.1, testing the model led to some changes and innovations that increased the performance of the model in different way. In this section, I outline some of those innovations, what informed them, and how each one interacts to improve the model.

Recall from section 4.1, that the model was developed using Frantz' *Blackfoot Grammar* first, and then further designed to fit the corpus. The corpus was designed and curated with this goal in mind, so I chose a variety of text types that would represent many different inflected and derived word-forms and orthographical forms. Doing so allowed me to control the allowable types of diversity in the corpus as informed by current variation documentation (see section 4.1). This helped me to achieve the goal of designing a model that conformed to a standard orthographic system, but that also accounted for the most common variations available within that system.

In this section, I first discuss the trigger system that I implemented in 4.4.2.1. Then, I describe a relaxer which accounts for common orthographic variations, and which is essential when running the model, and testing it on the corpus in 4.4.2.2. In 4.4.2.3, I discuss the innovations and techniques implemented to enhance the morphological model.

4.4.2.1 Phonological triggers

The phonological model that I utilized for this project is described in Schmirler, Arppe and Genee (Forthcoming). The phonological model designed by Schmirler et al. makes use of regular expressions just as explained in section 1.4 of this thesis. The paper explains how this phonological model was created using only the phonological rules listed in the appendix of *Blackfoot Grammar* (Frantz 2017:176-179). These rules describe the Blackfoot phonology system well enough that they could be replicated into a list of regular expressions with minimal changes made by the author and are already ordered accurately for feeding and bleeding rules. This model has proven to be effective and has worked with few changes. However, these rules alone do not account for significant features of Blackfoot morpho-phonology such as allomorphy. This section briefly outlines how I implemented these allomorphic triggers, and how I edited the phonological rules to make it perform better.

The common Blackfoot allomorphic patterns are described in Frantz (2017:84-89). I mainly focussed on what Frantz calls morpheme-initial variation (Frantz 2017:84-86). Each variation that he describes begins with a heading with the rule written in shorthand. The allomorphic triggers are shown in Table 28. I provide each trigger as an example stem as it is represented in the model with its trigger, alongside an inflected example. The bolded text in the allomorphic description are the shorthand headings indicating which rules correspond to the descriptions provided by (Frantz 2017:84-86), and the bolded text on the uninflected stems represent the triggers.

Table 26: Allomorphic trigger examples

Uninflected stem	**Allomorphic description**	**Example inflections**
hpóós	Words which take an allomorphic *oh-* only when prefixed. **Ø ~ oh**	póósa, nitohpóósiima
^SPitákkaa	Prefixes and independent nouns that take shortened person prefixes.	itákkaawa, nitákkaawa
i3mitáá	Initial /i/ which changes to /o/ when prefixed. **i ~ o**	imitááwa, nitómitaama
i2sskssíínaa	Initial breaking /i/. Sometimes non-permanent in initial position. **Ø ~ I**	isskssíínaawa, nitsssksíínaa
a4soyínnáápiooyis	Initial /a/ becomes breaking /i/.	asoyínnáápiooyisi, nitsisoyínnáápiooyisi
a3pisaan	Initial /a/ becomes /o/. **a ~ o**	apisaani, nitópisaani

The first four allomorphic triggers in the Table were the most common. The last two are not very common, and entirely absent among the verb stems. The two rules without a corresponding description from Frantz are the short prefix trigger ^SP, and the /a/ becomes breaking /i/ trigger. The short prefix variation for independent nouns is not described in Frantz' chapter on allomorphy, and the /a/ to breaking /i/ rule only occurs on a few stems that I observed and is not described by Frantz at all.

For every stem that displayed allomorphy but did not conform to the patterns in Table 28, I created a duplicate in the lexicon as a prefixed and non-prefixed form. There were very few such stems. An example of an irregular allomorphic pattern is shown in example (21) using the example stem *ponoka*, which becomes *innoka* when prefixed.

(21) a. ponokawa
ponoka-wa
elk.NA-AN.SG
'elk' (Blackfoot Online Dictionary https://dictionary.blackfoot.atlas-ling.ca/#!/results

b. siksinnokaiksi
sik-ponoka-iksi
black-elk.NA-SG
'black elks' (Blackfoot Online Dictionary https://dictionary.blackfoot.atlas-ling.ca/#!/results)

There are no other known stems which alternate with the same phonological pattern that *ponoka* does, so it is treated as an exception, and enumerated in the animate noun stem lexicon.

These triggers are handled in the phonology before the regular phonological rules. The regular phonological rules were implemented by Schmirler, Arppe and Genee (Forthcoming) using the appendix of Frantz' grammar (2017:176-179), which lists the rules in order, to account for feeding and bleeding. The allomorphic trigger rules and morpho-phonological rules operate independently from the regular phonology, following the design implemented by Dunham (2014:297-309). His phonological model of Blackfoot implemented allomorphic and morpho-

phonological rules using Frantz (1991) as a guide, and implemented allomorphic rules preceding the regular phonological rules, but they were applied without trigger rules. His reason for not using triggers was that the OLD's morphological parser was not a model of Blackfoot specifically but rather a general parser that was designed to parse words from many languages rather than a grammatical model of Blackfoot, so triggers could not be implemented in the same way.

4.4.2.2 Orthographic relaxer

Orthographic relaxers are simple, flexible graphological programs that use regular expressions to define acceptable variations in spelling. By acceptable variations I mean variations which can be accepted and recognized by the FST as defined by the designer. For example, in Blackfoot, it might make sense to say that grapheme <ee> could be used to write /i/ or /ii/ if orthographic representation is being influenced by English, but it would not make sense to say <ee> could also be used to write /o/ in the standard orthography. The <ee> grapheme example does not occur in the corpus I am rather using it to illustrate how one might account for the effect of English orthography on Blackfoot spelling. Designers of orthographic relaxers must therefore be able to identify the most common spelling variations for their purpose and create their relaxer accordingly.

I tailored the orthographic relaxer for this project specifically to fit the corpus. As explained in section 4.1, although the corpus was quite homogenous in terms of orthography, there was still some variation and there are some phonological phenomena such as pitch accent mobility that require more research, and so could not yet be modelled. An example of relaxation rules using regular expressions is shown in in Figure 19.

```
1   define relax [ s s s (->) s s,
2   s s (->) s s s,
3   m m (->) m,
```

Figure 19: Example of relaxation rules using regular expressions

Notice that the regular expression rules for orthographic relaxers are written in a similar way to other phonological rules, but with a couple of differences. First, by convention the rules are all defined and compiled as a single rule rather than several rules that can be ordered.[50] Second, the -> symbol used to signify 'becomes' is in parentheses. These are used to express 'can,' so 'becomes' in parenthesis could be read as 'can become.' This creates flexibility, so one orthographic symbol is not required to become another but can if needed. An example of implementing this model in an FST model of Blackfoot is shown in Table 29 (normative sequence is highlighted for comparison).

Table 27: Result of orthographic relaxer implementation

Surface form variation	**Normative surface form**	**Underlying form**
Sssksinoawa	**Ss**ksinoawa	ssksino+VTA+Ind+3Sg+4SgO
isspomookinnaana	isspó**mm**okinnaana	sspommo+VTA+Imp+3Sg+1PlO

The difference between the varied surface form and the normative surface form is that the lengthened /ss/ is represented as /sss/. The underlying form shows that the verb is represented as the normative *ssksino* in the model but can still parse the variation because of the relaxer.

The rules that were most useful for the model were lengthening and shortening alternations of all vowels, for /VVV/, /VV/, and /V/ patterns. The same was done for nasal consonants and /s/, (as shown in Figure 19). Flexibility for the acute pitch-accent symbol was the most important because many words in the corpus were written without any accents. It was also important to capture pitch accent mobility.

The last important variation was syllable deletion. Person prefixes are often deleted by speakers leaving traces of the person morphemes in initial position such as /t/ or /ts/. Syllables in final position are often either devoiced or left out entirely which influences orthography. The model allows suffixes *-wa* and *-yi* to become *-w,* and *-y* respectively, or to just be deleted. Final vowels in general can be deleted, also due to devoicing.

[50] This is the convention, but if feeding and bleeding needed to be accounted for, then the rules could be implemented differently.

4.4.2.3 Morphological modelling choices

Two choices related to morphological modelling that I made are worth mentioning. One is the structure of prenoun and preverb prefixes, and the other involves the relationship between the model and other technological innovations that helped to improve the performance of the model. I discuss these in this section.

At first, I was going to create a model where all preverbs and prenouns could combine freely. When implementing this model on the corpus, I found that doing so along with the orthographic relaxer was causing very unnecessary recursion and producing even more analyses for surface forms than before. An example is with the past tense marker *ii-*. In some words, this morpheme was analysed to appear repeatedly regardless of whether it was even present in the surface form at all. To fix this, I identified that in most cases, the tense, aspect and negation markers (excluding *saw-*) appeared only in the position immediately following person prefixes and could only combine with each other in specific ways. For example, future prefixes never appeared twice in the same word, and two different tense markers (I identified *yáak-,* 'future' *áyaak-,* 'immediate future' *á-,* 'present, durative' and *ii-* 'past' as tense markers),[51] could not co-occur in the same word. However, tense and aspect markers could co-occur in the same word, so it was useful to separate the tense markers from the perfective aspect morpheme *akaa-* and its variations *okaa-* and *Ikaa-*. My solution to this was to divide these prefixes from the other preverbs into a second lexicon and allow them to only appear once in a specific order. That order was negation, tense, and then aspect. I did this by implementing a counter using flags which counted these preverbs and allow them to either appear once in any given word, or to not be present by replacing them with a null prefix.

This technique not only improved the precision of the model, but also sped up the model computationally. This version of the model took only a few minutes to parse the entire corpus, while the previous version never completed parsing it on my machine. Besides its practical advantage, the specific order of negation, tense, and then aspect provided the best results when tested on the corpus. I tried other orders, but the most optimal quantitatively was tense followed

[51] I recognize that the tense preverbs listed here may best be described in other ways. I follow Frantz (2017:33-36) in referring to them as tense markers.

by aspect. Allowing tense to precede aspect only increased the This technique was also essential in the context of the other software I describe in section 4.4.2.3.

Constraining the prefixes was very helpful, but this version of the model still produced the results shown in Table 21 in section 4.3.2, so it still was not very precise. To address this, I developed a simple software program in Python that privileged the most likely parses. I designed the program to count the number of preverbs in each parse, place them in a dictionary and order them in ascending order.[52] It then converts the ordered dictionary into a tuple to keep the position of each parse. The program prints only the nine most likely parses. This method was effective, as proven in section 4.3.2. This program only counts prefixes marked with PV/ or PN/, which are used to mark the other preverbs and prenouns, not the negation, tense, and aspect prefixes. This works because the negation, tense, and aspect prefixes are the most likely to appear, and are already constrained by the morphological model to only appear once each. As a result, if it is ambiguous as to whether the prenoun is the future tense marker *yáak-* or the preverb *áák-* for example, the form with *yáak-* will appear first, and the form *áák-* will appear right after it. This way, all the most likely possible forms are captured, while still allowing for the orthographic flexibility needed to parse the entire corpus.

[52] This can also be done using a weighted model, but the preverb counter was useful to help create a more accurate weighted model than would have otherwise been less accurate if produced by an unconstrained model.

5. Trajectory for the future

In section 5.1, I lay out a possible trajectory for future technological developments using the Blackfoot noun and verb morphological model described in this thesis as a basis. In section 5.2, I discuss some research methods made possible by the model, and some of the gaps in Blackfoot language research which can be filled with the help of the computational model. In section 5.3, I discuss how the technology can be utilized to enhance current technology which is focussed on revitalization resources. Section 5.4 provides a brief conclusion to the whole project.

5.1 Future trajectory for Blackfoot technological development

The research and development for the technology described in this thesis represents the first steps in creating computational tools for Blackfoot. This section offers suggestions for improving the model based on technological innovations that have been implemented in other projects that used the same natural language processing methods described in this thesis. Section 5.1.1 outlines ways that the morphological model can be improved and expanded using the same morphological modelling methods that form the central project described in chapters 2 and 3 of this thesis. Section 5.1.2 provides suggestions for peripheral computational methods that will increase the functionality of the FST model.

5.1.1 Improving the current morphological model

The topics discussed in this section are based on work that has been undertaken by the Alberta Language Technology Lab, and on collaborative projects with other developers and researchers. These are all technologies which improve the functionalities of the computational model for the various purposes that it was designed for.

Improving the current model and expanding it to analyse other parts of speech should be the first step in ongoing development. The noun and verb model described in this thesis represents a significant achievement for Blackfoot technological development. As shown in chapter 3 of this thesis, nouns and verbs represent a huge proportion of the words used in Blackfoot, but there are still some words, and one significant word class that is not yet accounted

for in the model (demonstratives). Here I discuss improvements that can be made to the Blackfoot noun and verb model and the research required to make these improvements. I end the subsection by making recommendations on how the model can be expanded to include free words and demonstratives.

The research on Blackfoot grammar is extensive enough that this project to create a computational model of the language was possible. However, there are still gaps in the research which posed challenges to creating the computational model and could not be addressed in this thesis. The first issue is pitch accent mobility. This phenomenon is seen in some stems which contain a high pitch syllable when they are inflected. One example of this pattern is when the pitch accent moves between the stem and the person prefix. Some examples of this phenomenon are shown below in example (22), showing pitch accent variations for the transitive animate verb stem *ssksini*.

(22) a. Issksinímayi
(i)ssksini-ma-yi
know.TI-3-0
'He knows it.' (Blackfoot Online Dictionary https://dictionary.blackfoot.atlas-ling.ca/#!/results)

b. Nítssksinihpa komohtsspíyihpi
nit-ssksini-hp-wa k-omoht-ihpi-hp-yi
1-know.TI-THEME-SG 2-means-dance.AI-CONJ-IN.SG
'I know why you dance.' (Frantz 2017:158)

The examples demonstrate the currently unpredictable nature of this pitch accent mobility. The shifting might be due to a phonological condition, or a morpho-phonological condition, but if that is the case, the condition is undescribed. Researching this phenomenon with the goal of understanding its pattern would help to narrow the phonology of the model, thereby improving its ability to produce narrow, accurate parses of Blackfoot word forms. A recent study of the nature of pitch accent mobility is found in Weber and Shaw (2021), who argue that Blackfoot fits within a mobile pitch accent typological framework rather than a lexicalized pitch accent

framework. This study represents an early step in understanding the nature of pitch accent shift in Blackfoot.

If the shift is phonologically conditioned, then this can be accounted for in the phonology and applied universally to the rewrite rules. If it is a morpho-phonological issue, then this can be accounted for in one of two ways. The most ideal solution, if it were possible, would be to create a more specific morpho-phonological rewrite rule which only targets morphemes of a specific shape. The other solution would be to mark which stems undergo this shift with a computational trigger, similarly to how this was done regarding initial allomorphy (see chapter 4, section 4.4.2).

Besides these issues, there may be further phenomena discovered in the field of Blackfoot linguistics that may require modifications to the model. The tests undertaken in chapter four sections 4.2 and 4.3 have shown that the model is very effective at what it was designed to do (that is, to parse Blackfoot noun and verb word forms accurately and precisely). However, that does not mean that future researchers may not still find ways to further narrow the model or to make it more efficient. This process of improvement will undoubtedly be ongoing, so while that is being done, further development can be undertaken to expand the model.

For the development of a complete computational model of Blackfoot morphology, demonstratives should be considered to represent a third syntactic and inflectional class of stems in Blackfoot alongside nouns and verbs. The demonstrative stems are inflected for animacy and plurality in the same way that nouns are. The animacy and number of the demonstrative must agree with the animacy and plurality of the noun it precedes. The same animacy and plural suffixes used for nouns are also used for demonstratives. Example (23) demonstrates this.

(23) a. Amoistsi míínistsi iikááhsiiyiaawa

amo-istsi miin-istsi iik-aahsii-yi-aawa

DEM-IN.PL berry-IN.PL very-good-3/0PL-PRO

‘These berries are very good.’ (Frantz 2017:70)

b. Áaksowatayi ámoksi

yáak-Iowat-a-yi amo-iksi

FUT-eat-DIR-3/0PL DEM-AN.PL

‘We will eat these.’ (Frantz 2017:70)

Example (23a) shows a demonstrative agreeing with an inanimate plural noun, and example (23b) shows a demonstrative being used as a deictic expression.

Blackfoot demonstratives can also take a variety of other suffixes which surface following the animacy suffix. These can express many different pieces of information about the nouns. Schupbach evaluates some of these suffixes as *-ma,* for stationary nouns *-ya,* for nouns that are moving along a path, and *-hka,* for nouns that are not visible to the speaker or far from the speaker (Schupbach 2012:9-10).

The final step in creating a full model of Blackfoot would be to design a model that can parse Blackfoot particles and pronouns such as the first and second person pronouns, *nistoowa* and *kistoowa*. There are very few such particles in Blackfoot (Bliss and Wiltschko 2022:139-142).

5.1.2 Improving functionality using peripheral methods

There are many advantages to using Finite State Models of grammars that are designed using the LEXC formalism. One advantage is that advanced technologies can be built using the language models to improve their functionality. These can range from simple orthographic transliteration tools to machine learning tools. Most of the tools discussed in this section have been or are in the process of being implemented in various other language modelling projects such as the Plains Cree project. Some peripheral tools have already been implemented and described in this thesis and include the Python program that was designed to increase the precision of the model by counting preverbs and prenouns, and the weighted morpheme model. This section outlines further methods and techniques that will provide further improvements to the model.

The first functionality tool that can be implemented is a simple orthographic transliteration tool. These can be used for languages which utilize multiple orthographies and allows the technology to not privilege one orthography over another. For example, Plains Cree utilizes two standard systems, the Standard Roman Orthography (SRO; Okimâsis 2018:3-7), and the Syllabic system. The SRO is widely used by academics and is easier to learn for literate English speakers because it utilizes the same characters used in English's Latin alphabet, making it popular among second language learners. The Syllabic system was developed by missionaries

during the early colonization of North America and was adapted for use in several Indigenous languages including Cree, Inuktitut, Blackfoot, Ojibway, and some Dene languages.[53] Both Roman and syllabic systems are used to varying degrees in all the communities mentioned, so it is useful to have transliteration software that allows for text that is written in either orthography. Such software was developed by the Alberta Language Technology Lab for Plains Cree (Littell et al. 2018:2623). I suggest that the same methods used for Plains Cree and other languages can also be used for Blackfoot.

A model that can parse text in various orthographies can be accomplished by designing a simple program that uses REGEX rules to transliterate the standardized Roman orthography which is used in the morphological and phonological model to other systems such as the Syllabic system. Blackfoot speakers may utilize a variety of other orthographic systems, most of which utilize Roman characters, but don't conform to the orthographic conventions developed by Frantz (2017). Figure 21 below provides a simplified example of how Frantz's orthography can be converted to Syllabic script using regex rules.

```
1  {pi(i)} -> ᑯ
2  {ka(a)} -> ᑳ
3  {ni(i)} -> ᓀ
```

Figure 20: Frantz' Blackfoot orthography to Syllabics example

The script in Figure 21 can be used convert the name for the Blackfoot nation *Piikani* to ᑯᐦᓀ, the syllabic transliteration of the word according to this translation by Stocken (1953).[54] Other orthographic systems could be implemented using the exact same technique of mapping characters and common character pairs or clusters with their orthographic equivalents in the other system.

[53] Cree was the first. Methodist missionary Reverend James Evans is credited with developing it around 1840 (Mclean 1890:160–175). The Cree system had a remarkably high literacy rate early on, and it spread to the speakers of the Dene language Carrier and was further adapted for Carrier by Father Adrien-Gabriel Morice in 1885 (Poser 2007:1–2). Poser (2007) also reports that syllabic systems had already been developed for Inuktitut and Dogrib by that point.

[54] I acknowledge that there are other syllabic orthographies developed for Blackfoot. I use this one only to illustrate my example.

I see such technology as having two advantages. First, it would allow researchers to analyze corpora utilizing a variety of orthographies. The other is that it would allow Blackfoot learners and speakers the freedom to utilize whichever orthographic system they feel is best. However, for the purpose of orthographic standardization and development, it may not be ideal to design an orthographic transliterator, especially for the competing Latin-based orthographies, but I would leave that decision to the Blackfoot communities that utilize the technology.

Another technology that could be developed is a translator that would work in a similar way to the orthographic transliterator. Such a transliterator would transform the underlying analysis produced by the FST into something more user-friendly for people with training in standard linguistic glossing, such as a parsed text utilizing the Leipzig Glossing rules (Comrie, Haspelmath and Balthasar 2015) or, in a more advanced form, into English words and phrases. This could be done by designing an FST program that transforms the underlying forms of stems and morphemes into something else on a 1:1 basis. For example, the underlying stem *ooyi* could be matched with the English word *eat*, the prefix *yáak-* with *will,* and the prefix *nit-* with *I.* The result would be an FST with the following input and output shown in Table 31.

Table 28: Simple translational FST

Surface form	**Underlying Analysis**	**Translation**
Nitaaksoyi	Fut+ooyi+VAI+Ind+1Sg	'I will eat'

In this case, the translation is made by matching the components of the underlying analysis to English words, and defining the order in which they appear. The same techniques can be used to produce a variety of other, more complex grammatical phrases.

The third technology moves away from translational FST software and into more quantitative, statistically based development methods. In this project, I have already implemented a simple, supervised, weighted machine learning model that is integrated with the morphological model (described in detail in section 4.4.1). The weighted model can be used to produce search suggestion bars and spell checkers by determining the most likely word-forms being typed or searched by users and suggesting those word forms. Word-forms that are probable will be the first to be suggested, and less probable ones will appear further down. Search suggestion bars are commonly used in search engines to facilitate faster searches and can be utilized in tools such as dictionaries or keyboards. These tools are implemented for well-

resourced languages by using statistical and machine learning models that are trained on huge corpora, and as a consequence, can use more advanced techniques such as n-gram models which take into account the probability of two words cooccurring. More advanced techniques will refine the Blackfoot model but cannot be implemented until the corpus has been fully parsed and hand-checked by a linguist because those models require highly accurate parses to learn from. This technology is already in the process of being implemented for Plains Cree (Lane, Harrigan, and Arppe 2022).

The final technology is a machine learning model. Machine learning is a technique that is used for a variety of purposes in language technologies, particularly for well-resourced languages[55]. For example, machine learning models are being used to produce English language chatbots that can be used to answer routine questions about health, and even producing novel text on a variety of subjects. A specific example of this technology is OpenAI's ChatGPT software (OpenAI n.d.). Creating chatbots such as this requires very large databases of English to ensure they are as pragmatically accurate, and as syntactically and semantically coherent as possible.

Highly advanced machine learning technology such as Blackfoot chatbots will not be possible until huge amounts of corpus data is collected on a variety of subjects. However, smaller projects can be used to augment current technologies. A study was conducted by researchers at the University of Colorado on the Algonquian language Arapaho exploring how a supervised machine learning model could be used to predict inflected word forms and their underlying analyses, essentially creating a learning-based FST (described in Moeller et al. 2018). The model learned from a handmade Arapaho FST, like the one designed for Blackfoot described in this thesis. It used pairs of surface forms and their underlying analyses as parsed by the model to replicate those pairs when faced with new data. The authors argue that the model can be used to capture linguistic innovations and previously undocumented stems and morphemes, which may not be represented in the dictionary database or in the corpus.

Machine learning technology that has been developed for well-resourced languages can also be leveraged to increase the usability of Blackfoot technologies. For example, researchers developing Plains Cree technology used an Artificial Intelligence tool to retrieve semantically

[55] Well-resourced languages have large amounts of data and are well described in the linguistic literature. This term contrasts with under-resourced languages which have relatively smaller amounts of data that has been collected and are not as well described in the linguistic literature, and this term exists on a spectrum. For example, English would be considered a well-resourced language and Blackfoot would be considered a relatively less-resourced language.

associated terms in dictionary searches (Arppe et al. 2023, Harrigan and Arppe forthcoming). The program utilized a machine learning program that learned from English data that was used to expand the search capabilities of the Plains Cree online dictionary. The implementation of this technology allows English to Plains Cree dictionary searches to return lexical entries for semantically associated words or synonyms if the word being searched cannot be found in the dictionary database.

Phonetic technologies also provide opportunities for increased usability of Blackfoot technologies. For example, researchers at the ALT lab designed a speech to text synthesizer for Plains Cree. Speech synthesizers are useful for Algonquian technologies because, if the product is accurate and precise, it allows users to hear how inflected word forms should pronounced (Harrigan, Arppe, and Mills 2019).

5.2 The morphological model as a parsing tool

One of the important contributions of this thesis project was the curation of a digital corpus of modern, standardized Blackfoot texts. The corpus was gathered to be used in the testing stage, and to inform development choices (described in detail in chapter 4). This corpus represents the largest body of curated modern Blackfoot texts, and contains only content that was elicited from, translated, or produced by native Blackfoot speakers. These factors make it usable for further research purposes, in conjunction with the morphological model.

A morpho-syntactically tagged corpus of Plains Cree is described in Arppe et al. (2020). The researchers created this corpus using an inflectional morphological model alongside a second model designed to disambiguate by taking syntactic relationships between words into account (Arppe et al. 2020:3). These resources were more advanced than the Blackfoot model when the paper was written, so some of their methods will not be possible to implement for Blackfoot until later, but the general method of using a morphological model can be applied early on.

Chapter 4 of the thesis provides evidence that the Blackfoot morphological model can be used as a morphological parser. In fact, by using a corpus to test the model, it can be considered first and foremost, a parsing tool. This is its most important function in the early stages of development because creating larger corpora of Blackfoot text will help to improve the FST in the future, which in turn will help to improve its parsing capabilities and expand functionalities

(see section 5.1.2). In this section, I provide a method to parse the corpus described in section 4.1, with the hope that future researchers will use one of these methods.

In order to parse the text, one would run the model with the corpus so that each word is parsed in its context. A team of humans could then determine the correct parse given the context. The output of this technique would appear as shown in Figures 22 and 23, where Figure 22 is the original text from the corpus, and Figure 23 is the parsed text.

```
Nitaakohtsi'poyi ninna Akaohkitopiiwa. Onni anistayini Piaiikksspitowayi.
Oksisstsi anistayini Saaomitsikkanaayi. Otohkiimaani Issitaakii.
```

Figure 21: Example of original corpus text [from Russell and Genee (2014)]

```
"<Nitaakohtsi'poyi>"
        +?

"<ninna>"
        "ninna" ninna+NAD+Px3Sg+Obv ninna+NAD+Px3Sg+Non

"<Akaohkitopiiwa>"
        "+?"

"<.>"

"<Onni>"
        "ninna" ninna+NAD+Px3Sg+Obv ninna+NAD+Px3Sg+Non

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

"<Piaiikksspitowayi>"
        +?

"<.>"

"<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 22: Example of parsed text from the corpus

In this output, the wordform from the original text appears between the less-than and greater-than signs. Below that, in double quotes are the different lemmas that the model evaluated as

being present in the wordform. On the immediate right of the lemmas are the separate parses; each individual parse is separated with a space, and periods keep their position to mark sentence boundaries. The '+?' symbol marks a word-form that was not recognized by the parser.

To achieve this output, I used the same program that limits the number of acceptable parses per word and changed it, so it only returns the three parses with the least number of prenouns or preverbs. I did this to keep the example from becoming too large. In practice, I would suggest using a number of parses that has been proven to be likely to include at least one correct parse. For more computational utility, it may be useful to add features that more explicitly mark word, sentence, and text boundaries.

This human-based parsing method is adequate in theory, but in practice it may be unrealistic. Alternatively, or in addition to the human-based parsing method, a constraint grammar model could be used to determine the most likely parse for each word. Constraint grammar models take syntactic contexts into account to determine the probability that a given word is more likely to be parsed as a noun or a verb (for example). A constraint grammar model was applied to a parsed Plains Cree corpus by Schmirler (2022).

In section 5.1, I suggest the development of a translational FST technology that transforms the generated analyses into English translations or linguistic glosses. Such a tool could transform the computational underlying analyses shown in Table 32 into standardized linguistic glosses, making the parsed corpus more usable for linguists. An example of this transformation is shown in Table 32.

Table 29: Translational gloss FST

Surface form	**Underlying Analysis**	**Gloss**
nitaaksoyi	Fut+ooyi+VAI+Ind+1Sg	1-FUT-eat.VAI

The transformational software would work as described in section 5.1.2 for Table 31, but instead of producing an English translation, would produce a standardized linguistic gloss. This would make the job of corpus editors even easier, as the corpus would conform to linguistic glossing standards.

5.3 The use of technology in Blackfoot education and revitalization

The model's usability to the Blackfoot speaking nations was one of the most important considerations for the modelling project. The model is designed so that it can be used to aid revitalization efforts. In section 5.3.1, I discuss how the model might be used to augment current technology from the Blackfoot Online Dictionary project. In section 5.3.2, I discuss some other practical uses for the technology.

5.3.1 Integrating the model for digital Blackfoot lessons

Many digital resources have been developed by Blackfoot nations, and by other organizations and individuals, designed to aid learning, (see https://blackfoot.algonquianlanguages.ca/resources/apps/ for a list of some of these resources, and overviews of their functions and features). In this section, I discuss how educational resources linked to the Blackfoot Online Dictionary project may be improved with the help of the morphological model.

In section 1.3 above, I describe how this project supports the Blackfoot dictionary project by allowing us to integrate intelligent dictionary capabilities like the ability to recognize inflected word forms. This will help to aid dictionary users to find complex word forms that they might find in context, thus aiding them in their searches. Future developments may also allow users to search for English phrases and return inflected Blackfoot words which match the English phrase (see section 5.1.2 above).

Besides a dictionary search function, the Blackfoot Online Dictionary (Frantz and Genee 2016-2023) has several other features and capabilities. One feature is a digital activity-based lesson page. This page allows collaborators to design learning activities for students of the language by manipulating pre-set formulas to put together different types of activities. The concept is good, but the development progress is slow because each lesson must be designed using linguistic resources that are accessible. I would propose that a morphological component to this resource could be developed faster by utilizing the model.

Lesson developers could define a morphological concept that they might want to teach, and then use the model to parse a corpus, and find word forms that relate to that morphological

concept. For example, the developer could search for words inflected with the conjunctive paradigm and use what they find to create activities. Currently, creating a lesson like this requires research to collect and verify the validity of forms. For example, there are a comprehensive set of lessons using the stem *oowatoo* to help teach the most common verbal inflections. This lesson is made possible because there is already extensive documentation of *oowatoo* in the indicative paradigm on the site. See (Bontogon et al. 2018) for an example of computer-assisted learning in Plains Cree.

Once lesson materials have been collected (examples, stories, entries, etc.), there are several different types of activities that can be made using those resources. Some of the morphological activities made possible by the site include matching games, where users could match inflected words to their translations, or to their linguistic analysis. Another might be a multiple-choice quiz where someone can identify different morphological features present in inflected word forms such as what mode a verb has been inflected in. Figures 24 and 25 show an example of a multiple-choice question where the user must identify the mode that the verb is inflected in, and a matching game where the user must match the correct English translation to each stem within inflected verb stems.

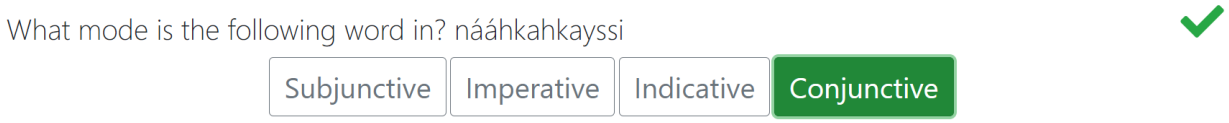

Figure 23: Example of a multiple-choice question informed by the corpus.

Figure 24: Example of a matching activity informed by the corpus.

I created the multiple-choice question in Figure 24 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 25, I 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. Based on these tests, 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.

5.3.2 Supporting the use of Blackfoot by individuals

I have discussed how the Blackfoot model can be used by researchers, and how it can be used to develop digital lessons using the Blackfoot lesson development tool. The last thing that I discuss is how the model can be implemented in everyday technology. The morphological model can also be used as the basis for spell checkers and next-word suggestions, or, in more accurately in the case of the Blackfoot model, next-morpheme suggestions. These can be implemented on smartphone keyboards and computer text-editing software. The spell checkers would use the model to generate normative surface forms and corresponding underlying forms. The software would check words as they are being typed by the user and use the generated surface forms to regularize the orthography. Just like with English spelling tools that one encounters on text editing and smartphone software, the words could be underlined in red, and spelling relaxers would be used to identify likely forms that conform to the standard orthography.

The suggestion bars would work in a similar way. For well-resourced languages such as English, statistical methods are used and applied to large corpora to find likely phrases and identify likely next words to the user. For Blackfoot, the morphological model will make this possible alongside simple machine learning implementations, such as the weighted modelling described in section 4.4.1. Since Blackfoot is an agglutinating language, next morpheme suggestions would be provided instead of next word suggestions. This feature could be applied to smartphone keyboards that could be made available to download and would more immediately be implementable on the dictionary website. This would make both texting and searching for terms in the dictionary faster.

5.4 Conclusion

To conclude, I developed a computational model of Blackfoot noun and verb morphology by drawing from computational and documentary linguistic works that came before, and from computational modelling work that was done for the related Plains Cree language as a basis. I would suggest that the methods I used in this thesis, from initial development to testing, could be used for other under resourced languages to begin the technological development process. This project is a starting point for new avenues in research and development in Blackfoot language technology.

In chapter 5, I provided some suggestions for some of the potential implementations of the model in the future, but I have no doubt that the model's uses could go beyond what I discussed. My hope is that the model will continue to be improved, and that it will be used to support the documentation, description, and revitalization of Blackfoot for years to come, and bring the study of the Blackfoot language into this age of digital technology.

Summary

The central goal of this thesis was to describe a project that made use of natural language processing techniques to aid in both the documentation and revitalization of the Blackfoot language. This section of the concluding chapter provides a less technical summary of the thesis to aid understanding for anyone who wishes to access the information in this thesis but does not have some of the required background knowledge to do so.

In chapter 1, I discuss the history and status of Indigenous languages in Canada. Using statistics found in the Canadian census, I argue that there is a general need for linguists to make an effort to take part in the revitalization of the languages which they study. This provides a context for this project as part of a larger project that is seeking to aid in language stabilization by providing digital tools that can be used for various purposes by those who are involved in research, revitalization, or both.

Section 1.1 gives a general introduction to Indigenous languages in Canada and section 1.2 provides a more specific discussion of the history, culture, and linguistic situation of the Blackfoot people. It provides a historical explanation for how the Blackfoot language began to decline in use, and situates it in its linguistic family, the Algonquian language family. This is followed by a brief overview of some of the common features of Algonquian grammar, and then a summary of the status of the Blackfoot language using statistics from the Canadian census.

Section 1.3 discusses the natural language processing method that I used in this project, that is, morphological modelling using a finite-state transducer. This method involves using software programming tools that were developed so that linguists could describe the internal structure of words (referred to generally as *morphology* throughout the text) in a given language and implement them into software. The tools operate as finite-state transducers, which are computational machines made up of a series of states and networks that can map two different sets of symbols to each other. In the case of this project, one set of symbols are the actual Blackfoot words as they would be written, and the corresponding set of symbols would represent the underlying internal structure of those written words.

Section 1.4 provides a detailed description about how finite-state transducers can be developed. It provides a simplified example of how Blackfoot verb inflection can be modelled using the LEXC formalism, and how Blackfoot phonological rules can be implemented and

combined with a morphological model to create a full computational model of Blackfoot word structure. The result is a small grammatical model that can generate acceptable words and can accurately analyze them.

Chapters 2 and 3 are devoted to describing the structure of Blackfoot noun and verb morphology respectively, and how I modelled them. This is the central component of this project, as the chapters describe the structure of the model and forms the basic technology that was designed to support the development of many advanced technological tools. I developed the models using descriptive resources of Blackfoot grammar as guides to provide the structure of the model.

Chapter 2 describes the model of Blackfoot noun structure. Section 2.1 provides a brief but detailed overview of Blackfoot noun inflection. It describes how nouns can be inflected for animacy, number, specificity, obviation and possession, and the prestems they can take. The section ends with a discussion of how verb stems can also be inflected as nouns using nominalizing suffixes. Section 2.2 compares Blackfoot noun morphology with Plains Cree noun morphology because the two languages are structurally very similar to each other, and the Blackfoot model was developed using the Cree model as a basis. Section 2.3 provides a description of the noun model by breaking it down into two paths, one that follows the regular noun morphology, and one indirect path that can derive nouns from verb stems.

Chapter 3 has the same structure as chapter 2 but focusses on the modelling of Blackfoot verb structure. Section 3.1 briefly details Blackfoot verb morphology, including person agreement, the types of prefixes and suffixes that verbs can take, and how verbs mark the roles that each of their arguments take. It also provides explanations for the Blackfoot verb modes and how they are used and describes how pronouns can appear attached to the ends of verbs in some environments. The section ends with an explanation of how different types of nouns can be derived from verb stems. Section 3.2 compares Blackfoot verb morphology to Plains Cree verb morphology. Section 3.3 describes how the information in sections 3.1 and 3.2 was implemented into the modelling software, first describing the direct path to inflect verb stems, and then the indirect path to derive verbs from noun stems.

Chapter 4 was devoted to the testing of the model described in chapters 2 and 3. I describe a corpus that I curated from various publicly available digital sources, and which was used extensively in the testing process in section 4.1. Sections 4.2 and 4.3 provide two different

but complementary methods for testing the model. 4.2 reports on a paradigm test where I manually entered full Blackfoot noun and verb paradigms into markup files. The results showed that the model was capable of accurately replicating about 99.3% of the underlying analysis-surface form pairs represented in the files when used with a simple pitch accent relaxer software. The relaxation component was needed to account for pitch accent mobility and discrepancies between the pitch accents' locations on stems represented in the morphological model files, and their locations in test files.

I describe the results of a different testing method in section 4.3. This test used the entire corpus to determine how many of the unique words found in the corpus could be parsed by the morphological model. The results were very encouraging with over 70% of the 13,786 unique word forms being parsed by the model. A sample was hand checked by me to identify whether the parses were accurate, and to determine the most common mistakes made by the model. The tests showed the model was not precise and was producing too many possible parses for each word, so I created a Python script that privileged parses with lower prefix counts, and that helped to increase the precision of the analyses produced by the model while ensuring that the privileged parses were accurate. I report on another hand checked sample that showed that 92% of the parses from a random sample of 300 parsed words had a correct parse in the top 9 results.

Section 4.4 describes other features of the model, some of which were informed by the corpus. This section describes the trigger system that I added to the phonological model to deal with allomorphic issues. It also describes the orthography relaxer that I created to be used with the model in order to deal with certain phonological phenomena and with common orthographic variations found in the corpus. It ends with a description of the morphological modelling choices that I made which were informed by testing the model on the corpus rather than just using documentation resources.

Chapter 5 provides a roadmap for the future of the project described in this thesis. In section 5.1, I describe the next step in Blackfoot modelling. The noun and verb models are good starts, but they still need to be refined, and more models for other parts of speech, particularly of demonstratives and pronouns, will be essential. 5.2 provides suggestions for some of the different digital technologies that the model can be used as a basis to create. Section 5.3 discusses specific uses for the technology in educational and revitalization efforts. The chapter ends with a summary of the entire thesis.

This project provides a contribution to Blackfoot language research and computational modelling. By providing a noun and verb model and by testing it on a large and orthographically homogenous corpus of modern Blackfoot, this project produced academic tools and applied resources for Blackfoot morphology.

Bibliography

Arppe, Antti, Lene Antonsen, Trond Trosterud, Conor Snoek, Dorothy Thunder, Atticus Harrigan, Jordan Lachler, Jean Okimâsis, and Arok Wolvengrey. 2014. 'Linguistic Insights from Computational Modeling of Plains Cree Morphology'. In *Papers of the Forty-Sixth Algonquian Conference*, edited by Monica Macaulay and Margaret Noodin. Vol. 46. Uncasville, Connecticut: MSU Press.

Arppe, Antti, Andrew Neitsch, Daniel Dacanay, Jolene Poulin, Daniel Hieber, and Atticus Harrigan. 2023. 'Finding Words That Aren't There: Using Word Embeddings to Improve Dictionary Search for Low-Resource Languages'. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, edited by Manuel Mager, Abteen Ebrahimi, Arturo Oncevay, Enora Rice, Shruti Rijhwani, Alexis Palmer, and Katherina Kann, 144–55. Toronto, Ontario: Association for Computational Linguistics.

Arppe, Antti, Katherine Schmirler, Atticus Harrigan, and Arok Wolvengrey. 2020. 'A Morpho-Syntactically Tagged Corpus for Plains Cree'. In *Papers of the Forty-Ninth Algonquian Conference*, edited by Monica Macaulay and Margaret Noodin, 49:1–16. East Lansing, Michigan: MSU Press.

Arppe, Antti, Katherine Schmirler, Miikka Silfverberg, Mans Hulden, and Arok Wolvengrey. 2019. 'Insights from Computational Modelling of the Derivational Structure of Plains Cree Stems.' In *Papers of the Forty-Eighth Algonquian Conference*, edited by Monica Macaulay and Margaret Noodin, 48:1–18. East Lansing, Michigan: MSU Press.

Beek, Shoukia van. 2016. 'Intersection: Indigenous Language, Health and Wellbeing'. Literature survey. British Columbia: First Nations Cultural Council.

Beesley, Kenneth, and Lauri Karttunen. 2003. 'A Gentle Introduction'. In *Finite State Morphology*, First Edition, 1–37. CSLI Publications.

Berman, Howard. 2006. 'Studies in Blackfoot Prehistory'. Edited by David Beck and Doris Payne. *International Journal of American Linguistics* 72 (2): 264–84.

Bliss, Heather, and Will Oxford. 2017. 'Patterns of Syncretism in Nominal Paradigms: A Pan-Algonquian Perspective'. In *Papers of the Forty-Sixth Algonquian Conference*, edited by Monica Macaulay and Margaret Noodin, 46:1–16. Uncasville, Connecticut: MSU Press.

Bontogon, Megan, Antti Arppe, Lene Antonsen, Dorothy Thunder, and Jordan Lachler. 2018. 'Intelligent Computer Assisted Language Learning (ICALL) for Nêhiyawêwin: An In-Depth User Experience Evaluation'. *Canadian Modern Language Review* 74 (3): 337–62.

Bowers, Dustin, Antti Arppe, Jordan Lachler, Sjur Moshagen, and Trosterud Trond. 2017. 'A Morphological Parser for Odawa'. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, 2:1–9. Honolulu, Hawaii: Association for Computational Linguistics.

Comrie, Bernard, Martin Haspelmath, and Balthasar Bickel. 2015. 'Leipzeig Glossing Rules: Conventions for Interlinear Morpheme by Morpheme Glosses'. Department of Linguistics of the Max Planck Institute for Evolutionary Anthropology & the Department of Linguistics of the University of Leipzig.

Czaykowska-Higgins, Ewa. 2009. 'Research Models, Community Engagement, and Linguistic Fieldwork: Reflections on Working within Canadian Indigenous Communities'. *Language Documentation and Conservation* 3 (1): 15–50.

Daschuk, J.W., Paul Hackett, and Scott MacNeil. 2012. ‘Treaties and Tuberculosis: First Nations People in Late-Nineteenth-Century Western Canada, A Political and Economic Transformation’. In *Aboriginal History: A Reader*, edited by Kristen Burnett and Geoffrey Read, First Edition, 71–80. Oxford University Press.

Dunham, Joel. 2014. ‘The Online Linguistic Database: Software for Linguistic Fieldwork’. Doctoral Dissertation, Vancouver, British Columbia: University of British Columbia.

Frantz, Donald. 1991. *Blackfoot Grammar*. First Edition. Toronto: University of Toronto Press.

———. 2017. *Blackfoot Grammar*. Third Edition. Toronto: University of Toronto Press.

Frantz, Donald, and Inge Genee. 2016. ‘Blackfoot Online Dictionary’. Blackfoot Online Resources. 2023 2016. https://dictionary.blackfoot.atlas-ling.ca/#!/results.

Frantz, Donald, and Norma-Jean Russell. 1989. *The Blackfoot Dictionary of Stems, Roots and Affixes*. Third Edition. Toronto: University of Toronto Press.

———. 2017. *The Blackfoot Dictionary of Stems, Roots and Affixes*. First Edition. Toronto: University of Toronto Press.

Genee, Inge. 2009. ‘From the Armchair to the Field and Back Again: C.C. Uhlenbeck’s Work on Blackfoot’. *Canadian Journal of Netherlandic Studies* 29 (2): 105–27.

———. 2020. ‘It’s Written Niisto, but It Sounds like KNEE STEW: Handling Multiple Orthographies in Blackfoot Language Web Resources’. Edited by Dorit Ravid, Elitzur Dattner, Terry Joyce, Beatrice Primus, Rachel Schiff, and Liliana Tolchinsky. *Written Language and Literacy* 23 (1): 1–28.

Genee, Inge, and Marie-Odile Junker. 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.

Gerdts, Donna. 1998. ‘Beyond Expertise: The Role of the Linguist in Language Revitalization Programs.’ In *Endangered Languages: What Role for the Specialist?*, edited by Nicolas Ostler, 2:13–22. Edinburgh, Scotland.

Greenwood, Davydd, and Morten Levin. 1998. *Introduction to Action Research: Social Research for Social Change.* Thousand Oaks, CA: Sage Publications.

Haig-Brown, Celia. 2012. ‘Always Remembering: Indian Residential Schools in Canada’. In *Aboriginal History: A Reader*, edited by Kristen Burnett and Geoffrey Read, First, 221–23. University of Toronto Press.

Harrigan, Atticus, and Antti Arppe. 2022. ‘Leveraging Majority Language Resources for Plains Cree Semantic Classification.’ In *Papers of the Fifty-Second Algonquian Conference (PAC52)*, edited by Monica Macaulay and Margaret Noodin, 52:1–20. Madison, Wisconsin: MSU Press.

Harrigan, Atticus, Antti Arppe, and Timothy Mills. 2019. ‘Preliminary Plains Cree Synthesizer’. In *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-3)*, edited by Antti Arppe, Jeff Good, Mans Hulden, Jordan Lachler, Alexis Palmer, Lane Schwartz, and Miikka Silfverberg, 3:64–73. Honolulu, Hawaii.

Harrigan, Atticus, Katie Schmirler, Antti Arppe, L Antonsen, Sjur Moshagen, T Trosterud, and Arok Wolvengrey. 2017. ‘Learning from the Computational Modelling of Plains Cree Verbs’. *Morphology* 27 (4): 565–98.

Heavy Shields Russell, Lena, and Inge Genee. 2014. *Ákaitsinikssiistsi: Blackfoot Stories of Old*. Regina: University of Regina Press.

Hulden, Mans. 2009. ‘Foma: A Finite-State Compiler and Library’. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session*, edited by Alex Lascarides, Claire Gerdent, and Joakim Nivre, 29–32. Athens, Greece: Association for Computational Linguistics.

Johnson, Ryan, Lene Antonsen, and Trond Trosterud. 2013. ‘Using Finite State Transducers for Making Efficient Reading Comprehension Dictionaries’. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, edited by Stephan Oepan, Kristin Hagen, and Janne Bondi Johannessen, 19:59–71. Oslo, Norway.

Junker, Marie-Odile, and Delasie Torkornoo. 2017. ‘Algonquian Dictionaries Project: Common Digital Infrastructure for Lexicography and Language Documentation’. In *Canadian Society for Digital Humanities*. Ryerson University, Toronto.

Kingston, Lindsey. 2015. ‘The Destruction of Identity: Cultural Genocide and Indigenous Peoples’. *Journal of Human Rights* 14 (1): 63–83.

Lane, William, Atticus Harrigan, and Antti Arppe. 2022. ‘Interactive Word Completion for Plains Cree’. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, edited by Smeranda Muresan, Preslav Nakov, and Aline Villavicencio, 1: Long Papers:3284–94. Dublin, Ireland.

Lindén, Krister, Eric Axelson, Senka Drobac, Sam Hardwick, Juha Kuokkala, Jyrki Niemi, Tommi Pirinen, and Miikka Silfverberg. 2013. ‘HFST - A System for Creating NLP Tools’. Edited by Cerstin Mahlow and Michael Piotrowski. *Communications in Computer and Information Science* 380 (Systems and Frameworks for Computational Morphology): 53–71.

Littell, Patrick, Anna Kazantseva, Rowland Kuhn, Aidan Pine, Antti Arppe, and Marie-Odile Junker. 2018. ‘Indigenous Language Technologies in Canada: Assessment, Challenges, and Successes’. Edited by Emily Bender, Leon Derczynski, and Pierre Isabelle. *Proceedings of the 27th International Conference on Computational Linguistics*, 2620–32.

Little, Carol-Rose. 2018. ‘Inanimate Nouns as Subjects in Mi’gmaq: Consequences for Agreement Morphology’. *Proceedings of WSCLA 21*, 127–41.

McIvor, Onowa. 2005. ‘The Contribution of Indigenous Heritage Language Immersion Programs to Healthy Early Childhood Development’. *Research Connections Canada: Supporting Children and Families* 12: 5–20.

Mclean, John. 1890. ‘The Syllabics System of the Cree Language’. In *James Evans: Inventor of the Syllabic System of the Cree Language*, 160–74. Toronto: Methodist Mission Rooms.

Miyashita, Mizuki, and Annabelle Chatsis. 2013. ‘Respecting Dialectal Variations in a Blackfoot Language Class’. Edited by J Reyhner, J Martin, L Loackard, and W.S. Gilbert. *Honoring Our Elders: Culturally Appropriate Approaches for Teaching Indigenous Students*, 109–16.

Moeller, Sarah, Ghazaleh Kazeminejad, Andrew Cowell, and Mans Hulden. 2018. ‘A Neural Morphological Analyzer for Arapaho Verbs Learned from a Finite State Transducer’. In *Proceedings of Workshop on Polysynthetic Languages*, edited by Emily Bender, Leon Derczynski, and Pierre Isabelle, 27:12–20. Santa Fe, New Mexico: Association for Computational Linguistics.

Moshagen, Sjur, Tommi Pirinen, and Trond Trosterud. 2013. ‘Building an Open-Source Development Infrastructure for Language Technology Projects’. In *19th Nordic Conference on Computational Linguistics*, edited by Stephan Oepan, Kristin Hagen, and

Janne Bondi Johannessen, 19:343–52. Oslo, Norway: Linköping University Electronic Press, Sweden.
Okimâsis, Jean. 2018. *Cree, Language of the Plains: Nēhiyawēwin, Paskwāwi-Pīkiskwēwin*. University of Regina Press.
OpenAI. n.d. ChatGPT, filed 2023. https://chat.openai.com.
Oxford, Will. 2019. 'Algonquian Languages'. In *The Routledge Handbook of North American Languages*, edited by Daniel Siddiqi, Michael Barrie, Carrie Gillon, Jason D. Haugen, and Éric Mathieu, 1st ed., 504–24. Routledge.
Poser, William. 2007. 'The Carrier Syllabics'. 1. Yinka Dene Language Institute Technical Report. University of British Columbia.
Prins, Samantha. 2019. 'Final Vowel Devoicing in Blackfoot'. Master's Thesis, Missoula, Montana: University of Montana.
Puchala, Chassidy, Anne Leis, Hyun Lim, and Raymond Tempier. 2013. 'Official Language Minority Communities in Canada: Is Linguistic Minority Status a Determinant of Mental Health?' *Canadian Journal of Public Health = Revue Canadienne De Santé Publique*, no. 104: 5–11.
Rice, Keren. 2011. 'Documentary Linguistics and Community Relations'. *Language Documentation and Conservation* 5: 187–207.
Schmirler, Katherine. 2022. 'Syntactic Features and Text Types in 20th Century Plains Cree: A Constraint Grammar Approach.' PhD Dissertation, Edmonton, Alberta: University of Alberta.
Schmirler, Katherine, and Antti Arppe. 2019. 'Modelling Plains Cree Negation with Constraint Grammar'. In *Proceedings of the NoDaLiDa 2019 Workshop on Constraint Grammar – Methods, Tools and Applications*, edited by Eckhard Bick and Trosterud Trond, 27–34.
Schmirler, Katherine, Antti Arppe, and Inge Genee. Forthcoming. 'Morphophonological Rule Development and Real-Time Rule Testing with XFST: A Model for Blackfoot'. In *Papers of the Fifty-Third Algonquian Conference*, edited by Monica Macaulay, Margaret Noodin, and Inge Genee. Vol. 53. East Lansing, Michigan: MSU Press.
Schupbach, Scott. 2012. 'Situational Demonstratives in Blackfoot'. *Coyote Papers: Working Papers in Linguistics*, no. 21: 1–22.
Silfverberg, Miikka, and Krister Lindén. 2009. 'Conflict Resolution Using Weighted Rules in HFST-TWOLC'. In *Proceedings of the 17th Nordic Conference on Computational Linguistics (NoDaLiDa 2009)*, edited by Kristiina Jokinen and Eckhard Bick, 174–81. Odense, Denmark: Northern European Association for Language Technology.
Snoek, Conor, Dorothy Thunder, Kaidi Lõo, Antti Arppe, Jordan Lachler, Sjur Moshagen, and Trond Trosterud. 2014. 'Modeling the Noun Morphology of Plains Cree'. In *52nd Annual Meeting of the Association for Computational Linguistics*, edited by J Good, J Hirschberg, and O Rambow, 52:34–42. Baltimore, Maryland.
Stocken, Harry. 1953. 'First ten chapters of Matthew's Gospel'. Internet Archive.
Taylor, Allan. 1969. 'A Grammar of Blackfoot'. Doctoral Dissertation, Berkley, California: University of California Berkley.
Truth and Reconciliation Commission of Canada. 2015. 'Honouring the Truth, Reconciling for the Future: Summary of the Final Report of the Truth and Reconciliation Commission of Canada'. www.trc.ca/assets/pdf/Honouring_the_Truth_Reconciling_for_the_Future_July_23_2015.pdf.

Weber, Natalie. 2021. ‘Blackfoot Words’. Database.
———. 2022. ‘How Synchronic Analysis Informs Subgrouping: Against Proto-Algonquian-Blackfoot’. Presentation presented at the 54th Algonquian Conference, University of Colorado Boulder.
Weber, Natalie, and Jason Shaw. 2022. ‘Situating Blackfoot within a Typology of (Mobile) Boundary Tone Grammars’. In *Proceedings of the 2021 Annual Meeting on Phonology*, edited by Noah Elkins, Bruce Hayes, Jinyoung Jo, and Jian-Leat Siah, 1–12. UCLA, Los Angeles.
Witschko, Martina, and Elizabeth Ritter. 2015. ‘Animating the Narrow Syntax’. *Linguistic Review* 32 (4): 869–908.
Yamada, Raquel-Maria. 2007. ‘Collaborative Linguistic Fieldwork: Practical Application of the Empowerment Model’. *Language Documentation and Conservation* 1 (2): 257–82.