CURATING A CORPUS OF BLACKFOOT NARRATIVE TEXTS

ALEXANDRA SMITH
***Iinisskimakii* (Buffalo Stone Woman)**
Bachelor of Arts, University of Lethbridge, 2022

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

MASTER OF ARTS

in

INDIGENOUS STUDIES

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

CURATING A CORPUS OF BLACKFOOT NARRATIVE TEXTS

ALEXANDRA SMITH

Date of Defence: July 28, 2025

Dr. Inge Genee Thesis Co-Supervisors	Professor	Ph.D.
Dr. Conor Snoek Thesis Co-Supervisors	Associate Professor	Ph.D.
Dr. Antti Arppe Thesis Examination Committee Member	Professor	Ph.D.
Shirlee Crow Shoe Thesis Examination Committee Member	Piikani Nation Elder	
Dr. Christopher Hammerly External Examiner University of British Columbia Vancouver, British Columbia	Assistant Professor	Ph.D.
Dr. Tabitha Spagnolo Chair, Thesis Examination Committee	Associate Professor	Ph.D.

Dedication

For Leo Alexander Jax,
Sakoinaamaahkaa, nohkowa, kitsiniiyimmo ki kitaaksskaahsawaakomimmo.
Takes the Last Gun, my son, I am grateful for you and I will always love you.

Abstract

This thesis documents the process of curating a corpus of Blackfoot narrative texts, referred to as the Blackfoot Narrative Text Corpus (BNTC) in the thesis. Blackfoot is a language spoken by four communities located in southern Alberta and in Montana of the United States, namely Apatohsipiikani, Kainai, Siksika and Aamskaapipiikani. The aim of this project is to develop a partially linguistically analyzed corpus of Blackfoot narrative texts to support the ongoing documentation and revitalization of the language. The corpus was compiled from published Blackfoot texts. Some texts are fully morphologically analyzed and glossed, while others were transliterated into the modern standard orthography from older spelling conventions but have not been further analyzed. After analysis and/or transliteration, the texts were integrated into the Korp corpus platform. The BNTC is an orthographically homogenous, searchable corpus currently containing 1,711 analyzed words and 8,681 unanalyzed words.[1] It is an open-ended flexible corpus to which new texts and/or additional analysis can continually be added. This project contributes to the broader field of Indigenous language documentation providing a corpus of Blackfoot narrative texts with partial linguistic analysis, an accessible resource for learners, teachers and researchers of Blackfoot.

[1] The BNTC is currently password-protected during development. Once finalized, it will be publicly accessible at https://korp.altlab.app/. In the meantime, access can be requesting by contacting alexandra.smith2@uleth.ca.

Acknowledgements

I would first like to acknowledge Iniskim, University of Lethbridge, for making this thesis possible. Iniskim is located on the traditional territory of the Siksikaitsitapi (Blackfoot Confederacy), which includes the Siksika, Kainai and Piikani Nations; Iniskim is also located on the lands of Treaty 7 territory.

This project was supported by funding from the *21st Century Tools for Indigenous Languages* Social Sciences and Humanities Research Council (SSHRC) Partnership Grant, as well as the *Documenting Variation in Niitsi'powahsin (Blackfoot Language)* SSHRC Insight Grant.

I am deeply grateful to my academic advisors and committee members Dr. Inge Genee, Dr. Conor Snoek, Dr. Antti Arppe, and Shirlee Crowshoe. I extend my sincere thanks to Inge for her unwavering support and encouragement throughout my Master's program. Her guidance was instrumental in developing my academic writing skills since my undergraduate studies. I am thankful to Conor, who first introduced me to Blackfoot linguistics through his teaching in the Community Linguist Certificate program, hosted by the Peigan Board of Education Society (PBOES) in partnership with the Canadian Indigenous Languages and Literacy Development Institute (CILLDI). Conor also generously shared his wealth of knowledge in corpus linguistics, for which I am grateful. Antti also contributed valuable guidance in corpus linguistics and computer programming, specifically in hosting the Blackfoot Narrative Text Corpus (BNTC) within Korp. I am appreciative to Shirlee Crowshoe for her ongoing support as a member of my committee, and of Dr. Christopher Hammerly who generously agreed to serve as the external examiner for this thesis.

I would also like to thank Dr. Felipe Bañados Schwerter, Dr. Natalie Weber, Dr. Katie Schmirler, Dominik Kadlec and Tait Hoyem for their intellectual contributions. I thank Felipe

for his assistance in uploading the Blackfoot texts into Korp and making the BNTC accessible online. I am grateful to Natalie for providing the digitized version of the *Original Blackfoot texts from the southern Peigans Blackfoot reservation, Teton County, Montana* (Uhlenbeck 1911). My thanks go to Dominik for developing the Uhlenbeck-to-Frantz orthography converter, and to Katie for taking the time to apply the converter to the Uhlenbeck texts. I am appreciative of Tait Hoyem for his work in making corrections and formatting the Uhlenbeck texts. I am also sincerely thankful to Caroline Russell, who also started at Iniskim when I began my Master's program. Her steady support and encouragement have helped me persevere through the hardships and challenges of graduate studies. I am grateful for her companionship, motivation and affirmation throughout my journey.

And, most of all, thank you to my family for their constant support throughout my academic journey, both undergraduate and Master's studies. Since the beginning of this journey, when I was a newly single mother and my son was just four months old, they have graciously helped care for him. They often stepped in to give me the time and space needed to focus on this thesis, even without being asked. Their encouragement and generosity have been instrumental in helping me stay grounded and committed throughout the challenges of academic life.

Lastly, I am eternally grateful to my son, whose presence has been my greatest source of strength and motivation. He has inspired me to persevere and complete this thesis with the hope that we may build a future where Blackfoot continues to thrive. It is my wish that, as he grows older, Blackfoot remains a living language – present in our lives, our home, and throughout Blackfoot territory.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Gloss abbreviation	Meaning
-	(hyphen) separates morphemes
.	(period) separator within a morpheme
=	Joins clitics
1	First person
2	Second person
21	First person plural inclusive
3	(Proximate) third person
4	Obviative third person
AI	Animate intransitive verb
AIO	Animate intransitive verb plus object
AN	Animate
COM	Comitative
CN	Conjunct nominal
CONJ	Conjunctive
DCT	Deictic preverb
DD	Distal demonstrative stem
DIR	Direct theme suffix
DM	Medial demonstrative stem
DP	Proximal demonstrative stem
DTP	Distinct third person pronouns
DUR	Durative aspect
FUT	Future
II	Inanimate intransitive verb
IMFUT	Immediate/imminent future
IMP	Imperative
IN	Inanimate
INCH	Inchoative
IND	Indicative
INS	Instrument
INT	Interior geometric configuration
INTS	Intensifier
INV	Inverse theme suffix
INVS	Invisible to the speaker
NAF	Non-affirmative endings
NAR	Narrative suffix
NEG	Negation
NMLZ	Nominalizer
NREF	Non-referring
MA	Motion away
MNR	Manner
MT	Motion towards
PERF	Perfective aspect

PL	Plural
POSS	Possessive suffix
PRO	Pronoun
PRF	{a'p} prefix
PST	Past
REFL	Reflexive
SBJV	Subjunctive
SG	Singular
STAT	Stationary
TA	Animate transitive verb
TH.TI	inanimate Transitive theme suffix
TI	transitive inanimate verb
VBLZ	Verbalizer
X>Y	X acts on Y

1. Introduction

This thesis describes the curation of a corpus of annotated Blackfoot narrative texts from existing sources. The Blackfoot Narrative Text Corpus (BNTC)[2] will be a resource for anyone wanting to know more about Blackfoot narratives, including teachers, learners, researchers, etc. The BNTC will be an important addition to Blackfoot resources for a language that is losing its speakers and needs more learning resources, as will be discussed further throughout this chapter.

A corpus is essentially a collection of texts; the term text refers to a wide variety of material, from a recorded conversation or lecture to an article from any source or a plain textbook (Barth & Schnell 2021, 7). Through the collection of existing naturalistic texts, we can see how the language is used in specific contexts. Corpora can be used for a multitude of tasks regarding the language itself, such as gathering contextualized language examples for language learning and teaching, obtaining frequency lists and statistical analysis of tokens, or doing a collocation analysis, which is finding words that frequently combine with a target word in a language. The types of tasks that can be done in a corpus depends on the interface that the corpus is hosted in. There are many corpora that exist for majority languages, like English: *Corpus of Contemporary American English* (COCA) (Davies 2008), *Coronavirus Corpus* (Davies 2019), *News on the Web* (Davies 2016), and the *Wikipedia Corpus* (Davies 2015); these are essentially large databases of specific registers of language. There are also several lesser known corpora available for Indigenous languages, for example, the Plains Cree corpus (Arppe et al. 2020) and *ChoCo*: a multimodal corpus of the Choctaw language (Brixey, Pincus & Artstein 2018). Differences in majority and minority language corpora are discussed in Section 2.2.

[2]The BNTC is currently password-protected during development. Once finalized, it will be publicly accessible at https://korp.altlab.app/. In the meantime, access can be requesting by contacting alexandra.smith2@uleth.ca.

The remainder of this chapter presents background information on the Blackfoot language. Section 1.1 gives general information about the Blackfoot language, including where it is spoken and who speaks it, followed by resources available as discussed in Section 1.2; Section 1.3 summarizes the educational programming for the language. Section 1.4 provides an overview of the remainder of the thesis.

1.1 The Blackfoot language

The Blackfoot language is part of the Algonquian branch of the larger Algic language family. Blackfoot is spoken by Blackfoot people who are from Apatohsipiikani, Kainai, and Siksika of Southern Alberta and Aamskaapipiikani of Montana, as shown in Figure 1. It has been said that

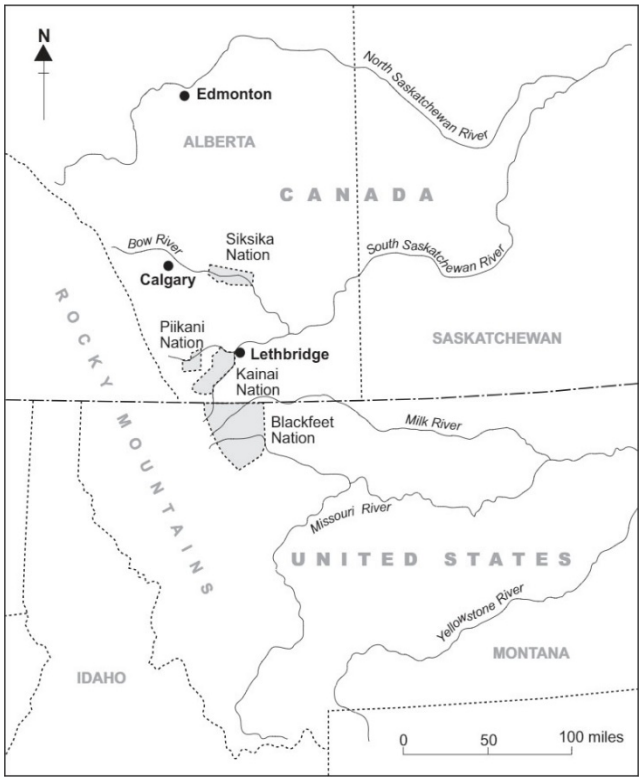
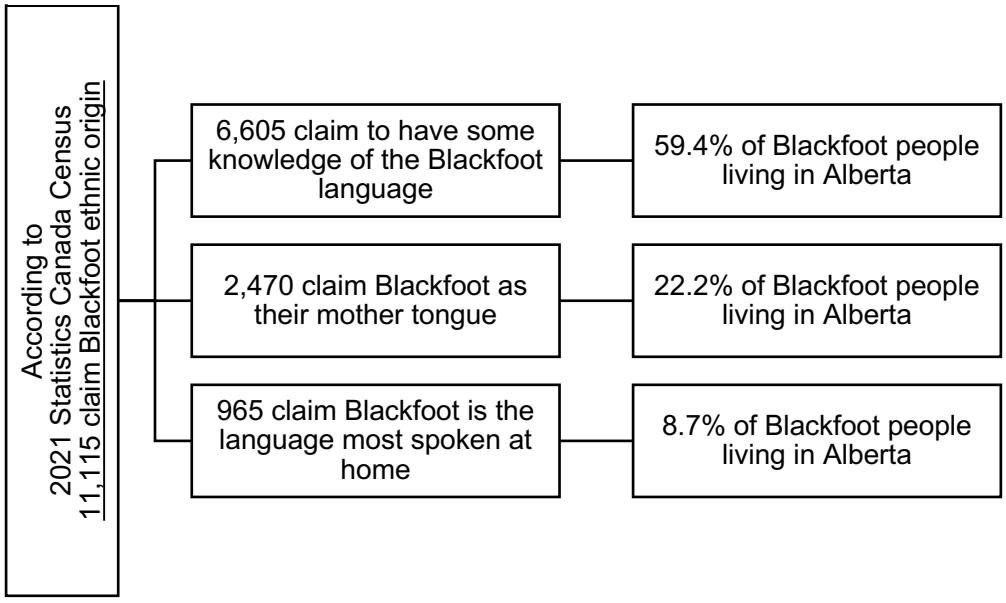


Figure 1. Map of present-day Blackfoot reserves. (Source: https://museum-companion.berghahnjournals.com/wp-content/uploads/2018/05/Ceremonies-fg_2.jpg)

the Blackfoot language is an endangered language, with its native speaker numbers on a consistent decline (Frantz 2009; Genee, 2009, 2020; Yellowhorn 2021). Recently, there has been a great push for revitalization of the Blackfoot language. According to the 2021 Statistics Canada Census, within Alberta, there were a total of 11,115 people who claimed to have Blackfoot ethnic origins (0.26% of the total Alberta population). Of those 11,115 people, 6,605 (59.4%) said they have some knowledge of the Blackfoot language, 2,470 (22.2%) said Blackfoot was their mother tongue, and 965 (8.7%) said that Blackfoot was the language that was spoken most often at home (Statistics Canada 2022). These statistics are summarized in Figure 2 below.

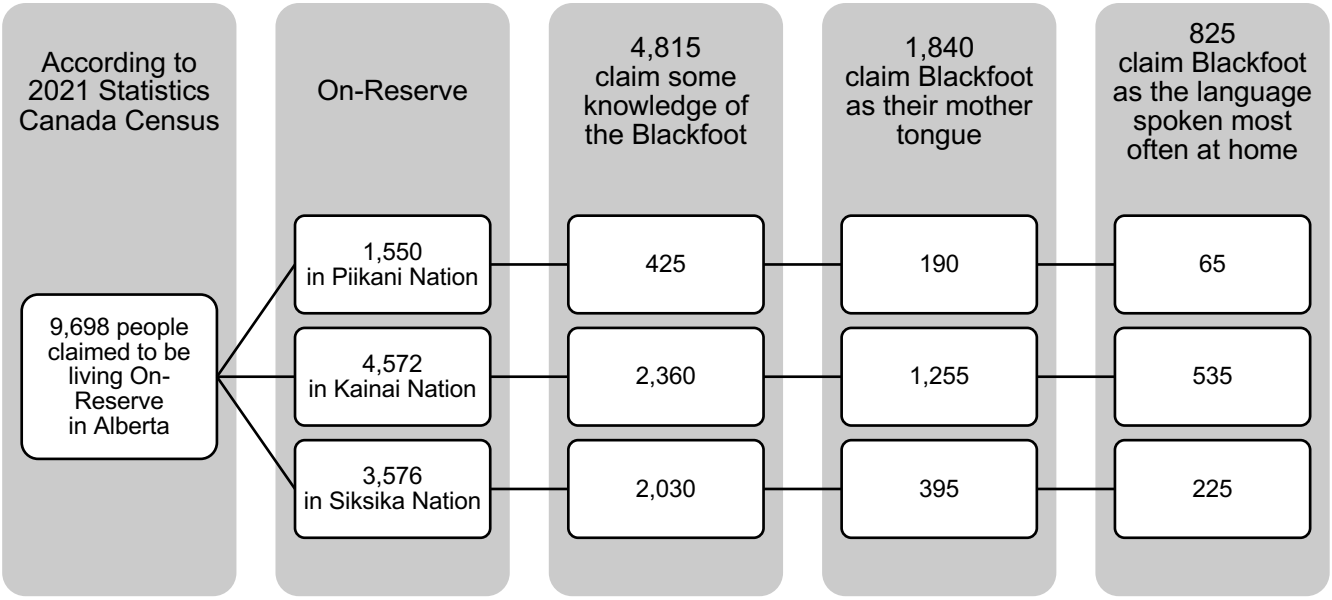
Figure 2. Statistics about the Blackfoot language among Blackfoot people within Alberta from the Stats Canada 2021 census.



The following numbers are individuals reported to be living on-reserve: 1,550 in Piikani, 4,572 in Kainai, and 3,576 in Siksika with a total of 9,698; 4,815 said they have some knowledge of the language (P-425; K-2,360; S-2,030), 1,840 said Blackfoot was their mother tongue (P-190; K-1,255; S-395), and 825 said that Blackfoot was the language that was spoken most often at

home (P-65; K-535; S-225) (Statistics Canada 2022). These statistics are summarized in the Figure 3 below. According to the 2020 census of the United States of America (USA), 34,810 people claim to be of the Blackfeet Tribe of the Blackfeet Indian Reservation of Montana.[3] The USA census did not have data related to the level of Blackfoot fluency.

Figure 3. Statistics about the Blackfoot language among Blackfoot people living on reserve in Alberta from Stats Canada 2021 census.



According to the Blackfoot Confederacy (2021) website,[4] at the time of writing, the Blackfoot confederacy has a total population of 34,000. Kainai has 13,000 members, Siksika has 7,800 members, Piikani has 3,500 members, and Aamskaapipiikani has 9,700 members.

It's important to note that the statistics mentioned in this section are not comparable, as they come from different sources resulting in different numbers. The sources are not measuring the same content nor are they asking the same questions, again, leading to different numbers.

[3]Source:https://data.census.gov/table/DECENNIALDDHCA2020.T01001?t=206:1671&d=DEC+Detailed+Demographic+and+Housing+Characteristics+File+A Accessed June 5, 2025

[4] https://blackfootconfederacy.ca/ last accessed in September 2023 (statistics no longer listed on website)

According to Genee and Junker (2018, 277), there were 23,456 registered Blackfoot members from the reserves in Alberta as of January 2017 and 10,938 people living on the Blackfeet reservation in 2011-2015. The numbers above suggest that there a quite a few people who are Blackfoot, but the data from Statistics Canada show that the level of Blackfoot fluency is low. Overall, the intergenerational transmission of Blackfoot is on a consistent decline; majority of Blackfoot speakers are elderly who are mostly first language speakers.

1.2 Blackfoot language resources

The Blackfoot language has limited learning resources, and the ones that do exist can be difficult to find. Blackfoot has a printed dictionary, *Blackfoot Dictionary of Stems, Roots, and Affixes* (Frantz & Russell 2017), which is in its third edition. This dictionary has been digitized and is a free online resource[5] (Genee & Frantz n.d.). Frantz (2017) also published a Blackfoot grammar book, titled *Blackfoot Grammar*, also in its third edition, explaining the grammar of the language; its usefulness for learners is somewhat limited due to its use of linguistic terminology. These sources are easily found in bookstores and can be purchased. Lena Russell (1996; 1997a; 1997b; 2001; 2002; 2003) wrote a Junior and High school Blackfoot curriculum (as a set of student and teacher books); these books are now out of print and no longer available. Vivian Ayoungman (1993a; 1993b; 1993c) also created a curriculum titled *Siksikai'powahsin/Siksika Language Series Kit*, which has three levels; this resource is also out of print and no longer available. The Galt Museum in Lethbridge sponsored the creation of *Blackfoot Language Workbook* and *Blackfoot Language Flash Cards* both of which have two editions. Both are

[5] Available at https://blackfoot.algonquianlanguages.ca/

available in print from the museum and available for download from their website[6] (Galt Museum n.d.). A more complete list of resources, last updated in 2021, can be found on the Blackfoot Language Resources (BLR) webpage, under the Resources tab.[7] Also listed on the BLR webpage are many scholarly articles about the Blackfoot language. The next section discusses various existing Blackfoot language programs.

1.3 Blackfoot language programming

Most of the Blackfoot-speaking reserves have Blackfoot classes in their schools, starting in elementary to Grade 12. Alberta Education has a Program of Studies (Alberta Education 2010) specifically for teaching Blackfoot from Kindergarten to Grade 12, which has learning outcomes for the teacher to create their own lessons.

According to the 2025/2026 Course Catalogues, as of June 2025 (unless otherwise noted), the following post-secondary institutions offer Blackfoot courses, outlined in Table 1 below.

Table 1. Blackfoot language programming available at post-secondary institutions.

Post-Secondary Institution	Blackfoot Courses Offered
University of Lethbridge (2025)	1. BKFT 1000: Introductory Spoken Blackfoot 2. BKFT 2000: Spoken Blackfoot II 3. BKFT 2210: Structure of Blackfoot Language 4. BKFT 3210: Blackfoot Morphology and Syntax
University of Calgary (2025)	1. INDL 301.01: Blackfoot Indigenous Language I 2. INDL 303.01: Blackfoot Indigenous Language II
Red Crow College (Mi'kai'sto 2025) on the Kainai Reserve	1. BKFT 100: Introductory Spoken Blackfoot I 2. BKFT 140: Language Structure: Roots, Prefixes, Suffixes 3. BKFT 155: The Blackfoot Sentence I: Declarative, imperative, interrogative

[6] Available at https://www.galtmuseum.com/blackfoot-language-workbook
[7] Available at https://blackfoot.algonquianlanguages.ca/

	4. BKFT 160: The Blackfoot Sentence II: Reflexive and Tense markers 5. BKFT 175: Vocabulary Expansion 6. BKFT 180: Using Blackfoot at Home 7. BKFT 189: Blackfoot for the Workplace 8. BKFT 185: Blackfoot for the Health Field 9. ILCD 114: Blackfoot Sign Language 10. ILCD 116: Blackfoot Storytelling
Lethbridge College (2023)[8]	1. BLK-1151: Blackfoot Language I 2. BLK-2251: Blackfoot Language II
Old Sun Community College (2023)[9] on the Siksika Reserve	1. SL 200: Introduction to Siksika Language (Siksikai'tsii'powahsin 1) 2. SL 204: Conversational Siksika (Siksikai'tsii'powahsin 2) 3. SL 260: Telling of Old Stories (Aikaitapiitsini'ksin, Siksikai'tsii'powahsin) 4. SL 300: A Study of Successful Language Immersion Approaches (1) (Siksikai'tsii'powahsin) 5. SL 304: A Study of Successful Language Immersion Approaches (2) (Siksikai'tsii'powahsin) 6. SL 308: Siksika Grammar (Siksikai'tsii'powahsin) 7. SL 312: Adult Language Immersion (Siksikaitsiipowahsin) 8. SL 324: Introduction to Indigenous Sign Language (A'pstaksini) 9. SL 416: Introduction to Ceremonial Language (Siksikai'tsii'powahsin)
Blackfeet Community College (2023)[10] in Montana, USA	1. NASX 141: Piikani Language Origins & Foundations 2. NASX 142: Intermediate Piikani Language 3. NASX 147: Plains Indian Sign Language 4. NASX 245: Advanced Piikani Language 5. PKNI 101: Piikani Language for Healthcare Professionals.

Despite the efforts that have been put forward for the language, there are very few people speaking it in the home. It's great for the schools to be putting forward their efforts for language revitalization, but schools alone cannot produce fluent Blackfoot speakers. One of the few

[8] Source: https://selfserv.lethbridgecollege.ca/Student/Courses/Search?keyword=blackfoot Accessed November 2023. The institution is now known as Lethbridge Polytechnic; the Blackfoot courses are no longer publicly available.

[9] Accessed November 2023. Courses are no longer publicly available.

[10] Accessed November 2023 Courses are no longer publicly available.

natural Blackfoot-speaking contexts that exist are in ceremony; I can attest to this as a Blackfoot woman. Most ceremonial practices are still carried out in fluent Blackfoot. There are some cases where English is used, when necessary, but for the most part, they are fluent environments. There are also fluent environments when speakers gather in a social setting and converse in Blackfoot.

1.4 Overview of the thesis

This chapter has laid the foundation for the rest of the thesis. The remainder of the thesis is organized as follows: Chapter 2 will discuss various aspects of building a corpus, including types and uses, differences between majority and minority language corpora, the different ways a corpus can be published, and other Blackfoot corpora, ending with a section on how the BNTC differs from other corpora. Chapter 3 presents the details of the curation of the BNTC itself, starting with the text selection, metadata, choices of orthography while choosing texts, followed by the linguistic analysis of texts within the BNTC, ending with its publication particulars. Chapter 4 describes the texts that are included in the BNTC. Chapter 5 presents the different features and search functions of Korp, the interface in which the BNTC is embedded. Finally, the thesis ends with a summary and conclusion in Chapter 6.

2. Building a corpus

A corpus can be used in a multitude of different ways to gain insight into a specific language. The process of building a corpus involves many different steps and considerations. This chapter discusses the use and structure of language corpora in general before addressing some of the differences between majority and minority language corpora and finally looking at Blackfoot corpora and the BNTC specifically. Section 2.1 will present the types of corpus-building situations and the various uses of the different types. Section 2.2 discusses the differences between majority and minority language corpora. Section 2.3 will highlight how and where corpora can be published. Section 2.4 looks at the other existing Blackfoot corpora and discusses how the BNTC complements these. Section 2.5 shows how the BNTC is different from other corpora.

2.1 Corpus types and their uses

Barth and Schnell (2021, 92) describe the following three corpus-building scenarios: Type 1: General corpus; Type 2: Language Documentation (LD) corpus; and Type 3: Research corpus; these are summarized in Table 2.

Table 2. Three types of corpus building scenarios (Barth & Schnell 2021, 92).

Type #	Corpus type	Research status	Typical features	Typical goals
1	General corpus	Well-researched	– Large – Range of text varieties – Written & spoken – Pre-existing texts – Digital texts available – Monitor, open-ended – Monolingual	– General reference – Wide range of focused analyses – Involvement in language planning – Other “non-linguistic” uses

2	LD corpus	Under-studied	– Small – Mostly spoken – Linked to primary media data – Collected texts – Requires processing/annotation – Monitor, open-ended – Mono-/multilingual	– General descriptive work – Dictionary work – Other "non-linguistic" uses
3	Research corpus	Either	– Small(er) – Written and/or spoken – Pre-existing or self-collected texts – Static – Mono-/multilingual	– Focused narrow research agenda

The general corpus type (Type 1) is representative of the larger, well-researched, majority languages. For example, the COCA (Davies 2008) would be categorized as a Type 1 corpus. The corpus contains more than one billion words, at the time of writing;[11] the data ranges from the years 1990-2019, with eight genres: spoken, fiction, popular magazines, newspapers, academic texts, TV and movie subtitles, blogs, and other web pages (Davies 2008). General corpora are normally very large, monolingual, and are representative of the language with a range of text varieties available. The texts that are included can be written or spoken, gathered from pre-existing and/or digital sources, which can be read or heard from within the corpus. General corpora are capable of monitoring a language because they are constantly growing with additional texts being added on a regular basis; this is also referred to as an open-ended corpus. With the addition of texts throughout the years, it is possible to see the shifts that are happening in the language, thus monitoring language change. With majority languages, their texts and recordings may need minor annotation and processing. Some annotation can be done by

[11] Information from https://www.english-corpora.org/coca/ Accessed May 13, 2025

automatic means, like with the use of a Part-of-Speech (POS) tagger, making it easier to add more texts to the corpus. A variety of research can be done using a general corpus. For example, Lin Hao (2023) used COCA to research the differences in the use of the nouns *street* and *road* to help English learners learn about these words in a more effective way through corpus use.

The language documentation (LD) corpus type (Type 2) is representative of minority languages that have not been extensively researched. LD corpora are often on the smaller side, usually bilingual with translations into a majority language. According to Barth and Schnell (2021), the LD corpus may contain mostly spoken texts linked with audio and/or video but may lack annotation. They can also contain pre-existing written texts, which may require processing and/or annotation. Smaller minority languages often lack readily available digital sources. Once relevant texts are located and collected, they may require scanning, Optical Character Recognition (OCR), or other forms of processing before they can be incorporated into a corpus. After texts are processed into a format suitable for corpus inclusion, they can be annotated. Minority languages often lack automatic annotation tools, resulting in texts being annotated manually, which is an approach that is both time consuming and resource intensive. Consequently, some texts in an LD corpus may lack annotation. Since LD corpora are on the smaller side, they are not usually representative of all aspects and uses of the language. However, they can be expanded over time and become more representative. LD corpora are generally used for descriptive work and/or dictionary building. The *Blackfoot Words* (BW) corpus (Weber 2022), discussed further in Section 2.4, has some features of an LD corpus. It was assembled from historical written sources which required major processing and manual annotation. The corpus could be used for descriptive work, such as tracing changes of specific Blackfoot words

over time. It can be difficult to categorize corpora based strictly on the types described above, as they may have features of multiple types.

The research corpus (RC) (Type 3) can be used for any language, minority or majority. The material in a RC is gathered for a specific research goal. Therefore, it will often be on the smaller side, and is usually static, meaning that it does not continue to grow. Once the research project is completed, these corpora are no longer attended to and thus, are not usually accessible to the public. The Kadlec corpus (Kadlec 2023), further discussed in Section 2.4, is an example of a research corpus, as it was used for one purpose only: to test his Blackfoot computational noun and verb model.

The BNTC is a mixture of Types 2 and 3, language documentation corpus and research corpus, respectively. As mentioned above, it's difficult to categorize this corpus as a single type. The BNTC is on the smaller side, with approximately 10,300 Blackfoot words, which is characteristic of a RC and an LD corpus. It is bilingual with pre-existing Blackfoot texts with English translations, a characteristic of both Types 2 and 3 corpora. The texts required major processing and annotation/analysis, which is a characteristic of an LD corpus. The intention is for this corpus to be dynamic rather than static, with additional analyzed texts added continuously, a characteristic of Type 2. Although the BNTC has some characteristics of Type 3, it is not a research corpus in the sense of responding to a specific question.

In conclusion, there are three types of corpora as described by Barth and Schnell (2021). It is often difficult to classify specific corpora, especially those of minority languages, clearly into one of these existing types. The BNTC has features of Types 2 (research corpus) and 3 (language documentation corpus). The next section discusses the differences between majority and minority language corpora.

2.2 Differences between majority and minority language corpora

There are many differences between majority and minority language corpora. The main differences include size, availability, representativeness (dialect, genre, age, gender, etc.), annotation and metadata, orthography and standardization, purpose and usage. These main differences will be discussed in the following paragraphs.

Majority and minority language corpora differ in terms of size and availability. Majority language corpora can be very large and are widely available and easily accessible. Majority languages have a lot more material that can be gathered for data input. For example, with a language such as English, the internet can be scraped for data input, and a plethora of books/texts is available to retrieve data from. These corpora are usually published on the internet where they're widely and easily accessible. Since majority languages have lots of financial, institutional and other supporting resources, these corpora can be very well maintained, where they won't go out of service from lack of funding or other reasons. Some widely available corpora require subscription access, where users create an account and pay for access. Corpora from *English-Corpora.org*, like COCA, are the most widely used corpora for English[12] and through their paid subscriptions they can be continuously maintained and won't become obsolete. In contrast, minority language corpora are usually on the smaller side due to lack of data sources available. This often means that additional processing is required, for example, text digitization, Optical Character Recognition (OCR), OCR verification and correction, and translation, before the sources can be used as viable data inputs. The limited availability of resources often results in lack of funding, leading to poorly maintained corpora that may eventually become inaccessible or obsolete, like the OLD corpus, further discussed in Section 2.4.

[12] Information from https://www.english-corpora.org/pdf/english-corpora-overview.pdf Accessed May 20, 2025

In terms of representativeness, corpora of majority languages have a significant advantage over those of minority languages. Since there is an abundance of material available for data, majority languages are usually very well represented in a corpus. In terms of dialectal variation, including data from a wide range of speakers from diverse backgrounds helps to ensure a good representation of the different spoken dialects. Likewise, regarding genres, a corpus is considered balanced when it includes categories with equivalent representation of the different genres occurring in the language. For example, COCA has eight different genres that are evenly distributed ranging from 119 to 128 million words in each category over 30 years. The *Corpus of Historical American English*[13] (COHA) contains data from the 1820s-2010s amounting to over 475 million words in 5 different genres (TV/movies, fiction, popular magazines, newspapers, and non-fiction books), further adding to the representation of the English language throughout corpora. Minority language corpora are usually less representative of the language, in terms of both dialect and genre. Since there are less data sources available, they often include any available text, which usually doesn't result in a balanced representation of the language. For example, the published sources available for Blackfoot are mostly in the Kainai dialect. The BNTC therefore contains more texts written by people who are from Kainai, leading to under-representation of Siksika and both Piikani dialects. Blackfoot doesn't have enough sources to create a corpus like COHA adding to the time depth and historical representation of the language. Although there are some older Blackfoot sources, there are large gaps in the documentation of the language, consequently, we are unable to see the variation over a range of years. The BW corpus (Weber et al. 2023) tries to fill this gap by including available Blackfoot historical sources, but they aren't necessarily balanced, in terms of dialect or type, nor are they

[13] Available at https://www.english-corpora.org/coha/

always reliable sources. Due to the limited sources in Blackfoot, there is a lack of representation of speakers from different age groups. This lack of representation can also be attributed to the language's declining fluency among younger generations.

Annotations are linked to the data within corpora. Annotation refers to the special types of information linked to individual words within corpora encoding interpretative linguistic information, e.g. phonetics, morphology, syntactic and semantic information (Barth & Schnell 2021, 110-125). Annotation also allows for more focused searches; for instance, Parts-of-Speech (POS) annotation can assist a user in searching for whether certain POS are being used less in specific contexts. Some annotation can be done automatically with available digital tools, like POS taggers. Annotations are usually interpreted by a human, and the different annotations allow the corpus to be searched. Majority languages often have extensive descriptions allowing for numerous annotations. In contrast, minority languages may not be as extensively described, thus annotations may not be as thorough or complete when compared to those of the majority language corpora. For instance, in the BNTC, only a portion of the texts have full analysis/annotation. There are also morphemes in Blackfoot that aren't fully understood. For instance, the morpheme {*a'p-*} serves different functions depending on the context it is used in; in the BNTC it is therefore simply glossed as PRF for "prefix" for consistency purposes.

Metadata is also linked to the data within a corpus; it allows the user of the corpus to understand what type of data is represented in the corpus. This includes properties like situational features of the individual texts, e.g. year of publication, source, author's name, register and genre. Metadata allows for better searches to be done within a corpus, e.g. asking questions about genre-based or generation-based speech patterns. With majority languages, the initial datasets used to curate corpora are usually large and the metadata can be very extensive.

For example, in COCA, a search can be narrowed with the selection of various metadata terms, like the selection of a year range or within specific genres. Searches within COCA can reveal variation in English.[14] On the other hand, minority language sources can lack metadata to begin with, specifically information about speakers or how a specific source was developed. The lack of metadata can make it difficult to use a corpus for various research topics or to compare data within a corpus.

Orthography refers to the spelling system of a language. Standardization of orthography refers to the process of developing and implementing a consistent writing system for a language. With majority languages, orthography tends to be standardized and consistent throughout a corpus, except in the case of some historical corpora that preserve historical orthographies. An example is the Helsinki Corpus of English texts,[15] which contains historical sources from the period 870-1710 in their original orthographies (Helsinki Corpus TEI XML Edition 2011). Minority languages don't always have a standard orthography, especially when the language isn't well-documented or well-described. Therefore, the sources could be in different orthographies. Orthography standardization and consistency throughout a corpus is useful, for instance to facilitate searches for specific words, but there are also drawbacks. There may be cases where it may be desirable to leave the orthography unstandardized. For example, in the absence of standardization, it's possible to observe how a word's pronunciation has changed over time or how it varies across different dialects, assuming that people wrote what they thought they heard. An example of a minority language corpus that preserves historical orthography is the *Blackfoot Words* (BW) corpus. In this corpus the original orthography from all sources is maintained, where each word is linked to a standardized lemma, so that all words can also be found by

[14] Information from https://www.english-corpora.org/pdf/english-corpora.pdf Accessed June 15, 2025
[15] Available at https://helsinkicorpus.arts.gla.ac.uk/display.py?fs=120&what=index

searching one consistent form (Weber et al. 2023, 1229). The BW corpus will be further discussed in Section 2.4. If there is standardization within a corpus, this could make it easier for a user to do searches, although this also depends on what the user is looking for. The BNTC is standardized to the Frantz orthography, while still retaining the original orthography from its original sources, to make it more usable for Blackfoot teachers and learners to find examples easily; the orthography of the BNTC will be further discussed in Section 3.3.

Corpora can be used for many different purposes, as was mentioned briefly in Section 2.1. The main uses for corpora include linguistic research, natural language processing (NLP), language learning and teaching, lexicography and language documentation. Linguistic researchers can make use of corpora by searching for specific examples of language use; for example, they can look for specific morphemes and analyze how they are used in different contexts. NLP needs large datasets for which corpora can also be used. NLP learns directly from corpora, specifically referred to as "training corpora" (Jurafsky & Martin 2025), which can also be used to train Artificial Intelligence (AI) programs. Similar to linguistic research, language learning and teaching can benefit from the use of corpora by searching for specific word usage. That data can support learners in understanding the appropriate context and functions of words, thus providing more accurate and effective language use. In a classroom, a teacher can only produce a limited number of examples spontaneously; thus, having access to a corpus can deliver a lot more examples in a shorter amount of time, especially in cases where the language is being taught by a non-native speaker. Language documentation aims at recording the observable use of language in a given society as much as possible (Barth & Schnell 2021, 182). Through the language documentation process, a plethora of language data is gathered, which is essentially an unformatted corpus. The data can be reorganized into small digital texts and have annotations

and metadata added to become a usable corpus. Corpora can be used by researchers, teachers and learners to look at real-life examples of language use.

As mentioned above, majority languages have existing large corpora. English has multiple corpora that are publicly accessible and either free or accessible by subscription, for example the *Corpus of Contemporary American English* (COCA) (Davies 2008), *Coronavirus Corpus* (Davies 2019), *News on the Web* (Davies 2016), and *Wikipedia Corpus* (Davies 2015), which are databases of specific registers of language. Other majority languages like Spanish and French also have corpora, like *Corpus del Español* (Davies 2016), *Corpus of 21st-Century Spanish (CORPES XXI)* (ROYAL SPANISH ACADEMY: Database [online] 2014), *Contemporary French Study Corpus* (Benzitoun, Debaisieux & Deulofeu n.d.), *Processing Oral Corpus in French* (TCOF).[16] There are also Indigenous language corpora, for example the Plains Cree corpus (Arppe et al. 2020) and *ChoCo* (Brixey, Pincus & Artstein 2018), both of which are leaning more towards a general (Type 1) corpus because of their size, but they are both bilingual with English. There are differences between majority and minority language corpora including size, availability, representativeness, annotation and metadata, orthography and standardization, purpose and usage. The next section will discuss the various methods whereby corpora can be published.

2.3 Publishing a corpus

Corpus publication is the last step in corpus creation. There are multiple ways that corpora can be published, including online publication, print, physical distribution and data repositories. Publication is understood as publishing online in the 21st century (Barth & Schnell 2021, 108).

[16] Available at https://tcof.atilf.fr/index.php

Corpora don't necessarily require interactive user interfaces to be published, what really matters is that the corpus can be used by those who are interested in the data. The most basic way to publish is by organizing the data in flat text files (.txt). But this has some disadvantages, the main one being that it doesn't allow for much functionality. For more advanced search features, one would have to use a concordancer, such as AntConc, which is able to analyze flat text files. AntConc (Anthony 2023) is a computer software that includes a "corpus analysis toolkit for concordancing and text analysis" (Anthony n.d.).

Corpora in the past were published by printing the data. An example of this is *Computational analysis of present-day American English* (Kučera & Francis 1967) and *Word Frequencies in Written and Spoken English* (Leech, Rayson & Wilson 2001). The latter is a reference book based on the *British National Corpus*, explained as looking "like a cross between a dictionary and a telephone directory" (Leech, Rayson & Wilson 2001, ix). Printing corpora as word lists, frequency tables and concordances was done prior to the availability of the internet and digital tools. Another way of corpus publication in the past was by CD-ROM; data would be loaded onto the CD-ROM to be distributed more widely. For example, the *Texas Instruments/Massachusetts Institute of Technology corpus* (TIMIT) was published this way (Garofolo et al. 1993).

Uploading corpus datasets into data repositories is another example of publication. For example, "the *Endangered Languages Archive* (ELAR) is a digital repository for preserving multimedia collections of endangered languages from all over the world, making them available for future generations" ("Endangered Languages Archive" 2002). ELAR allows for users to upload their data to make it publicly accessible. It also has capabilities to set certain protection parameters, for example only allowing specific data available with permission from the original

author. The *Linguistic Data Consortium* (LDC) is another example of a data repository. The site allows for users to upload their datasets and then they can also pay a fee for specific licenses to obtain the datasets included in the LDC (University of Pennsylvania 1992). The next section will discuss existing published Blackfoot corpora.

2.4 Other Blackfoot corpora

Some corpora do exist for the Blackfoot language. For example, Weber (2022) published *Blackfoot Words* (BW), which is a relational database of Blackfoot lexical forms. In the curation of BW, 63,493 individual lexical forms from 30 different sources from the years 1743–2017 have been digitized that represent the four reserve dialects. Version 1.1 of BW contains lexical forms from nine of these sources including:

1. Wordlist from *The North American Indian: Vol. 6. The Piegan. The Cheyenne. The Arapaho* (Curtis 1911),
2. Wordlist from *Narrative of a journey to the shores of the Polar Sea in the years 1819, 20, 21, and 22 (*Franklin 1823),
3. Wordlist from the *"Introduction" of Indians of North-west America, and Vocabularies of North America, with an Introduction* (Gallatin 1848),
4. Wordlist from the *"Report on the Blackfoot Tribes" of Report of the Fifty-Fifth meeting of the British Association for the Advancement of Science, held at Aberdeen in September 1885* (Hale 1886),
5. Wordlist from *Observations on Hudsons Bay, 1743, and Notes and observations on a book entitled A voyage to Hudsons Bay in the Dobbs Galley 1749* (Isham 1949/1743),
6. Wordlist from "Miscellaneous contributions to the ethnography of North America" in *Proceedings of the Philological Society 2(28)* (Latham 1846),
7. "Initial change in Blackfoot" in *Contributions to Anthropology: Linguistics I (Algonquian)* from Taylor (1967)
8. *A grammar of Blackfoot* from Taylor (1969)
9. Wordlist in *The present state of Hudson's Bay: containing a full description of that settlement, and the adjacent country, and likewise fur trade, with hints for its improvement, & c. & c.: to which are added, remarks and observations made in the inland parts, during a residence of nearly four years, a specimen of five Indian*

languages, and a journal of a journey from Montreal to New York from Umfreville (1790)

With plans to continually add more forms, each abstract lemma is linked to an individual token it is derived from thus allowing a grouping of the lexical forms, regardless of its original orthography and source (Weber et al. 2023, 1208-1209). Each item in the database is tokenized at several levels in its original orthography—the word, the stem, and the morpheme—which allows for building a hierarchical structure into the database (Weber et al. 2023, 1227). Tokens that have the same stem or morpheme are linked to an abstract standardized lemma given in Frantz's orthography, capturing all phonemic contrasts, allowing for searchability and comparison of all forms available in the database (Weber et al. 2023, 1229-1230). The fact that each lexical form in the corpus is tokenized down to the morpheme implies a degree of morphological analysis. Keeping the forms in their original orthography, combined with the large time span of the collection, can allow searches for a single form from different times and sources. This type of search can allow one to see if the form has changed over time, in pronunciation, usage or between dialects. The BW corpus is currently under restricted access, according to the website.[17] Something similar applies to Dunham's OLD corpus, also discussed further below. Future plans for the BW corpus include updating the corpus software for the data to be viewed online in a safe manner, keeping it under restricted access to keep it shielded from data scraping (Natalie Weber, May 2025, p.c.).

Dominik Kadlec (2023) also curated a Blackfoot corpus by extracting Blackfoot text from these seven different sources,

1. *Blackfoot Grammar* (Frantz 2017),
2. *Blackfoot Dictionary of Stems, Roots, and Affixes* (Frantz & Russell 2017),

[17] Available at https://www.blackfootwords.com/ Accessed May 2025

3. webpages from the Jehovah's witness website,[18]
4. *Niitsitapiisini* website (Glenbow 2001),
5. transcribed stories from the Blackfoot Language Resources Story Archive,[19]
6. transcribed words and phrases from the Blackfoot Language Resources Conversations tab,[20]
7. *Ákaitsinikssistsi: Blackfoot Stories of Old* (Heavy Shields Russell & Genee 2014).

This corpus amounts to approximately 13,784 unique word forms (Kadlec 2023, 63-64), and was assembled specifically to test his Blackfoot noun and verb computational model. Kadlec's corpus is, to date, is the largest orthographically standardized Blackfoot corpus; the corpus consists only of texts in the Frantz orthography (Kadlec 2023, 62). There is some overlap between the Kadlec corpus and the BNTC, but the latter includes linguistic analyses. There are plans to publish this corpus as a flat text file on the *Blackfoot Language Resources* webpage (Kadlec & Genee p.c.; Kadlec 2023, 61).

Another example of a Blackfoot corpus was created by Joel Dunham (2013), which was initially referred to as the Blackfoot Language Database (BLD). The BLD was created through the use and manipulation of the Online Linguistic Database (OLD) open-source software (Dunham 2013, 76). The BLD is a web-based application designed to facilitate collaboration between Blackfoot researchers. It provides a secure platform for storing linguistic fieldwork data, such as text, images, audio and video, which can also be accessed by other researchers (Dunham 2013, 76). In Dunham's dissertation he further describes the OLD and a Blackfoot dataset. The Blackfoot dataset is a collection of 23,708 forms, consisting of data from the BLD, which contained original fieldwork of multiple contributors, content of the *Blackfoot Grammar* (Frantz 1991) and other sources; the other portion of the dataset includes the morphemes, example words

[18] Available at https://www.jw.org/en/library/?contentLanguageFilter=bla
[19] Available at https://stories.blackfoot.atlas-ling.ca/#/stories
[20] Available at https://blackfoot.algonquianlanguages.ca/conversations/

and sentences from the *Blackfoot Dictionary of Stems, Roots, and Affixes* (Frantz & Russell 1995) (Dunham 2014, 213-214). From this dataset a corpus of morphemes was extracted that was used to create Blackfoot morphological parsers, which amounted to 4,267 morphemes (Dunham 2014, 217-218). I was not able to access this data in the OLD nor BLD at the time of writing as the webpages are inaccessible; this is very unfortunate, as it would have been beneficial to have access to the morphological parsers that were developed to further expand the analysis in the BNTC. In conclusion, three different Blackfoot corpora that have been developed, although with differing accessibility issues. The following section describes how the BNTC differs from the previous corpora discussed.

2.5 How the BNTC differs from other corpora

The BNTC differs from other corpora in several ways that will be discussed in this section. The BLD (Dunham 2013) includes a lot of elicitation materials from linguistic fieldwork (Inge Genee, p.c.), which differs from the BNTC. As mentioned above, the content of the BLD has example words and sentences from the first edition of the *Blackfoot Dictionary* (Frantz & Russell 1995). The BNTC has 19 narrative texts that are fully analyzed, which are listed in Table 3 (Chapter 4). A similarity our corpora have is the element of morphological analysis. The OLD has some analysis done by the researchers who uploaded their work. The main difference between the BNTC and the BLD is that the former contains full narrative texts.

The BNTC differs from the Kadlec corpus as the latter does not have linguistic analysis of its data. The Kadlec corpus does contain a few full texts, and there is some overlap with the BNTC: both our corpora contain the texts from the *Blackfoot Grammar* (Frantz 2017), *Niitsitapiisini* website (Glenbow 2001), transcribed stories from the Blackfoot Language Resources Story

Archive and *Ákaitsinikssistsi: Blackfoot Stories of Old* (Heavy Shields Russell & Genee 2014). The Kadlec corpus also contains individual Blackfoot words and sentences and is only available as a monolingual flat text file.

The BNTC differs from the BW corpus as the first version of the latter contains individual words from eight different published Blackfoot wordlists, which are listed above in the previous section. They are similar in having linguistic analysis, the BNTC contains linguistic analysis of full narrative texts. Whereas, BW has each lexical form tokenized to the morpheme implying a degree of morphological analysis, but it's not laid out as a 4-line-analysis as it is done in the BNTC. Another similarity is that both corpora are published on interactive platforms.

The BNTC is similar to the Plains Cree corpus in terms of containing full texts and both corpora are similarly bilingual with English and an Indigenous language. A similarity both corpora share is they're both published on the Korp platform with linguistic analysis. The BNTC and *ChoCo* (the Choctaw corpus) also have some similarities and differences. A main difference is that *ChoCo* is a multimodal corpus containing audio, video and textual data. Another difference is that the BNTC is published on the Korp platform, and *ChoCo* is published as text files; some files have English translations, but a small portion of the files are only in Choctaw. This is a similarity in that both the BNTC and *ChoCo* are bilingual, although the BNTC is entirely bilingual. *ChoCo* also has a portion of narrative texts, as does the BNTC.

To summarize, this chapter has demonstrated that corpora are valuable tools for gaining insight into languages. It has discussed the numerous steps and important considerations involved in curating and publishing language corpora. Also discussed are the general use and structure of language corpora and differences between majority and minority language corpora. This chapter also highlighted the previously curated Blackfoot corpora and compared them to the

BNTC, while also drawing comparisons to the Plains Cree and Choctaw language corpora. The BNTC contributes to the growing field of digital tools and technologies for Indigenous languages aimed at supporting revitalization. The next chapter will delve into the steps that went into the curation of the BNTC.

3. The Blackfoot Narrative Text Corpus

This chapter will discuss the various steps that went into curating the Blackfoot Narrative Text Corpus (BNTC), Section 3.1 discusses the considerations of text selection for the corpus. Section 3.2 presents the metadata and what information was collected. Section 3.3 addresses the considerations of Blackfoot orthography when choosing which Blackfoot texts are included in the corpus. Section 3.4 explains the linguistic analysis applied to the corpus data. Section 3.5 discusses the process of publishing the BNTC.

3.1 Text Selection

In this section, I discuss text selection for the BNTC. Blackfoot is an understudied language, as described above in Section 1.1. The language is mostly spoken, but more recently with Indigenous language revitalization efforts, more written sources are being published. Most of the basic Blackfoot language description exists in the form of the *Blackfoot Dictionary of Stems, Roots and Affixes* (Frantz & Russell 2017), the *Blackfoot Digital Dictionary* (Genee & Frantz n.d.) and the *Blackfoot Grammar* (Frantz 2017).

As discussed above in Section 2.2, there are differences in gathering and preparing texts for majority language corpora compared to minority language corpora. With regard to the BNTC, upon embarking on gathering Blackfoot texts, multiple problems arose. The main issues were the lack of suitable sources, both digital and hard copy, and orthographic variation. The issues with orthographies will be discussed in greater detail in Section 3.3.

It is impossible to create a corpus that is representative of the entire Blackfoot language with the lack of Blackfoot sources. In total, I was able to locate 20 published sources that had substantial written Blackfoot text. The corpus includes 13 of these (some of them include more

than one story). The corpus includes 54 stories, ranging from 21-1326 Blackfoot words, with approximately 10,300 words in total. Compared to majority language corpora, this is a small corpus. Most of the texts that are included are in the Kainai and Siksika dialects, with a few in the Aamskaapipiikani dialect, and three stories in Apatohsipiikani dialect. I attempted to include more stories in the Apatohsipiikani dialect, but it was difficult to find such texts; this was personally disappointing as an Apatohsipiikani member. Even though the three stories mentioned above are narrated by an Apatohsipiikani person, they were translated with the help of Kainai individuals. As a result, these texts may not be strictly representative of the Apatohsipiikani dialect. Although each text in the corpus is noted as belonging to a specific dialect, it's important to note that strict separation between the dialects is often complicated. Blackfoot speakers usually have multiple affiliations to other Blackfoot nations, such as being born in one place and growing up or marrying in another place. For example, in the Uhlenbeck (1911) collection he mentions that his interpreter is not from Aamskaapipiikani, but the texts are categorized as being Aamskaapipiikani dialect; these will be discussed further in Section 4.12. Each of the texts included in the corpus will be described in detail in Chapter 4.

Published and publicly available Blackfoot texts were chosen to include in the BNTC. Working with published texts help avoids copyright and/or ethical issues. However, where possible I tried to contact authors or rights holders to ask for additional permission. For instance, I had intended to include several Blackfoot graphic novels published by USAY (Urban Society for Aboriginal Youth), but when I contacted them, they were unable to assist me to obtain permission from the original speakers/translators of the texts, so I decided not to include these texts. I looked for texts that had any form of Blackfoot written in Roman orthography, regardless of specific spelling conventions. An exception to this is the collection of stories in Uhlenbeck

(1911), which uses an idiosyncratic phonetic transcription. However, this transcription was consistent enough to allow for transliteration into the Frantz orthography, as described in Section 4.12. The Blackfoot texts also needed to be cohesive pieces longer than just a few sentences, therefore highly repetitive materials, such as children's books, were excluded. One of the primary goals of the corpus is to examine how the Blackfoot language behaves in context; thus, isolated sentences were insufficient and could potentially skew the overall corpus content. For this reason, the example sentences from the Blackfoot dictionaries (Frantz & Russell 2017; Genee & Frantz n.d.) were likewise excluded.

Another goal of the corpus is to include as much morphological annotation and analysis as possible. Texts that included interlinearizations with their publication are included with their analyses, like *Aakiipisskani 'the women's buffalo jump'* (Áistainskiaakii Many Feathers et al. 2013) and "Ikasskini" (Frantz 2017). With these types of texts, glossing abbreviations were updated to match a standard set of glossing conventions developed for this project, ensuring all morpheme glossing is consistent throughout the corpus. Where an existing analysis was changed, the original is conserved in a note section included with the text. This will be discussed further in Section 3.4. The next section discusses the process of metadata collection.

3.2 Metadata

I attempted to collect a standard set of metadata for each text. Not all the sources had the same type of metadata available and was in different sections throughout each source. To collect data systematically, the data was entered into a spreadsheet with specific headings. The following list shows those headings along with the type of information I looked for in each source:

1. **Year** – What year was the work published?
2. **Title1** – What is the title of the whole source?

3. **Format** – What format is Title1 in? (book/article/webpage/pdf/etc.)
4. **Author1** – Who is the author of the whole source? (Last name, First name)
5. **Story collection: year** – What year was the Blackfoot text (Title2) written?
6. **Story collection: format** – How did the Blackfoot text (Title2) come to its present form? (i.e. initially recorded, then transcribed, then analyzed, then published)
7. **Author2** – What is the name of the person who produced the Blackfoot text (Title2)? (Last name, First name)
8. **Age** – What is the Author2's age at the time of production?
9. **Gender** – Is Author2 Male or Female?
10. **Fluency** – What is Author2's Blackfoot fluency? (mother tongue speaker (L1)/ second language learner (L2))
11. **Interpreter** – What is the name of the interpreter? (if one was used) (Last name, First name)
12. **Title2** – What is the title of the Blackfoot text?
13. **Number of Blackfoot words** – How many Blackfoot words are in the Blackfoot text (Title2), including the title (if in Blackfoot)?
14. **Dialect** – Where is Author2 (and/or interpreter) from? (Apatohsipiikani, Kainai, Siksika, Aamskaapipiikani)
15. **Genre** – What type of text is Title2? (e.g. narrative, legend, Napi story, procedural, prayer, contemporary text, argumentative, song)
16. **Interlinearized** – Is there linguistic analyses available? (yes or no)
17. **Frantz Orthography** – Is the Blackfoot in Frantz orthography? (yes or no)

Metadata from the included texts are presented in Chapter 4. The above categories like the citation of Title1 (#2), along with Title2 (#12), Author2 (#7), year of the story collection (#5) will be included with each text in the BNTC; other parts of the metadata will be linked to the corpus platform on an external website.

It's important to note that, from my research, there do not appear to be Blackfoot terms for specific Blackfoot story genres. Therefore, the texts described in this thesis are categorized by means of western labels that best match the Blackfoot texts, which are included under the heading genre (#15) in the list above. Although Blackfoot people refer to certain "genres" of stories, these titles—such as Napi stories, origin stories, traditional stories, personal stories, or lesson stories—are also used to describe the texts in the BNTC. Napi stories recount Napi's

adventures; legend refers to stories about Blackfoot "heros" like Scar-Face and Katoyissa; personal stories are narratives drawn from the storyteller's life; procedural texts describe the process of making or doing something; contemporary texts are more recent creations; narrative serves as a category that do not clearly fit into any of the above genres. This is unlike Plains Cree, which have specific Cree words for their story genres as described by Schmirler (2022, 122-125). The next section discusses the orthography of the Blackfoot texts that are included in the BNTC.

3.3 Orthography

The Blackfoot people did not create an alphabet for their spoken language in a written form; rather they would record events using pictographs in winter counts or rock paintings (Genee 2020, 4). There were several early missionaries and other Europeans who attempted to write the Blackfoot language according to the rules of their native languages, even experimenting with and developing syllabaries, including Father Constantine Scollen, John Maclean, John Williams Tims, C.C. Uhlenbeck and Jack Holterman (Genee 2020, 4-6). The most widely accepted alphabet is Donald G. Frantz's orthography that was developed during the 1960s and 1970s. This orthography was adopted by the Canadian Nation schoolboards in 1975 and is used in principle in most educational materials in Canada (Frantz 2017, 185; Genee 2020, 6). This orthography will be referred to as the "Frantz system" and/or the "Frantz orthography." Prior to the Frantz system, and even after it was made, there were numerous other ways Blackfoot was recorded. In a lot of instances, people spell Blackfoot words as they hear them using conventions of the English orthography, meaning one sound could be written multiple ways or if a speaker doesn't pronounce the grammatical endings, they were left out.

Several considerations went into the decision whether to include a text in the corpus. I wanted to keep it as straightforward as possible. Thus, I included any source with Blackfoot text written in Roman orthography, whether it uses the Frantz system fully or makes an attempt at using it. Sources without the standard spelling are transliterated into the Frantz system, while also retaining the original orthography. When there is a transliteration, the original spelling stays at line 1 of the standard Leipzig glossing rules and the corrected Frantz spelling is given in line 1a. This is shown in example (1), from Katoyissa (Glenbow 2001), below (details of the glossing format are discussed in Section 3.4 below):

(1) Line 1: *Ota'kotsi mi aohkii iitoohtoyiihkiawa moisk iinaksipokayini aawaasainiinayi.*

1a:	Otaakotssi	mi	aohkiiyi	iitoohtoyiihkiaawa …
2:	ot-saakotsi-hsi	om-yi	aohkii-yi	iit-yoohto-yiihk-yi=aawa
3:	3-boil.over.II-CONJ	DD-IN.SG	water-IN.SG	DCT- hear.TA-NAR-3PL=PRO.3PL
1:	… amoysska	i'naksipokaayini	awaasai'niyinayi	
2:	am-o-yi-hka	i'naksipokaa-yini	a-waasai'ni-yini=ayi	
3:	DP-4SG-INVS	baby-4SG	DUR-cry.AI-4SG=DTP.SG	

4: 'As the water boiled, they heard a crying baby.' ("Katoyissa," Glenbow 2001)

Texts that are not written in the Frantz orthography are included as long as they are written in a consistent manner that allows for easy interpretation. For example, the stories from the Siksika Curriculum, *Siksika Old Stories level 2* (Many Guns 1994) and *Siksika Old Stories level 3* (Bad Boy & Poor Eagle 1994) are not in the Frantz orthography. But they appear to be written by the same person who used the same spelling conventions throughout the stories, this is discussed further in Sections 4.13 and 4.14. The Uhlenbeck (1911) collection of stories are written in a personalized phonetic transcription. Both collections mentioned above are included in the BNTC; the way they are transcribed allows for a fairly straightforward transliteration into the Frantz spelling.

Transliteration for corpus inclusion means taking the spelling system that is used in the sources and applying the Frantz orthography to it. The process of the transliteration is done by means of partial or full morphological analysis using the Blackfoot dictionaries (Frantz & Russell 2017; Genee & Frantz n.d.) for reference. Transliteration for the Uhlenbeck (1911) collection was done with the help of a spelling converter script that was developed by Dominik Kadlec; more information is given on this in Section 4.12 with the discussion of the Uhlenbeck collection. The choice to standardize all the forms to the Frantz orthography was for the purposes of corpus searchability. Barth and Schnell (2021, 102) mention, regarding spoken transcriptions, that corpus queries become more complicated when there are phonetic variants of the same word form that are orthographically different, which is exactly what happens when spoken Blackfoot is written down, as discussed by Genee (2020). The standardization of the Blackfoot texts is kept in line 1, namely line la, of the 4-line gloss structure. The other lines of the gloss structure are further described in the next section.

3.4 Linguistic analysis

The linguistic analysis of the Blackfoot texts includes 4-line interlinearizations. The 4-line morphological analyses are glossed according to the Leipzig glossing format (Comrie, Haspelmath & Bickel 2015), which is a standardized set of rules and terms used by linguists for interlinearizations. Example (2) below outlines what each line is for. The texts that have analysis were analyzed according to this format, elaborated below. All information is displayed in Korp, which will be shown further in Section 5.2.

(2) Line 1: Blackfoot orthography as it appears in original source
Line 1a: standard Blackfoot orthography
Line 2: morpheme breakdown
Line 3: morpheme analysis
Line 4: ‘Idiomatic English translation’

The texts need to be machine-readable to be collated and investigated with the help of computers (Barth & Schnell 2021, 7). The stories needed to be properly formatted before any of them can be correctly displayed on the Korp platform. Some stories, for instance from *Ákaitsiniksiistsi* (Heavy Shields Russell & Genee 2009), were in Microsoft Word documents with their 4-line analyses, which I received from Inge Genee. Other texts were in PDFs with their interlinearizations, and yet others first needed to be scanned and digitized. For the texts in PDFs and those that were scanned as PDFs, I made use of the Adobe Acrobat function of Optical Character Recognition (OCR) to streamline the digitization process. First, I put the story content into Plain Text files by copying and pasting the content after running it through OCR. Once all the stories were digitized in this way, I manually double checked that the digitized versions matched the original versions, in terms of spelling and the formatting. Some texts had multiple line breaks throughout a single phrase/sentence. I grouped the Blackfoot text by the sentence if the original sources had breaks throughout the sentences. This ensures that the Blackfoot is as coherent as possible, giving as much context as possible.

After the stories were double checked, they were put into individual Microsoft Excel files. Column A is reserved for the Blackfoot phrase as it is in the original source, corresponding to line 1 above. The phrase from column A is split into its individual words with each word on its own row in column B (corresponding to line 1a above). Column C (line 2) shows the morpheme breaks for the word in column B; column D (line 3) has the morpheme gloss corresponding to the breaks in column C. Column E (line 4) is the entire English phrase as it is in the original source, in the same row as the original Blackfoot phrase in column A. Column F is used for notes about words in column B. Each phrase is separated by a single empty row. An example is shown in Figure 4. Having the stories organized in Excel sheets makes for a smooth transition into the back end of the Korp.

A	B	C	D	E	F
Omi stohkanai'naksstssim otanowawayi anniyai aimsimiootoyiisoyiihk otsitapiimiksi.	Omi	om-yi	DD-4SG	'Their youngest daughter prepared a meal for her people.'	
	isstohkanai'nakstssim ma	isstohkana-i'nakstssi-wa	most-young.AI-3SG		Note: Morpheme-Final Allomorphy (Frantz, 2017, p. 88) Original: isstohkana-i'nakstssi-m most-be.young.AI-NOM
	otanoaawayi	w-itan-oaawa-yi	3-daughter-3PL.POSS-4SG		
	anniayi	ann-yi-ayi	DM-OBV-VBLZ		
	otaisimiotoyiisookiihki aawa	ot-a-isimi-mato-yiiso-ok-yiihk-yi=aawa	3-DUR-secretly-go-feed.TA-INV-NAR-3PL=PRO.3PL		Note: The original transcript has "aimsimiootoyiisoyiihk" the spelling here is consistent with the audio; transcript also contains "otsitapiimiksi" which is not in audio.
**empty row between phrases					

Figure 4. Excel Screenshot of linguistic analysis from "Katoyissa" (Glenbow 2001).

The linguistic analysis of the analyzed texts was done manually. The texts that had existing analysis were double checked to ensure their morpheme breakdowns (line 2) were as complete as possible, and their morpheme glosses (line 3) were standardized to match the list of glossing abbreviations given in the List of Abbreviations on pages xii-xiii and expanded in Appendix 1. The morpheme breakdowns were done following the principles laid out in the *Blackfoot Grammar* (Frantz 2017) and using the entries in the *Blackfoot Dictionary of stems, roots and*

affixes (Frantz & Rusell 2017) and the *Online Blackfoot Dictionary* (Genee & Frantz n.d.). Frantz's *Blackfoot Grammar* describes predictable interactions between roots and affixes, including allomorphy (2017, 84-89) and phonological rules (2017, 28-32, 176-179). The Blackfoot dictionaries give most of the morphemes that exist in Blackfoot, their definitions and some diagnostic forms that also help with analysis along with showing how different roots interact with affixes. Using these resources in conjunction, the morpheme breakdowns and standard spelling were established. In uncertain cases of analysis, such as if a morpheme could not be identified, it is marked by a question mark.

Identifying morphemes, establishing morpheme breaks and assigning analysis can be challenging. In the beginning, I did a trial with the Blackfoot FST developed by Dominik Kadlec (2023) in hopes of speeding up analyzing the texts. In the testing of the FST, I found that it was not developed enough to use the results as an aid to establish the morpheme breakdowns and glosses. Since the texts were not all orthographically homogenous, the FST couldn't accurately assign an analysis, and it generated too many potential analyses. The results of the FST were difficult to decipher and in the end, I decided against using it, as I would have to reference the Blackfoot dictionaries to assess the FST results anyway; it was faster to do the analysis manually. Although the Blackfoot FST as developed by Kadlec (2023) was not developed enough to use for my project, there have been major recent developments to advance the computational model (e.g. Schmirler & Arppe, under development; Schmirler, Arppe & Genee 2024; Genee et al. 2023; Schmirler et al. 2024) which are promising for future work.

A list of standard glossing abbreviations used in the corpus was compiled specifically for this project. The list of glossing abbreviations is an amalgamation of various resources' abbreviations including the Leipzig glossing rules (Comrie, Haspelmath & Bickel 2015), conventions that are

followed by Algonquian linguists, the *Blackfoot Grammar* (Frantz 2017) and Shupbach's (2013) MA thesis which provides a novel analysis of the Blackfoot demonstrative system. This is shown in detail in in Appendix 1.

There are also several abbreviations I introduced that differ from those commonly used. For example, the prefix that Frantz (2017, 102) refers to as the "associative" linker prefix: {omohp- ~ iihp- ~ ohp-}, which has been previously glossed in various ways, like "ASSOC" or "with," depending on context. In this thesis, I adopt the abbreviation "COM" for comitative, following the Leipzig rules, as it precisely captures the function of this prefix. Another prefix with a similar function is {omoht- ~ iiht- ~ oht-} which is commonly glossed as "source," "means," "content," "path" or "INSTR," depending on context. I have chosen to gloss it as "INS" for instrument as a consistent gloss for this prefix, including its allomorphs. A third linker prefix denotes temporal or spatial information is {it- ~ ist-}; this is commonly glossed as "then" or "there." I have chosen to adopt the gloss "DCT" for "deictic preverb" reflecting its context-dependent meaning, following Schupbach (2013). There are three prefixes that Frantz (2017, 101) refers to as "degree" prefixes, which are {iik-} glossed as "very," {sska'-} glossed as "extra(ordinary)," and {sstonnat-} glossed as "extreme." In other literature, they may all be glossed as "very," but I chose to use the gloss "INTS" for "intensifier" to reflect their intensifying properties. All the analyzed texts follow the list of abbreviations to make the search functions easier. The next section will discuss the publication of the corpus on the Korp platform.

3.5 Publishing the BNTC

The overall goal of the BNTC is to assist in the revitalization of Blackfoot. The corpus is intended for language learners and teachers. Initially, the plan was to publish the BNTC as a

collection of text files or one large text file which would have been publicly available on the *Blackfoot Language Resources* website,[21] similar to the planned publication of the Kadlec corpus mentioned in Section 2.3. The hope was to eventually have a user interface implemented on the BLR website where the corpus could be searched and utilized by language learners and teachers. It was decided against publishing this way because it has many disadvantages. One of the disadvantages, as mentioned in Section 2.3, is this type of publication greatly lowers the usability of the corpus. The user will have to know which software makes use of flat text files, like AntConc (Anthony 2023) or WordSmith (Scott 2024).

The BNTC is now published on the Korp[22] interface (Borin, Forsberg & Roxendal 2012), available through the Alberta Language Technology Laboratory (ALTLab). Korp was originally developed in Sweden as part of *Språkbanken* (The Language Bank of Sweden). *Språkbanken*[23] has multiple free, accessible digital language tools. Korp makes use of the IMS Open Corpus Workbench (CWB) (Evert & Hardie 2011). Korp is also used for the morphosyntactically tagged Plains Cree corpus (Arppe et al. 2020) and in future, Ojibwe texts will also be available (Felipe Bañados Schwerter, p.c.). As mentioned in Arppe et al. (2020, 11), Korp "was adapted by [their] Norwegian collaborators, the Giellatekno and Divvun research teams at UiT Arctic University of Norway, for the morphologically rich indigenous Sámi language," which makes it suitable for the Blackfoot language as well because, like Sámi, it is also morphologically rich.

The process for uploading the Blackfoot texts into the BNTC has two steps. Once the Blackfoot texts are complete, I upload the files into a share drive. Felipe Bañados Schwerter,[24]

[21] Available at https://blackfoot.algonquianlanguages.ca/
[22] Available at https://korp.altlab.dev/#?cqp=%5B%5D&corpus=blackfoot
[23] Available at https://xn--sprkbanken-35a.se/om-oss
[24] I am grateful to Felipe for the work he's done in processing the Blackfoot text files for the inclusion in the BNTC. Without Felipe, the BNTC would not be as accessible as it is.

the computer programmer working with the ALTab, converts the files into the appropriate format and uploads them to Korp.

4. The Blackfoot Narrative Text Corpus: description of included texts

Each source included in the corpus is described in this section. The main information about each text is summarized in Table 3 and Table 5 below, including story title, the year it was published, total Blackfoot word count, dialect and the speaker's name. Table 3 presents the texts that have full analysis included in the corpus and Table 5 lists the texts that have partial analysis. This section will also include a basic description of the Blackfoot Narrative Text Corpus.

Table 3 below lists the texts that have full analysis in the Blackfoot narrative corpus. Full analysis means that the text is entirely morphologically analyzed in an interlinear glossed format. Each word of the text has an analysis attached to it in the corpus. Some of the texts had analyses that were originally included with them, others I have analyzed myself. For the ones that came with analyses, I used these as a basis for my own analysis. I changed the glossing abbreviations so that they are all consistent and if there was an analysis that I didn't agree with, it would be retained in the notes section, and my analysis would be the main analysis shown in the corpus.

Table 3. Summary of Blackfoot texts included in the BNTC with full analysis.

Title of Blackfoot Text	Year	Total Bkft words	Dialect	Speaker
A finger bone and a rag doll	2008	50	Kainai	Lena Heavy Shields Russell
A spirit	2008	56	Kainai	Lena Heavy Shields Russell
An old woman left behind	1969	58	Siksika	Matthew Many Guns
Blood Clot	2001	113	?	?
Cold weather	2008	25	Kainai	Lena Heavy Shields Russell
Ikasskini	1965	214	Siksika	Matthew Many Guns
In Flanders Fields	2018	56	Kainai	Lena Heavy Shields Russell
Making Bannock or fry bread	2017	111	Kainai	Beverly Little Bear Hungry Wolf

Making berry soup	2017	55	Kainai	Beverly Little Bear Hungry Wolf
My father, Rides-Many-Horses #1	2008	70	Kainai	Lena Heavy Shields Russell
My father, Rides-Many-Horses #2	2008	57	Kainai	Lena Heavy Shields Russell
Naaahsa is an artist!	2023	350	Kainai	Translators: Faye Heavy Shield & Norma Russell
Napi and the Bullberries	2001	59	?	?
Olden days	1983	126	Aamskaapipiikani	Annie Mad Plume
Prayer	2009	32	Kainai	Lena Heavy Shields Russell
Rattlesnakes	2012	43	Kainai	Lena Heavy Shields Russell
The Lord's prayer	?	21	Siksika	Interpreter: Paul Bird
The women's buffalo jump	2013	242	Kainai	Sandra Many Feathers, Brent Prairie Chicken, Wes Crazy Bull
Why the Blackfoot language is important to preserve	2009	28	Kainai	Lena Heavy Shields Russell
Wolf Trail	2001	71	?	?
Total Blackfoot Words 1837				

Table 4 below shows the basic descriptions of various parts of speech within the fully analyzed texts. The following statistics were found using the extended search within Korp, which will be fully explained in Section 5.1. It should be noted that the noun count may not be exactly accurate as the search was for any analysis that contains "-AN" or "-IN" pertaining to the animate or inanimate singular or plural suffixes, which occur on nouns but also on demonstratives.

Table 4: Parts of speech statistics based on analyzed texts within the BNTC

Parts of Speech Statistics			
Total Nouns: 504			
Animate Nouns (NA): 239	Inanimate Nouns (NI): 265		
Total Verbs: 776			
Intransitive Inanimate Verbs (II): 86	Intransitive Animate Verbs (AI): 408		
Transitive Inanimate Verbs (TI): 108	Transitive Animate Verbs (TA): 184		
Total Derived nouns: 83			
From TA: 6	From TI: 10	From AI: 50	From II: 4
Total Demonstratives: 253			
{*om*}: 73	{*ann*}: 104	{*am*}: 75	

Table 5 below includes the texts that have partial analysis in the Blackfoot narrative corpus. Each of the texts listed have the bare minimum of English to Blackfoot phrase alignment. Other variations of partial analysis include a loose translation of each Blackfoot word to a few phrases and/or words that have a full gloss.

Table 5. Summary of Blackfoot texts included in the BNTC with partial analysis.

Title of Blackfoot Text	**Year**	**Total Bkft words**	**Dialect**	**Speaker**
A woman sacrificed to a butte	1910	38	Aamskaapipiikani	Boys & Tatsey
An old woman left on a camp-ground	1910	66	Aamskaapipiikani	Boys & Tatsey
Bear-chief's songs	1910	99	Aamskaapipiikani	Bear-chief & Tatsey
Belly-fat	1910	897	Aamskaapipiikani	Boys & Tatsey

Blue-face	1910	390	Aamskaapipiikani	Boys & Tatsey
Buffalo calling stone	2001	74	?	?
Clot-of-blood	1910	1326	Aamskaapipiikani	Boys & Tatsey
Horses found on an island	1910	96	Aamskaapipiikani	Boys & Tatsey
Names of clans	1910	251	Aamskaapipiikani	Boys & Tatsey
Napi and the black birch	2001	60	?	?
Napi and the bullberries	2001	59	?	?
Napi and the bullberries	1993	125	Siksika	Matthew Many Guns
Napi and the coyote eyes	1993	595	Siksika	Matthew Many Guns
Napi and the coyote race	1993	184	Siksika	Beatrice Poor Eagle
Napi and the mice	1993	231	Siksika	Matthew Many Guns
No more buffalo	1983	530	Aamskaapipiikani	Jenny Running Crane
Old man/Napi	2001	60	?	?
Scar-face	1910	556	Aamskaapipiikani	Boys & Tatsey
Small number and the basketball tournament	2012	261	Piikani/ Kainai	Speaker: Eldon Yellowhorn Translation help: Connie & Andy Crop Eared Wolf
Small Number and the Kit Foxes	2015	392	Piikani/ Kainai	Speaker: Eldon Yellowhorn Translation help: Connie & Andy Crop Eared Wolf
Small Number counts to 100	2010	136	Piikani/ Kainai	Speaker: Eldon Yellowhorn Translation help: Connie & Andy Crop Eared Wolf
The bear and the thunder woman	1994	268	Siksika	Margaret Bad Boy
The leader-buffalo	1910	317	Aamskaapipiikani	Boys & Tatsey
The origin of the buffaloes	1910	555	Aamskaapipiikani	Bear-chief & Tatsey
The origin of the buffalo-stones	1910	96	Aamskaapipiikani	Boys & Tatsey
The people living in the north	1910	84	Aamskaapipiikani	Boys & Tatsey
The six neglected boys	2001	72	?	?
The story of Blackfoot Ridge	1983	131	Aamskaapipiikani	Mae Calf Boss Ribs
The two buffalo-lodges	1910	183	Aamskaapipiikani	Boys & Tatsey

The wolf trail	2001	71	?	?
The wolverine	1910	113	Aamskaapipiikani	Boys & Tatsey
Thunder	2001	103	?	?
Two adventures of the Old Man	1910	114	Aamskaapipiikani	Boys & Tatsey
Whom-the-buffalo-inquires-after	1910	22	Aamskaapipiikani	Boys & Tatsey
Total Blackfoot Words		8555		

In total, there are 54 texts included of which 19 are analyzed and 35 are unanalyzed. 10 texts from the Glenbow *Niitsitapiisini* webpage are told by an unknown male voice, but the texts themselves were a collaborative effort, which also means that these cannot be assigned to a specific dialect. The 17 texts from the Uhlenbeck collection are noted as Aamskaapipiikani dialect, all interpreted by Joseph Tatsey, whom was from Kainai, living in Aamskaapipiikani. There are three additional texts that are of Aamskaapipiikani dialect. Eight texts are of Siksika dialect, three texts are of Apatohsipiikani dialect (with translation assistance from individuals from Kainai) and 13 texts are of Kainai dialect. 17 texts were told by women, 23 texts were told by men and/or interpreted by a man and 14 were a collaborative effort between men and women. The story tellers ages range from those of the young boys in the Uhlenbeck collection to that of an 81-year-old woman.

4.1 *Ákaitsinikssiistsi: Blackfoot Stories of Old*

A book called *Ákaitsinikssiistsi: Blackfoot Stories of Old* told and written by Lena Heavy Shields Russell and edited by Inge Genee (2014) includes eight short Blackfoot texts. Lena Heavy Shields Russell is from Kainai. All stories, except “Rattlesnakes,” were told throughout 2008-2009, when Heavy Shields Russell was 75-76 years old. “Rattlesnakes” was told in 2012

when she was 79 years old. The following stories are included in the book and in the corpus, all are narratives, unless otherwise noted:

1. "Why the Blackfoot language is important to preserve" (Heavy Shields Russell & Genee 2014, 3-5) is an argumentative text,
2. "Prayer" (Heavy Shields Russell & Genee 2014, 7-9) is a prayer text,
3. "My father, Rides-Many-Horses #1" (Heavy Shields Russell & Genee 2014, 11-13),
4. "My father, Rides-Many-Horses #2" (Heavy Shields Russell & Genee 2014, 15-17),
5. "A finger bone and a rag doll" (Heavy Shields Russell & Genee 2014, 19-21),
6. "A spirit" (Heavy Shields Russell & Genee 2014, 23-25),
7. "Cold weather" (Heavy Shields Russell & Genee 2014, 27-29),
8. "Rattlesnakes" (Heavy Shields Russell & Genee 2014, 31-33).

Dr. Inge Genee gave me the unpublished analyses that were used in the writing of the stories and preparation of the glossary, and I used these as a basis to prepare my own analysis. Genee (p.c.) mentioned that some of the stories were first audio recorded and then transcribed, but others were written first by Heavy Shields Russell, then audio recorded and transcribed thereafter. All eight stories are fully analyzed in the corpus.

4.2 Blackfoot Language Resources Story Archive

The *Blackfoot Language Resources* website includes a story archive[25] with audio recordings of stories told in Blackfoot. The website includes 20 stories in the Siksika dialect and 5 stories in the Kainai dialect. Two procedural texts are included as told by Beverly Little Bear Hungry Wolf, titled

1. "Making Bannock or Fry Bread,"
2. "Making Berry Soup"

Little Bear Hungry Wolf, a female from Kainai, speaks Blackfoot as her mother tongue language. Both these stories were collected by audio recording in 2017, transcribed and analyzed

[25] Available at https://stories.blackfoot.atlas-ling.ca/#/stories

in 2021. Both texts are examples of very natural Blackfoot speech, and as a result include a few filler words and hesitations/repetitions. The disfluencies, hesitations and false starts made some words difficult to interpret, and these are not included in the analyses. I originally analyzed "Making Bannock or Fry Bread," while Inge Genee analyzed "Making Berry Soup," in 2021. I then corrected both analyses for inclusion in the corpus. Both stories are fully analyzed in the corpus.

4.4 *Aakiipisskani 'the women's buffalo jump'*

Aakiipisskani 'the women's buffalo jump' (Áistainskiaakii Many Feathers et al. 2013) is a Napi story, written in English, Blackfoot and Blackfoot syllabics; it has several authors listed: Sandra Many Feathers, Brent Prairie Chicken, Wes Crazy Bull, and David Osgarby. Many Feathers firstly told the story in English, and then it was transcribed by Osgarby. Many Feathers, Prairie Chicken, and Crazy Bull, all from Kainai, collaborated to translate the story into Blackfoot from the English transcription. "Osgarby transcribed and interlinearised the Blackfoot recording sessions" (Áistainskiaakii Many Feathers et al. 2013, 1-2). No further information was given nor found about the authors. This story has 242 analyzed Blackfoot words in the Frantz orthography; it is also given in syllabics, also included in the corpus. The original interlinear glosses are given in 86 different lines in Section 6, the numbers correspond to the Blackfoot syllabics in Section 3, the Blackfoot orthography in Section 4, the English in Section 5. There was one correction I made while putting the data into the spreadsheet; the English, Blackfoot and gloss labelled (64) corresponds to the Blackfoot syllabics labelled (65) and the English, Blackfoot and gloss labelled (65) corresponds to the Blackfoot syllabics labelled (64). In the spreadsheet, I switched the syllabics from (64) and (65) so that they match the correct English,

Blackfoot and gloss – this change is noted in the "notes" column and will be visible in the corpus. This story is fully analyzed in the corpus.

4.3 "An old woman left behind"

A narrative text called "An old woman left behind" told by Matthew Many Guns is included as an appendix in Genee (2009, 936-939). Many Guns is a male from Siksika. The story was told in 1969 at the age of 48. This story is fully interlinearized, it has 58 Blackfoot words and is in the Frantz orthography. I used the analyses in the article as a basis for my own analysis. Genee mentions in the article that the story was collected by means of a recording by Don Frantz and was transcribed and analyzed initially by him as well (2009, 936). This story is fully analyzed in the corpus.

4.5 "Ikasskini"

A legend text called "Ikasskini" retold by Matthew Many Guns is included in Frantz (2017, 187-197) as an appendix. Many Guns is from Siksika; he told the story in 1965 at the age of 44. The story is fully interlinearized, has 214 Blackfoot words, and is in the Frantz orthography. In the book it states that "events described herein are said to have taken place in 1843. Story recorded as told by Jack Big Eye of the Siksika reserve in 1965. Retold the same year by Matthew Many Guns after listening to the recording by JBE." (Frantz 2017, 187). Therefore, we can assume that the stories were collected by means of audio recording then, retold in another recording and then transcribed into writing by Frantz. In the analysis, there are some textual notes that Frantz includes relating back to the version told by Big Eye, which are not included in

the corpus. I used the analyses as a basis for my own analysis. The story is fully analyzed in the corpus.

4.6 *Naaahsa aisinaki! Naaahsa in an artist!*

A children's book called *Naaahsa aisinaki! Naaahsa is an artist!* is a contemporary narrative text written by Hali Heavy Shield. It was published in 2023 and is written in Blackfoot with English translations. The Blackfoot translations were done by Faye Heavy Shield and Norma Jean Russell, both females from Kainai. At the time of publishing, Faye Heavy Shield was 70 years old. The story is given in Frantz orthography, which needed only a few small spelling corrections. The interlinearization is by me. The story itself contains 350 Blackfoot words. I personally obtained permission from Hali Heavy Shield to include the story in the corpus via email. The story is fully analyzed in the corpus.

4.7 "The Lord's Prayer"

A prayer text called "The Lord's Prayer" is included in Ermineskin and Howe (2005). In the article, it's stated that the prayer was found handwritten in Tims' notebook in syllabics, no information is given as to when the prayer was written. This prayer is fully interlinearized in the article (Ermineskin & Howe 2005, 2-4), I used these analyses as a basis for my own analysis. It has 21 Blackfoot words and is in the Frantz orthography. His interpreter was a male named Paul Bird; no further information about the interpreter and collection were included. The prayer is also given in syllabics, although the syllabics will not be included in the corpus at this time. The prayer is fully analyzed in the corpus.

4.8 "In Flanders Fields'"

The famous poem "In Flanders Fields" has been translated into Blackfoot by Lena Heavy Shields Russell, originally published in 2018 (Kalinowski 2018), officially published in 2021 (Pulido-Guzman 2021). Kalinowski (2018) mentions that Heavy Shields Russell started working on the translation in 2003, meaning she was between the ages 70-85 while she was working on the translation. Pulido-Guzman (2021) mentions that Glenn Miller put the official poster together with the Blackfoot and the English translations. The official poster is where the Blackfoot and English came from for inclusion in the corpus. It contains 56 Blackfoot words. I worked on the interlinearizations of the text, the full analysis is included in the corpus.

4.9 Glenbow Traditional Stories

The Glenbow Museum had an exhibition called "Niitsitapiisini: Our Way of Life," which was open from November 2001 to November 2020 ("Glenbow" n.d.). This exhibition has a companion website called *Niitsitapiisini, Our Way of Life.*[26] The website contains eight traditional stories. The stories include:

1. "The Six Neglected Boys/Miohpokoiksi," a legend,
2. "Buffalo Calling Stone/Iinisskimm," a legend,
3. "Blood Clot/Katoyissa," a legend,
4. "Thunder/Aksisstsikomma," a legend,
5. "The Wolf Trail/Makoyoohsokoyi," a legend,
6. "Napi And the Black Birch/Napi ki Siikokiinis," a Napi legend,
7. "Napi and the Bullberries/Napi ki Mi'ksinittsiimiksi," a Napi legend,
8. "Old Man/Napi," a narrative text.

[26] Available at https://www.glenbow.org/blackfoot/BL/html/traditional_stories.htm

These stories are not in the Frantz orthography and do not have analyses included. The website also contains audio recordings, with a male voice, and a written English translation. On the Glenbow website, it mentions several names that collaborated in the *Niitsitapiisini: Our Way of Life* exhibition before it opened in 2001. From this, I'm assuming the recordings were done prior to 2001; there is no specific information on whose voice it is in the recordings. I wrote an email to the addresses available on the website, to find out further information about the site contents, but I did not hear anything back. All the stories, not including "Blood Clot/Katoyissa," "The Wolf Trail/Makoyoohsokoyi," and "Napi and the Bullberries/Napi ki Mi'ksinittsiimiksi," are included in the corpus with Blackfoot to English phrase alignment. The "Blood Clot" story was originally analyzed by Heather Bliss (p.c.) in 2010 and updated in 2024; "The Wolf Trail" analyzed in 2024; "Napi and the Bullberries" was also analyzed in 2024. With Bliss's permission I used her analyses as a basis to prepare my own analysis. These texts are fully analyzed in the corpus.

4.10 Small Number collection

Simon Fraser University has a collection of animated stories about a character "Small Number," told in a multitude of languages; there are three stories that are narrated in Blackfoot by Eldon Yellow Horn. The copyright for the Small Number stories belongs to the Math Catcher Outreach Program;[27] I received permission from Veselin Jungic as the coordinator of the Math Catcher Outreach Program to use the stories in the corpus. The three stories in Blackfoot are all contemporary short stories:

1. "Small Number Counts to 100,"[28] published in 2010,

[27] Available at https://www.sfu.ca/~vjungic/Small-Number/book-1.html
[28] Available at https://www.sfu.ca/~vjungic/Small-Number/sec_SN_01_B.html

2. “Small Number and the Basketball Tournament,”[29] published in 2012,
3. “Small Number and the Kit Foxes,”[30] published in 2015.

“Small Number Counts to 100” and “Small Number and the Kit Foxes” were written in English by Jungic and Mark MacLean; “Small Number and the Basketball Tournament” was written in English by Jungic. The three stories were then translated into Blackfoot by Eldon Yellow Horn, a male from Piikani Nation in his early 50s, with help from Connie Crop Eared Wolf, a female from Kainai Nation in her mid 50s, and Andy Crop Eared Wolf, a male from Kainai Nation in his early 60s, between 2010 and 2015. Yellow Horn, C. Crop Eared Wolf and A. Crop Eared Wolf grew up as bilingual speakers. The three stories will be included in the corpus with a Blackfoot to English phrase alignment at minimum.

The story “Small Number Counts to 100” was inspired by a story that Jungic and co-author Dr. Mark MacLean heard from Ms. Rina Sinclair, a Siksika elder. Ms. Sinclair told a story about her childhood and her encounter with a “beautiful black cat” that turned out to be a skunk. With Ms. Sinclair’s permission, they combined her story with a well-known mathematical puzzle. The story “Small Number and the Basketball Tournament” was inspired by Jungic’s experience as a volunteer at the Vancouver Friendship Centre. The story “Small Number and the Kit Foxes” was inspired by a story that Jungic and MacLean heard from Ms. Kathy Scout-Bastien of the Piikani Nation. Ms. Scout-Bastien gave permission to include parts of her story in the Small Number story (Jungic, p.c., email dated January 2025)

[29] Available at https://www.sfu.ca/~vjungic/Small-Number/sec_SN_03_B.html
[30] Available at https://www.sfu.ca/~vjungic/Small-Number/sec_SN_08_B.html

4.11 *Stories of Our Blackfeet Grandmothers*

A book called *Stories of our Blackfeet Grandmothers*, originating from the Heart Butte Bilingual Program (1984), includes three narrative Blackfoot texts:

1. "No More Buffalo" (Heart Butte Bilingual Program 1984, 1-14) told by Jenny Running Crane,
2. "Blackfoot Ridge" (Heart Butte Bilingual Program 1984, 15-18) told by Mae Calf Boss Ribs,
3. "Olden Days" (Heart Butte Bilingual Program 1984, 19-22) called told by Annie Mad Plume.

At the telling of these stories Running Crane was 69 years old, Calf Boss Ribs was around 63 years of age, Mad Plume was 69 years old; these three ladies are from Aamskaapipiikani. In the editor's notes (Heart Butte Bilingual Program 1984, iv) it states that "the stories are from a recording of Blackfeet Elders telling stories to students at Heart Butte School during the Spring of 1983." Therefore, we can assume that the stories were collected by means of audio recording then it was transcribed into writing by Norma Russell with the assistance of Donald Frantz. All stories are written in the Frantz orthography. Norma Russell is listed as the author of the book; she was solely responsible for preparing the Blackfoot text with the help of Don Frantz. All three stories are included in the corpus; "Olden Days" is fully interlinearized and the other two stories are included with Blackfoot to English phrase alignment.

4.12 *Original Blackfoot texts from the southern Peigans Blackfoot reservation, Teton County, Montana*

A book of Blackfoot texts called *Original Blackfoot texts from the southern Peigans Blackfoot reservation, Teton County, Montana* was written and published by C.C Uhlenbeck in 1911. The stories are all narrative texts, unless noted otherwise:

1. “Names of Clans” (Uhlenbeck 1911, 1-4), a descriptive text,
2. “The people living in the north” (Uhlenbeck 1911, 5),
3. “The origin of the buffaloes” (Uhlenbeck 1911, 6-12),
4. “The origin of the Buffalo-stones” (Uhlenbeck 1911, 12-13),
5. “The leader-buffalo” (Uhlenbeck 1911, 13-18),
6. “Blue-face” (Uhlenbeck 1911, 18-23),
7. “Belly-fat” (Uhlenbeck 1911, 23-34),
8. “Clot-of-blood” (Uhlenbeck 1911, 34-50), a legend text,
9. “Scar-face” (Uhlenbeck 1911, 50-57), a legend text,
10. “Horses found on an island” (Uhlenbeck 1911, 57-58),
11. “The two buffalo-lodges” (Uhlenbeck 1911, 58-60),
12. “The wolverine” (Uhlenbeck 1911, 60-61),
13. “An old woman left on a camp-ground” (Uhlenbeck 1911, 62),
14. “A woman sacrificed to a butte” (Uhlenbeck 1911,62-63),
15. “Two adventures of the Old Man” (Uhlenbeck 1911, 63-65), a Napi legend,
16. “Whom-the-buffalo-inquires-after” (Uhlenbeck 1911, 65-66),
17. “Bear-chief’s songs” (Uhlenbeck 1911, 66-68), a song in text form

The stories were collected between May 11 and August 15, 1910, while Uhlenbeck was staying in Montana (Uhlenbeck 1911, v). His interpreter was Joseph Tatsey of “Blood Descent,” staying in Aamskaapipiikani, where his mother comes from (Uhlenbeck 1911, iv). Uhlenbeck mentions that these stories were collected in the afternoons and evenings from younger boys, except for two, “the origin of the buffalo” and “Bear-chief’s songs” were told by Bear-chief himself, but that the texts were all verified by Tatsey (Uhlenbeck 1911, v). Therefore, we can assume that the texts are in the Aamskaapipiikani dialect, taking into account that the interpreter came from the Kainai dialect area. No further details were given about Tatsey.

Multiple steps were taken before these texts could be included in the corpus. The first was that the series had to be digitized. I am grateful to Natalie Weber who let me use her digitized version that was originally done for her corpus *Blackfoot Words*. Secondly, there were some mistakes in the digitized transcription that needed to be corrected. In the orthography system that Uhlenbeck created, there are many accents used, but some of these were misconstrued as glottal stops during

transcription; an example of this correction is shown in (3) below, with the specific corrections bolded. After the correction of the texts, they were also formatted into spreadsheets.

(3) Original: Kɛ́nni**χ'**kaiei**χ'**tsístapanistàini**χ'**kataiau amóksi Isksínaitapìks.
Edited: Kɛ́nni**χ**kaiei**χ**tsístapanistàini**χ́**kataiau amóksi Isksínaitapìks.
English: 'Then these were called Bug-people.' ("Names of clans" Uhlenbeck 1911, 2)

These texts are not in the Frantz orthography, which did not exist at the time; rather they are written in a semi-phonetic transcription system created by Uhlenbeck for the purpose. This orthography needed to be converted to the Frantz system prior to the inclusion of these texts in the corpus. The third step was therefore to determine which of Uhlenbeck's sounds match with Frantz' sounds in the orthography. This was done by matching Uhlenbeck's sounds to the International Phonetic Alphabet (IPA), then matching the IPA to the characters used by Frantz. Table 6 shows what that process looked like. Fourthly, Dominik Kadlec created an orthography converter from Table 6. Then Katherine Schmirler ran the texts through the orthography converter, with some small changes to the converter itself. Lastly, the texts were included in the corpus, with the basic transliteration into the Frantz orthography with Blackfoot to English alignment. It's important to note that the texts that are included in the corpus are the automatically converted versions that will eventually need to be hand-checked against the Frantz orthography and the dictionaries.

Table 6. Comparison of Uhlenbeck and Frantz orthography to IPA.

	Uhlenbeck	**IPA**	**Frantz**
Vowels	A	ʌ / a / aː	a / aa
	ă	ʌ	A
	ä	ɛː	Ai
	å	ɔː	Ao
	A	ʌ	A
	E	e / eː	ai(i) / ii
	ɛ	ɛ	Ai

	i	i / iː / ɪ	i / ii
	o	o / oː / ʊ	o / oo
	u	ʊ	o / oo
	ai/ɛ	æ	Ai
	au	aʊ	aw(a)
Fricatives	χ	x	H
	(i)χ́	(i)ç	(ii)h
Consonants / Sem-vowel	k	k	k / kk
	m	m	m / mm
	n	n	n / nn
	p	p	p / pp
	s	s	s / ss
	t	t	t / tt
	w	w	W
	y	j	Y
	’	ʔ	’

4.13 *Aakaitapitsinniksiists: Siksika Old Stories Level 2*

The curriculum called *Siksikai’powahsin/Siksika Language Series Kit* developed by Vivian Ayoungman (1993) includes a book titled *Aakaitapitsinniksiists: Siksika Old Stories Level 2* (Many Guns 1994). The book includes three Napi stories told by Matthew Many Guns:

1. “Napi and the Mice”
2. “Napi and the Coyote Eyes”
3. “Napi and the Bullberries”

Many Guns is from Siksika; his life spanned from 1921-1993 (Many Guns 1994, 2); thus, the book was published after Many Guns already passed. The stories were transcribed from an audio tape recording, further details about the collection of the stories were not included. Before each story, there is a note that states “The text is not written word-for-word as the taped story. Each time a story is told orally, it changes slightly depending on what the storyteller wants to emphasize” (Many Guns 1994, 18, 31, 46). These stories are in an orthography that appears to be

making an attempt at the Frantz orthography; it is updated into the Frantz orthography in the corpus as shown in (4) below. All three stories are included in the corpus with Blackfoot to English alignment.

(4) Niitsohtakoyihk, kanaiskiinaiks itsapomahksipaskayihkiya omiim otokaani.
Niitsohtakoyiihk kaanaisskiinaiksi ita'pomahksipasskayiihkiaawa omima o'tokaani.
'There was a loud sound, some mice were having a pow-wow in the elk's head.'
("Napi and the Mice," Many Guns 1994)

In this book of stories, there are two versions of each story, a shorter illustrated version and a longer unillustrated version. I chose the longer versions to include in the BNTC, this will be elaborated on in the following section.

4.14 *Aakaitapitsinniksiists: Siksika Old Stories Level 3*

The curriculum called *Siksikai'powahsin/Siksika Language Series Kit* developed by Vivian Ayoungman (1993) includes a book titled *Aakaitapitsinniksiists: Siksika Old Stories Level 3* (Bad Boy & Poor Eagle 1994). The book includes two stories:

1. A Napi legend called "Napi and the Coyote Race," told by Beatrice Poor Eagle,
2. A narrative text called "The Bear and The Thunder Woman," told by Margaret Bad Boy.

Poor Eagle was 81 and Bad Boy was 93 at the time of publishing, both from Siksika, both females. The stories were transcribed from an audio tape recording, further details about the collection of the stories were not included. Before each story, there is a note that states "The text is not written word-for-word as the taped story. Each time a story is told orally, it changes slightly depending on what the storyteller wants to emphasize" (Bad Boy & Poor Eagle 1994, 24, 43). These texts are written in a non-standard orthography, but they also appear to be written by the same person as the stories in Level 2 (described in Section 4.13 above), from the various

spelling conventions used. The orthography, like Level 2 above, makes an attempt at the Frantz orthography; all words use the letters of the Blackfoot alphabet as laid out by Frantz, as shown in (5) below.

(5) Original: Ki omaaka apiisi ihtsitotomahkayihka.
Frantz orthography: Ki omaka aapi'siwa ihtsito'toomaahkaayiihka.
English translation: 'And a coyote came along.'
("Napi and the Coyote Race," Poor Eagle 1994)

In this book, as mentioned above, there are two versions of each story; one version, I assume to be the original recorded story which is longer and includes more details; the other version includes illustrations with coinciding excerpts from the longer version of the story. I took the longer versions or each story because they contain more detail and have a better representation of natural Blackfoot speech. The stories have been transliterated into the Frantz orthography in the corpus with a Blackfoot to English alignment.

5. Using the corpus

This chapter will present the different functions and types of searches that can be done with the Blackfoot Narrative Text Corpus. The BNTC can be beneficial for teachers, learners, linguists and others by providing examples of Blackfoot phrases and/or words that are standardized to the Frantz orthography.

The BNTC can assist teachers in practical ways to support their teaching and lesson development within the classroom. Teachers can utilize the corpus to find real-life vocabulary examples in natural contexts with specific grammar points to make lessons about. Finding realistic examples ensures that acceptable forms are being taught to the students in an appropriate context. It's important to note that the BNTC cites where all content comes from, so the teacher can also cite exactly when, where, and who the examples come from.

Learners can use the corpus to find how native speakers use specific words in multiple contexts, also showing different inflections of roots. Another way the corpus can support learners is by illustrating how different morphemes interact at the morphophonological level. In Blackfoot, there are multiple phonological rules that apply to words in different contexts; these are described by Frantz in Chapter 5 (2017, 28-31) and all rules are listed in an appendix (2017, 176-179). Native speakers do not have to consciously think about such rules –
they are applied automatically in speech, whereas a second language learner must be explicitly taught these rules and must actively consider how and where the rules apply, both orally and orthographically. An example is the rule called *ih-Loss* which Frantz lists in a rewrite rule formation: $ih \rightarrow \emptyset \,/\, s\_s$ (2017, 177). *ih-Loss* is where an <ih> is lost when it is both preceded and followed by an <s>. An example is shown below in (6), with the <ih> combination italicized in the second line, which does not appear in the first line due to *ih-Loss*.

(6) piókska'ssini
pi-okska's*i*-*h*sin-yi
long-run.ai-nmlz-in.sg
'a long run' (Frantz 2017, 130)

Linguists can use the corpus to find morphological and syntactic patterns. This type of analysis might include studying frequencies and usage patterns. The BNTC is capable of compiling statistics of specific searches, which would help the linguist in their analysis of patterns. Since the BNTC consists of only narrative texts, this can be useful in analyzing morphological and syntactic patterns and their relationships that arise in the texts. An example of this is analyzing discourse topicality in regard to verbal and morphology derivation within Blackfoot texts, as was done by Genee (2009). The BNTC displays information in a KWIC format, explained below in Section 5.1, which is useful for these analyses.

Other researchers could use the corpus for a multitude of reasons, like training data for a Blackfoot noun or verb computational model. Since the BNTC is a larger corpus of homogenous orthography it can be useful for training FSTs, similar to the one described by Kadlec (2023). Another way the BNTC can be used is to look for different spelling conventions that people use in Blackfoot. Since the original orthography from each source was preserved, one can look for a specific root and compare it against the original orthography. Section 5.1 will describe the basic search functions within Korp, and specific examples of the searches will be shown in Section 5.2.

5.1 Korp search functions

The Korp interface has three ways of indexing and searching the corpus data: simple, extended and advanced. In the *simple* search, complete Blackfoot words or Blackfoot character sequences with the choice of additional criteria to be checked (initial part, medial part, final part, case-insensitive, diacritic-insensitive) can be searched; this is shown in the screenshot in Figure 5; an actual search using will be shown in Section 5.2.

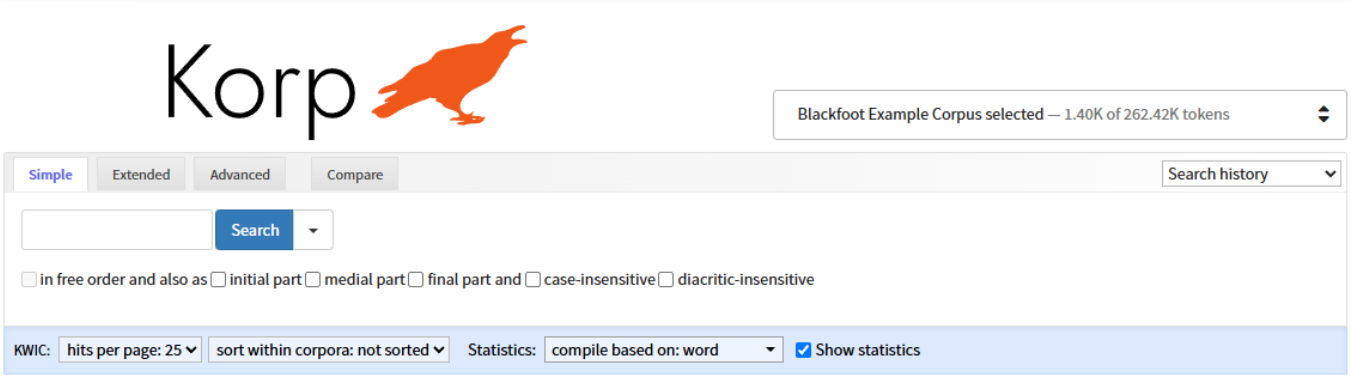


Figure 5. Korp screenshot of what the simple search function looks like.

In the *extended* search, there are multiple features that can be explicitly searched. The dropdown menu allows a user to search for different features of words throughout the corpus, e.g. by word, word attributes or text attributes as shown in Figure 6. Actual search examples using these functions are shown in Section 5.2. These options allow the user to refine their searches to exactly what they want to find.

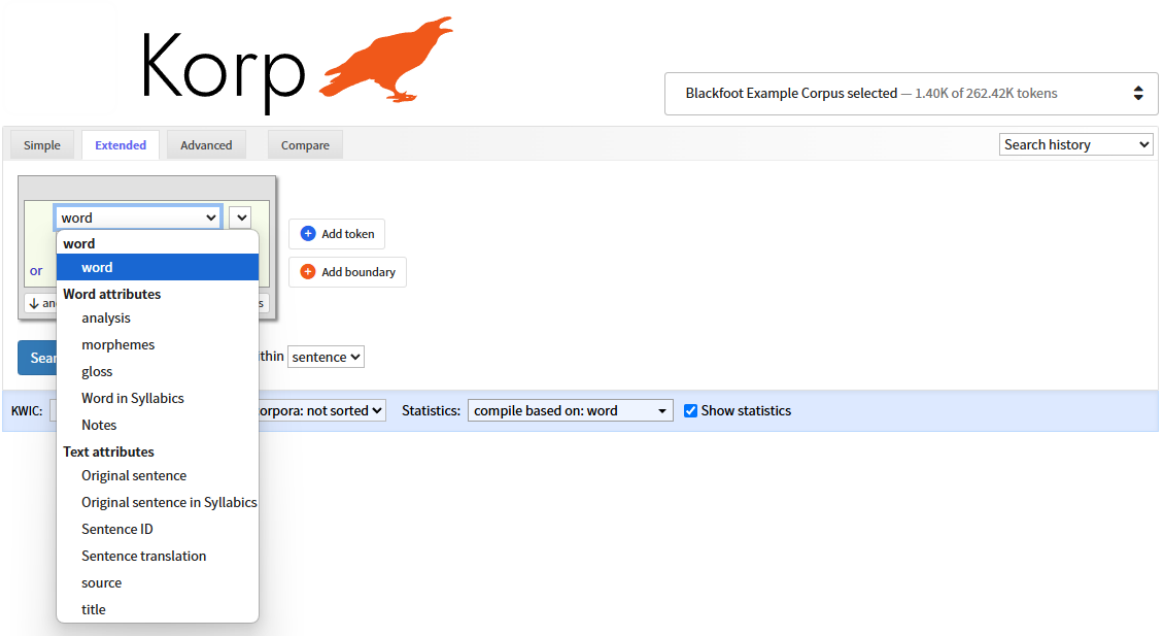


Figure 6. Korp screenshot showing what the extended search function with options dropdown menu extended looks like.

Within each search box, there is a choice to make the search case-sensitive or case-insensitive; by default, the setting is case-sensitive. The "Aa" can be clicked on to make the necessary choice, which is circled in red in Figure 7, when case-insensitive is chosen, the "Aa" turns from black to a grey color.

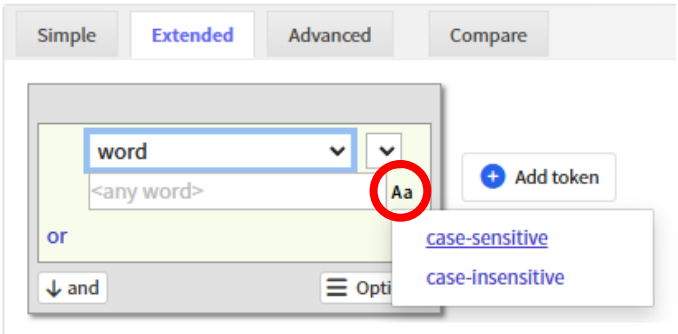


Figure 7. Korp screenshot with the case sensitivity option shown.

In the search box, there is an “or” hyperlink, highlighted with a red circle below in Figure 8, that allows for more content to be searched for. Clicking the “or” link, adds another search criteria, shown in Figure 8. The same dropdown function can be used in the second search criteria as the first; the “or” link can be clicked as many times as the user wishes. When the “or” is clicked, only one of the search criteria applies to the search results. There is also an “and” link, pointed at with the red arrow in Figure 8 in the search box, that can be clicked for additional criteria to be added to the search. When the “and” is clicked, all search criteria are applied to the search results.

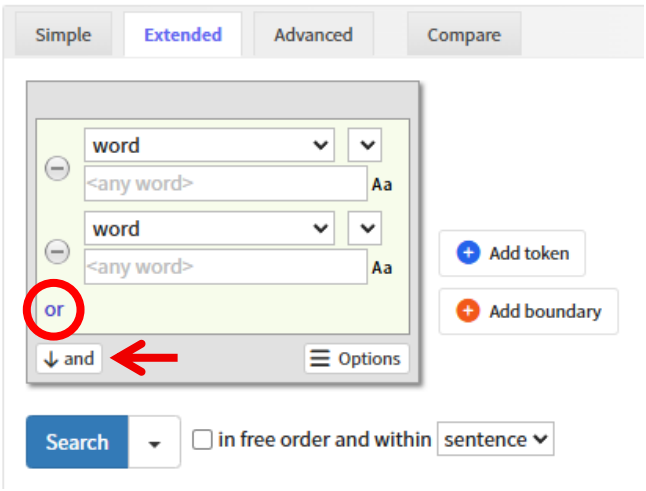


Figure 8. Korp screenshot showing the extended search function after clicking the “or” link, highlighted with a red circle.

The searches can be in more depth by adding additional tokens and/or boundary units. To add additional tokens to a search, the blue plus sign with “Add token” (on the right side below) can be clicked which adds an additional search box, which are the larger boxes shown in Figure 9. To add additional boundary units, the orange plus sign with “Add boundary” (also on the right side in Figure 9) can be clicked, an additional option is shown to choose either “first” or “last”—both examples are shown in Figure 9—once an option is chosen, the boundary unit search criteria appear (with a blue background). When any search is done, the results are shown in a

KWIC (Key Word In Context) format. With any search, there is a checkbox that allows the user to choose whether statistics are drawn up for the search. A search with the search results will be shown in Section 5.2, with both the KWIC and statistics view.

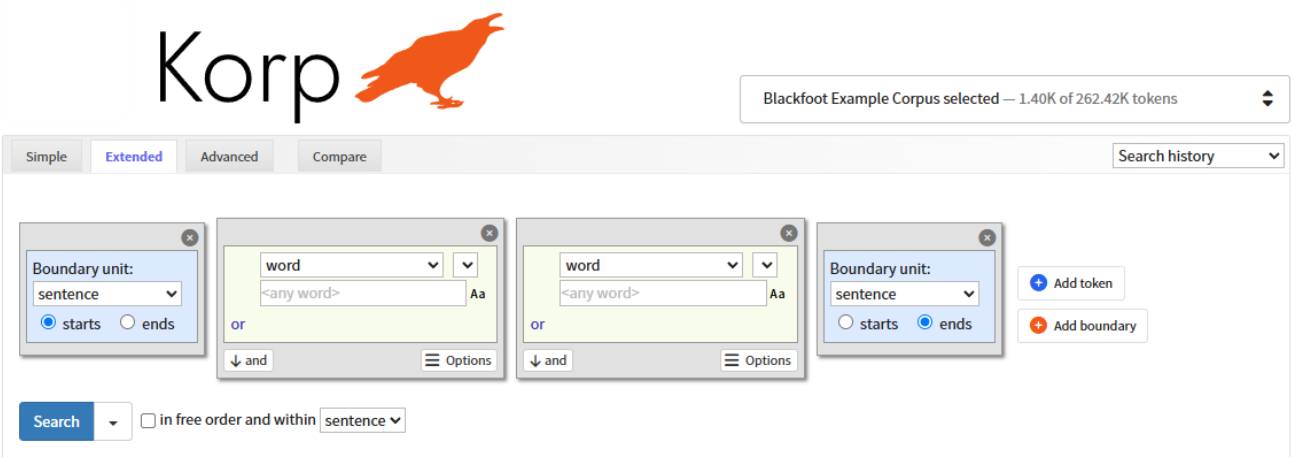


Figure 9. Korp screenshot showing additional token and boundary search boxes.

The last search function is the *advanced* type. Figure 10 shows what this looks like on the Korp interface. To utilize this search function, the user must have some basic knowledge of regular expression (RegEx) syntax. The user can also click on the "manual" link under the "Custom CQP query:" search box to get an idea of which RegEx can used to search the BNTC. Using the advanced search allows a user to tailor their searches quite a bit more than the other search functions allow for. The following section will show actual examples querying the BNTC using multiple search functions discussed in this section.

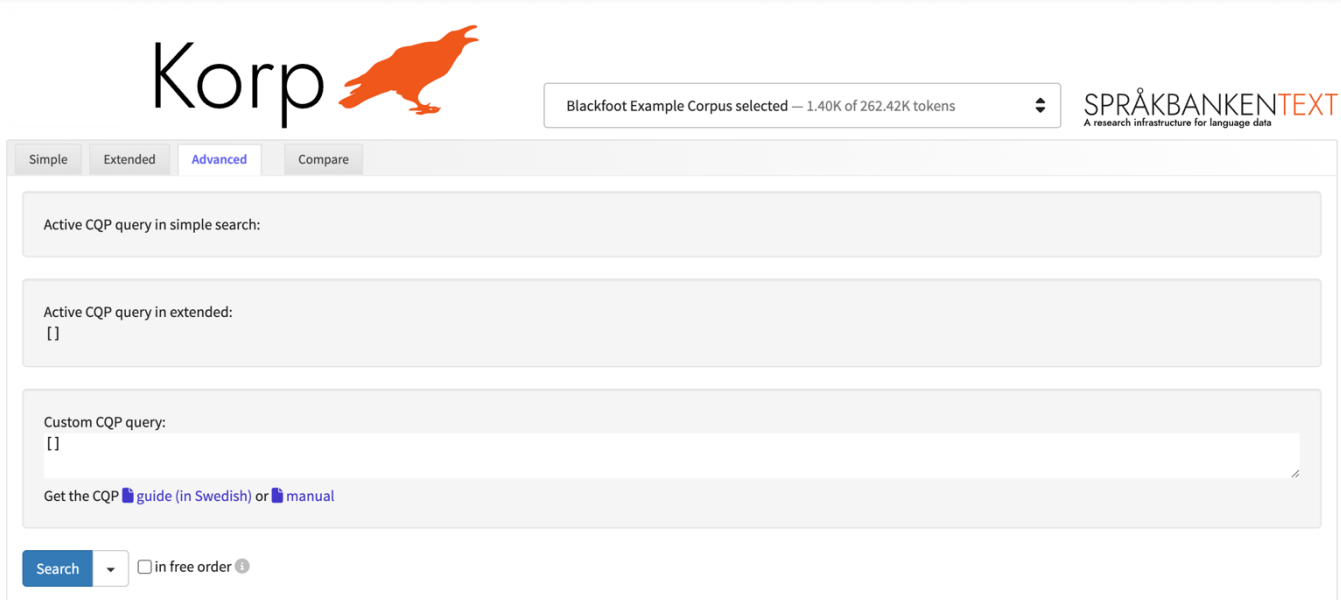


Figure 10. Korp screenshot showing what the advanced search function looks like.

5.2 Search examples

This section will present actual examples of searches that can be done within the BNTC. The first search will explain the details of the KWIC results, statistics view, and the reading view. For example, if a linguist wanted to use the corpus to search for demonstratives, Figure 11 shows an extended search for analyzed demonstratives in the corpus.

There are 17 texts that are fully analyzed, which were mentioned in Table 3 of Chapter 4. The demonstrative glossing follows the system as laid out in Schupbach (2013). The search criteria includes any analysis containing any "DD – distal demonstrative stem" for the root {*om*-} or "DM – medial demonstrative stem" for the root {*ann*-} or "DP – proximal demonstrative stem" for the root {*am*-}. The linguist would find these abbreviations listed in the "List of abbreviations" in this thesis on pages xi-xii. All search results are shown in a KWIC format, where any word that contains the search criteria is shown in bold down the center with its context on either side of the word. This particular search yields 253 results (marked by the red

arrow in Figure 11), meaning that there are 253 words in the BNTC that contain any of the demonstrative glosses in their analyses.

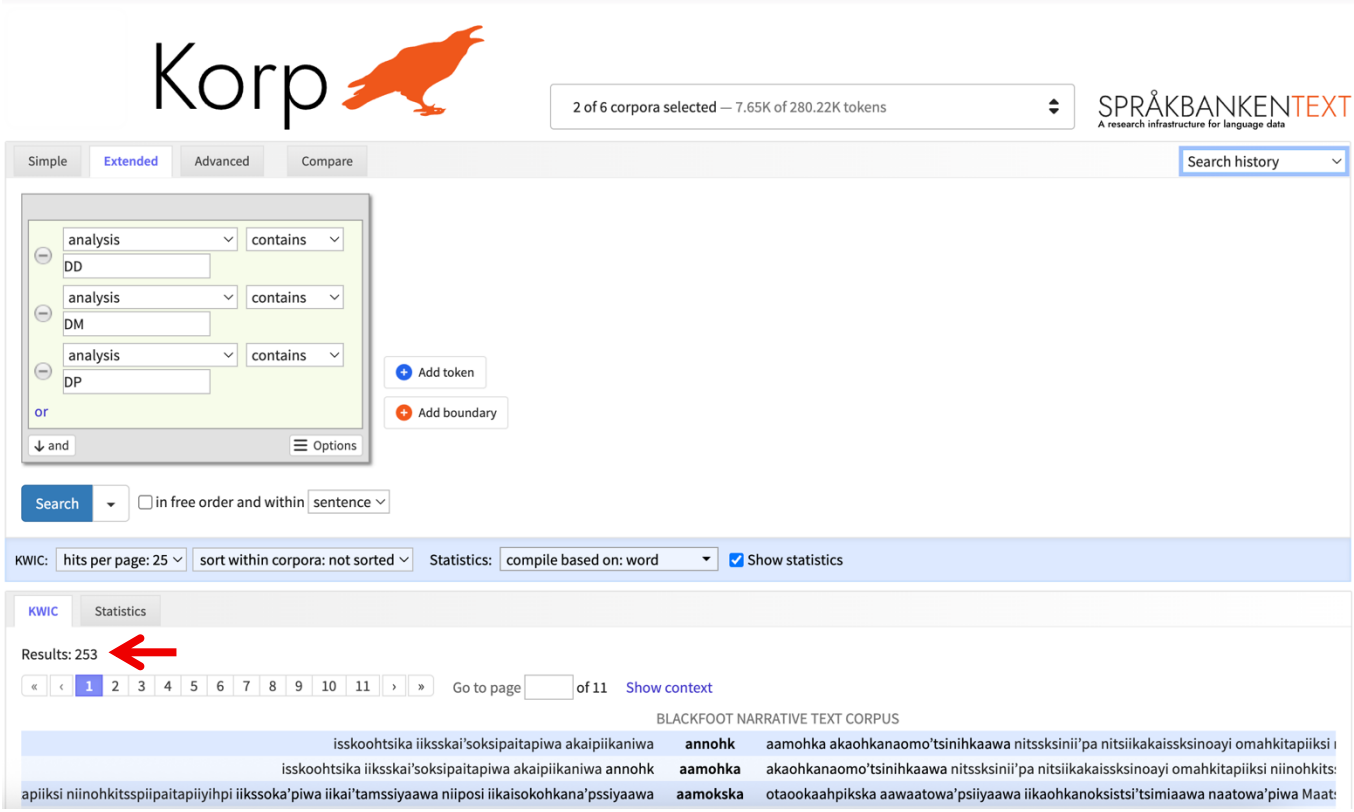


Figure 11. Korp screenshot of an extended search example for analysed demonstratives.

When any KWIC is clicked on (or any word in the context), a sidebar on the right shows up, as shown in Figure 12; the word that has been clicked on becomes highlighted–*annííksi* as shown in this screenshot. Every KWIC has standardized spelling in the Frantz orthography.

In the side bar, the following information is shown:

- **Corpus** shows the corpus title,
- **Read this text in Korp** is a link to read the text,
- **Text attributes** of the text that the word comes from is shown, including
    - its **title** with the storyteller(s) in brackets along with the year the story was told;

- the **source**;
- the **sentence translation**, which is the whole sentence that the word comes from;
- **sentence ID**, which is information that has to do with the back end of Korp, where each word within the corpus is tagged to the sentence that they belong to and each of those sentences are assigned a sentence ID to keep everything in the correct order;
- **original sentence**, which is the original Blackfoot sentence in the original orthography of the source;

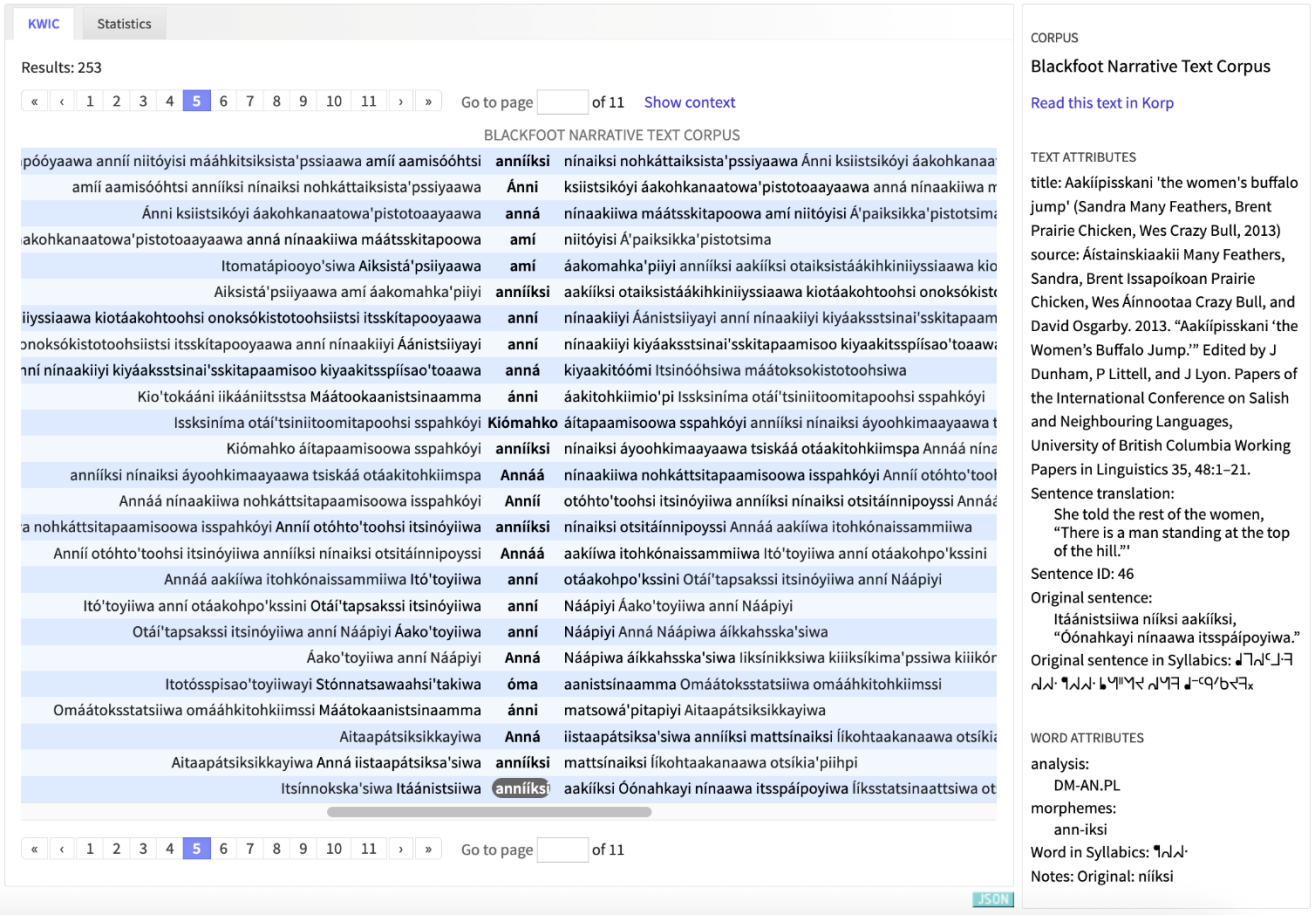
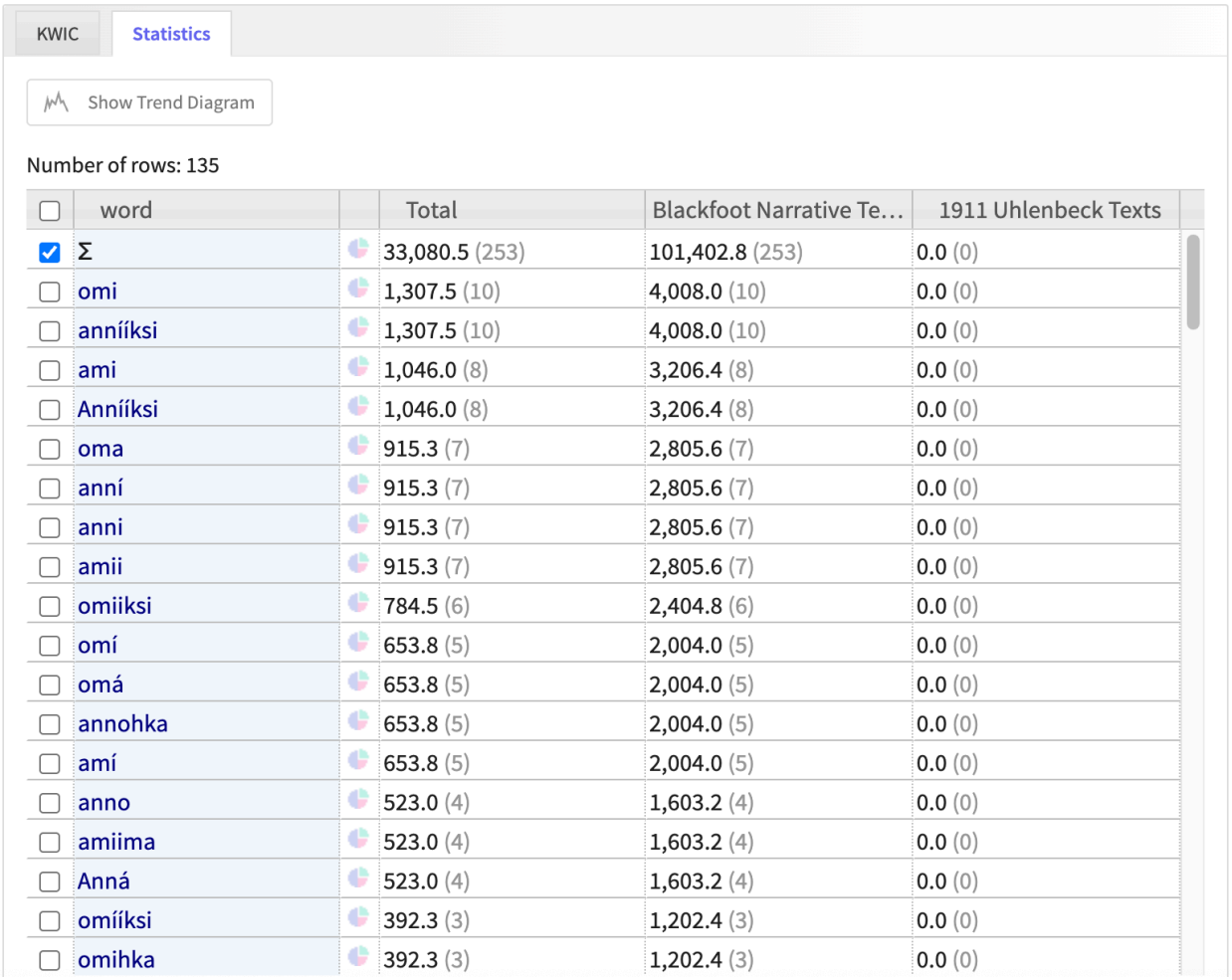


Figure 12. Korp screenshot showing the KWIC view of the search from Figure 11.

        - **original sentence in syllabics** shows the syllabics corresponding to the original phrase above (syllabics appear only if the story originally came with syllabics, if not the field will read *[empty]*).
- **Word attributes** of the KWIC chosen
    - **Analysis** shows the morpheme breaks where these are available ,
    - **Morphemes** show the word's morphemes,
    - **Word in syllabics** shows the word in Blackfoot syllabics, if the original text came with syllabics; if the spelling of the original word was corrected to match the Frantz spelling then this also shows a matching corrected syllabic version
    - **Notes** show any notes (marked by *Note:*) and/or original analyses (marked by *Original:*) that go along with the word. It can also go along with the sentence in some cases, if the word is on the same line as the phrase in the original Excel spreadsheets.
        - *Original:* preserves any analyses that came from the original texts, the analysis under the morpheme label is the analysis that I provided.
        - *Note:* marks something significant, like a change or mistake, in the original texts about the word and/or phrase.

Figure 13 shows the statistics view of the demonstrative search above in Figure 12. This view lists the individual word types that appear in the KWIC results in an ordered fashion of decreasing frequency. The left column, with the header "word," lists each individual type. The two middle columns, with headers "Total" and "Blackfoot Narrative Text Corpus" have the exact same numbers because only the texts within the Blackfoot Narrative Text Corpus category

contain analyzed texts. The column on the right with the header “1911 Uhlenbeck Texts” do not contain any analysis at the time of writing.



word	Total	Blackfoot Narrative Te…	1911 Uhlenbeck Texts
Σ	33,080.5 (253)	101,402.8 (253)	0.0 (0)
omi	1,307.5 (10)	4,008.0 (10)	0.0 (0)
anníiksi	1,307.5 (10)	4,008.0 (10)	0.0 (0)
ami	1,046.0 (8)	3,206.4 (8)	0.0 (0)
Anníiksi	1,046.0 (8)	3,206.4 (8)	0.0 (0)
oma	915.3 (7)	2,805.6 (7)	0.0 (0)
anní	915.3 (7)	2,805.6 (7)	0.0 (0)
anni	915.3 (7)	2,805.6 (7)	0.0 (0)
amii	915.3 (7)	2,805.6 (7)	0.0 (0)
omiiksi	784.5 (6)	2,404.8 (6)	0.0 (0)
omí	653.8 (5)	2,004.0 (5)	0.0 (0)
omá	653.8 (5)	2,004.0 (5)	0.0 (0)
annohka	653.8 (5)	2,004.0 (5)	0.0 (0)
amí	653.8 (5)	2,004.0 (5)	0.0 (0)
anno	523.0 (4)	1,603.2 (4)	0.0 (0)
amiima	523.0 (4)	1,603.2 (4)	0.0 (0)
Anná	523.0 (4)	1,603.2 (4)	0.0 (0)
omíiksi	392.3 (3)	1,202.4 (3)	0.0 (0)
omihka	392.3 (3)	1,202.4 (3)	0.0 (0)

Figure 13. Korp screenshot of the statistics view of the search from Figure 11.

The black number in the right columns is the relative frequency of the type and the grey number in brackets is the absolute frequency of the type, which is the total number of times that specific token appears in the BNTC that fits the specific search criteria (which are only tokens that have analysis for this search). Figure 13 is the default view of the statistics, whereas Figure 14 shows the statistics compiled with the case-insensitive option selected showing the top 10 tokens clipped in the screenshot. In order to have the statistics compiled, the “Show Statistics”

must be checked, located above the red arrow in Figure 14. Also shown is where to find the case options in the drop-down menu by the red arrow, which is done in the same fashion as explained in Section 5.1 in Figure 7. It should be noted that the difference in the "number of rows" is 21 between Figure 13 and 14, which means there is 21 less tokens in Figure 14 (has 114 rows) than in Figure 13 (has 135 rows). The sigma (Σ) row (the top row) is included in the total "number of rows." Which means that there are 134 individual types in Figure 13 and 113 individual types in Figure 14. We can see that the total results match from Figure 12, 13, and 14; the total number of tokens is shown above the KWIC view beside "Results" in Figure 12 is 253, and for Figure 13 and 14 in the top row with the sigma symbol has 253 in brackets.

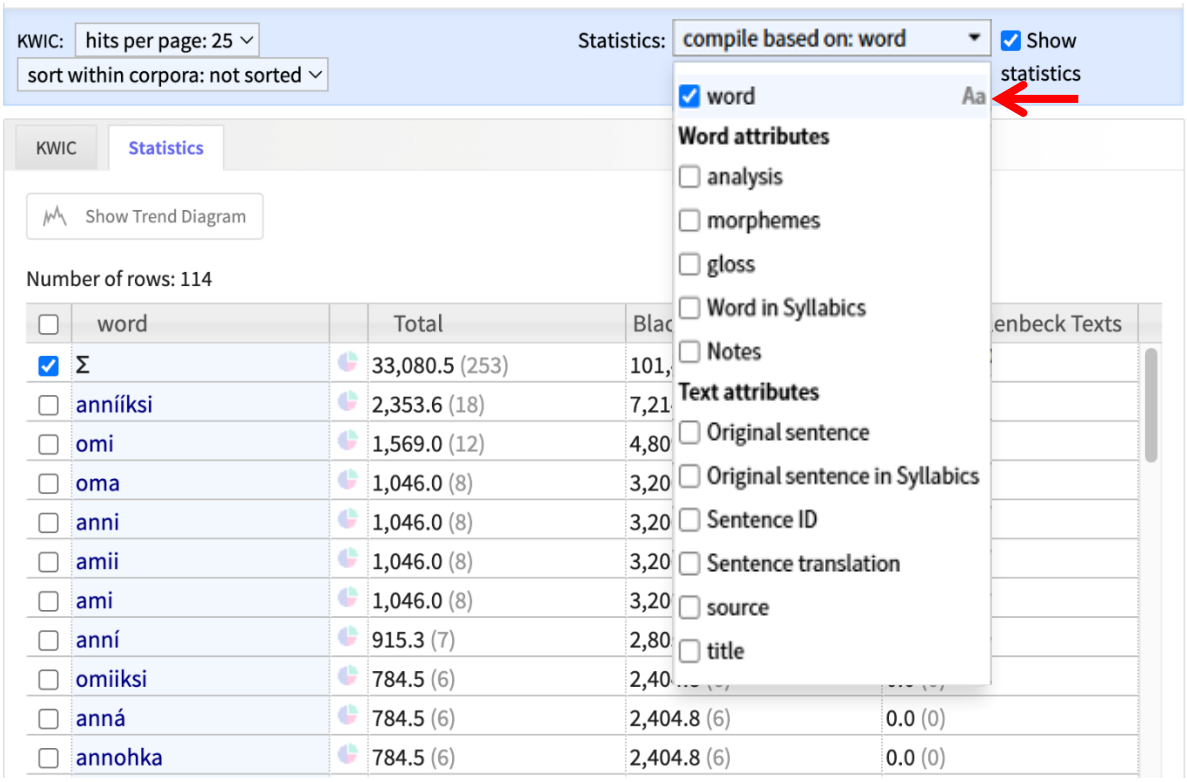


Figure 14. Korp screenshot showing case-insensitive view of the statistics of the search from Figure 12 with the case sensitivity option highlighted with a red arrow.

The linguist could infer from these results that the most frequent demonstratives in the BNTC are *anniiksi, omi, oma, anni* and *amii/ami*. This would then also allow the user to search the unanalyzed part of the corpus for additional examples. Looking at the KWIC, it can also be seen that most of the demonstratives precede nouns. Some demonstratives also precede verbs, but the inflection on the demonstrative also changes depending on what follows it, which can be further investigated using the corpus.

Figure 15 shows the reading mode view, which shows up when you click the link “Read this text in Korp.” This allows the user to read the text in the Korp interface. Any word within the story can also be clicked and its word attributes can be seen in the right sidebar, which is explained above with Figure 12.

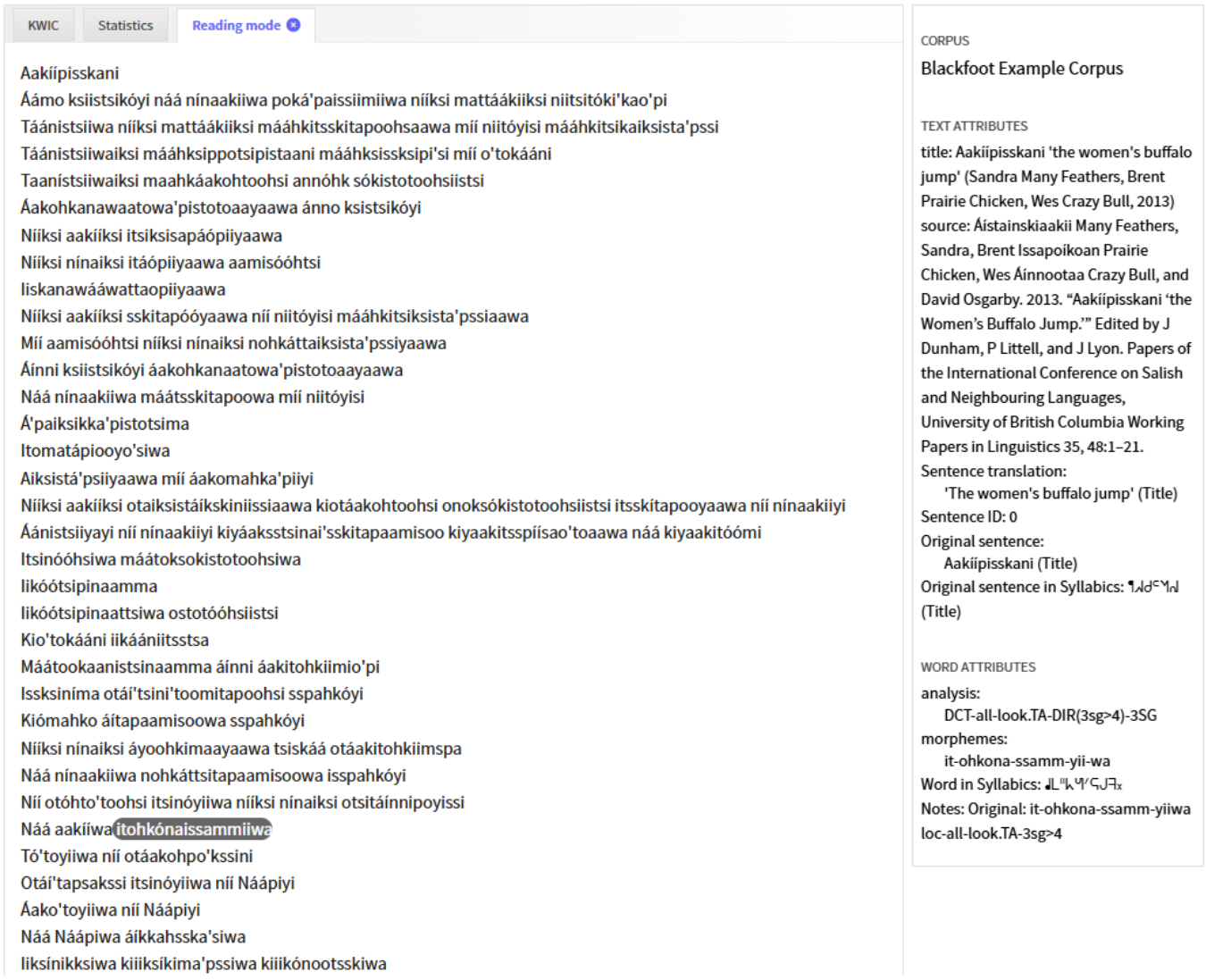


Figure 15. Korp screenshot of the reading view with a word clicked on expanding the right side bar showing the attributes of the highlighted word.

The next paragraph will discuss a situation if a teacher were teaching Blackfoot and wanted to teach about the animate transitive verb (VTA). In Blackfoot, the VTAs are more complex than

the inanimate transitive verb (VTI) in terms of their structure and agreement patterns and therefore, more difficult to teach. There are a lot more inflection possibilities for the VTAs than for the VTIs. Using the list of abbreviations (pages xi-xii), they would know that the abbreviation for a VTA is marked by "TA" – this can be searched for to find all analyzed VTAs. Figure 15 shows the search and results of all analyzed VTAs contained within the BNTC. Once the teacher conducts the search, (shown in Figure 16) they can see that there are 208 analyzed VTAs in the BNTC. Again, if any word is clicked on in the results, the sidebar on the right would appear to view information about the word clicked on. These results would allow the teacher to find real examples that are grammatically correct and standardized to the Frantz orthography. It would also allow the teacher to teach about other post-inflectional suffixes that occur on VTAs.

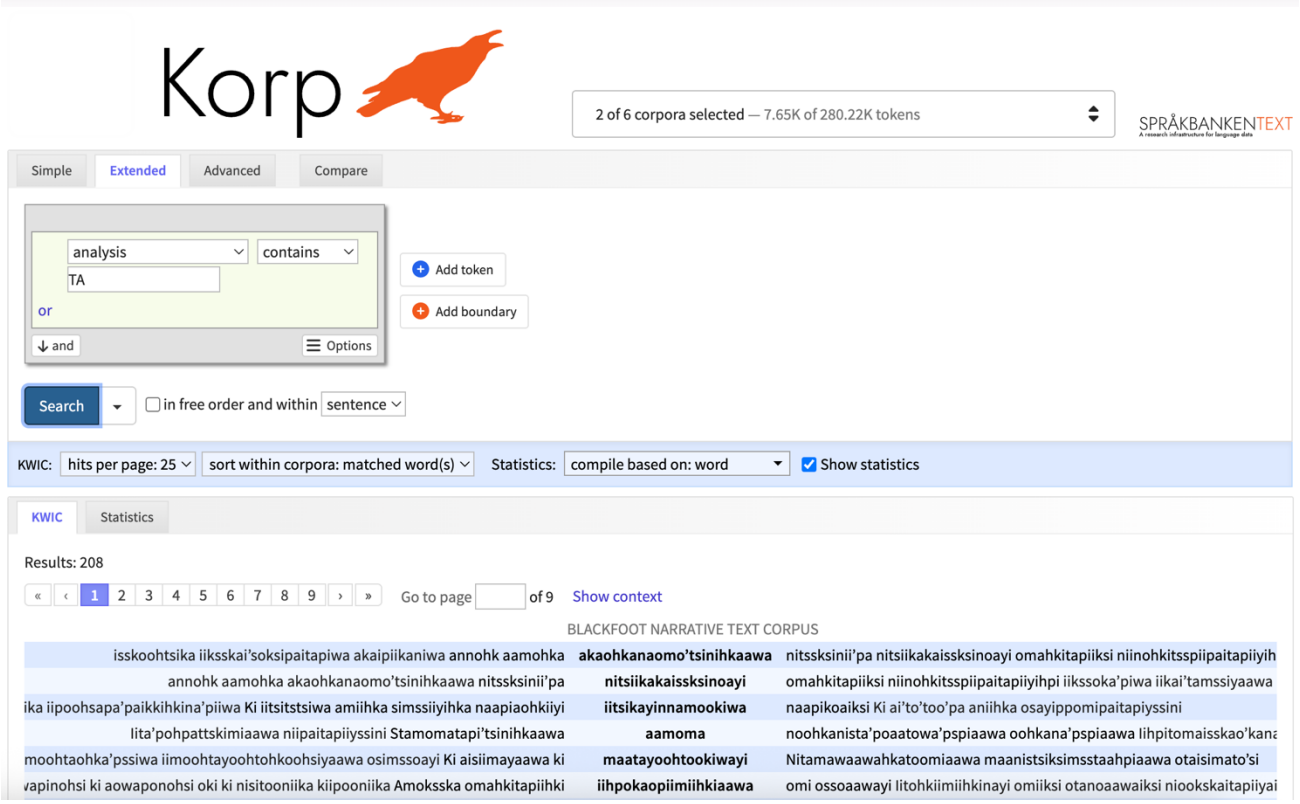


Figure 16. Korp screenshot of an extended search for analysed VTAs.

If the teacher wanted to narrow their lesson and the search results for VTAs with the direct theme suffix (DIR), they can use the search shown in Figure 17. The search for VTA and DIR yields 105 results. This search shows all contexts within the BNTC that have both TA and DIR in their analyses. The results show how different VTA roots interact with the DIR variants, so that the teacher can create a lesson on how the different phonological rules apply in different contexts because depending on how the root ends and how the suffix begins, there are interactional rules that apply and change the phonetics of Blackfoot words.

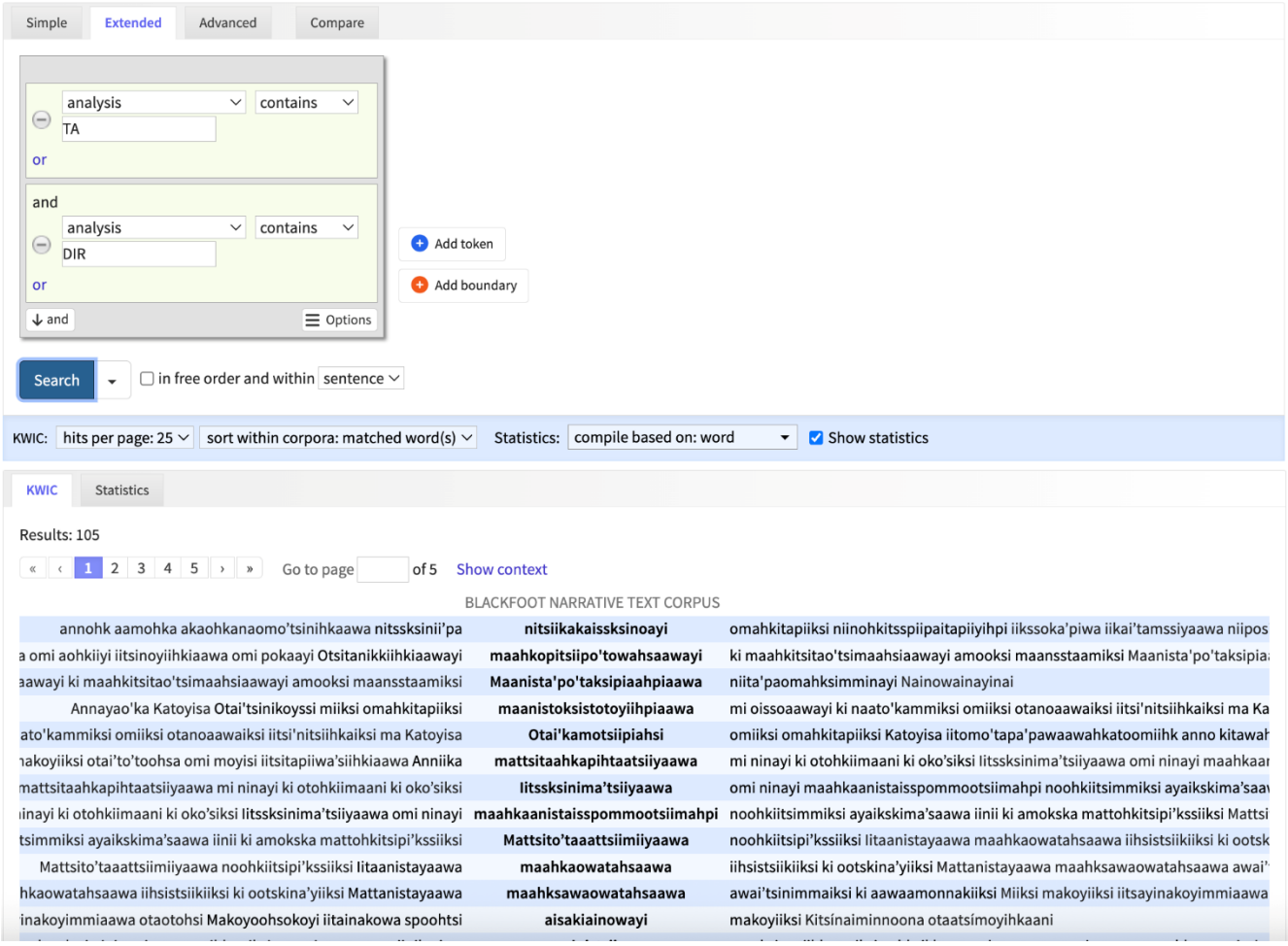


Figure 17. Korp screenshot showing an extended search for VTA and DIR.

Another example of a search in the BNTC could be for a search for first person long prefix. If a researcher wanted to see what type of morphemes follow the long variant of the first person

prefix, Figure 18 shows a simple search for *nit* in the initial part of any word, with the case-insensitive and diacritic-insensitive options chosen to widen the search results. These results show that there are 318 words that have {*nit-*} as the initial part of the word within the BNTC. The researcher could use the simple search function, as shown in Figure 18 below which shows an overview of the words that have {*nit-*} at the beginning. After sifting through these results, they can go into the extended and/or advanced search tabs to narrow their search. An interesting feature of Korp, is when there is any search done in the simple or extended search features, the search criteria are also automatically shown in the advanced search tab in RegEx. This is shown

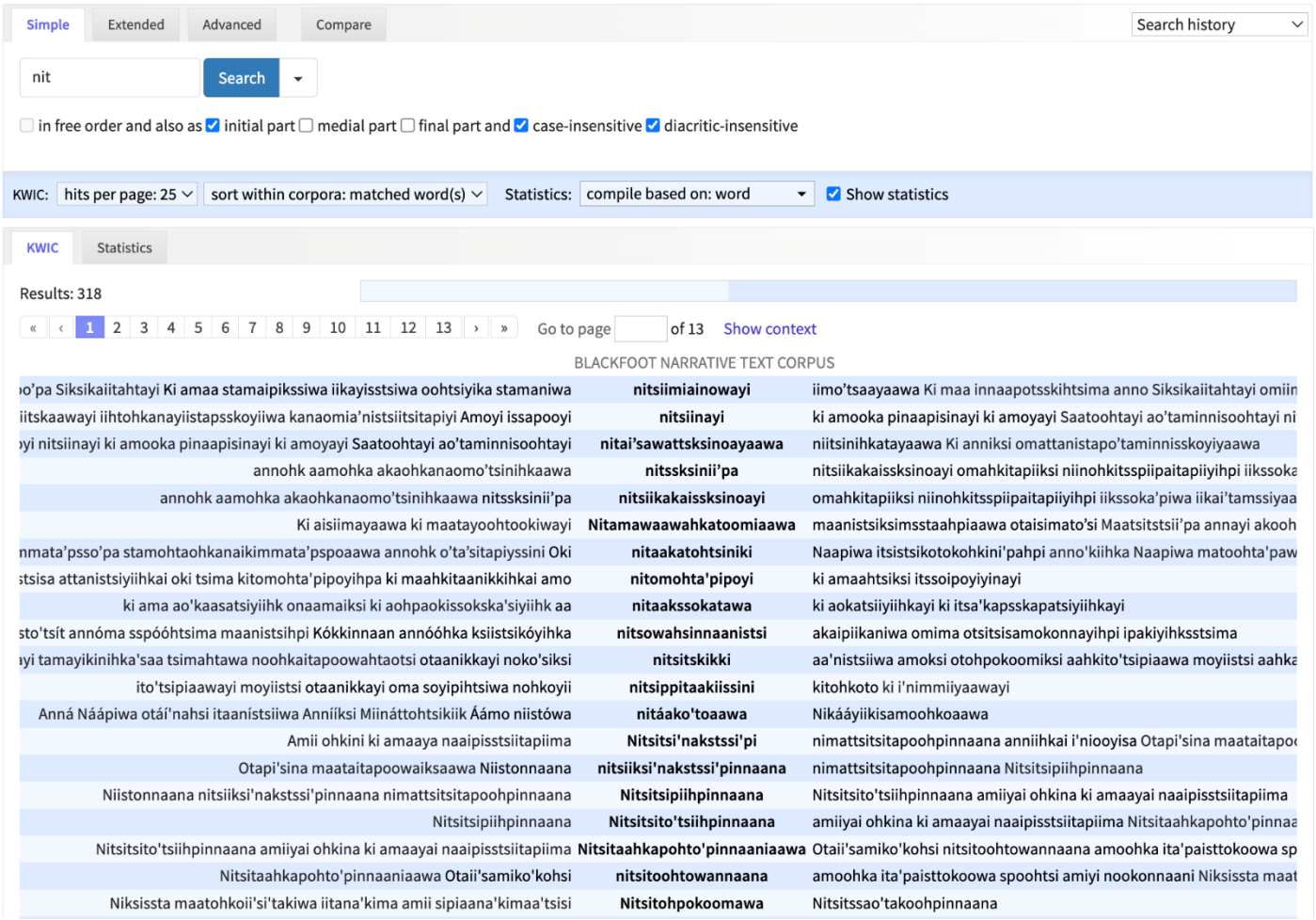


Figure 18. Korp screenshot showing a simple search for {nit} (long first person prefix).

in Figure 19 for the search shown in Figure 18. This feature can also help its users to learn more about RegEx and how to form searches in Korp.

This section has shown multiple examples of searches within the BNTC using the Korp interface. The different searches that Korp is capable of are simple, extended and advanced. This section showed actual search examples of the simple and extended kind, with a view of what the advanced search can look like. Different scenario examples were drawn using hypothetical people that would be most likely to use the BNTC.

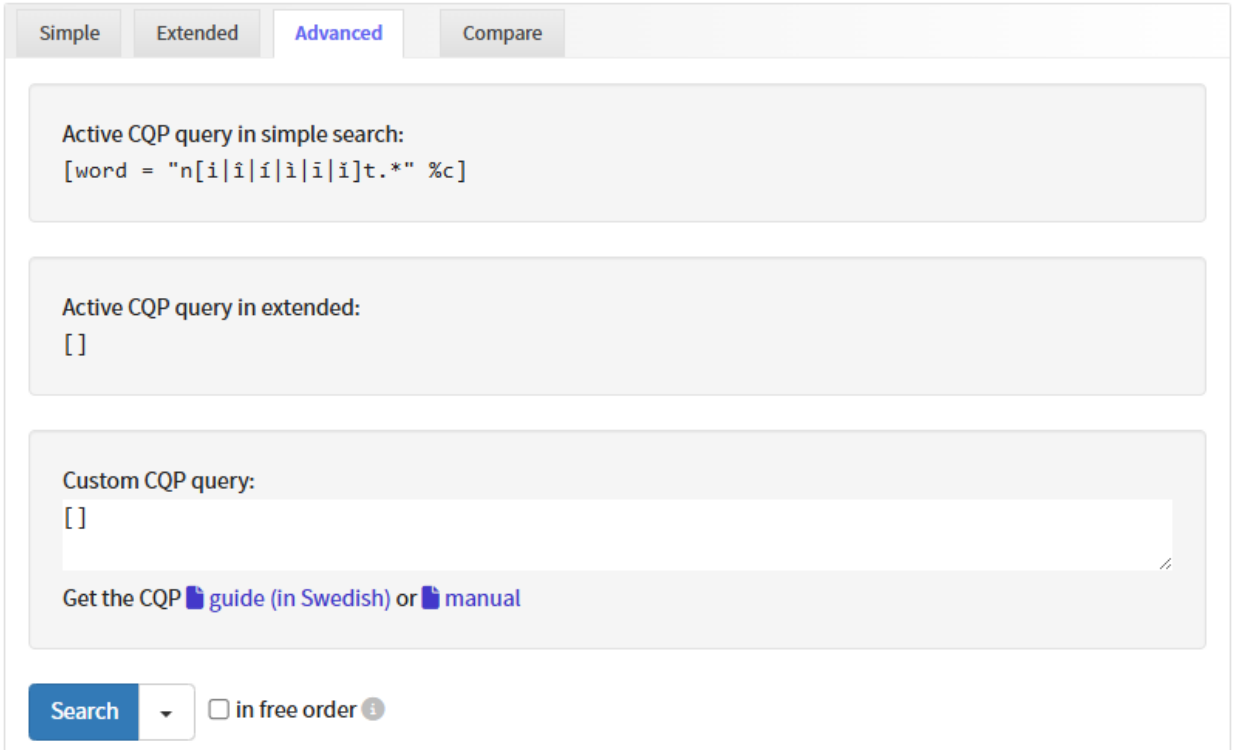


Figure 19. Korp screenshot showing the simple search (from Figure 18) shown in RegEx in the advanced tab.

6. Summary and conclusion

The purpose of this thesis was to describe the process of curating the Blackfoot Narrative Text Corpus from written texts to aid in the revitalization and documentation of the Blackfoot language. The BNTC is housed on the Korp interface with capabilities to query the texts. There are approximately 1,800 analyzed words and 8,500 unanalyzed words.

Chapter 1 laid the foundation, presenting background knowledge of the Blackfoot language. Chapter 2 discussed various aspects of a corpus in general, including points on minority and majority language corpora, publication of a corpus, other Blackfoot corpora and how the BNTC compares to other corpora. Chapter 3 presented the choices and decisions that went into the curation of the BNTC. Chapter 4 discussed all the Blackfoot texts that are included in the BNTC. Chapter 5 introduced the Korp interface and illustrated some actual search examples.

In conclusion, the BNTC is the first of its kind for Blackfoot resources. It makes previously unavailable sources widely accessible that are orthographically homogenous and are enhanced with linguistic analysis. The BNTC is a great tool for teaching and learning Blackfoot, which can be used in various ways, as shown in Section 5.2. The BNTC is featured on a very user-friendly, open access, searchable corpus platform.

Future plans for the BNTC include expanding the corpus by incorporating additional texts and enhancing the existing ones with more linguistic analysis. Additionally, once the Blackfoot FST is further developed, it could be used to support and expedite linguistic annotation process. Linking morpheme breaks to the corresponding lemma entries in the *Online Blackfoot Dictionary* would enhance the usability of the BNTC. Adding links to the dictionary lemmas will provide more context for the user. The online dictionary also includes some audio that would help learners with their pronunciation as well. In addition to linking the corpus entries to the

Online Blackfoot Dictionary audio, linking existing audio recordings to corresponding texts within the corpus would further enhance the BNTC's usefulness for language learners.

References

Áístainskiaakii Many Feathers, Sandra, Brent Issapóíkoan Prairie Chicken, Wes Áínnootaa Crazy Bull & David Osgarby. 2013. Aakíípisskani 'the women's buffalo jump.' (Ed.) J Dunham, P Littell & J Lyon. *Papers of the international conference on salish and neighbouring languages* (University of British Columbia Working Papers in Linguistics 35) 48. 1–21.

Alberta Education. 2010. Blackfoot Language and Culture Twelve-Year Program Kindergarten to Grade 12. https://education.alberta.ca/media/563920/blackfoot-k-12.pdf.

Anthony, Laurence. 2023. AntConc (Version 4.2.4). Computer Software. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software.

Anthony, Laurence. Laurence Anthony's Website. https://www.laurenceanthony.net/software/antconc/. (21 December, 2023).

Arppe, Antti, Katherine Schmirler, Atticus G. Harrigan & Arok Wolvengrey. 2020. A Morphosyntactically Tagged Corpus for Plains Cree. In Monica Macaulay & Margaret Noodin (eds.), *Papers of the Forty-Ninth Algonquian Conference*, 1–16. Michigan State University Press. https://doi.org/10.14321/j.ctvv417gp.5.

Ayoungman, Vivian. 1993a. *Level 1* (Siksikai'powahsin/Siksika Language Series Kit.). Gleichen, Alberta: The Siksika Nation.

Ayoungman, Vivian. 1993b. *Level 2* (Siksikai'powahsin/Siksika Language Series Kit.). Gleichen, Alberta: The Siksika Nation.

Ayoungman, Vivian. 1993c. *Level 3* (Siksikai'powahsin/Siksika Language Series Kit.). Gleichen, Alberta: The Siksika Nation.

Bad Boy, Margaret & Beatrice Poor Eagle. 1994. *Aakaitapitsinniksiists = Siksika old stories* (Siksikai'powahsin/Siksika Language Series Kit. Level 3). Siksika, Alta., Edmonton: Siksika Nation ; Duval House Pub.

Barth, Danielle & Stefan Schnell. 2021. *Understanding Corpus Linguistics* (Understanding Language). New York: Routledge.

Benzitoun, C., J.-M. Debaisieux & J. Deulofeu. The ORFÉO project: a study corpus for contemporary French. In *Corpus of Spoken French and Spoken French of Corpus.*

Blackfeet Community College. 2023. Blackfeet Community College Course Catalogue. *Blackfoot Community College 2021-2024 Course Catalogue*. https://bfcc.edu/course-catalog/. (21 November, 2023).

Borin, Lars, Markus Forsberg & Johan Roxendal. 2012. Korp - the corpus infrastructure of Språkbanken. In *Proceedings of LREC 2012*, 474–478. Istanbul: ELRA.

Brixey, Jacqueline, Eli Pincus & Ron Artstein. 2018. Chahta Anumpa: A multimodal corpus of the Choctaw Language. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, et al. (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). https://aclanthology.org/L18-1532/.

Comrie, Bernard, Martin Haspelmath & Balthasar Bickel. 2015. The Leipzig Glossing Rules. Max Planck Institute for Evolutionary Anthropology. https://www.eva.mpg.de/lingua/resources/glossing-rules.php.

Davies, Mark. 2008. The Corpus of Contemporary American English (COCA). Available online at https://www.english-corpora.org/coca/.

Davies, Mark. 2016. Corpus del Español. *Corpus del Español: Web/Dialects*. Available online at http://www.corpusdelespanol.org/web-dial/.

Dunham, Joel Robert William. 2013. The Blackfoot Language Database. In *Papers of the 41st Algonquian Conference*, 75–80. Albany, NY: State University of New York Press.

Ermineskin, Rachel & Darin Howe. 2005. On Blackfoot Syllabics and the Law of Finals. In *37th Algonquian Conference*. Ottawa. http://hdl.handle.net/1880/112322.

Evert, Stefan & Andrew Hardie. 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceeedings of the Corpus Lingusitics 2011 Conference*. University of Birmingham, UK.

Frantz, Donald G. 2009. *Blackfoot grammar*. 2nd ed. Toronto: University of Toronto Press.

Frantz, Donald G. 2017. *Blackfoot grammar*. Third edition. Toronto ; Buffalo ; London: University of Toronto Press.

Frantz, Donald G. & Norma Russell. 2017. *Blackfoot dictionary of stems, roots, and affixes*. Third edition. Toronto ; Buffalo: University of Toronto Press.

Galt Museum. Blackfoot Language Workbook. https://www.galtmuseum.com/blackfoot-language-workbook.

Garofolo, John S., Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett & Nancy L. Dahlgren. 1993. *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM*. Gaithersburg, MD: U.S. Department of Commerce: Technology Administration, National Institute of Standards and Technology; Computer Systems Laboratory; Advancy Systems Division.

Genee, Inge. 2009. What's in a morpheme? Obviation morphology in Blackfoot. *Linguistics* 47(4). 913–944. https://doi.org/10.1515/LING.2009.032.

Genee, Inge. 2020. “It’s written *niisto* but it sounds like *knee stew* .”: Handling multiple orthographies in Blackfoot language web resources. *Written Language & Literacy* 23(1). 1–28. https://doi.org/10.1075/wll.00031.gen.

Genee, Inge & Donald G. Frantz. n.d. Online Blackfoot Dictionary. *Blackfoot Dictionary*. https://dictionary.blackfoot.atlas-ling.ca/.

Genee, I., Kadlec, D., Schmirler, K., & Arppe, A. (2023). A computational model for Blackfoot demonstratives. Paper presented at the 55th Algonquian Conference at the University of Alberta in Edmonton, Alberta, October 19-22, 2023.

Genee, Inge & Marie-Odile Junker. 2018. The Blackfoot Language Resources and Digital Dictionary project: Creating integrated web resources for language documentation and revitalization. *Language Documentation & Conservation* 12. 298–338.

Heavy Shields Russell, Lena Ikkináínihki & Inge Piitáákii Genee. 2014. *Ákaitsinikssiistsi: Blackfoot Stories of Old* (First Nations Language Readers Blackfoot). Regina, Saskatchewan: University of Regina Press.

Jurafsky, Daniel & James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd edn. https://web.stanford.edu/~jurafsky/slp3/.

Kadlec, Dominik M. 2023. *A computational model of Blackfoot noun and verb morphology*. Lethbridge, AB: University of Lethbridge Master’s Thesis. https://hdl.handle.net/10133/6635.

Kalinowski, Tim. 2018. “In Flanders Fields” honoured by Elder. *Lethbridge Herald*. Lethbridge, AB. https://www.pressreader.com/canada/lethbridge-herald/20181203/281483572451608?srsltid=AfmBOoq0hpSc4srhB5Ja4XtFV6SR_MOlA6MhrNfXnsABTstMqG5yH4IR.

Kučera, Henry & W. Nelson Francis. 1967. *Computational analysis of present-day American English*. Providence: Brown University Press.

Leech, Geoffrey N., Paul Rayson & Andrew Wilson. 2001. *Word Frequencies in Written and Spoken English: based on the British National Corpus*. Hoboken: Taylor and Francis.

Lin Hao, Xu Shujuan. 2023. A Study of Synonyms Based on COCA Corpus Road and Street as Examples. *Lecture Notes on Language and Literature* 6(3). https://doi.org/10.23977/langl.2023.060305.

Many Guns, Matthew. *Aakaitapitsinniksiists = Siksika old stories* (Siksikai’powahsin/Siksika Language Series Kit. Level 2). Siksika, Alta.: Siksika Nation.

Mi’kai’sto, Red Crow Community College. 2025. Mi’kai’sto Red Crow Community College Indigenous Language and Culture Diploma (ILCD).

https://www.redcrowcollege.com/program/indigenous-language-and-culture-diploma-ilcd. (8 June, 2025).

Old Sun Community College. 2023. Old Sun Community College Course Catalogue. *Siksika Knowledge Program: Siksikai'tsii'powahsin Siksika Language Courses*. http://oldsuncollege.ca/index.php/siksika-knowledge-program-siksikaitsiipowahsin-siksika-language-courses/. (21 November, 2023).

Pulido-Guzman, Alejandra. 2021. Blackfoot translation of "In Flanders Fields" honours Indigenous veterans. *Lethbridge Herald*. Lethbridge, AB. https://lethbridgeherald.com/news/lethbridge-news/2021/11/09/blackfoot-translation-of-in-flanders-fields-honours-indigenous-veterans/.

ROYAL SPANISH ACADEMY: Database (CORPES XXI) [online]. 2014. Corpus of 21st-Century Spanish (CORPES). Available online at http://www.rae.es. (13 May, 2025).

Russell, Lena. 1996. *Blackfoot 10: Niitsi'powashsini*. Edmonton Alberta: Duval House Publishing.

Russell, Lena. 1997a. *Blackfoot 20: Niitsi'powashsini*. Duval House Publishing.

Russell, Lena. 1997b. *Blackfoot 30: Niitsi'powashsini*. Edmonton Alberta: Duval House Publishing.

Russell, Lena. 2001. *Blackfoot 9*. Edmonton Alberta / Standoff, Alberta: Duval House Publishing / Kainaiwa Board of Education.

Russell, Lena. 2002. *Blackfoot 8*. Edmonton Alberta / Standoff, Alberta: Duval House Publishing / Kainaiwa Board of Education.

Russell, Lena. 2003. *Blackfoot 7*. Edmonton Alberta / Standoff, Alberta: Duval House Publishing / Kainaiwa Board of Education.

Schmirler, Katherine. 2022. Syntactic Features and Text Types in 20th Century Plains Cree: A Constraint Grammar Approach. University of Alberta Library. https://doi.org/10.7939/R3-PZ87-YE25.

Schmirler, Katherine, Antti Arppe & Inge Genee. 2024. Morphophonological Rule Development and Real-Time Rule Testing with XFST: A Model for Blackfoot. In Inge Genee, Monica Macaulay & Margaret Noodin (eds.), *Papers of the Fifty-Third Algonquian Conference / Actes du cinquante-troisième Congrès des Algonquinistes*, 253–268. Michigan State University Press. https://doi.org/10.14321/jj.9345418.18.

Schmirler, Katherine, and Antti Arppe. Under development. Finite-State Transducer-Based Computational Model of Blackfoot Morphology. https://github.com/giellalt/lang-bla/tree/main/src/fst.

Schmirler, K., Genee, I., Arppe, A., & Kadlec, D. (2024). Ongoing development of a finite-state transducer based morphological model for Niitsi'powahsin. Paper presented at the 56th Algonquian Conference at the First Americans Museum (cohosted by the Sam Noble Museum & Endangered Language Fund) in Oklahoma City, Oklahoma, October 24-27, 2024.

Schupbach, Shannon Scott. 2013. *The Blackfoot demonstrative system: Function, form, and meaning*. Missoula, MT: University of Montana Master's Thesis. https://scholarworks.umt.edu/etd/964.

Scott, M., 2024, WordSmith Tools version 9 (64-bit version) Stroud: Lexical Analysis Software.

Statistics Canada. 2022. (table). Census Profile. 2021 Census of Population. Statistics Canada Catalogue no. 98-316-X2021001. https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/index.cfm?Lang=E. (17 January, 2023).

Uhlenbeck, C.C. 1911. *Original Blackfoot texts from the southern Peigans Blackfoot reservation, Teton County, Montana. With the help of Joseph Tatsey, collected and published with an English translation.* Amsterdam: J. Muller.

University of Calgary. 2025. University of Calgary: Official Calendar: Indigenous Langauges INDL. *University of Calgary Official Calendar (2025-2026)*. https://calendar.ucalgary.ca/courses?subjectCode=INDL&page=1&cq=. (8 June, 2025).

University of Lethbridge. 2025. *University of Lethbridge Undergraduate Calendar*. Webpage. https://www.ulethbridge.ca/sites/ross/calendar/ug/topic.htm#t=Topics%2FCourse_Catalogue-Blackfoot__BKFT.htm. (6 June, 2025).

University of Pennsylvania. 1992. Linguistic Data Consortium (LDC). *Linguistic Data Consortium (LDC)*. https://www.ldc.upenn.edu/. (14 May, 2025).

Weber, Natalie. 2022. Blackfoot Words. Zenodo. https://doi.org/10. 5281/zenodo. 5774980.

Weber, Natalie, Tyler Brown, Joshua Celli, McKenzie Denham, Hailey Dykstra, Rodrigo Hernandez-Merlin, Evan Hochstein, et al. 2023. Blackfoot Words: a database of Blackfoot lexical forms. *Language Resources and Evaluation* 57(3). 1207–1262. https://doi.org/10.1007/s10579-022-09631-2.

Yellowhorn, Eldon. 2021. Blackfoot Blogs and Boutique Languages. In Lisa Crowshoe, Inge Genee, Mahaliah Peddle, Joslin Smith & Conor Snoek (eds.), *Sustaining Indigenous Languages: Connecting Communities, Teachers, and Scholars*, 173–182. Northern Arizona University.

2001. Niitsitapiisini: Our Way of Life. *Glenbow Museum: Where the world meets the west*. https://www.glenbow.org/blackfoot/EN/html/index.htm.

2002. Endangered Languages Archive. *Endangered Languages Archive*. https://www.elararchive.org/. (14 May, 2025).

2021. Blackfoot Confederacy. https://blackfootconfederacy.ca/. (28 September, 2023).

Appendix 1: Table of linguistics glosses, their meaning, corresponding morphemes and sources

Gloss	Meaning	(morpheme) [gloss]	Source
-	(hyphen) separates morphemes		Leipzig glossing rules
.	(period) separator within a morpheme		Leipzig glossing rules
=	Joins clitics		Leipzig glossing rules
1	First person		Leipzig glossing rules
2	Second person		Leipzig glossing rules
21	First person plural inclusive		Algonquian
3	(Proximate) third person		Leipzig glossing rules
4	Obviative third person		Frantz
AI	Animate intransitive verb		Algonquian
AIO	Animate intransitive verb plus object		Algonquian
AN	Animate	(-wa [AN.SG] / -iksi [AN.PL])	Algonquian
COM	Comitative	(omohp- ~ iihp- ~ ohp-)	Leipzig / Smith (2025)
CN	Conjunct nominal	(-hp ~ -'p / -o'p [21.CN])	Frantz (2017)
CONJ	Conjunctive	verb paradigm	Frantz (2017)
DCT	Deictic preverb	(it- ~ ist-)	Schupbach (2013)
DD	Distal demonstrative stem	(om)	Schupbach (2013)
DIR	Direct theme suffix	(-a: / -ii ~ -yii)	Frantz (2017)
DM	Medial demonstrative stem	(ann)	Schupbach (2013)
DP	Proximal demonstrative stem	(am)	Schupbach (2013)

DTP	Distinct third person pronouns	(=aistsi [=DTP.IN.PL] / =aiksi [=DTP.AN.PL] / =áyi [=DTP.SG])	Frantz (2017)
DUR	Durative aspect	(á-)	Leipzig glossing rules
FUT	Future	(yáak-)	Leipzig glossing rules
II	Inanimate intransitive verb		Algonquian
IMFUT	Immediate/imminent future	(áyaak-)	Algonquian
IMP	Imperative		Leipzig glossing rules
IN	Inanimate	(-yi [IN.SG] / -istsi [IN.PL])	Algonquian
INCH	Inchoative	(á'-)	Frantz (2017)
IND	Indicative		Leipzig glossing rules
INS	Instrument	(omoht- ~ iiht- ~ oht-)	Leipzig glossing rules
INT	Interior geometric configuration	(-o) in the demonstratives	Schupbach (2013)
INTS	Intensifier	(iik- / sska' / sstonnat-)	Smith (2025)
INV	Inverse theme suffix	(-ok)	Frantz (2017)
INVS	Invisible to the speaker	(-hka) in the demonstratives	Frantz (2017) / Schupbach (2013)
NAF	Non-affirmative endings	(-waatsiksi [SG] / -waistsaawa [IN.PL] / -waiksaawa [AN.PL] / -hpa)	Frantz (2017)
NAR	Narrative suffix	(-yiihk)	Frantz (2017)
NEG	Negation	(máát- ~ Imá:t- / kátá'- ~ Ikátá'- / miin- ~ piin- / sta'- / say- ~ saw-)	Leipzig glossing rules
NMLZ	Nominalizer	(-n ~ -hsin / -a'tsis [INS.NMLZ])	Leipzig glossing rules
NREF	Non-referring	(-i)	N- Leipzig glossing rules -REF Smith (2025)
MA	Motion away	(-ya) in the demonstratives	Schupbach (2013)
MNR	Manner	(aanist- / niit-)	Schupbach (2013)
MT	Motion towards	(-ka) in the demonstratives	Schupbach (2013)
PRF	Perfective aspect	(ákaa- ~ Ikaa-)	Leipzig glossing rules
PL	Plural	(-istsi [IN.PL] / -iksi [AN.PL])	Leipzig glossing rules
POSS	Possessive suffix	(-m)	Leipzig glossing rules

PRO	Pronoun	(=aawa [=PRO.3PL])	Frantz (2017)
PST	Past	(ii-)	Leipzig glossing rules
RECIP	Reciprocal	(-o:tsiiyi ~ -tsiiyi [AI] / -o:tsiim ~ -tsiim [TA])	Leipzig glossing rules
REFL	Reflexive	(-o:hsi)	Leipzig glossing rules
SBJV	Subjunctive		Leipzig glossing rules
SG	Singular		Leipzig glossing rules
STAT	Stationary	(-ma) in the demonstratives	Schupbach (2013)
TA	Animate transitive verb		Algonquian
TH.TI	Inanimate Transitive theme suffix	(-hp ~ -'p / -m)	Frantz (2017)
TI	Transitive inanimate verb		Algonquian
VBLZ	Verbalizer	(-(a)o'ka / -ayi ~ -yi)	Algonquian
X>Y	X acts on Y		Leipzig glossing rules