Syntactic Features and Text Types in 20th Century Plains Cree:

A Constraint Grammar Approach

by

Katherine Margaret Schmirler

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

Doctor of Philosophy

Department of Linguistics
University of Alberta

Abstract

This dissertation describes the creation of a morphosyntactically tagged corpus of Plains Cree (*nêhiyawêwin*), an Indigenous language of North America, and demonstrates three ways in which this corpus can be used to explore morphosyntactic variation in the language on a larger scale than previously feasible. The corpus includes ~152,000 words of Plains Cree drawn from several published volumes of transcribed oral text, collected in the 1920s and 1980s-1990s, offering a variety of text types and time periods to consider. These texts are divided by time period into two subcorpora, the Bloomfield subcorpus and the Ahenakew-Wolfart subcorpus.

The tagged corpus is created using two main tools: 1) a Finite State Transducer-based morphological model for Plains Cree, and 2) a Constraint Grammar-based syntactic parser. Though the initial development of the former predates the dissertation, manual validation of the morphological analyses produced by the model undertaken as part of this work has contributed to its ongoing development. The latter is a core component of the present work, and aims to disambiguate ambiguous wordforms and assign basic syntactic functions. Chapter 2 describes the morphosyntactic features needed to build the syntactic parser, as well as some that are not currently implemented but will contribute to an improved model in the future. Chapter 3 describes the creation of the syntactic parser using the Constraint Grammar formalism, including improvements from an earlier iteration. Chapter 4 evaluates the effectiveness of the syntactic parser and describes the corpus in more detail, including the texts it contains and the morphosyntactic feature tags assigned by the models. Chapter 5 focuses on the argument tags, exploring the variation of where and when arguments occur in a language with flexible word order. Even with only a morphosyntactically tagged corpus, the pragmatic influences on argument realisation and word

order can be observed. In both Chapters 4 and 5, variation between the subcorpora is explored as well, demonstrating the ways in which the subcorpora differ, and how these differences are obscured when the corpus is examined as a whole—different verb classes, noun classes, persons, word order patterns, etc. Chapter 6 then offers an example of how variation between the subcorpora and the different text types they contain can be explored, using Principal Component Analysis (PCA) to undertake a text type analysis. Differences between the subcorpora are apparent, though similarities in narratives are also demonstrated; the primary contrasts are found between narratives and speeches and, within narratives, between dialogue and non-dialogue narrative.

Among the first large corpora for under-resourced Indigenous languages, this automatically tagged corpus allows for the exploration of oral language use on a large scale. The corpus serves as the basis of a searchable online corpus for academics and community members, as a tool for research and as a supplement to language education.

Preface

This dissertation is an original work by Katherine Margaret Schmirler. The research conducted for this dissertation and underlying the supporting tools has been undertaken as part of the "21st Century Tools for Indigenous Languages" project directed by Dr. Antti Arppe at the University of Alberta, with collaborators from across Canada and around the world—academics, Indigenous language speakers, and students (https://21c.tools/). The morphological model used in this work to form the basis of the morphological tags in the tagged corpus, which in turn make the creation and application of the syntactic parser possible, is an ongoing collaborative effort as part of the 21st Century Tools project. The syntactic parser described herein also has its roots in the same project: preceding the development of the parser carried out by the author of this dissertation beginning in 2016, initial modelling was undertaken by Drs. Antti Arppe, Lene Antonsen, and Trond Trosterud (the latter two of UiT The Artic University of Norway). The morphological gold standard of validated forms for the Ahenakew-Wolfart corpus, though now substantially expanded, owes its initial form to collaboration between the author of this work and Atticus Harrigan.

Parts of the manuscript reproduced in Appendix A have previously been published, and an edited version of Chapter 6 has been accepted for publication in the *Papers of the 53rd Algonquian Conference*:

- Schmirler, K., Arppe, A., Trosterud, T., & Antonsen, L. (2018). Building a Constraint Grammar Parser for Plains Cree Verbs and Arguments. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

- Schmirler, K., & Arppe, A. (forthcoming). Plains Cree textual analysis with PCA: Across the Bloomfield and Ahenakew-Wolfart subcorpora. In M. Macaulay & M. Noodin (Eds.), *Papers of the 53rd Algonquian Conference*. MSU Press.

I acknowledge the funding that made this dissertation possible, a Social Sciences and Humanities Research Council of Canada (SSHRC) Partnership Development Grant (890-2013-0047), a Kule Institute for Advanced Study (KIAS) Research Cluster Grant, a SSHRC Connections Grant (611-2016-0207), a SSHRC Doctoral Fellowship (752-2017-2105), and a SSHRC Partnership Grant (895-2019-1012).

Examining committee:

Antti Arppe, Supervisor
Anja Arnhold, Supervisory Committee
Arok Wolvengrey, Supervisory Committee
Evangelia Daskalaki, Examiner
Amy Dahlstrom, External Examiner

Dedication

To Nana and Papa, who showed me that I could make anything I set my hands to. And to Jessie, who kept me going day after day for fourteen years, and will never be forgotten.

Acknowledgements

A year ago, never mind seven, I could not have imagined what this dissertation would become, and none of it would be possible without the family, friends, and mentors who were with me at every step.

I must start at the beginning, and thank Arok Wolvengrey of the First Nations University of Canada at the University of Regina, who taught my first linguistics course and started me on this path, and his colleagues, Jan van Eijk and Olga Lovick, whose support has extended well beyond my undergraduate degree. Their support has taken me across Canada and back, through over eight years of graduate school. I would of course not be where I am today without my Cree instructors, Solomon Ratt, Doreen Oakes, and Jean Okimâsis, also of the First Nations University of Canada. Little did I know that when I chose to try a language other than French as part of my degree requirements that it would lead to relationships with these wonderful mentors and shape the rest of my life. I am grateful to Aaron Dinkin, who led me on a deep dive into historical linguistics and guided me through the absurdity of a one-year Masters degree at the University of Toronto. I am also indebted to the many other wonderful students and instructors I met during my time there. The big city will never be for me, but it was an experience I wouldn't trade for anything.

In the Alberta Language Technology Lab in the University of Alberta's Department of Linguistics, I have worked with many wonderful linguists and learned much from our resident computer scientists: Jordan Lachler, Atticus Harrigan, Megan Bontogon, Erin McGarvey, Lex Giesbrecht, Eddie Santos, Matt Yang, and Daniel Dacanay, and every other member, however brief their time with us. I have to acknowledge the many amazing folks we worked with in Maskwacîs, collecting recordings of Cree words for an online dictionary. Through the lab, I also had the pleasure of collaborating with academics from across Canada and around the world: Trond Trosterud, Lene Antonsen, Sjur Moshagen, Marie-Odile Junker, Arden Ogg, and Miikka Silfverberg, among many others. In other areas of the department, I was glad to meet Matthew Kelley, Ivy Mok, Brian Rusk, Izzy Hubert, Nik Toler, and Scott Perry, along with everyone else who made the department a fantastic community to be a part of. I am so grateful to my committee (Anja Arnhold, Arok Wolvengrey, and Antti Arppe), examiners (Amy Dahlstrom and Evangelia Daskalaki), and my examination chair (Johanne Paradis) for all their support and guidance and a wonderful defense.

I of course cannot express how grateful I am to my supervisor, Antti Arppe. He welcomed me to the Alberta Language Technology Lab despite my minimal computational knowledge and taught me to embrace statistical analysis and computational modelling. Thank you for letting me exercise my existing skills and to grow into new ones; historical linguistics is my first love and I was more than able to use and share my knowledge of historical Algonquian in lab projects, but I was never pigeonholed and able to explore every aspect of Cree I could. It was an honour to be involved with all your projects and to see how I can use my skills and knowledge to support language revitalisation and Cree communities.

I can never repay the support of my family and friends over the years: my parents, Randy and Sharon, who have supported me in every way through all my years in school and helped me moved across the country more than once; my sister, Kim, with our secret sibling language; and the many friends I can still turn to no matter the time and distance: Sandi, Jon & Kelly, Falene, Melissa, and Leah & Leonard. Finally, thank you to Daniel for your support and love. I wouldn't be here without you.

Contents

List of Tables

List of Figures

List of Abbreviations

Persons and person interactions

0	inanimate person, unspecified for number
0′PL	inanimate obviative plural person
0′SG	inanimate obviative singular person
0PL	inanimate plural person
0SG	inanimate singular person
1	first person, unspecified for number
1<3	proximate third person acting on first person, unspecified for number
1>3	first person acting on proximate third person, unspecified for number
1G	first person goal, unspecified for number
1PL	exclusive first person plural
1PL>2PL	exclusive first person plural acting on second person plural
1PL-G	exclusive first person plural goal
1SG	first person singular
1SG>2SG	first person singular acting on second person singular
1SG>3SG	first person singular acting on proximate third person singular
1SG-G	first person singular goal
2	second person, unspecified for number
21PL	inclusive first person plural
21PL-G	inclusive first person plural goal
2G	second person goal, unspecified for number
2PL	second person plural
2PL>1SG	second person plural acting on first person singular
2PL-G	second person plural goal
2SG	second person singular
2SG>1SG	second person singular acting on first person singular
2SG>3PL	second person singular acting on proximate third person plural
2SG>3SG	second person singular acting on proximate third person singular

2SG-G	second person singular goal
3	proximate third person, unspecified for number
3′	obviative third person
3″	further obviative third person
3″G	further obviative third person goal
3′>3″	obviative third person acting on further obviative third person
3′>3PL	obviative third person acting on proximate third person plural
3′>3SG	obviative third person acting on proximate third person singular
3′-G	obviative third person goal
3-G	proximate third person goal
3PL	proximate third person plural
3PL>1SG	proximate third person plural acting on first person singular
3PL>3′	proximate third person plural acting on obviative third person
3PL-G	proximate third person plural goal
3SG	proximate third person singular
3SG>1	proximate third person singular acting on first person, unspecified for number
3SG>1SG	proximate third person singular acting on first person singular
3SG>2PL	proximate third person singular acting on second person plural
3SG>2SG	proximate third person singular acting on second person singular
3SG>3′	proximate third person singular acting on obviative third person
3SG-G	proximate third person singular goal
X	unspecified actor
X>3	unspecified actor action on a proximate third person, unspecified for number

Other linguistic glosses

A	animate
AI	animate intransitive (verb), see also VAI
CNJ	conjunct
COND	conditional
D	dependent (inalienably possessed)
DEM	demonstrative
DIR	direct
DIST	distal
DUB	dubitative
EXCL	exclusive
FOC	focus
FUT	future
I	inanimate
IC	initial change
II	inanimate intransitive (verb), see also VII
IMP	imperative
INAN	inanimate actor/goal (used in tables)
INCL	inclusive
IND	independent
INTER	interrogative
INV	inverse
IPC	particle (“indeclinable particle”)
IPH	particle phrase
IPL	locative particle
IPN	indeclinable nominal (denominal particle)
IPT	temporal particle
IPV	preverb
LOC	locative
MED	medial

N	noun
NA	animate noun
NDA	animate dependent (inalienably possessed) noun
NDI	inanimate dependent (inalienably possessed) noun
NEG	negative
NI	inanimate noun
NOM	nominal
OBL	oblique
OBV	obviative
PERS	personal (pronoun)
PL	plural
PRON	pronoun
PROP	proper noun
PROX	proximate
PRS	present tense
PST	past tense
QST	question particle
REL	relational
RFLX	reflexive
SG	singular
TA	transitive animate (verb), see also VTA
TI	transitive inanimate (verb), see also VTI
V	verb
VAI	animate intransitive verb
VAI-O	animate intransitive verb (morphologically), that takes an object (equivalent to VAI-t)
VAI-t	animate intransitive verb (morphologically), that behaves like a transitive verb (equivalent to VAI-O)
VII	intransitive inanimate verb
VTA	transitive animate verb
VTI	transitive inanimate verb

Morphosyntactic features and syntactic tags (for modelling & reporting results)

@<ACTOR	actor syntactic function tag, following the verb
@<GOAL	goal syntactic function tag, following the verb
@<N	nominal dependent syntactic function tag, following the nominal
@A	any actor syntactic function tag (shorter form used in tables)
@A/G	any actor or goal syntactic function tag (shorter form used in tables)
@ACTOR/GOAL	any actor or goal syntactic function tag
@ACTOR>	actor syntactic function tag, preceding the verb
@G	any goal syntactic function tag (shorter form used in tables)
@GOAL	any goal syntactic function tag
@GOAL>	goal syntactic function tag, preceding the verb
@INS	any instrument syntactic function tag
@IPL	any locative particle syntactic function tag
@IPL-V	any locative particle syntactic function tag, associated with a verb
@IPT	any temporal syntactic function tag
@IPT-V	any temporal particle syntactic function tag, associated with a verb
@LOC-V/IPL	any locative noun associated with either a verb or a locative particle
@N	any nominal dependent syntactic function tag
@N>	nominal dependent syntactic function tag, preceding the nominal
@NEG	any negative particle syntactic function tag
@NEG-IPC	syntactic function tag for a negative particle associated with a particle
@NEG-N	syntactic function tag for a negative particle associated with a noun
@NEG-V	syntactic function tag for a negative particle associated with a verb
@OBL	any oblique syntactic function tag
@P	any adpositional phrase syntactic function tag
@Pos>	possessor preceding possessum syntactic function tag
@PRED	any predicate syntactic function tag
@PRED-AI	animate intransitive predicate syntactic function tag
@PRED-II	inanimate intransitive predicate syntactic function tag
@PRED-TA	transitive animate predicate syntactic function tag

@PRED-TI	transitive inanimate predicate syntactic function tag
@Quant	any quantifier syntactic function tag
@Quant-N	any quantifier syntactic function tag, associated with a noun
@Quant-V	any quantifier syntactic function tag, associated with a verb
@Quot	any quotative predicate syntactic function tag
@Quot-AI	animate intransitive quotative predicate syntactic function tag
@Quot-TA	inanimate intransitive quotative predicate syntactic function tag
@Quot-TI	transitive inanimate quotative predicate syntactic function tag
4Sg	inanimate obviative singular
4Sg/Pl	obviative third person (4 used for morphological and syntactic modelling, cf. 3′)
4Sg/PlO	obviative third person goal (4 used for morphological and syntactic modelling, cf. 3′)
5Sg/PlO	further obviative person goal (5 used for morphological and syntactic modelling, cf. 3″)
IIZ	zero-place inanimate intransitive
Prop	proper noun
PrtHT	ht-preterit
PV/	introduces a preverb
Px	possessed by, unspecified for person
Px1Pl	possessed by first person plural
Px1Sg	possessed by first person singular
Px21Pl	possessed by inclusive first person plural
Px2Sg	possessed by second person singular
Px3Pl	possessed by third person plural
Px3Sg	possessed by third person singular
Px4Sg/Pl	possessed by obviative third person (“4” used for morphological and syntactic modelling, cf. 3′)
PxA	animate possessed noun
PxI	inanimate possessed noun
Quant	quantifier

Quot	quotative verb
w/@A	denotes a nominal with an @ACTOR tag
w/@G	denotes a nominal with an @GOAL tag
w/@Loc-V	denotes a locative particle with an @LOC-V tag
w/@N	denotes a demonstrative or numeral with an @N tag
w/@P	denotes a locative noun with an @P tag
w/@Pos	denotes a nominal with an @Pos tag
w/A	denotes a verb with an overt actor
w/G	denotes a verb with an overt goal
w/overt	denotes a verb with at least one overt participant
w/P	denotes a verb with an overt patient
w/S	denotes a verb with an overt subject

Other abbreviations

A	actor, agent
AGV	actor-goal-verb word order
ALTLab	Alberta Language Technology Lab
AV	actor-verb word order
AVG	actor-verb-goal word order
A-W	Ahenakew-Wolfart
BT	Bloomfield texts
CG	Constraint Grammar
CG-1	Constraint Grammar v.1
CG-2	Constraint Grammar v.2
CG-3	Constraint Grammar v.3
CM	Cecilia Masuskapoe
EM	Emma Minde
FST	Finite State Transducer
G	goal
GAV	goal-actor-verb word order
GB	Glecia Bear
GV	goal-verb word order
GVA	goal-verb-actor word order
iCALL	intelligent computer-assisted language learning
JD	Joe Douquette
JK	Jim Kâ-Nîpitêhtêw
MGS	Morphological Gold Standard
MW	Mary Wells
non-SAP	non-speech act participant (i.e., neither first nor second person)
NP	noun phrase (i.e., a nominal and any associated modifiers)
O	object
OCR	optical character recognition
OVS	object-verb-subject word order

P	patient
PAS	Preferred Argument Structure
PC	Principal Component
PC1	Principal Component 1
PC2	Principal Component 2
PCA	Principal Component Analysis
PCT	*Plains Cree Texts* (Bloomfield, 1934)
postV	postverbal
preV	preverbal
PV	Peter Vandall
S	subject
SAP	speech act participant (i.e., first or second person)
SGS	Syntactic Gold Standard
SSSC	*Sacred Stories of the Sweetgrass Cree* (Bloomfield, 1930)
SVO	subject-verb-object word order
SW	Sarah Whitecalf
TTR	type-token ratio
V0	verb with zero animate participants (VII)
V1	verb with one animate participant (VAI, VTI)
V2	verb with two animate participants (VTA)
VA	verb-actor word order
VAG	verb-actor-goal word order
VG	verb-goal word order
VGA	verb-goal-actor word order
VS	verb-subject word order
VSO	verb-subject-object word order
VT	transitive verb

Glossary of Terms

Here I expand upon some of the abbreviations given above, offering brief explanations of how the terms they represent are used in this work, especially within Algonquianist tradition and the Constraint Grammar formalism. I also include some terminology that is not represented by abbreviations.

actor/goal:	within Algonquianist tradition (and used herein as part of this tradition), these terms are used to represent the morphosyntactic actor and undergoer of an action; 'actor' is roughly equivalent to subject and 'goal' to object
Algonquian Person Hierarchy:	a topicality hierarchy that accounts for morphosyntactic patterns, wherein second persons are more topical than first persons, which are together more topical than third persons; within third persons, proximate is more topical than obviative; all animate entities are more topical than inanimate
animacy:	system of noun classification; *animate* nouns include all people, animals, and trees, as well as some other idiosyncratic nouns, and *inanimate* nouns are all others
clause:	within a Constraint Grammar parser, defined by clause boundaries as labelled during the programming of the parser (e.g. clause boundary punctuation such as periods, commas, etc.)
corpus/subcorpus:	throughout, the Plains Cree corpus or full corpus refers to the collection of both the Bloomfield texts and the Ahenakew-Wolfart texts. Individually, these are the Bloomfield (BT) subcorpus and the Ahenakew-Wolfart (A-W) subcorpus

direction:	within a hierarchical alignment language, *direct* verbal morphology indicates a more topical entity acting on a less topical entity and *inverse* verbal morphology indicates a less topical entity acting on a more topical entity
local paradigm:	verbs involving interactions between speech participants
mixed paradigm:	verbs involving interactions between speech participants and non-speech act participants
nonlocal paradigm:	verbs involving interactions between non-speech participants
obviation:	discourse-based topicality within animate third persons; the most topical entity is proximate and any others are obviative
speech act participants:	first and second persons, in contrast to third persons, which are termed non-speech act participants
transitivity classes:	Algonquian transitivity classes are traditionally divided by 1) their transitivity and 2) the animacy of their participants; inanimate intransitive (VII, inanimate actor), animate intransitive (VAI, animate actor), transitive inanimate (VTI, inanimate goal, animate actor), transitive animate (VTA, animate goal, animate actor)
word:	orthographic word delineated by whitespace, following the Standard Roman Orthography for Plains Cree

Chapter 1

Plains Cree, digital resources for Indigenous languages, and the creation and use of a tagged corpus

This dissertation details the creation of a morphosyntactically tagged corpus for Plains Cree (*nêhiyawêwin*), an Indigenous language of North America, and illustrates its use through three studies on the morphological features, syntactic patterns, and text types found therein. The corpus, containing ~152,000 Plains Cree words, is built using a selection of published and digitised Plains Cree texts, a Finite State Transducer-based morphological analyser with manually validated results, and a Constraint Grammar (CG)-based syntactic parser. In this dissertation, I describe the process of creating the Plains Cree parser. I then apply the parser to the corpus and present the results in three stages. First, I look at the accuracy of the parser, and the frequency of various morphosyntactic tags. Second, I look at the word order patterns in the corpus, with a focus on whether or not arguments are overtly realised as nominals and, if so, where they occur relative to the verb. Third, I explore the morphosyntactic features with respect to the text types in the corpus, using a register analysis approach.

1.1 Plains Cree

Plains Cree (*nêhiyawêwin*; ISO 639-3: crk) is an Algonquian language spoken primarily in Alberta and Saskatchewan. It is a member of the Cree-Montagnais-Naskapi dialect continuum that stretches from the Rocky Mountains in Western Canada to Quebec and Labrador in Eastern Canada. The Cree dialects proper are shown in the map in Figure 1.1; further east are the dialects of Innu-aimun, (often called Montagnais), Naskapi, and East Cree. The dialects of Cree are generally differentiated through two key sound correspondences: one regarding Proto-Algonquian **r* (also reconstructed as **l*) and the other regarding Proto-Algonquian **s* and **š* (**ʃ*). The former occurs as *y* (/j/) in Plains Cree, *th* (/ð/) in Woods Cree, *n* in Swampy Cree, *l* in Moose Cree, and *r* (/ɾ/) in Atikamekw. Reflexes of **s* and **š* provide a broad classification between "western" and

"eastern" dialects; these fall together to /s/ (which may vary freely with /ʃ/) in the westernmost dialects including Western Swampy Cree, but they remain distinct from Eastern Swampy Cree eastwards, before falling together again in Naskapi (MacKenzie, 1980; Rhodes & Todd, 1981).

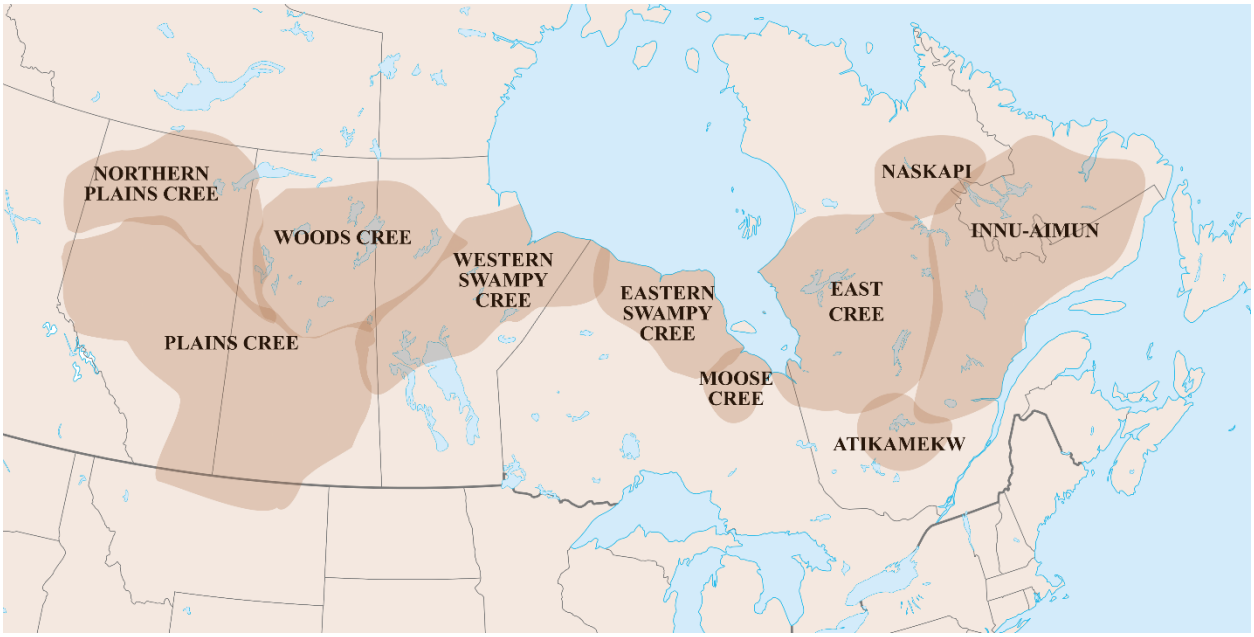


Figure 1.1: Map of the Cree-Montagnais-Naskapi dialect continuum[1]

The Algonquian language family is one of the largest and most widespread language families of North America. The family ranges from Blackfoot and Plains Cree in Alberta in the west to Eastern Algonquian languages such as Mi'kmaq and Maliseet-Passamaquoddy in Eastern Canada and the Northeastern United States; dialects of Cree are spoken as far north as the Northwest Territories and Plains Algonquian languages such as Arapaho are spoken as far south as Colorado and Oklahoma. Like other Algonquian languages, Plains Cree is known for complex, polysynthetic verbal morphology, a noun classification system based on animacy, a system of hierarchical

[1] This map is adapted from "Map of Cree dialects" (https://commons.wikimedia.org/wiki/File:Cree_map.svg by Noahedits, licensed under CC BY-SA 4.0) and edited for names and range of dialects based on About the Innu Language (n.d., https://www.innu-aimun.ca/english/about/the-innu-language/), the Algonquian Linguistics Atlas (n.d., https://www.atlas-ling.ca/), East Cree Dialects (n.d., https://www.eastcree.org/cree/en/grammar/east-cree-dialects/), Grammar, (n.d., https://www.eastcree.org/cree/en/grammar/), and Wolvengrey (2011, p. 4), as well as further personal communication with A. Wolvengrey.

alignment with direct and inverse marking on verbs, and flexible word order, wherein the subject and object can occur in any position relative to the verb, and are frequently not realised as overt nominals. While word order does not factor greatly in the modelling of Plains Cree morphosyntax, the detailed morphological information features heavily.

Recent Statistics Canada data (*Mother Tongue by Geography, 2021 Census*, n.d.) reports 6,870 speakers for Plains Cree; this is on the higher end for speakers of Cree dialects, which range from ~250 for Moose Cree to ~10,000 for Innu-aimun. Plains Cree is still learned as a first language in some communities, as well as being taught in some schools, and efforts are underway to develop immersion programs. A number of Plains Cree resources are available, including dictionaries (e.g., Arppe et al., under development; LeClaire et al., 1998; Wolvengrey, 2001), textbooks (e.g., Okimâsis, 2021; Ratt, 2016, 2022), children's books (e.g., Florence, 2019; Harris, 2018; Lavallee & Silverthorne, 2014; Miso, 2020; Sainte-Marie, 2022), published volumes of narratives, poetry, speeches, and prayers (A. Ahenakew, 2000; E. Ahenakew, 1995; Bear et al., 1998; Demers et al., 2010; Kâ-Nîpitêhtêw, 1998; Masuskapoe, 2010; McLeod & Wolvengrey, 2016; Minde, 1997; Rockthunder, 2021; Vandall & Douquette, 1987; Whitecalf, 1993; Whitecalf & Whitecalf, 2021; Wolvengrey, 2007), as well as television and radio broadcasts (e.g., https://www.aptntv.ca/hockey/videos/, https://www.nfb.ca/playlist/wapos-bay-series-cree/, https://www.cfwe radio.ca/on-air/conversational-cree/).

1.2 Digital resources for Indigenous languages of the Americas

Compared to majority languages like English, Indigenous and other under-resourced languages of the world have not been the focus of digital resource development on a large scale. However, ongoing efforts have resulted in a number of digital resources for Indigenous languages, including online dictionaries, part of speech taggers, morphological analysers, syntactic parsers, spoken corpora, and written corpora, as well as tools for optical character recognition (OCR), spell checking, speech recognition, speech synthesis, and machine translation. The existence and stage of development of these tools varies, of course, by language. I focus on Indigenous languages of the Americas here, though many similar resources exist for minority languages around the world. Despite this apparently lengthy list, consider that the number of dictionaries or corpora listed here for several dozen Indigenous languages is still smaller than the number that exist for English alone.

1.2.1 Dictionaries

Online dictionaries and word lists are available for a number of Indigenous languages. These may be either bilingual dictionaries, offering translations of words into majority languages (usually English or Spanish in the Americas, occasionally French in Canada, and Portuguese in Brazil) or, more rarely, monolingual dictionaries, defining words in the language itself. Dictionaries may also be "intelligent", with underlying morphological models to analyse inflected forms entered into the search bar, return the entry for the lemma, and show the morphological analysis. Some dictionaries allow for relaxed search with variation in spelling or orthography, even if they do not use morphological analysis. In the search for online dictionaries of Indigenous languages, it also became apparent that ongoing development and, very likely, financial support for these projects and tools is not consistent: in a decade-old blog post listing 21 online dictionaries for Indigenous languages of the Americas (Oppenneer, 2013), eight of these (38%) were no longer accessible either by the URL provided or through a separate search for the same resource at a new domain (though newer resources, of uncertain connection to the old, could sometimes be found).[2]

A selection of available online dictionaries for three Indigenous language families of Canada is given in Table 1.1 through Table 1.3.[3] URLs marked with an asterisk (*) are intelligent dictionaries, i.e., these dictionaries can take inflected forms as search terms and analyse them using a morphological model to link them to the appropriate dictionary entry. Table 1.1 lists a selection of Algonquian language dictionaries; many of these are created through the Algonquian Linguistic Atlas (Algonquian Dictionaries and Language Resources Project, n.d., https://www.algonquianlanguages.ca/) and there is a particular focus on the Cree-Montagnais-Naskapi dialect continuum. Table 1.2 includes a selection of Dene language dictionaries and Table 1.3 includes a selection of Siouan language dictionaries.[4]

[2] For a more detailed survey of available online dictionaries for Indigenous languages in Canada, see Pankratz, Arppe, & Lachler (2022).

[3] These are but a small selection of the language families found in Canada, and not all the languages included are spoken (only) in Canada.

[4] The formats of the dictionaries referenced in this section can vary considerably: some are wordlists, lexical databases, or downloadable PDFs, in addition to those that look more familiar to users of bilingual dictionaries for majority languages. The list is limited to resources that are freely accessible online: they were easy to find with a search for the language name (both exonyms and endonyms were used) followed by "dictionary" or "wordlist", they were not behind a paywall, and, if PDFs, they were searchable.

Table 1.1: A selection of online Algonquian language dictionaries

Language	Dictionary link(s)
Plains Cree	*https://itwewina.altlab.app/ (Arppe et al., under development) https://dictionary.plainscree.atlas-ling.ca/ (*Plains Cree Dictionary*, n.d.) https://www.creedictionary.com/ (*Online Cree Dictionary*, n.d.)
East Cree	https://dictionary.eastcree.org/ (Eastern James Bay Cree Dictionary, n.d.)
Fort Severn Cree	https://fortsevern.atlas-ling.ca/ (Fort Severn Online Dictionary, n.d.)
Innu-aimun	https://dictionary.innu-aimun.ca/Words (*Innu Dictionary*, n.d.)
Moose & Swampy Cree	https://dictionary.moosecree.atlas-ling.ca/ (*Moose & Eastern Swampy Cree Online Dictionary*, n.d.) https://www.spokencree.org/glossary (*Spoken Cree*, n.d.) https://moosecree.ca/ (*Cree – English Online Dictionary*, n.d.)
Naskapi	https://dictionary.naskapi.atlas-ling.ca/ (*Naskapi Online Dictionary*, n.d.)
Atikamekw	https://www.langueatikamekw.ca/ (*Online Atikamekw Dictionary*, n.d.)
Michif	https://dictionary.michif.atlas-ling.ca/ (*Michif Online Dictionary*, n.d.) https://www.metismuseum.ca/michif_tools.php (*Heritage Michif Dictionary*, n.d.; *Northern Michif Dictionary*, n.d.)
Blackfoot	https://dictionary.blackfoot.atlas-ling.ca/ (*Blackfoot Online Dictionary*, n.d.)
Ojibwe dialects	https://dictionary.nishnaabemwin.atlas-ling.ca/ (*Nishnaabemwin Online Dictionary*, n.d.) https://ojibwe.lib.umn.edu/ (*The Ojibwe People's Dictionary*, n.d.)
Passamaquoddy-Maliseet	https://pmportal.org/browse-dictionary (*The Passamaquoddy-Maliseet Dictionary*, n.d.)
Proto-Algonquian	https://protoalgonquian.atlas-ling.ca/ (*Proto-Algonquian Online Dictionary*, n.d.)
Abenaki	http://westernabenaki.com/dictionary/A (*Western Abenaki Dictionary*, n.d.)
Myaamia	https://mc.miamioh.edu/ilda-myaamia/dictionary (*Myaamia-Peewaalia Dictionary*, n.d.)

Mi'gmaq	https://www.mikmaqonline.org/ (*Mi'gmaq Mi'kmaq Micmac Online Talking Dictionary*, n.d.)
Lenape	https://www.talk-lenape.org/ (*The Lenape Talking Dictionary | Home*, n.d.)
Arapaho	https://homewitharapaho.wordpress.com/ (Cowell, 2012) https://verbs.colorado.edu/arapaho/public/view_search (*Arapaho Lexical Dictionary*, n.d.)

Table 1.2: A selection of online Dene language dictionaries

Language	Dictionary link(s)
Tsuut'ina	*https://gunaha.altlab.dev/ (demo version)
Tłı̨chǫ Yatıì	http://tlicho.ling.uvic.ca/users/mainview.aspx (*Tłı̨chǫ Yatıì Multimedia Dictionary*, n.d.)
Tolowa (Siletz Dee-ni)	https://siletz.swarthmore.edu/ (*Siletz Talking Dictionary*, n.d.)
Dëne Sųłiné	http://www.ssdec.nt.ca/ablang/ (*Aboriginal Language Resources*, n.d.)
Dene Dháh	http://www.ssdec.nt.ca/ablang/ (*Aboriginal Language Resources*, n.d.)

Table 1.3: A selection of online Siouan language dictionaries

Language	Dictionary link(s)
Lakota	*https://nldo.lakotadictionary.org/ (*NLD Online v.5*, n.d.)
Iowa-Otoe-Missouria	http://www.iowayotoelang.nativeweb.org/dictionary.htm (*IOM Dictionary*, n.d.)
Nakoda (Stoney)	https://dictionary.stoneynakoda.org/ (*Stoney Nakoda Dictionary*, n.d.)
Dakota	https://fmp.cla.umn.edu/dakota/ (*Dakota Dictionary Online*, n.d.) https://dictionary.swodli.com/ (*Dakota-English Dictionary*, n.d.)

Online dictionaries can also be found for many other languages of the Americas, such as for Tlingit, Tsimshian, and Haida at https://www.sealaskaheritage.org/institute/language/resources (*Language Resources*, n.d.), a demo version of an intelligent Haida dictionary at https://guusaaw.

altlab.dev/ (*Gúusaaw Northern Haida dictionary*, n.d.), Inuktut varieties at http://www.labradorvirtualmuseum.ca/english-inuttut.htm (*English-Inuttut Dictionary*, n.d.) and https://tusaalanga.ca/dialect (*Inuktut Tusaalanga*, n.d.), Kalaallisut at https://ordbog.gl/ (*Greenlandic Dictionaries*, n.d.), and Lillooet at https://lingpapers.sites.olt.ubc.ca/pnwll-volumes/lillooet-english-dictionary/ (van Eijk, 2013). Further south, in Mexico, dictionaries are available for a number of Zapotec varieties at https://talkingdictionary.swarthmore.edu/zapotecs/ (*Zapotec Talking Dictionaries*, n.d.) and Mixtec varieties at https://www.sil.org/resources/archives/56285 (*Acatlán Mixtec Dictionary*, 2014) and http://mixtec.nativeweb.org/ (*Tu Idau*, n.d.), as well as for Páez (an isolate) at https://talkingdictionary.swarthmore.edu/paez/ (*Nasa Yuwe Talking Dictionary*, n.d.) and Chamaccoco (Zamucoan) at https://talkingdictionary.swarthmore.edu/chamacoco/ (*Chamacoco Talking Dictionary*, n.d.) in Central and South America.

1.2.2 Morphological and syntactic computational models

For Indigenous languages of the Americas, with their frequently complex, polysynthetic morphology, morphological models have been a priority in natural language processing. Combined with established dictionaries and grammars, morphological models can underlie intelligent dictionaries, tagged corpora, spell checkers, and autocomplete functions. In Table 1.4, I include a limited list of morphological models for Indigenous languages of the Americas. A number of these have been created under the 21st Century Tools for Indigenous Languages Project (https://21c.tools/) and are based on Finite State Transducers (FSTs), while others use other programming languages or machine learning approaches. Syntactic models for Indigenous languages are much rarer. A Constraint Grammar-based Plains Cree model (Schmirler et al., 2018; this work) is under development and Universal Dependency treebanks have been undertaken for other Indigenous languages of the Americas, with an eventual goal of parsing (e.g., Park et al., 2021 for Yupik; Pugh et al., 2022 for Nahuatl; Rueter et al., 2021 for Apuriña; Thomas, 2019 for Guaraní; Tyers & Henderson, 2021 for K'iche'; Wagner et al., 2016 for Arapaho).

Table 1.4: A selection of morphological models for Indigenous languages of the Americas

Language	Citation(s)
Arapaho	Kazeminejad et al. (2017); Moeller et al. (2018)
Michif	Davis et al. (2021)
Kwak'wala	Littell (2018)
Kanyen'kéha (Mohawk)	Assini (2013); Kazantseva et al. (2018)
Plains Cree	Harrigan et al. (2017); Snoek et al. (2014)
Odawa	Bowers et al. (2017)
Inuktitut	Micher (2017)
Kalaallisut	Resources (n.d.), https://oqaasileriffik.gl/en/resources/
Tsuut'ina	Arppe, Cox, et al. (2017); Holden et al. (2022)
East Cree	Arppe, Junker, et al. (2017)
Upper Tanana	Lovick et al. (2018)
Nahuatl	Martínez-Gil et al. (2012); Pugh & Tyers (2021)
Yine	Torres et al. (2021)
Guaraní	Kuznetsova & Tyers (2021)

1.2.3 Corpora

Corpora for Indigenous languages of the Americas can vary considerably in size, format, and content. They can be large or small, ranging from under 10,000 tokens to several million. They can consist of recordings of spoken speech, transcribed spoken speech, or written text, either originally in the Indigenous language or translated. They can be monolingual or presented in parallel with a majority language translation, especially when built from translated materials. They may be tagged to varying degrees, whether for parts of speech, for morphological features, for syntactic relationships, or not at all. Some corpora are drawn only from word lists, such as for Blackfoot, where extensive word lists gathered from various sources and glosses and analyses have been added (N. Weber, personal communication). Corpora can also be created from written texts either created in the language or translated from a majority language, such as Wikipedia pages, published books, and translations of the Bible or other religious material including prayers or

songs. In this section, I briefly describe several corpora for different Indigenous languages of the Americas.

A corpus for K'iche' consisting of approximately 10,000 tokens has been annotated for morphosyntactic features using the Universal Dependencies annotation scheme. These are drawn from sources such as dictionaries, pedagogical materials, linguistic fieldwork, folk tales, Bible translation, Wikipedia pages, and legal texts. Most of the corpus is paired with English or Spanish translations (Tyers & Henderson, 2021). For Cherokee, there is a corpus consisting of approximately 40,000 tokens, though this corpus is not yet tagged. The texts in this corpus are all translations into Cherokee from English stories, which are combined for a parallel corpus (Frey, 2018). A corpus for Choctaw contains approximately 50,000 tokens, also not tagged. This is a multimodal corpus, consisting of written text and recordings of speech drawn from sources such as pedagogical materials, linguistic fieldwork, short stories told in Choctaw, Bible translation, dictionaries (Brixey & Artstein, 2020). A corpus for Ayuuk, containing 6,000 phrases, is aligned with Spanish for the purposes of machine translation research. It contains a variety of translated resources, including the Bible, songs, poetry, the Mexican constitution, personal writings, fables, and social media (Zacarías & Meza, 2021). For Guaraní, there are two sizeable corpora, one of approximately 100,000 tokens tagged by a morphological model, drawn from the Bible and Wikipedia pages, already aligned to Spanish translations (Kuznetsova & Tyers, 2021), and one of approximately 230,000 tokens with sentence-level alignment to Spanish translations, drawn from news articles and blogs (Chiruzzo et al., 2020). Among the largest corpora for Indigenous languages are those for Inuit languages. For example, an Inuktitut corpus contains approximately eight million Inuktitut words with parallel English translations, drawn from the proceedings of the Legislative Assembly of Nunavut. These have been used alongside the English translation to build machine translation tools (Joanis et al., 2020; Martin et al., 2003; Roest et al., 2020). Similarly, Kalaallisut has a corpus of approximately 18 million words (Resources, n.d.; https://oqaasileriffik.gl/en/langtech/corpus/; "Oqaasileriffik har indsamlet," 2021). Unquestionably, these last three languages, Guaraní, Inuktitut, and Kalaallisut have much larger corpora than other languages on this list. These languages share one important feature that the others do not—they are official languages, of Paraguay, Nunavut, and Greenland respectively. Such an observation further highlights the importance of official status or federal government support for Indigenous

language revitalisation, such as those outlined by the Truth and Reconciliation Commission's Calls to Action (*Truth & Reconciliation*, 2015).

In addition to the Plains Cree corpus described in this dissertation (~152,000 words), which has been available in some form since 2017 for research purposes, there is also another corpus of Plains Cree, with approximately 50,000 tokens. This corpus is aligned to English translations, but not tagged for morphosyntactic features, and has been constructed for the purpose of machine learning research. This corpus is composed of materials such as religious songs, pedagogical materials, social media content, and children's stories (Teodorescu et al., 2022). Though the Plains Cree corpus described in this dissertation is not at present parallel, translations exist to make this possible in the future. Additionally, the online interface for the Plains Cree corpus described herein, found at https://korp.altlab.app/, has linked each wordform to its lemma in the online Plains Cree dictionary at https://itwewina.altlab.app/, so both a sentence-level parallel translation and individual word glosses will be available.

As this sample of Indigenous language corpora demonstrates, the availability of natural language data from Indigenous languages to form a corpus varies by language. Many have works in translation, but fewer have texts originally in the Indigenous language. While some are aligned with translations into majority languages, many are still not yet tagged for morphological or syntactic information. Any addition to corpora for Indigenous languages offers valuable opportunities for research and community use, whether for natural language processing, typology research, or teaching materials.

1.2.4 Other tools

Other tools, such as digital text tools, speech tools, and translation tools are also under development. Digital text tools include Optical Character Recognition (e.g., Cordova & Nouvel, 2021 for Quechua, Hubert et al., 2016 for Haida), keyboards (e.g., Santos & Harrigan, 2020 for Plains Cree syllabics), and transliteration tools, such as between ASCII and UTF-8 (Pine & Turin, 2018 for Heiltsuk) or between writing systems, such as the Roman alphabet and syllabic systems (e.g., Inuktitut Transcoder, n.d., https://www.inuktitutcomputing.ca/Transcoder/index.php for Inuktitut; integrated into the dictionary at https://itwewina.altlab.app/ for Plains Cree). Speech tools for speech recognition and synthesis are also underway for some languages (e.g., Harrigan

et al., 2019 for Plains Cree; Liu et al., 2022 for Hupa; Rehrig, 2017 for Kalaallisut). Machine translation resources exist, at least in demo form, for several Indigenous languages of the Americas. These include Cherokee (Zhang et al., 2020), Inuktitut (Joanis et al., 2020; Knowles et al., 2020; Le & Sadat, 2020; Roest et al., 2020), Kalaallisut (Jones, 2022; Resources, n.d.; https://nutserut.gl/), Guaraní (Gasser, 2018), Quechua (Ortega et al., 2020; Rios, 2015), Aymara (Coler & Homola, 2014), and Ayuuk (Zacarías & Meza, 2021). Machine translation relies heavily on morphological analysis for polysynthetic languages, to relate these features to those of a fusional majority language (e.g., English or Spanish), as well as corpora for testing, and so machine translation is generally limited to languages for which analysers and corpora already exist.

1.3 Contributions of this dissertation

The contributions of this dissertation involve three main components. First, a morphosyntactically tagged corpus is created using published Plains Cree texts, a morphological analyser, and a syntactic parser. Second, the use of this corpus is demonstrated—this includes a description of the corpus and its morphosyntactic features, an investigation of word order patterns, and a look at text types. Third, the tools and methods presented herein can be readily extended to other texts, thus contributing to the ongoing creation of new resources.

1.3.1 Creating a morphosyntactically tagged corpus for Plains Cree

1.3.1.1 The texts of the corpus

The Plains Cree corpus is divided into two subcorpora on the basis of time period: the Ahenakew-Wolfart (A-W) subcorpus (Ahenakew, 2000; Bear et al., 1998; Kâ-Nîpitêhtêw, 1998; Masuskapoe, 2010; Minde, 1997; Vandall & Douquette, 1987; Whitecalf, 1993) and the Bloomfield subcorpus (BT, "Bloomfield texts"; Bloomfield, 1930; 1934). Together, these volumes total 241,922 tokens (34,115 types), including Cree, English, French, numerals, punctuation, and metadata. When the non-Cree forms are trimmed, 152,405 Plains Cree tokens (31,616 types) remain. The wordforms are tagged for morphological features using a morphological model, the output of which is included in a manually-verified gold standard, and for syntactic relationships using a Constraint Grammar-based parser (introduced in §1.3.1.2). For more about the volumes, the speakers, and the content of the texts, see Chapter 4.

1.3.1.2 The morphological analyser and syntactic parser

The morphological feature tags in the corpus are assigned based on a Finite State Transducer-based morphological model (Snoek et al., 2014; Harrigan et al., 2017).[5] As part of the ongoing development of this model and the corpus, a morphological gold standard (MGS) for each subcorpus has also been created. Each MGS involves first applying the morphological model to each unique wordform type in the texts, then manually examining each type to verify, correct, and add analyses as necessary; these analyses retain morphological ambiguity, so many forms have multiple analyses. Thus, each morphological analysis in the corpus has been manually checked for accuracy. This process has also identified 1) issues in the morphological model to improve development and 2) stems to be added to the dictionary that underlies the model. The validated analyses are used in the current corpus, rather than automatically generated analyses from the model. For more details on the validation process and how this process increased the proportion of tagged forms, see Chapter 4.

The syntactic model for Plains Cree is a Constraint Grammar-based parser that uses lists of context-dependent constraints to 1) disambiguate ambiguous forms and 2) assign syntactic functions, based on the morphological features of wordforms and their relationships to each other. The syntactic model described in chapter 3 and applied to the corpus is the second iteration, the first being described in Schmirler et al. (2018) and its underlying manuscript, reproduced here in Appendix A.

1.3.2 Tagged corpora: Beyond descriptive studies

Studies of languages like Plains Cree generally take a detailed approach, perhaps with descriptive statistics, and consider a smaller number of examples in great detail. This is by necessity, as large-scale studies have generally not been possible without laborious hand-coding, counting, and calculating (e.g., Wolvengrey, 2011). Digital corpora, especially once tagged for parts of speech, morphological features, and syntactic relationships, are an invaluable tool in research, making it possible to complement detailed, descriptive research with large-scale statistical methods.

[5] Substantial revisions to the model have been undertaken since these publications.

In its present state, the Plains Cree corpus described herein can be used to discuss the relative occurrence of different parts of speech and morphological features in the corpus, as well as within and between the subcorpora, the occurrence of overt arguments, and various other features within different text types. How is the language changing over time? How "free" is free word order in practice? How do personal stories differ from legends, from speeches? The results that arise from these questions can be compared to not only previous descriptions of Plains Cree, but to descriptions in other Algonquian languages, or typologically similar languages around the world. This corpus is also an opportunity to explore these features in an oral language with a shorter history of literacy, as the current corpus consists only of transcribed spoken text.

1.3.2.1 Corpora and language typology

In my exploration of linguistic typology in this Plains Cree corpus, I focus on three broad typological features—verbal argument indexing, hierarchical alignment, and flexible word order—and how these interact with the overt realisation of arguments. In Plains Cree, both agent and patient arguments are marked on the verb, and they may also be represented by a separate noun, pronoun, or noun phrase, though these are often omitted. This verbal marking is described as a system of hierarchical alignment, as the morphological marking is determined by the relative topicality of the participants. This detailed morphology then co-occurs with word order flexibility, such that the order of the verb and arguments does not determine the meaning of the sentence. The frequency of these features among the world's languages varies: WALS indicates that the marking of both the agent and patient on a verb is the most common type of verbal person marking—including no overt marking—(Siewierska, 2013b), hierarchical alignment the least common alignment system (Siewierska, 2013a), and flexible word order falls somewhere in between (Dryer, 2013).

Though the marking of both agent and patient on the verb is common, the majority of languages with such a system are under-resourced and endangered minority languages, and thus are in many cases without large corpora to investigate their typological patterns. This observation, alongside the relative infrequency of hierarchical alignment, and to some extent flexible word order, makes the Plains Cree corpus an excellent tool for examining such typological features on a larger scale. Though few languages share all of these features with Plains Cree, comparisons can still be

undertaken with the corpus investigations that exist. In addition to looking at languages with very similar features, I also explore analogous patterns in majority languages, such as null subjects in Spanish and Portuguese, and the occurrence of nouns vs. pronouns in English (see Chapter 7).

1.3.2.2 Cross-linguistic text type comparison

There are several text type distinctions traditionally defined for Plains Cree: sacred and non-sacred narratives, recent and distant past narratives, funny stories, personal stories, and lectures (for more on these types, see Chapter 6). The content of these has been described, though the morphosyntactic features that occur within them have not. Using a register analysis approach, I explore the text types of Plains Cree, and then consider their similarities to text types and morphosyntactic features in other languages, such as lectures and sermons, and dialogue narratives and plays. Contrary to the traditional labels in Plains Cree, a primary distinction that emerges is that between dialogue and non-dialogue narrative, and in line with the labels, a distinction between narratives and lectures.

1.3.3 Ongoing resource creation

The Plains Cree corpus, though a considerable addition to the available digital resources for Indigenous languages, is still limited in a number of ways. Unlike majority languages, where the texts that comprise a corpus can be sampled from various sources and form a representative sample of the language spanning many types (news, fiction, nonfiction, diary entries, social media posts, as well as both written and spoken components), the Plains Cree corpus consists of a considerable portion of the extant texts which do not represent a great variety of text types, as further detailed in Chapter 6.[6] However, the current tagged corpus is far from the limit—with validated morphological analyses for over 31,616 unique types, and ever-improving morphological and syntactic models, new texts can be added with relative ease, as long as the orthography used is the same (or has regular variations that can be readily identified and adapted). With more tagged text, variation can be explored in even more detail, finding similarities and differences within and across text types, among different communities, and in different time periods. A number of existing works

[6] This might instead be termed an "exemplary corpus", following Bungarten (1979, pp. 42–3, as cited in Leech, 2007, p. 137), providing a suitable example of the language for investigation, though still not a representative sample.

are already earmarked to be added to the corpus, such as Rockthunder (2021), Whitecalf & Whitecalf (2021), and Wolvengrey (2007). The resources need not stop with the Plains dialect: similarities between Plains Cree and Woods Cree offer the opportunity for relatively straightforward adaptation of the Plains morphological and syntactic models to Woods Cree. Additionally, further research may demonstrate at least some syntactic similarities (in addition to morphological similarities) between Plains Cree and other Algonquian languages, thus lessening the development time of syntactic models for related languages.

Every digital resource that can be added to the limited list presented in §1.2 is another stepping stone on the path to language revitalisation and use in the 21st century and beyond. These tools must be developed with and for the communities and speakers of the languages—whatever the academic merit of these resources, community priorities must always be at the forefront of development. Digital tools and the resources they create can underlie community-focused and pedagogical outcomes. The models and tagged corpus can support language learning both in and out of the classroom, such as language practice with intelligent computer-assisted language learning (iCALL) applications, or examples of language use in context that can be linked directly to dictionary entries or grammatical descriptions. In urban communities, where access to fluent teachers and elders may be limited, digital tools can support language use and learning for children and adults alike. Like the morphological model, which can be implemented in tools such as a spell checker or word completion system, the parser can be expanded to form a grammar checker and can feed into iCALL applications; such tools, ubiquitous for majority languages in the 21st century, offer more opportunities for digital language use in a digital world.

1.4 Dissertation overview

In Chapter 2, I describe the morphosyntactic features of Plains Cree that I take into consideration in the creation of the Plains Cree parser. For verbs, these features include transitivity classes, person, number, order, and direction. For nominals, these features include animacy classes, person, number, possession, and their occurrences as locative and oblique elements. Particle classes are also briefly explored. This chapter also addresses features not yet implemented, but that will be useful in future parser development.

The development of the CG-based parser for Plains Cree is described in Chapter 3. A CG-based parser is an ordered list of constraints—if the context of a word demonstrates a particular pattern, the constraint acts on that word, whether to select or remove a set of features for an ambiguous word, or to label its syntactic function. The Plains Cree parser currently focuses on disambiguating some of the most common lexical ambiguities and on the assignment of actor and goal tags (Algonquianist labels roughly corresponding to subject and object).

In Chapter 4, I describe the tagged corpus in detail. First, I examine the efficacy of the parser described in the previous chapter, including the degree of ambiguity removed and the accuracy of the syntactic function tags. I then describe the texts of the corpus: the speakers, the editors, and other notes on the works. Finally, I present the frequency of the various morphosyntactic tags applied to the corpus, as the first full description of the morphosyntactic features of a Plains Cree corpus.

In Chapter 5, I explore the word order variation in the corpus. As a language with flexibility in the ordering of verbs and participants, as well as rich morphology that results in frequent unrealised arguments, both the occurrence of overt actors and goals and their position relative to the verb and to each other is explored.

In Chapter 6, I undertake a register analysis of the text types in the corpus. I first explore the corpora together, finding that while the corpora are in some ways very different in their morphosyntactic features, they are also in many ways alike. I then explore each subcorpus separately for a deeper look at both traditionally defined Plains Cree text types and a distinction between the presentation, rather than content, of a narrative.

Chapter 7 includes a discussion of the results in Chapters 5 and 6 with respect to languages with similar typological features, a look at text types and their features across languages, and a general conclusion. The typological comparisons look at languages with some combination of verbal argument indexing, hierarchical alignment, and flexible word order. The text type discussion looks at differences within narratives and between narratives and lectures.

Together, these chapters summarise the features of Plains Cree used to build a syntactic model, describe a Plains Cree corpus and the model's application, and undertakes two small-scale

investigations using the first morphosyntactically-tagged corpus for the language. With a tagged corpus, these investigations can go beyond previous descriptions of Plains Cree typology and text features, using statistics to explore the variation in word order and text types on a larger scale than previously possible.

Chapter 2

What goes into a syntactic parser? Plains Cree morphosyntax

2.1 Introduction

The Plains Cree corpus is built, essentially, of three components: the digitised texts from various published volumes, the morphological feature tags provided by a Finite State Transducer (FST)-based morphological model and a manually validated "gold standard" (cf. Chapter 4), and the syntactic function tags provided by a Constraint Grammar (CG) parser (cf. Chapter 5). The parser, in short, is a series of constraints that take the morphological feature tags of individual words and relate those features to other words in a sentence in order to identify syntactic relationships between them—thus, the morphosyntactic features to be referenced throughout the parser, and by extension this dissertation, must be introduced. This chapter thus describes the major morphosyntactic categories of Plains Cree and how these interact to form larger constituents.

Plains Cree demonstrates a number of morphosyntactic phenomena that have posed, and continue to pose, problems for computational modelling. Harrigan et al. (2017) discuss, for example, the issue of discontinuous person morphology, or long-distance dependencies between morphemes within a single word, in the morphological modelling of Plains Cree. From the syntactic perspective, a similar issue of flexible word order arises—where a language like English relies heavily on word order to establish relationships between words, Plains Cree does not, and morphological features must instead be relied on. Semantics and pragmatics are also highly relevant to thorough modelling of Plains Cree syntax and many of the issues that arise are often the result of semantic or pragmatic relationships not yet implemented in the parser; these are noted throughout, often with suggestions for how these might be implemented in future development.

Section 2.2 introduces the major subcategories of nouns, pronouns, and verbs, the morphological categories expressed on each of these, and how these words and their features interact in sentences.

Section 2.3 explores other word classes and syntactic relationships, especially those where word order plays a greater role. Section 2.4 highlights more complex issues, especially of semantics and pragmatics, that, while not necessarily considered at all in the current iteration of the parser, would greatly benefit future development. The chapter concludes in §2.5.

2.2 Identification of core arguments

In order to identify the core arguments of verbs in Plains Cree, several features of arguments and the verbs they are associated with must be referenced in the parser. These include at minimum nominal animacy classes, verbal transitivity classes, and their person and number morphology. In this section, I introduce the nominal and verbal classes of Plains Cree and the morphological categories expressed on these elements. Word order is not used to determine the role of an argument; arguments are often not lexically expressed and, when they do occur, they can occur in any order. Further discussion of word order patterns in the literature and in the Plains Cree corpus are left to Chapter 5.

2.2.1 Nominal classes and morphological categories

Nominals (nouns and pronouns) in Plains Cree are divided into two main subclasses: animate and inanimate. Animate nouns include people, animals, and trees, as well as some more idiosyncratic items, while inanimate nouns include everything else. While there appears to be a considerable degree of semantic motivation behind these subclasses, the idiosyncrasies indicate that this is a grammatical classification system (i.e., grammatical gender), rather than a purely semantic distinction (e.g., Ahenakew, 1987, pp. 17–9). The animacy of a noun determines the morphology, verbs, and pronouns with which it may co-occur (see §2.2.2). Various pronoun subclasses exist, though herein I focus on personal and demonstrative pronouns; personal pronouns refer to human beings and are always animate, while demonstrative pronouns may be either animate or inanimate (Wolfart, 1973, pp. 33–8).

Before a discussion of person and number, I will first introduce obviation, an important category for animate nominals in Plains Cree. Obviation refers to the relative topicality of a third person entity in the discourse: the more topical entity is proximate (with no overt marking, 3SG or 3PL), and the less topical entity is obviative (with the suffix *-(w)a*, 3′, "fourth person"). Whenever more

than one animate third person entity occurs in a discourse, one must be proximate and the other(s) obviative; this includes when one acts on another, or cases of possession. In (1)a., both Johnny and Mary are animate, and thus one is proximate and the other obviative. In this case, some previous context would either establish Johnny as the proximate entity, or this sentence alone would indicate that Johnny is more topical to the situation, for whatever reason. In (1)b., *mitten* is animate, and thus must be obviative when possessed by an animate third person. This situation differs slightly from (1)a., as here 'mitten' must be obviative and the possessor is always considered more topical, whereas either Johnny or Mary could be proximate or obviative, depending on the context (Wolfart, 1973, pp. 12–20).

(1) Obviation (Okimâsis, 2004, p. 140)[7]

a. Two animate participants

câniy kî-wîcihêw mêrîwa

câniy	kî-	wîcih-	**-ê**	-w	mêriy	-wa
NA	IPV	VTA	DIR[8]	3SG	NA	3′
Johnny	PST	help	3>3′		Mary	

'Johnny helped Mary.'

b. Animate possessum

otastisa

o(t)-astis	-a
3 NA	3′
mitten	

'his/her mitten(s)'

Note that in (1)a., it may appear that the proximate and obviative mark the subject and object, as in case-marking languages. However, this is not the case for Plains Cree, as is further explained in

[7] The glossing system used herein is adapted from Wolvengrey (2011). Following the morphological breakdown in line two, line three offers the grammatical category of each element and line four gives the gloss for more lexical elements and further information for more grammatical elements. Throughout, Plains Cree forms are given in the Standard Roman Orthography (e.g. Okimâsis & Wolvengrey, 2008).

[8] Explanation of the direct/inverse system of Plains Cree is given in §2.2.2.3.

§2.2.2.3. In addition to the obviative (3′), there is also another category of further obviative (3″, "fifth person") which occurs when two obviative entities interact in a discourse. These are relatively rare (see Schmirler, in press); they are not distinguished by nominal marking (both with the suffix *-(w)a*), but differences in verbal marking are seen instead.

Nouns and pronouns may be singular or plural, demonstrated in examples (2) and (3), and in Table 2.1 and Table 2.2 below. Additionally, these tables demonstrate the persons marked in Plains Cree: personal pronouns include first, second, and third persons, as well as a distinction between inclusive (21PL) and exclusive (1PL) first person plural. The third person singular personal pronouns may be used for either proximate or obviative,[9] though disambiguation between these readings is not yet attempted in the parser. Simple forms (e.g., 'I', 'you') contrast with emphatic or additive-focal forms (e.g., 'I too', 'you too') (Wolfart, 1973, pp. 33–8; Wolvengrey, 2011, pp. 207–8).[10] The demonstrative pronouns occur only for non-speech act participants and include both animate and inanimate pronouns (unlike personal pronouns, which are only animate). Inanimate persons are represented with 0SG or 0PL, to distinguish from animate third persons. Animate demonstrative pronouns occur in both proximate and obviative forms. Personal pronouns are relatively infrequent, as they are used with verbs for emphasis, while demonstrative pronouns may occur alongside a noun to further specify it, or they may occur as the arguments of verbs themselves. Demonstrative pronouns are additionally either proximal, medial, or distal, but these features do not influence the relationships between verbs and core arguments (Wolfart, 1973). Possession is also expressed on nouns, though as possession is not used to determine relationships between nouns and verbs in the parser, it is instead discussed in §2.3.1.3.

Examples of animate and inanimate nouns are given in (2) and (3) respectively, alongside their plural forms and, for the animate nouns, obviative forms, to demonstrate morphological differences. Note that the obviative *-(w)a* is identical to the inanimate plural. This syncretism extends to the demonstrative pronouns as well; this ambiguity is addressed in the parser.

[9] Per Wolfart (1973, p. 38), though a cursory exploration of the corpus does not offer any obvious examples.
[10] As personal pronouns are not required and are relatively infrequent (see Chapter 4 for more details), they in general already serve an emphatic purpose when they are included; thus Wolvengrey (2011) prefers to distinguish between what is here termed simple as emphasising the referent, and the emphatic as additive-focal to bring "attention to an additional referent" (pp. 207–8).

(2) Animate nouns (Okimâsis, 2004, pp. 10–11)

a.	*nâpêw*	‘man’	*nâpêwak*	‘men’	*nâpêwa*	‘man/men’
b.	*atim*	‘dog’	*atimwak*	‘dogs’	*atimwa*	‘dog/dogs’
c.	*asiniy*	‘stone’	*asiniyak*	‘stones’	*asiniya*	‘stone/stones’

(3) Inanimate nouns (Okimâsis, 2004, p. 11)

a.	*masinahikan*	‘book’	*masinahikana*	‘books’
b.	*astotin*	‘hat’	*astotina*	‘hats’
c.	*têhtapiwin*	‘chair’	*têhtapiwina*	‘chairs’

Table 2.1: Personal pronouns (Wolfart, 1973, p. 38)

Singular			Plural		
	Simple	Emphatic		Simple	Emphatic
1SG	niya	nîsta	1PL	niyanân	nîstanân
			21PL	kiyanaw	kîstanaw
2SG	kiya	kîsta			
			2PL	kiyawâw	kîstawâw
3SG	wiya	wîsta	3PL	wiyawâw	wîstawâw

Table 2.2: Demonstrative pronouns (Wolfart, 1973, p. 33)

	Animate			Inanimate	
	3SG	3PL	3′	0SG	0PL
Proximal	awa	ôki	ôhi	ôma	ôhi
Medial	ana	aniki	anihi	anima	anihi
Distal	nâha	nêki	nêhi	nêma	nêhi

The categories expressed on nouns and pronouns (animacy, person, number, obviation) may be used to establish participant relationships between nouns and verbs in the Constraint Grammar parser. These same categories are also used to identify relationships between nouns and pronouns.

In the next section, I describe the verbal features and how these are used in conjunction with nominal features to establish syntactic relationships.

2.2.2 Verbal classes and morphological categories

Central to the modelling of Plains Cree syntax are, unsurprisingly, verbs; namely, the parser must make reference to their transitivity and animacy classes and person morphology. The literature regarding Algonquian languages traditionally identifies four verb classes, determined by the animacy of the participants they govern, and by their transitivity, which indicates the number of core arguments we can potentially identify (generally, one argument for intransitive and two for transitive). Hence, there are four classes: inanimate intransitive verbs (VII), animate intransitive verbs (VAI), transitive inanimate verbs (VTI), and transitive animate verbs (VTA). VIIs include all verbs that take inanimate actors,[11] which include verbs that denote action, manner (e.g., sitting, standing, rolling) and attributes (as Plains Cree does not have a class of adjectives), but also includes impersonal verbs such as weather terms (e.g., *yôtin* 'it is windy') and temporals (e.g., *tipiskâw* 'it is night'), among others. VAIs include again verbs of motion, manner, and attributes, as well as intransitive actions, but are used to describe the actions or attributes of animate entities. VTIs are transitive verbs where the actor is animate and the goal is inanimate. VTAs are transitive verbs where two participants are animate (e.g., Dahlstrom, 1991; Wolfart, 1973, 1996; Wolvengrey, 2011).[12,13] Examples of each of these verb types are given in (4).

[11] Note that Algonquianist tradition generally prefers the terms actor and goal over subject and object (e.g., Bloomfield, 1946; Wolfart, 1973). In this work, actor and goal are the terms used unless otherwise specified. These terms do not align precisely with more widely used terms for either syntactic (e.g., subject, object, indirect object) or semantic roles (e.g., agent, patient, experiencer), but may encompass a variety of such roles.

[12] For VTAs, there is generally overt person morphology for both the actor and goal, while VTIs consistently only mark the actor. This is not so in all Algonquian languages, where VTI goals can also be marked on the verb (e.g., Ojibwe; Valentine, 2001, p. 311).

[13] These descriptions do not lay out or limit the total number of semantic or syntactic roles that can occur with each of these verb types. For example, VTAs include ditransitive benefactive verbs, which have an animate goal, e.g., a recipient, and a theme, which may be animate or inanimate and is not morphologically specified on the verb. Similarly, a verb stem itself may indicate the means by which an action is performed (e.g., 'by hand', 'by tool'), or an instrument may be lexically specified alongside the verb. Such three-place predicates can be identified in the parser in order to model these patterns. Similarly, impersonal VIIs, or zero-place predicates, can be identified and the parser can be blocked from assigning actors to these verbs. This has now been undertaken by A. Wolvengrey in the dictionary that underlies the morphological model, and this classification can be adapted for parsing.

(4) Verbal morphology (adapted from Okimâsis, 2004, 2021)

a. *wâpiskâw*

wâpiskâ	-w
VII	0SG
be.white	

'it is white'

b. *mîcisow*

mîciso	-w
VAI	3SG
eat	

's/he eats'

c. *wâpahtam*

wâpaht-	-am
VTI	3SG
see	

's/he sees it'

d. *niwâpamâw*

ni-	wâpam	-â	-w
1	VTA	DIR	3SG
	see	1>3	

'I see him/her'

There may be disagreement or uncertainty when it comes to the classification of VAIs and VTIs: are they to be classified semantically or morphologically? A further subclass, called VAI-t or VAI-O (a "transitive VAI" or a "VAI that takes an object"), is sometimes specified, and refers to verbs that are semantically transitive but, morphologically, behave like VAIs. Though VAI-t was used in the Plains Cree dictionary (Wolvengrey, 2001), on which the earliest iterations of the

morphological model were based, both the underlying dictionary database and the current morphological model label these as VTIs, but direct them to the part of the model that applies VAI morphology. Additionally, there also exist a smaller number of semantically intransitive verbs that are morphologically identical to the typical VTI; these are coded as VAIs that are directed to VTI morphology. For these reasons, Wolvengrey (2011) explores a three-way classification system that, rather than referring to the transitivity of verb, indicates the number of animate entities associated with the verb. V0, then, has no animate participants (i.e., VII), V1 has one animate participant (i.e., VAI and VTI), and V2 has two animate participants (i.e., VTA). While an elegant system, a classification that does not distinguish the transitivity of a verb class causes several problems when it comes to efficient modelling of Plains Cree syntax, and so the traditional four-way classification is maintained herein, appealing to semantics rather than morphology.[14] VAIs, then, can be expected to take one animate actor, while VTIs can take one animate actor and one inanimate goal regardless of their morphology.

2.2.2.1 Person

Plains Cree verbs are inflected for various persons. Alongside those persons mentioned above with respect to pronouns, unspecified actor[15] (X) is also marked. These persons occur in both the VAI and VTI paradigms, while different persons and person combinations are found in VII and VTA paradigms. The VII paradigm includes inflections for singular and plural inanimate (0SG and 0PL respectively), as well as an obviative for each (0′SG and 0′PL). These latter forms are considerably less frequent and are used when an animate third person is also present in the context, such as when an inanimate entity is possessed by an animate third person (e.g., Dahlstrom, 1991, p. 12).

The VTA paradigm is by far the most complex, as it includes inflection for nearly every combination of persons as well as the direction of the action, discussed in the following subsection.

[14] This four-way system that focuses on syntax and semantics over morphology is also often employed teaching grammars, such as Okimâsis (2004, 2021), making the morphological model and any tools that stem from it, such as the online dictionary, more accessible for students using the tools. The subclasses of VTIs are also directly based on Okimâsis's usage (A. Wolvengrey, personal communication).

[15] For VAIs and VTIs, the unspecified actor forms indicate that an action is going on, but do not indicate who is doing the action. An animate actor is implied and present, but simply unspecified. For some VTIs, the unspecified actor is derived to a VII using the suffix *-ikâtê*, and then behaves morphologically as a VII. For VTAs, the unspecified actor can be translated as "someone" or as a passive: *nikî-itikawin* 'someone said to me', 'I was told'. The morpheme *-ikawi* takes the place of the direction morphology for unspecified actor VTAs and the following goal morphology is then often more reminiscent of the VAI paradigm.

The VTA paradigm also displays considerably more forms for the unspecified actor; while VAI and VTI forms have only one form per order, though with multiple variants, the unspecified actor can act on every other person in the VTA paradigm. The VTA paradigm also includes the inanimate actor paradigm, called a rare or minor paradigm (e.g., Ahenakew, 1987; Schmirler, in press; Wolfart, 1973). It resembles a number of VAIs derived from VTAs, namely reflexive and reciprocal forms, and is used when an inanimate force acts upon an animate entity, e.g., *ê-itohtah**iko**yan* 'it takes you there', cf. reflexive *ê-itohtah**iso**yan* 'you take yourself there' (Wolfart, 1973; Wolvengrey, 2011). However, in VTA clauses, unlike for the other three verb classes, it is not sufficient to only know the verb class and the person features in order for a parser to assign syntactic function tags, and thus direction morphology is also relevant (§2.2.2.3).

2.2.2.2 Order

To fully describe verbal person marking, another aspect of verbal morphology must first be introduced: verbal orders. There are three major verbal orders in Plains Cree: independent, conjunct, and imperative; each of these is distinguished by different morphological patterns. The imperative is the most straightforward category, from an English-speaking perspective, while independent and conjunct forms are more complex in their semantics and pragmatics.[16] While often simply described as verb forms used in matrix and subordinate clauses respectively, conjunct verbs frequently occur in matrix clauses (Cook, 2014; Okimâsis, 2004, 2021; Wolfart, 1973, 1996). The complexity of the situation is far beyond the scope of the present work, but the interested reader may look to Cook (2014) and Harrigan (forthcoming) for detailed investigations of the pragmatic aspects of independent and conjunct forms. For the present purposes, the key fact is that independent and conjunct verbs mark person differently. Independent verbs use suffixes for non-speech act participants (third persons) and circumfixes for speech act participants (first and second persons), so person is marked both by prefixes and suffixes. VAI independent person affixes are presented in Table 2.3 (Okimâsis, 2004, 2021; cf. Wolvengrey, 2011, p. 111).[17] Person prefixes

[16] Each order consists of various subcategories or (sub)modes (e.g., Wolfart, 1973, pp. 41–6). However, as orders do not correspond to modes or moods in general linguistic terminology (e.g., Wolvengrey 2011, pp. 44–5), *order* is used to describe these categories.

[17] For more examples of Plains Cree verbal person morphology, see Okimâsis (2004, 2021) and Wolvengrey (2011, Chapter 2).

have two allomorphs: the forms shown below, that occur before consonants, and a variant ending in *t* before vowels (*nit-*, *kit-*). This pattern is also seen for possession (see §2.3.1.3).

Table 2.3: Independent VAI inflections

Person	Prefix	Suffix	Example (*nipâ-* 'to sleep')
1SG	ni-	-n	**ni**nipâ**n** 'I sleep'
2SG	ki-	-n	**ki**nipâ**n** 'you (sg.) sleep'
3SG	—	-w	nipâ**w** '(s)he sleeps'
1PL	ni-	-nân	**ni**nipâ**nân** 'we (excl.) sleep'
21PL	ki-	-naw	**ki**nipâ**naw** 'we (incl.) sleep'
2PL	ki-	-nâwâw	**ki**nipâ**nâwâw** 'you (pl.) sleep'
3PL	—	-wak	nipâ**wak** 'they sleep'
3′	—	-yiwa	nipâ**yiwa** '(s)he/they (obv.) sleep'
X	—	-(nâ)niwiw	nipâ**niwiw** 'someone sleeps'

Table 2.4: Conjunct VAI inflections

Person	Conjunct preverb	Suffix	Example (*nipâ-* 'to sleep')
1SG	ê-	-yân	**ê-**nipâ**yân** 'I sleep'
2SG	ê-	-yan	**ê-**nipâ**yan** 'you (sg.) sleep'
3SG	ê-	-t	**ê-**nipâ**t** '(s)he sleeps'
1PL	ê-	-yâhk	**ê-**nipâ**yâhk** 'we (excl.) sleep'
21PL	ê-	-yahk	**ê-**nipâ**yahk** 'we (incl.) sleep'
2PL	ê-	-yêk	**ê-**nipâ**yêk** 'you (pl.) sleep'
3PL	ê-	-cik	**ê-**nipâ**cik** 'they sleep'
3′	ê-	-yit	**ê-**nipâ**yit** '(s)he/they (obv.) sleep'
X	ê-	-hk	**ê-**nipâ**hk** 'someone sleeps'

Conjunct verbal morphology marks person only in suffixes, though these forms frequently occur with conjunct prefixes, which fall into the broader category of preverbs (see §3.3.2 and §3.4.2); the most common conjunct preverb is *ê-*. Various other conjunct preverbs with different

grammatical and semantic functions may be used, though these are not explored herein (cf. Cook, 2014). A subset of conjunct is often labelled as a separate order, the future conditional, which occurs as conjunct suffixes followed by *-i*. The VAI conjunct suffixes are demonstrated in Table 2.4 (Okimâsis, 2004, 2021, cf. Wolvengrey, 2011, p. 112).

Similar to the conjunct, imperatives are marked only using suffixes. There are immediate imperatives ("do *x* now") and delayed imperatives ("do *x* later"). For VAIs and VTIs, these occur in the second person singular, the second person plural, and the first person plural inclusive ("let's do *x*"—perhaps more accurately a hortative than an imperative). For VTAs, the person marking is slightly more complex, as a speaker can direct a second person to act on the speaker or on a third person. Examples of VAI imperatives are given in Table 2.5 (Wolvengrey, 2011, p. 113).

Table 2.5: Imperative VAI inflections

	Person	Preverb/prefix	Suffix	Example (*nikamo-* 'to sing')
Immediate	2SG	—	-Ø	nikamo 'sing!"
	2PL	—	-k	nikamo**k** 'all of you, sing!'
	21PL	—	-tân	nikamo**tân** 'let's sing!'
Delayed	2SG	—	-hkan	nikamo**hkan** 'sing later!"
	2PL	—	-hkêk	nikamo**hkêk** 'all of you, sing later!'
	21PL	—	-hkahk	nikamo**hkahk** 'let's sing later!'

2.2.2.3 Direction

In the VTA paradigm, the person marking interacts considerably with direction morphology. Both the person morphology and the direction features must be referenced in order to accurately assign actor and goal tags to nominals in a clause. The direction system of Plains Cree indicates whether a more topical participant acts on a less topical participant (direct) or a less topical participant acts on a more topical participant (inverse). This is often summarised in the Algonquian Person Hierarchy, given in (5). Speech act participants (SAPs, first and second persons) always rank above non-speech act participants (non-SAPs, third persons), and animate participants always rank above inanimate participants, indicated by the double >>. Within the local paradigm (SAPs acting on SAPs), second persons are considered "more topical" than first persons; this is argued on the basis

of morphology, as the independent order second person prefix is used when first and second person interact (and when they co-occur in the inclusive). In the mixed paradigm (SAPs and third persons), speech act participants are more topical than third persons. Within the non-local paradigm (third persons only), topicality is determined by discourse topic—proximate is more topical than obviative (Wolvengrey, 2011, pp. 173–6). Select examples of each paradigm are given in Table 2.6 through Table 2.8.

Direction marking is also evidence that the proximate and obviative categories expressed on nouns are not like case: neither category inherently expresses actor/subject or goal/object, but instead their syntactic roles are indicated by the direction on the verb. If the verbal morphology is direct, the proximate is the actor, and if the verbal morphology is inverse, the obviative is the actor. Note that for the independent paradigms in Tables 2.6 through 2.8, the person marking remains unchanged and indicates only which persons are involved in the action. The direction of the action is indicated by the direction morphology. However, it is not always the case that persons and direction are so clearly marked, as some conjunct verbs make use of portmanteau morphemes,[18] such as the examples given in Table 2.9.

The direct and inverse morphology of Algonquian languages is generally not considered an active/passive distinction; key evidence for this includes the fact that second person acting on first person is always direct and first person acting on second is always inverse (e.g., Wolvengrey, 2011). The intricacies of verbal person marking, especially within VTA paradigms, are much more thoroughly discussed and exemplified in works such as Dahlstrom (1991; 1995), Okimâsis (2004, 2021), Wolfart (1973), and Wolvengrey (2011).

(5) Algonquian person hierarchy (adapted from Wolvengrey, 2011, p. 57)

SAPS			>>	non-SAPs				
2	>	1	>>	3	>	3′	>>	0
				animate entities			>>	inanimate entities

[18] For the purposes of morphological modelling, morphemes need not be segmented as they are in detailed linguistic examples; instead, they can be identified as "chunks": the suffix *-êw* is interpreted as a single unit that corresponds to `3Sg+4Sg/PlO`—the morphological model's way of expressing "third person singular (subject/actor) and third person obviative (object/goal)".

Table 2.6: Local independent VTA paradigm: select examples

		Affixes[19]			Example (*wîcih* 'help')
Persons	Direction	Prefix	Theme	Suffix	
2SG>1SG	direct	ki-	-i	-n	**ki**wîcih**in** 'you (sg.) help me'
1SG>2SG	inverse	ki-	-iti	-n	**ki**wîcih**itin** 'I help you (sg.)'
2PL>1SG	direct	ki-	-i	-nâwâw	**ki**wîcih**inâwâw** 'you (pl.) help me'
1PL>2PL	inverse	ki-	-iti	-nân	**ki**wîcih**itinân** 'we help you (pl.)'

Table 2.7: Mixed independent VTA paradigm: select examples

		Affixes			Example (*wîcih* 'help')
Persons	Direction	Prefix	Theme	Suffix	
1SG>3SG	direct	ni-	-â	-w	**ni**wîcih**âw** 'I help him/her'
2SG>3PL	direct	ki-	-â	-wak	**ki**wîcih**âwak** 'you (sg.) help them'
3PL>1SG	inverse	ni-	-ik[20]	-wak	**ni**wîcih**ikwak** 'they help me'
3SG>2PL	inverse	ki-	-iko	-wâw	**ki**wîcih**ikowâw** '(s)he helps you (pl.)'

Table 2.8: Non-local independent VTA paradigm: select examples

		Affixes			Example (*wîcih* 'help')
Persons	Direction	Prefix	Theme	Suffix	
3SG>3′	direct	—	-ê	-w	wîcih**êw** '(s)he helps him/her/them (obv.)'
3PL>3′	direct	—	-ê	-wak	wîcih**êwak** 'they help him/her/them (obv.)'
3′>3SG	inverse	—	-ik	—	wîcih**ik** '(s)he/they (obv.) help(s) him/her'
3′>3PL	inverse	—	-ik	-wak	wîcih**ikwak** '(s)he/they (obv.) help(s) them'

[19] Person marking (prefixes and suffixes) are discussed above for VAIs. As there are minor differences for the VTA paradigms, these are given here as well, though the focus is the direction theme signs (direct or inverse). The person prefixes *ni-* and *ki-* for first and second persons are familiar, as is *-n* for singular speech act participants, *-nân* and *-nâwâw* for 1PL and 21PL, *-w* for third person singular, and *-wak* (*-w* + plural *-ak*) for third person plural.

[20] Each of these inverse theme signs containing <ik> are analysed as underlying *-ik(w)*, where the *Cw* cluster surfaces differently in different contexts: <w-w> may surface as <w> or <ow> and *Cw* surfaces as *C* word-finally. For more details see Wolfart (1973, 1996).

Table 2.9: Mixed conjunct VTA portmanteaus: select examples

		Affixes		Example (*wîcih* 'help')
Persons	Direction	Preverb	Suffix	
1SG>3SG	direct	ê-	-ak	**ê-**wîcih**ak** 'I help him/her'
2SG>3SG	direct	ê-	-at	**ê-**wîcih**at** 'you (sg) help him/her'
3SG>1SG	inverse	ê-	-it	**ê-**wîcih**it** '(s)he helps me'
3SG>2SG	inverse	ê-	-isk	**ê-**wîcih**isk** '(s)he helps you (sg)'

For the purposes of syntactic modelling, the person and direction marking on Plains Cree verbs, combined with the transitivity class of verb stems, is generally sufficient for identifying actors and goals within a clause (e.g., Schmirler et al., 2018). For example, for a VTA with first person singular, third person plural, and direct morphology with an animate plural noun in the clause, the animate noun will be recognised as the goal and no overt actor will be recognised. Aside from verbs and their core arguments (actors and goals), several other syntactic relationships can be readily identified in Plains Cree clauses. The majority of these relationships, discussed in §2.3, involve nouns, though other word classes are also relevant.

2.3 Other syntactic relationships

As arguments can generally be determined by grammatical marking such as direction and obviation, word order is not widely used in the Plains Cree parser to identify actors. However, other relationships in Plains Cree sentences can be modelled using the relative positions of words in the clause; many of these are discussed in the following section, which primarily concerns relationships between nouns and word classes other than verbs.

2.3.1 Noun phrases

2.3.1.1 Demonstratives

Though demonstrative pronouns can occur without an adjacent noun, as the arguments of verbs and in other positions, phrases consisting of demonstratives and nouns are also frequent. In a CG-based parser, the constraints can make use of adjacent nouns to determine the correct reading for an ambiguous demonstrative. For example, plural inanimate demonstratives are formally identical

to obviative animate demonstratives; if a noun with particular features occurs next to a demonstrative, the reading with the matching features can be selected.[21] The relative order of the noun and demonstrative changes the meaning in isolation, especially in teaching grammars: for example, when before the noun it is translated like a noun phrase, *ôma masinahikan* 'this book', but as a predicate when the demonstrative follows the noun, *masinahikan ôma* 'this is a book' (Ahenakew, 1987, p. 17).[22]

These phrases can then be associated with verbs as their arguments, on the basis of the features of the noun. However, in previous applications of the parser (e.g., Schmirler et al., 2018), it is not uncommon to find two or more instances of the same argument on one side or both sides of the verb. When these occur on the same side of the verb, there is generally some material intervening between a noun and a demonstrative; these may be particles, additional demonstratives, numerals, or other elements. Of these, numerals appear to be good candidates for inclusion within a noun phrase, e.g., *awa pêyak nâpêw* 'this one man'. When the two matching arguments are on either side of a verb, these are frequently a noun on one side of the verb and a demonstrative on the other, or a demonstrative on both sides. At present, these are all simply marked as the actor or goal, on the basis of their features, with no attempt to link the nominals to each other. Examples are given in (6) and (7); the last lines give the syntactic function tags supplied by the parser.

(6) Multiple goals (Douquette, 1987, pp. 72–3)

iyisâc awa nitatamiskawâw awa kisêyiniw.

iyisâc	awa	ni-	atamiskaw-	-â	-w	awa	kisêyiniw
IPC	DEM.A.SG	1	VTA	DIR	3SG	DEM.A.SG	NA
reluctantly	this		shake.hands.with	3>3′		this	old.man
	@GOAL>		@PRED-TA			@N>	@<GOAL

'I reluctantly shook hands with this old man'

[21] See again Table 2.2, where the syncretism in demonstrative pronouns is clear. The syncretism between the animate obviative and the inanimate plural can also be seen in the nominal suffix *-a*; the syncretism in the pronouns appears to be found in Proto-Algonquian (Proulx, 1988).

[22] Though these word orders are considered semantically distinct in isolation, both appear to occur as noun phrases, e.g., 'this book' in the corpus, at least on the basis of the English translations, and as such they are both modelled as noun phrases (a demonstrative dependent on a noun) and can be identified as actors and goals. Further scrutiny of their semantic and pragmatic patterns is required in future research, which may then influence the syntactic model.

(7) Multiple actors (Bear, 1998, pp. 130–1)

ahpô êtikwê awa ê-wî-kiskinohtahikoyâhk awa.

ahpô_êtikwê	awa	ê-	wî-	kiskinohtah-	-iko	-yâhk	awa
IPH	DEM.A.SG	IPV	IPV	VTA	INV	1PL	DEM.A.SG
maybe	this			show.the.way	3>1		this
	@ACTOR>			@PRED-TA			@<ACTOR

'maybe it [the owl] is going to show us the way'

While these examples show the more extreme cases of multiple demonstratives (with or without a noun), cases where a noun and demonstrative occur divided by a verb are not uncommon. For these cases of multiple instances of the same argument, particularly on both sides of a verb, there are a number of analyses in the literature, such as offering further clarification of a participant, adding an afterthought, etc. (e.g., Wolvengrey, 2011, pp. 340–64). Additionally, a subset of demonstratives can also serve as focus markers, indicating that a preceding noun or pronoun is focused. This allows constructions such as those in (8) to occur, wherein *awa*, the animate counterpart to *ôma*, has two different functions in otherwise identical contexts—as such, these ambiguities continue to defy context-based disambiguation (cf. Chapters 3 and 4).

(8) Demonstratives and focus (Wolvengrey, 2011, p. 296)

aw âwa mahihkan

i.

awa	awa	mahihkan
DEM.A.3SG	DEM.A.3SG	NA
this	this	wolf

'This is the wolf.'

OR

ii.

awa	awa	mahihkan
DEM.A.3SG	IPC	NA
this	FOC	wolf

'this here wolf'

Wolvengrey (2011) constructs extreme examples wherein forms such as *ôma* or *awa* can be used in a single sentence with all three functions, such as *ôm ôm ôma* 'this is the one here (inanimate)', where the functions of *ôma* are demonstrative pronoun, focus marker, and predicating demonstrative respectively (pp. 296–7). The disambiguation of pronoun uses from focus uses has continued to present issues for syntactic modelling (Schmirler et al, 2018; see also Chapter 3 and Appendix A).

2.3.1.2 Nominal predicates

As seen above, demonstrative pronouns can be used to predicate a noun. Personal pronouns can also predicate nouns in this manner. As with demonstratives, no copula is used for predication. However, a verbal construction can be used instead; a VAI stem is derived from a noun via the suffix *-iwi* and inflected as usual. Both options are exemplified in (9).

(9) Copular constructions (Wolvengrey, 2011, pp. 294–5)

a. *mahihkan niya*.
   mahihkan niya
   NA PRON.1SG
   wolf
   'I am Wolf (a name).'/'I am a wolf.'

b. *nimahihkaniwin*.
   ni- mahkihkaniwi- -n
   1 VAI 1SG
   be.a.wolf
   'I am a wolf.'

Personal pronouns and demonstratives may also occur together in such phrases, with differing meanings dependent upon the animacy of the demonstrative, demonstrated in (10). In (10)a., the pronoun indicates possession of the demonstrative; in (10)b., the same is true, but there is also a nominal antecedent for the third person pronoun, *wiya*. For (10)c., the form *awa* focuses 'wolf', which is possessed by the first person pronoun, the position of which predicates 'wolf'.

(10) Demonstratives and pronouns (Okimâsis, 2004, p. 20; Wolvengrey, 2011, p. 295)

a. *niya anima*.

niya	anima
PRON.1SG	DEM.I.SG
	that

'That (inanimate thing) is mine.'

b. *mahihkan ôma wiya*.

mahihkan	ôma	wiya
NA	DEM.I.SG	PRON.3SG
wolf	this	

'This (inanimate thing) is Wolf's.'

c. *mahihkan awa niya*.

mahihkan	awa	niya
NA	IPC	PRON.1SG
wolf	FOC	

'This wolf (here) is mine.'

The syntactic relationships between nouns, personal pronouns, and demonstrative pronouns allow for at least some disambiguation between the various functions of demonstrative pronouns and determine the semantic relationships within phrases involving nominal and pronominal elements. Whole noun phrases can then be identified as the arguments of verbs where relevant. The capabilities of the parser are detailed in Chapters 3 and 4.

2.3.1.3 Possession

Syntactic relationships between nominals may also be seen in cases of possession. When a possessed noun is preceded by a noun with appropriate features, that noun may be identified as its possessor, as in (11). Following nouns may also be considered possible possessors.

(11) Possession

câniy okâwiya

câniy o- -kâwiy- -a

NA 3 NDA 3′

Johnny mother

‘Johnny’s mother’

Animate and inanimate nouns may be possessed by all animate persons; the possessive morphology is nearly identical to the person morphology on independent verbs (compare Table 2.10 with Table 2.3 above). The possessum may be singular or plural, which is marked on the noun using the same plural morphology as non-possessed nouns. When an animate noun is possessed by an animate third person, the possessum is marked as obviative, with no number distinction as is usual for obviative nouns (e.g., Ahenakew, 1987; Dahlstrom, 1991; Okimâsis, 2004, 2021; Wolfart, 1973).

Table 2.10: Possessive morphology, possessed inanimate noun

Possessor			Possessum	
Person	Prefix	Suffix	Example (*maskisin* ‘shoe’)	Number
1SG	ni-	—	**ni**maskisin ‘my shoe’	SG
			nimaskisina ‘my shoes’	PL
2SG	ki-	—	**ki**maskisin ‘your (sg) shoe’	SG
			kimaskisina ‘your (sg) shoes’	PL
3SG	o-	—	**o**maskisin ‘his/her shoe’	SG
			omaskisina ‘his/her shoes’	PL
1PL	ni-	-inân	**ni**maskisin**inân** ‘our (excl) shoe’	SG
			nimaskisin**inân**a ‘our (excl) shoes’	PL
21	ki-	-inaw	**ki**maskisin**inaw** ‘our (incl) shoe’	SG
			kimaskisin**inaw**a ‘our (incl) shoes’	PL
2PL	ki-	-iwâw	**ki**maskisin**iwâw** ‘your (pl) shoe’	SG
			kimaskisin**iwâw**a ‘your (pl) shoes’	PL
3PL	o-	-iwâw	**o**maskisin**iwâw** ‘their shoe’	SG
			omaskisin**iwâw**a ‘their shoes’	PL
3′	o-	-iyiw	**o**maskisin**iyiw** ‘his/her (obv) shoe’	SG
			omaskisin**iyiw**a ‘his/her (obv) shoes’	PL

Within both animate and inanimate nouns, further subclasses can be identified: dependent (inalienable) vs. independent (alienable) nouns. Dependent nouns must be possessed to occur as free wordforms.[23] Dependent nouns include body parts, some articles of clothing, and kinship terms. When dependent noun stems begin with a vowel, rather than taking the otherwise expected *nit-*, *kit-*, or *ot-* variant, like for verbs and independent noun stem possession, the prefix is simply a consonant: *n-*, *k-*, *w-* (~ *o-*) (Wolfart 1973, pp. 15, 28–9).

2.3.2 Locatives

Locative nouns are formed with suffixes, simple locative *-ihk* and distributive locative *-inâhk*, and can be translated as 'at, in, on', etc. The distributive locative is generally translated as 'among', used especially for humans and animals. Examples are given in (12) and (13) (Wolfart, 1973, p. 31; 1996, p. 421). Locative nouns do not occur as core arguments, that is, as actors or goals (Wolvengrey, 2011, p. 39). They may occur simply specifying the locative setting of an action, or they may occur with particles that further describe the location, which may be likened to adpositions, creating adpositional phrases. For Plains Cree, this class of adpositional particles would include both prepositions and postpositions. The latter group is much smaller, consisting of forms that contain relative roots, a root class in Algonquian languages that occur with antecedents; hence, the locative noun occurs before and adposition like *isi* (Okimâsis, 2004, pp. 26–7; Wolvengrey, 2011, pp. 246–8). Examples of such phrases are given in (14).

(12) Simple locatives

a.	*sâkahikan*	'lake'	*sâkahikanihk*	'at the lake'
b.	*pihko*	'ashes'	*pihkohk*	'in the ashes'

(13) Distributive locatives

a.	*nêhiyaw*	'Cree person'	*nêhiyânâhk*	'in Cree country'
b.	*mostos*	'buffalo'	*mostosonâhk*	'in buffalo country'

[23] Dependent stems may also be included in derivations, such as in verb stems (as can free noun stems) (e.g., Wolfart, 1996).

(14) Adpositional phrases

a. *ohpimê wâskahikanihk*

ohpimê	wâskahikan	-ihk
IPC	NI	LOC
off	house	

'off to the house'

b. *misatimokamikohk isi*

misatimokamikw-	-ihk	isi
NI	LOC	IPC
horse barn, stable		towards

'towards the horse barn'

In a syntactic model, two distinct sets of prepositions and postpositions are labelled for identifying adpositional phrases, but can otherwise be treated as a single class of "locative particles", discussed in §2.3.4.

2.3.3 Obliques

Oblique nouns, for the purposes of modelling, include all other non-locative nouns associated with verbs that cannot be identified as actors and goals. These include themes, both animate and inanimate, that occur as the third arguments of ditransitive VTAs, such as in example (15). Here, the goal is 'his dog(s)' and the theme, 'the fish'.[24]

(15) Actor, goal, oblique (Wolvengrey, 2011, p. 215)

ana nâpêw kî-asamêw otêma kinosêwa.

ana	nâpêw	kî-	asam	-ê	-w	o-	-têm-	-a	kinosêw	-a
DEM.A.SG	NA	IPV	VTA	DIR	3SG	3	NA	3′	NA	3′
that	man	PST	feed	3>3′			dog		fish	

'That man fed his dog(s) fish.'/'That man fed fish to his dog(s).'

[24] For more on this example, see Chapter 5, examples (3) and (4).

For the purposes of modelling, oblique constraints can simply look for any noun in a clause that does not receive another syntactic tag (the current state of the Plains Cree parser), or they can label obliques only in the context of verbs that semantically allow for obliques. These would include ditransitive VTAs, as mentioned above (Plains Cree has, for example, a class of benefactive verbs derived using *-(st)amâw*, making identification of many such verbs straightforward, to a point), but also some VAIs, which, while morphologically intransitive, do allow for some specification of an "object", such as *minihkwê-* 'to drink'. To accurately model oblique nouns, some classification based on verbal semantics will be required; the current, more general constraints serve as a starting point for these improvements.

2.3.4 Particle classes

Particles in Plains Cree are a large heterogeneous class of words with many different functions. Ogg (1991) explores Plains Cree connective and temporal particles in texts; Wolvengrey (2011) looks at locative and temporal settings, labelling locative and temporal particles in the process. For Innu-aimun, a language closely related to Plains Cree, Oxford (2008) conducted a thorough particle classification, identifying a considerable range of classes and subclasses of particles. These include adnominal particles (adjectival, quantifying), prepositions, adverbs (circumstantial, degree, modal), focus particles, question particles, negating particles, conjunctions (coordinators, subordinators), and interjections, along with further subcategories for many.

For the current Plains Cree parser, only some of these categories are explored and modelled. These included negative particles, adpositional/locative particles, quantifiers, and degree adverbs, though little is done to explore how these interact with different word classes. The present work does not attempt any in-depth particle function classification, but does construct some basic lists for implementation in the parser. For a fuller description of these and their place in the Plains Cree parser, see Chapter 3.

2.4 Remaining issues

Beyond the above-mentioned patterns and functions, several issues remain for syntactic modelling, though they have not yet been seriously implemented in the parser. These include verbal orders

and the relationships of matrix and subordinate clauses, questions, lexical semantics, and the issue of the "clause".

2.4.1 Verbal orders

Verbal orders, as introduced in §2.2.2.2, have different syntactic, semantic, and pragmatic functions that are thus far not explored for the purposes of automatic parsing. Classifying independent clauses as matrix clauses and conjunct as subordinate would be a good first step for modelling interclausal relationships, though this is not entirely straightforward, as many conjunct clauses are matrix clauses (e.g., Cook, 2014). Though there are over 6,000 independent verbs and over 13,000 conjunct verbs in the Ahenakew-Wolfart subcorpus (see Chapter 4 for more details), very few of these occur in the same "clause" (as defined for the parser, see §2.4.4), so modelling these relationships requires a broader look at interclausal relationships, which are not yet accounted for in the parser.

Similarly, the issue of verbal orders can be linked to lexical semantics, discussed further in §2.4.3. A clearer picture of verbal semantics may aid in identifying verbs that are more likely to occur as matrix verbs, such as verbs of thinking, so structures such as "I think that *x*…", "I know that *x*…" may be more readily identified. Independent and conjunct verbs are also used differently in questions (e.g., Cook, 2014), so modelling the syntax of interrogatives would also be beneficial.

2.4.2 Questions

Research into both polar and content questions is available for Plains Cree (e.g., Blain, 1997; Cook, 2014), though computational modelling has not yet been undertaken. This is for a key reason: the Ahenakew-Wolfart subcorpus, consisting primarily of narratives and lectures, has very few questions on which to base the applicable constraints. However, looking at these in the corpus, many uses of *tân*-words do not occur in questions, and many questions in general occur as simple one- or two-word, verbless constructions, for the purposes of clarification or back-channelling on the part of the interviewer or the speakers. Despite these issues, the modelling of questions will be undertaken in future development of the parser, taking existing research as a starting point and building upon constraints on the basis of questions in the A-W subcorpus.

2.4.3 Lexical semantics

The inclusion of lexical semantics, particularly of verbs and nouns, will allow for considerable improvement to the coverage and accuracy of the Plains Cree parser. Identifying verbal semantics, for instance, may help in determining which nouns are more likely to be core arguments rather than obliques (as in (15) above), which will also require some nominal classification, such as [±human], or [±pet] for dog vs. fish—for more on the semantic and pragmatic issues here, see Chapter 5, §5.2. Verbal semantics will also allow for labelling of semantic role rather than simply syntactic role, which can allow for semantic as well as syntactic analyses in tagged a corpus.

The greatest challenge here, perhaps, still lies in particle classification and disambiguation, especially as a number of particles have multiple functions. The coordinator *êkwa*, for example, can be used not only as 'and', but also frequently as 'then', or 'now', and still yet can foreground added information (Ogg, 1991, pp. 26–8). Similarly, many particles can modify different word classes, though because of the less restricted word order of Plains Cree, proximity is not necessarily the best means of identifying which wordform a particle is best associated with. The meanings of particles, however, can still be used to identify some relationships to other words (e.g., locative particles as adpositions), and to give more details in syntactic relationships (e.g., a locative particle describing the location of the action described by the verb). Particle classification is further discussed in Chapter 3.

2.4.4 Beyond "clauses"

There are two main windows the Constraint Grammar parser uses to delimit syntactic relationships: the sentence and the clause. A sentence is delineated by the beginning of a sentence (as defined in the processed text) and sentence-ending punctuation, while a clause fits within this, delineated by any clause-boundary punctuation, including commas—that is, a Constraint Grammar clause is not equivalent to a syntactic clause. A CG clause is only a textual entity, e.g. delineated by punctuation, and as such imposed by the writer of a text (here, the editors transcribing it), in contrast to whatever linguistic unit a clause might be in spoken language. The Plains Cree parser makes extensive use of clause boundary punctuation to avoid overapplication of constraints, but this also results in many forms not labelled for any syntactic relationship, simply because there is no verb in the CG clause. Many of these wordforms are particles, which can modify a verb or an

entire (linguistic) clause, though no attempt to model these has yet been made in the Plains Cree parser. Some of these forms may be pre-clausal orienting information, or post-clausal afterthoughts or other resumptive information, as extensively discussed in Wolvengrey (2011).

Looking beyond the clause will be beneficial for a number of outstanding issues in the Plains Cree parser. A consideration of relationships between clauses may further increase the accuracy of actor and goal recognition. The identification of quotations, which at present are delineated as clauses simply with the clause boundary punctuation, could be connected to the quotative verbs that introduce them. In both identifying quotative verbs, and perhaps making use of quotation marks to identify quoted material, quotations may be modelled in future versions of the parser. For a fully syntactically annotated corpus, these elements and more will need to be accounted for. While some research exists (e.g., Cook, 2014; Wolvengrey, 2011), a partially annotated corpus can be an asset to this type of research as well: by exploring what can and cannot be tagged by the parser, the missing constraints can begin to take shape.

2.5 Conclusion

This chapter has briefly introduced the morphological features and syntactic relationships present in Plains Cree, for the purposes of automatic syntactic parsing using a Constraint Grammar parser. Individual features of nouns, verbs, pronouns, and particles added by a morphological model and manually validated can be supplemented in the parser with additional features and categories, before the relationships between words are identified and used for disambiguation and the assignment of syntactic functions. The next chapter details how the features introduced in this chapter are implemented in the parser, how some of the issues have been managed, and which issues still remain.

Chapter 3

Building a Constraint Grammar parser for Plains Cree

3.1 Introduction

In this chapter, I introduce the Constraint Grammar formalism and describe its application to create a syntactic parser for Plains Cree. The Constraint Grammar formalism, with a focus on morphological features and sentential context, lends itself well to the complex morphology and flexible word order of Plains Cree, as described in the previous chapter. Although some may consider Constraint Grammar an out-of-date formalism, this chapter demonstrates its effectiveness when modelling a less-resourced, morphologically rich language. The ongoing development of Constraint Grammar has also continued to improve its capabilities, many of which are utilised in the Plains Cree parser (Bick & Didriksen, 2015).

In §3.2, I summarise the Constraint Grammar formalism: §3.2.1 presents advantages of using the formalism for modelling Plains Cree syntax, §3.2.2 discusses the history and underlying principles of Constraint Grammar, §3.2.3 details some features of the latest version of Constraint Grammar that is used for the Plains Cree parser, and §3.2.4 lays out the basic structure of a Constraint Grammar parser and the syntax of constraints. In §3.3, I describe the Plains Cree parser; in §3.3.1, I summarise the features modelled in a previous version of the parser, then detail the present version in §3.3.2, which continues the development of constraints based solely on morphological features, but also introduces some lexical semantic features to the parser. Remaining issues and future development are discussed in §3.4.1 and §3.4.2 respectively. The chapter concludes in §3.5.

3.2 Constraint Grammar

3.2.1 What is Constraint Grammar?

A Constraint Grammar (CG) parser is a parser designed to disambiguate and assign syntactic functions using the Constraint Grammar formalism, which is a language-independent formalism

that makes use of morphological information in its context (surface syntax) to parse text. The parser is comprised of a list of context-based constraints that `ADD`, `REMOVE`, and `SELECT` tags and map syntactic functions with reference to the morphological features of words (such as those supplied by a morphological model) and the features of those words around them, that is, their context (e.g., Bick & Didriksen, 2015; Karlsson, 1990; Karlsson et al., 2011).

Though an admittedly older formalism, CG has several advantages for the modelling of Plains Cree. First, extensive reliance on morphological information works well for a morphologically rich language with flexible word order like Plains Cree, where word order has little bearing on syntactic relationships, especially the relationships between nouns and verbs. Second, CG is composed of explicit rules (constraints) that are written and tested by a linguist familiar with the structure of and literature on a language, which allows a parser to be developed and tested using only a small hand-annotated corpus. This contrasts with machine learning approaches, which require many tens or hundreds of thousands of words in an annotated corpus. For a less-resourced language like Plains Cree, with an available morphologically verified corpus of over 152,000 Plains Cree words, of which just 6,000 words are syntactically annotated,[25] the CG formalism is invaluable for the creation of a parser. Third, in conjunction with the previous advantage, the constraints in the a CG-based parser are written quite plainly, and with some basic training a linguist can begin to develop a parser with little or no programming knowledge. Again, this is invaluable when working with less-resourced languages.

Why is the development of such a parser worthwhile? In an academic context, a parsed corpus can be used to explore syntactic patterns, allowing for larger-scale analyses than previously possible, but other advantages exist for classrooms and communities. CG can easily adapt a parser to a grammar checker which, along with the spell checker made possible by a morphological model, can be used by speakers to communicate in Plains Cree digitally, with the same ease with which it is currently possible to communicate in imposed majority languages such as English or French. For example, CG has been used to great effect for grammar checkers for Swedish (e.g., Arppe,

[25] Indeed, this 6,000-word development corpus is only partially annotated, including just those features presently included in the parser. Fine details of Plains Cree syntax, such as the functions of its various particles, are not well-described, and so the development of the parser is both a means to learn about the syntax of the language and a tool to allow for future syntactic investigations and descriptions.

2000; Birn, 2000) and various Sámi dialects (e.g., Domeij et al., 2019; Trosterud & Moshagen, 2021; Wiechetek, 2018). The constraints can be used in iCALL (intelligent computer-assisted language learning) programs; while fill-in-the-blank activities are made possible by a morphological model (e.g., Bontogon et al., 2018 for Plains Cree), a syntactic model can expand the possibilities to more involved activities, such as writing full free-form sentences in response to generated questions. As the parser improves, so too will its advantages for linguists, students, and speakers of Plains Cree.

3.2.2 History

Karlsson (2011a) describes the origins of the CG formalism with reference to 24 goals. Here, I discuss only a few of these and their implications and theoretical underpinnings. For one, it should be noted that CG is designed for the parsing of unrestricted text, that is, it should give an analysis to any input as well as it can, rather than differentiating "good" (grammatical) input from "bad". While it is certainly a goal of a parser to yield the best possible analysis for a given language, and in doing so the constraints represent what is possible and impossible in a given context, it is not necessarily a goal to rule out ungrammatical constructions (cf. a goal of generative analysis, to define all and only all well-formed sentences). Though parsing "unrestricted" text, the parser can still make use of punctuation, such as for delineating clauses or identifying questions or quotations. Indeed, considerable preprocessing of text is required to be parsed with the CG formalism. This reduces the processing load on the parser, allowing for faster results with less computing power (Karlsson, 2011a, pp. 2–5). The CG formalism was designed to be language independent. While each individual parser is written for a particular language, the overall formalism was not meant to be biased to any particular type of syntactic structure. Thus, CG was initially applied to Finnish, which relies heavily on morphological information, but also English, where linear order plays a greater role in syntactic relationships. The application of CG to Sámi dialects and Basque (Antonsen & Trosterud, 2011; Wiechetek & Arriola, 2011), among other morphologically complex and less-resourced Indigenous languages, has also been successful.

Another key goal of Constraint Grammar is to resolve lexical ambiguity, such that forms that can have multiple morphological analyses are reduced, as much as possible, to a single analysis on the basis of context. The disambiguation in the current iteration of the Plains Cree parser has not

changed greatly from the 2017 version, as many of the persistent ambiguities can not be straightforwardly disambiguated from context. Rather than dedicate time and resources to heuristic constraints, the focus of the current development has instead been on syntactic function mapping.[26] However, context is key not only to resolving ambiguities, but context is also the basis of syntactic mapping: making use of context to identify not only which reading is appropriate, but also to identify syntactic relationships. Thus, the relationships between words are a major focus in both cases; if the syntactic relationship between words can be determined by another word in context, the best reading of an ambiguous word can likely also be identified Karlsson (2011a, pp. 11–4). As syntactic functions are explored, the relationships between words can also be fed into disambiguation where possible.

Other goals of the CG formalism tie in well with its advantages for Plains Cree, discussed in §3.2.1. The formalism was designed such that a parser can be written based on extant grammars, or a linguist's knowledge and observations of a language based on texts, and the parser allows for ready testing and debugging, without intensive programming experience on the part of the developer. Constraints can also be added incrementally, so the parser can be developed in stages (Karlsson, 2011a, pp. 14–8). This allows for straightforward, relatively fast development of a parser without the need for large, manually tagged corpora, and each stage of development can be used to explore a corpus to inform the next stage. Karlsson (2011a) sums up CG with two main points:

> (i) Language is an open-ended system where there is no strict demarcation line between grammatical and ungrammatical sentences. Therefore all grammars are bound to leak.
> (ii) The cornerstone of syntax is morphology, especially the language-particular systems of morphological features. Syntactic rules are generalizations telling (a) how word-forms, conceived as complexes of morphological features, occur in particular word order configurations, and (b) what natural classes, "syntactic functions", can be isolated and inferred in such configurations. (p. 37)

For languages such as Plains Cree, where word order is not a reliable indicator of syntactic functions, the morphological features of wordforms and their relationships to other wordforms' morphological features can be used to great effect in the development of a CG-based parser.

[26] See §3.3 of this chapter for further discussion of remaining ambiguities.

3.2.3 VISL CG-3

The version of Constraint Grammar used for the Plains Cree parser is the third iteration, CG-3 or VISL[27] CG-3 (Bick & Didriksen, 2015). The principles laid out in Karlsson et al. (2011) developed into CG-1. A second version, CG-2, introduced features such as `BARRIER` ("do not look beyond a particular element to find the stated context") and the ability to create `SET`s (see below) (e.g., Tapanainen, 1996). Finally, VISL CG, an open-source version compatible with CG-2 was released. CG-3 is compatible with VISL CG and implements more features than either VISL CG or CG-2 (Bick & Didriksen, 2015).

Unlike earlier versions, CG-3 does not emphasise a flat, non-hierarchical structure; while such a structure is inherent in the constraints, the CG-3 formalism directly allows for bracketing of constituents and for transformation to nested tree structures. The ability to implement constituency or tree-based structures allows for more straightforward mapping between a CG-based parser and a generative account of a language's syntax; therefore, linguists who prefer to work with generative theories can still make use of CG in building a parser. Bracketing, or "chunking", is used to identify frequent phrases made up of interrelated elements, and label them as phrases accordingly. This function is used to some extent in the Plains Cree model for identifying particle phrases, discussed and exemplified in §3.3.2 below. Along these lines, CG-3 also allows for the use of templates, such as grouping a determiner-adjective-noun sequence in English into an NP. Templates have not been used in the Plains Cree parser, but could be an option for simplifying some constraints, such as those for nouns phrases or prepositional phrases (see §3.3.1) (Bick & Didriksen, 2015).

A particular advantage of CG-3 utilised frequently in the Plains Cree parser is the implementation of regular expressions. These are used to allow for different spellings of the same word (e.g., *kêkway* as an alternative, perhaps dialectal or archaic, form of *kîkway*), different forms of certain pronouns (e.g., singular vs. plural), the correction of ambiguities introduced by an overzealous morphological model, and the selection of the form with the fewest number of preverbs. This last option is a straightforward heuristic representing a preference for lexicalised forms (where the preverb is part of the lemma) over derived forms recognised by the morphological model. This is demonstrated in example (1), where the constraint in (1)a. selects the reading with two hyphens in

[27] Visual Interactive Syntactic Learning Project (https://visl.sdu.dk/).

the lemma (hyphens are used to separate preverbs from stems) and the constraint in (1)b. selects the reading with one hyphen. The constraints must occur in this order to select the reading with the most preverbs already lexicalised and included as an entry in the dictionary, as in (1)c., where the semicolon marks the removed reading in the "trace" mode used for testing. Unification of features has also proven useful; demonstratives and nouns can be linked when they share the same animacy and number features, rather than writing constraints for all feature pairings. Here, a `SET` is created to group animacy and number features ((2)a.), then constraint in (2)b. instructs the parser to remove the IPC (particle) reading of a demonstrative when the features of the demonstrative and the preceding noun "match" (`$$NUMBERGENDER`). For more on the syntax of constraints see §3.2.4.

(1) Lexicalised preverbs

a. `SELECT:PV2 ("[^<]+[-].+[-][^>]+"r) ;`

b. `SELECT:PV1 ("[^<]+[-][^>]+"r) ;`

c.
```
"<nitati-miyw-âyân>"
    "ati-miyw-âyâw" V AI Ind Prs 1Sg SELECT:PV2
;   "miyw-âyâw" PV/ati V AI Ind Prs 1Sg SELECT:PV2
```

(2) Unification

a.
```
SET NUMBERGENDER = (A Sg) OR (A Pl) OR (I Sg) OR (I Pl)
OR (A Obv) ;
```

b.
```
REMOVE:DemNotIPC IPC (-1 N + $$NUMBERGENDER)(0 Pron +
Dem + $$NUMBERGENDER) ;
```

CG-3 offers a number of other advantages not yet used in the Plains Cree parser, though many could be implemented. For example, CG-3 allows for the inclusion of statistical information, so that, for example, a more frequent part of speech could be selected, or selecting the class of goal a verb with certain semantic features is likely to have. The former is hard-coded in some constraints in the Plains Cree parser; if a word can be both a particle and a noun/verb, the particle reading is considered more likely and is thus selected, unless the context strongly suggests otherwise: for example, the particle reading is removed if the wordform is also a noun and is adjacent to an

agreeing demonstrative. CG-3 also allows for the constraints to scope over larger sections of text than just clauses or sentences; this would be useful in identifying longer-distance anaphoric relationships in Plains Cree, such as tracking participants across sentences and paragraphs. CG-3 is also more adaptable to the features of a particular text type (Bick & Didriksen, 2015). As Plains Cree text types are already known to display some different syntactic patterns (cf. Schmirler & Arppe, 2020; Schmirler et al., 2018), patterns more common in narratives vs. speeches could be identified and then implemented in a future version of the Plains Cree parser.

3.2.4 Basics of a CG parser

In the CG formalism, each word has its *reading* or, for an ambiguous form, *readings*, that make up its *cohort*, exemplified in (3). The first line is the wordform as it occurs in the text. Each reading consists of the lemma in quotation marks followed by morphological tags, such as those produced by a morphological model, which can be referenced by the constraints.

(3) `"<kîkway>"` (“what”)
`"kîkway" N I Sg`
`"kîkway" N A Sg`

Key to CG is *context*: the morphological features and the linear order of various elements are used to relate one word to another. For example, many English words are ambiguous for part of speech, such as noun or verb. In a sentence like *the bears dance*, both *bears* and *dance* have noun and verb readings. However, constraints would select a noun reading for *bears* because it is preceded by *the*, and as there are no further candidates for a finite verb, and there is agreement between a plural noun and a bare verb in English, a verb reading would be selected for *dance*. Following disambiguation, syntactic functions can then be determined; in this same example, *the* would be dependent on *bears* as an article, and *bears* dependent on *dance* as a subject. A considerable number of constraints may be required to disambiguate forms with several readings, such as English *that*, but this is not considered detrimental in CG, as compactness is not a requirement for practical modelling (Karlsson, 2011a, pp. 25–7).

The parser consists of several elements. First, there must be an appropriately preprocessed file includes each token (wordform, punctuation, etc.) of a text or corpus, each with its cohort of

analyses. For Plains Cree, this is produced from the A-W and BT subcorpora, which have been analysed with the Plains Cree morphological model and manually verified (Harrigan et al., 2017; Schmirler et al., under development; see Chapter 4). For some languages, syntactic functions can be included at this point, if a form has only one possible function (e.g., *the* in English will be dependent on a following noun, though some intervening material such as adjectives is permitted; this can be labelled in the preprocessing stage to streamline the parser). Then there are the constraint files (which may be conflated into a single file) (Karlsson, 2011b). For Plains Cree, the disambiguation and syntactic function assignment constraints are separated into two files, which are applied in that order to the texts. In the future, these files could be combined, and their constraints interleaved where necessary, to better parse the language.

Within the constraint file(s), there are various sections. First, all the tags and mappings that will be referenced in the constraints must be specified: sentence delimiters and clause boundaries are listed, as are all the morphological tag features that appear in the texts as supplied by the morphological model. The syntactic function tags must also be specified. `SET`s, or groups of `LIST`s (described in more detail below), are also defined. The sentence delimiters define the default window over which constraints can apply, while clause boundaries (for Plains Cree, these include sentence-delimiting punctuation such as periods and question marks, but also commas and semicolons) are made use of with the `BARRIER` function; rather than a constraint's window being the entire sentence, it is a "clause" as delimited by certain punctuation marks. Following these, the constraints begin. For Plains Cree, these are given some basic ordering so that earlier constraints do not rule out key relationships laid out in later constraints: for example, the constraints first try to associate a demonstrative with a noun, before it can be labelled as an actor or goal; actors and goals are labelled before obliques are identified. In a more fully implemented CG parser, the ordering of larger sections is also important: e.g., blanket constraints should apply before heuristic constraints (Karlsson, 2011b).

Disambiguation constraints and syntactic function mapping constraints differ slightly in their syntax. Disambiguation constraints `REMOVE` or `SELECT` readings, or `ADD` tags to certain readings to further specify them and provide opportunities for more specific contexts in future constraints. In (4), the constraint `REMOVE`s the reading with the tags `A` (animate) and `Obv` (obviative) when

the context is as follows: the word immediately before (`-1`) is an inanimate (`I`) plural noun and the form itself (`0`) also has another reading, inanimate plural demonstrative. This constraint could also be expressed as selecting the inanimate plural reading.

(4) A disambiguation constraint

```
a. ACTION:name            <target>  <context>
b. REMOVE:DemINnotANObv   A + Obv   (-1 N + I + Pl)(0 Dem +
                                    I + Pl) ;
```

Syntactic function constraints `MAP` a function tag to a `TARGET` in a particular context. In (5), the constraint `MAP`s the function tag `@<N` ("dependent on a noun to the left"; note that all syntactic tags begin with `@`) on a `TARGET`, the same inanimate singular demonstrative in (4), `IF` the context applies, same as the above example. These constraints apply to a situation such as that in (6). The demonstrative is disambiguated because it is adjacent to a noun with features that match only one of its readings; the syntactic function is applied in the same context.

(5) A function constraint

```
a. ACTION:name    <tag> TARGET  <target>      IF  <context>
b. MAP:DemNINPlR  @<N   TARGET  Dem + I + Pl  IF  (-1 N + I
                                                  + Pl) ;
```

(6) Disambiguation and function mapping ("these little shoes of mine")

```
"<nimaskisinisa>"                                  ("my little shoes")
   "maskisinis" N I Pl Px1Sg
"<ôhi>"                                            ("these")
   "ôma" Pron Dem Prox I Pl @<N MAP:DemNINPlR
;  "awa" Pron Dem Prox A Obv REMOVE:DemINnotANObv
```

In the next section, I summarise the previous development of the Plains Cree CG-based parser, then describe the development of the current version and what remains to be considered in future

iterations. While examples are kept to a minimum, as plain text often describes the content and effects of constraints more efficiently, the constraints precede much like the above examples.

3.3 The CG-based parser for Plains Cree

This section details the previous and ongoing development of the Plains Cree CG-based parser (Schmirler et al., under development b.). The source code can be found at https://github.com/giellalt/lang-crk/tree/main/src/cg3.

3.3.1 Previous development

The previous iteration of the Plains Cree parser, as presented in Schmirler et al. (2018), focused on the implementation of a number of core relationships between nominals and verbs, making reference only to the morphological features available in the cohorts provided by the Plains Cree morphological model and subsequent manual validation.[28] Using the animacy, number, person, and obviation features of nouns, these were associated with verbs on the basis of their verb class, person, and number features. In this way, VIIs could take inanimate actors, VAIs animate actors, VTIs animate actors and inanimate goals, and VTAs animate actors and animate goals. For example, a singular animate noun will be labelled as an actor, if in the same clause there is a VAI with a 3SG (third person singular) feature tag. More detailed constraints were required for actors and goals of VTAs, though as the verbal features give explicit person information, the constraints were still straightforward.

A number of nominal elements were identified as actors or goals in this version of the parser. Nouns alone could be marked as actors, as could demonstrative or personal pronouns. Nouns could also enter into noun phrases; the noun, as the head of such phrases, was then marked as an actor or goal if the context allowed. Basic morphological features were also used to identify these noun phrases, sequences of demonstratives and nouns that can act as units when associated with other elements, namely verbs. If an inanimate singular demonstrative preceded or followed an inanimate singular noun, these were considered a single phrase, with the noun as its head. Such contexts were also used for disambiguation of ambiguous demonstratives: animate obviative and inanimate plural

[28] Of course, some of these morphological features do give some indication of semantics, such as animacy or verb transitivity, but otherwise no further attempts were made to implement semantic information.

demonstratives are formally identical, and so these were disambiguated based on the proximity to nouns of the relevant features. The parser laid out in Schmirler et al. (2018) also included first attempts at labelling a number of other relationships, such as associating locative nouns with preceding particles and labelling nouns without actor or goal tags as obliques. These rules, however, did not take into account the semantics of particles, verbs, or nouns, and as such were not discussed in the evaluation of the parser and were not considered in the analysis. Further development of these rules is discussed below with respect to the present stage of parser development.

Aside from these largely untested constraints for locatives and obliques, the weakest section of the Plains Cree parser in Schmirler et al. (2018) was the disambiguation module. While some ambiguities could be resolved through context (e.g., demonstrative pronouns), many constraints were written based on more likely patterns[29] or were based on individual issues arising from the FST tags in the corpus. For example, the former type included the selection of particles over nouns or verbs, based on the observation that particles are as a whole the most frequent word class. A number of idiosyncratic constraints were implemented to resolve ambiguity introduced by the morphological model, not necessarily inherent to the wordforms. The morphological model has the option of ignoring vowel length in order to increase the likelihood of offering some analysis, regardless of a possible discrepancy in spelling or transcription; this introduced ambiguity between first- and second-person conjunct forms, which differ only in vowel length. Thus, the Plains Cree parser would select the "correct" analysis on the basis of the actual spelling of the wordform—long vowels prompted the selection of the first person reading, short vowels the second person reading. Such constraints are not useful—and are in fact detrimental—outside of the context of current corpus, such as in a grammar checker, where accuracy in spelling cannot be assumed. However, for the purposes of Schmirler et al. (2018) and the present work—to analyse the order of nouns and verbs in a manually verified corpus—these constraints are not out of place.

Schmirler et al.'s (2018) parser left a number of major issues to tackle. For example, overapplication of actor constraints to *ôma* (which may be an inanimate singular demonstrative or

[29] Heuristic constraints are made possible by the VISL CG-3 formalism; these were not implemented in Schmirler et al. (2018) and are beyond the scope of the present work as well.

a focus particle) in the context of impersonal VIIs was a persistent issue. Recall that semantic information was not used previously, and so impersonal VIIs were not labelled any differently from other VIIs; these VIIs should not have actors, and yet due to the ambiguity of *ôma*, actors were identified. The parser also did not allow for more complex noun phrases, such as those with numerals between a demonstrative and noun, or a noun specified by a numeral with no demonstrative. The former situation resulted in a number of instances where two actors, both the demonstrative and the noun, were identified for the verb in that clause. Fortunately, the parser itself and the remaining issues identified in Schmirler et al. (2018) formed a solid base for further development laid out below.

3.3.2 Current development

In this section, I describe the development of the parser that has been undertaken since Schmirler et al. (2018), both with respect to morphological and lexicosemantic features, and to word order. As part of this process, a development file is used to test the constraints ("Syntactic Gold Standard"; see Chapter 4). As this development file was also used for the earlier iteration of the parser, further tagging has been undertaken for testing new constraints: previous omissions or errors in the development file were added and corrected, and on a number of occasions were corrected in response to tags supplied by the parser, which also helped identify coding errors. However, the labelling produced by the parser was never used as grounds for coding in the development text, only to identify areas that could be checked in the text and added on the basis of the text and its translation.

3.3.2.1 Morphological and lexicosemantic feature tags

To further develop the CG-based parser for Plains Cree, a number of changes and additions to the constraints were required. These constraints include possessive relationships, noun phrases, and obliques. Other aspects required more extensive modelling, as `LIST`s were created to allow for these to be tagged beyond the tags available in the morphological model: the fuller implementation of the relationships between locatives and adpositions, the recognition of particle phrases, disallowing actors for zero-place predicates, and the initial stages of particle classification into negative, locative, temporal, and quantifying particles.

Possessive relationships were far easier to implement than might be expected from the rich system of possession discussed in the previous chapter. While there are nearly 3,000 possessed nouns in the A-W subcorpus and over 5,000 in the BT subcorpus, only 33 and 91 nouns respectively are labelled as a possessor by the parser. In my initial foray into the implementation of possessive relationships, I expected to tag both nouns and pronouns as possessors in noun phrases (i.e. `@Pos>`, much like `@N>`, associates a demonstrative with a following noun). However, a cursory exploration of the corpus showed that when personal pronouns preceded possessed nouns, they were in almost all instances behaving as actors or goals in the phrases in question. Of course, these pronouns are still the possessors in a broader, pragmatic sense, but their syntactic function appeared to be as an actor or goal, and so these roles take precedence.[30] Thus, only three constraints were implemented to recognise these relationships. These constraints all allow for animate possessors: a singular animate noun may possess a noun with a `Px3Sg` tag, a plural animate noun may possess a noun with a `Px3Pl` tag, and an obviative noun may possess a noun with a `Px4Sg/Pl` tag.

The current parser has also implemented constraints for more detailed nominal phrases than previously allowed. Beyond the possessive phrases discussed above, the current constraints now recognise phrases with numerals and phrases with resumptive pronouns; numerals were already tagged by the morphological model and gold standard, while other pronoun types were tagged within the parser. Previously, the parser recognised noun phrases that consisted of nouns and demonstratives. Now, phrases consisting of demonstratives, numerals, and nouns; demonstratives and numerals; and numerals and nouns are all recognised. In cases where a noun is present, both a demonstrative and a numeral will be marked `@N` ("dependent on a nominal"). However, where there is no noun, the numeral is dependent on the demonstrative, which may in turn be an actor or goal.[31]

[30] Further development of a parser may find a way to appropriately tease out such pragmatic relationships, though only syntactic functions are currently marked. Given the (non-)observation of personal pronouns as possessors in the corpus, this may also serve as a potential heuristic constraint in future work.

[31] In phrases such as *awa pêyak* 'this one', the numeral is associated with the demonstrative, which can then be marked as an actor or goal, though this the numeral might be interpreted the head of the phrase instead. However, the current solution allows for more precision in associating the phrase with a verb, as the demonstrative and not the numeral carries the marked animacy and number features. Future implementation may allow for more detailed constraints that will allow numerals to behave as actors and goals.

Oblique constraints have existed in the Plains Cree parser in some form since before the work undertaken in Schmirler et al. (2018). At the time of that work, obliques, like adpositions and locatives, were not taken into account in the analysis. The pre-existing constraints, drawn from parsers built for other languages to exemplify their use, did not allow for pronouns to be oblique elements; this avoided the problem of overapplication of oblique constraints especially to *ôma*, which has posed an ongoing problem in both disambiguation and function assignment. However, there are certainly instances were *ôma* is an oblique nominal, and so further consideration is needed.

For the basic implementation of lexical semantics beyond tags drawn from the manually validated morphology analyses, groups were drawn from dictionaries and textbooks of Plains Cree. The dictionary database underlying Wolvengrey (2001) often contained information that was not implemented in the morphological model tags, such as particle phrases and zero-place VIIs. Dictionary entries and corpus investigation were used in tandem to identify elements such as negatives—these were forms in the dictionary glossed as 'no, not' or similar, and a brief search through the corpus demonstrated that these were used as negators before verbs. Additionally, textbooks could be used for lists of adpositions that occur with locative nouns (e.g. Okimâsis, 2004, 2021; Ratt, 2016, 2022).

When these groups were identified, some of them include phrases of two or more words, which must be linked together. Particle phrases are sequences of two or more particles that are considered lexicalised and can function as a single unit. This is achieved through 164 pairs of disambiguation constraints (or sets, if the phrase consists of three or more words), one for each phrase. An example of such a pair is given in (7). The first of these constraints is a `SUBSTITUTE` constraint, which in this case finds the lemma *namôya* 'not' and replaces this with *namôya wîhkâc* 'never', targeting only those instances of *namôya* that are followed by *wîhkâc*. Then, a `REMCOHORT` ("remove cohort") constraint removes the whole cohort (all readings plus the wordform) for *wîhkâc*, but only in those instances where it is preceded by the newly created particle phrase. In some cases, such as this example, additional information is added with the tag `Neg`; for single-word items these tags are added by means of `LIST`s.

(7) Constraint pair for the particle phrase *namôya wîhkâc*

a.
```
SUBSTITUTE ("<namôya>" "namôya" IPC Neg)
   ("<namôya_wîhkâc>" "namôya_wîhkâc" IPH Neg) TARGET
   ("<namôya>" IPC) IF (1 ("<wîhkâc>" IPC)) ;
```

b.
```
REMCOHORT ("<wîhkâc>") (-1 ("<namôya_wîhkâc>" IPH)) ;
```

The CG formalism allows for the inclusion of `LIST`s and `SET`s. The former is a list of morphological tags or wordforms/lemmas that are given some label in the parser, and the latter is a group of lists that may be subgrouped or added to/subtracted from each other. For example, lists are used to group all preverbs together, and to indicate that forms with tags like `N`, `V`, and `IPC` are all `WORD`s. These `LIST`s and `SET`s can be referenced in constraints or, as is frequent in the current iteration of the Plains Cree parser, can be used to add additional tags that allow for more precise reference in disambiguation and function mapping. This approach is used in the implementation of zero-place predicates and their appropriate lack of actors was achieved in three main steps. First, a `LIST` was created of known zero-place VIIs (Wolvengrey, 2001). Second, this `LIST` was used to add the tag `IIZ` ("II zero") to all readings for these verbs. Third, a `SET` of VIIs minus `IIZ`s was created; this `SET` was then referenced in the constraints that assign actors to VIIs, rather than all VIIs, as was implemented previously. Additions to such `LIST`s in the future will continue to improve the accuracy of the parser.

Negative particles and particle phrases (e.g., *namôya wîhkâc* above) have been listed and tagged as such so they may be associated with verbs, nouns, or other particles. Currently, negative particles are associated with other syntactic material using three constraints. First, a negative particle is associated with any verb later in a clause. Second, if there is no verb in that context, the negative particle is associated with an immediately following noun or, third, an immediately following particle. Further refinement will be required for relationships between negative particles in other contexts, as well as for the association of the appropriate negative particle before different verb types. While the latter is not required for the current parser, which is used to analyse syntactic patterns in texts produced by fluent speakers, a grammar checker will require that only certain negative particles occur with independent, conjunct, or imperative verbs, for example (Schmirler

& Arppe, 2019a). As the current constraints are applied to the full corpus, the existing patterns can be explored more readily.

`LIST`s were also effectively used to fully implement relationships between adpositions and locative nouns. The previous Plains Cree parser included a broad constraint that allowed a locative noun to be associated with any immediately preceding particle; as this ignored the function of the particle, the constraint wildly overapplied and thus adpositional phrases were not included in the phrase order analysis presented in Schmirler et al. (2018). For the current parser, particles with spatial/adpositional functions were identified using textbooks (Okimâsis, 2004, 2021; Ratt, 2016, 2022) and a dictionary (Wolvengrey, 2001). Those which are known to occur with postpositional function were separated out in their own `LIST`. Thus, two constraints could then be implemented; one to associate locative nouns with preceding preposition and another for those with following postpositions.[32] Once adpositions were accounted for, remaining locative nouns were associated with nearby verbs.

Additional locative elements were listed on the basis of dictionary entries and, along with adpositions, marked as `IPL` (indeclinable locative particle; abbreviation drawn from Wolvengrey, 2011). These `IPL`s could then also be associated with other adpositions, namely *ohci* 'from', such as *êkota ohci* 'from there'; *ohci* would then be associated with a nearby verb, as were other `IPL`s. An impressionistic corpus exploration of `IPL`s such as *êkota*, *ékotê*, *ôta*, etc. (i.e. those glossed as 'here' or 'there') noted that these elements can be predicated with a following demonstrative[33]; thus, *êkota ôma* 'it is there [e.g., that something took place]' or *êkota ôma ohci* 'it is from there [e.g., that someone came]'. These relationships were also modelled; the predicating demonstrative was associated with the `IPL` as its head, then the `IPL` would be associated with *ohci* if present, or a nearby verb, first looking immediately adjacent, then two words way, then three and beyond. This pattern of constraints was developed in response to discrepancies between the development text and the parser for the `@IPL-V` and `@IPT-V`[34] function tags; while the relationships could go in any direction, they nearly always occurred with the closer verb (on the basis of translation),

[32] Note that locative nouns are not required to be found in adposition phrases.
[33] Or particle (e.g., Wolvengrey, 2001).
[34] These are read as "locative particle associated with a verb" and "temporal particle associated with a verb" respectively; the direction of the verb is indicated with < and > as for other function tags.

regardless of whether it preceded or followed the particle. Similarly, a `LIST` of temporal particles (`IPT`, cf. Wolvengrey, 2011) was created on the basis of dictionary content. These generally included particles that referenced time (now, later, early, later), time of day or year (morning, evening, winter), as well as other adverbials such as 'suddenly' or 'again'. These are associated with nearby verbs in the same manner as locative particles. While these constraints have allowed for the parser to cover considerably more of a Plains Cree sentence, the existing `LIST`s of locative and temporal particles are certainly incomplete, and the identification of further functional subcategories will also allow for more precise and accurate analysis.

Another `LIST` was created to specify quotative verbs, to explore the ways in which they pattern differently (or otherwise) to non-quotative verbs. Thus far, the verbs included as quotative verbs are *itwêw* 's/he says' (VAI), *itam* 's/he says to/about something' (VTI), and *itêw* 's/he says to/about someone' (VTA). In addition to receiving `@PRED` tags, these also receive `@Quot` tags, with the transitivity class specified (e.g., `@Quot-AI` for *itwêw*).

Finally, a `LIST` of quantifying particles (`Quant`) was also constructed, similar to those for locative and temporal particles. Quantifiers include numerals, though these are primarily handled with the noun phrase constraints discussed for the previous development (§3.3.1). Currently, the parser makes no attempt to differentiate quantifiers that are used with different word classes; a cursory corpus investigation suggests that many can occur with nouns, verbs, and particles. As such, this is a rather heterogeneous list that is comprised of forms that might be considered degree adverbs as well as quantifiers proper.[35] Quantifiers in general are presently only associated with an immediately following noun or an immediately adjacent verb, with some idiosyncratic constraints for specific wordforms that can occur with specifically plural nouns or pronouns. Thus, *kahkiyaw* 'all' and *âtiht* 'some' are both associated with a following nominal before a verb; *iyikohk* 'to such an extent' will be associated with any verb in its clause before a noun, and *mitoni* 'really' and *mistahi* 'a lot' will both be associated with any following verb before an adjacent noun. In future development, these could be better handled using heuristic constraints.

[35] Oxford (2008) observes a similar pattern for Innu and notes that cross-linguistically "degree adverbs are also analysed as involving quantificational semantics" (p. 137).

A number of issues present themselves when working with quantifiers: some quantifiers, such as *pêyak* 'one' and *kahkiyaw* 'all', may also act as actors and goals, so constraints must be included to allow for these as well. Additionally, floating quantifiers are a documented phenomenon in Plains Cree (e.g., Dahlstrom, 1991; Wolvengrey, 2011), so determining the element a quantifier is associated with is not always possible with the present constraints; even when immediately adjacent to a noun or verb, this may not be the appropriate interpretation. Therefore, both overapplication and underapplication of `@Quant` function tags, like those for other particles, is unavoidable at this time, though frequencies cannot be precisely determined. Investigations of these tags in the corpus can be used in the future to refine the constraints and improve their application, as well as improving the classification of quantifiers and degree adverbs.

This attempt at particle classification is but the barest foray into the complexities that are to be found there. Furthermore, works such as Ogg (1991), Oxford (2008), and Wolvengrey (2011) tease out shades of meaning and functions that are not yet recognised in the Plains Cree parser, and indeed may not be functions that can or need to be modelled automatically. However, this may act as a starting point for a fuller classification for Plains Cree particles along the lines of Oxford (2008) for Innu, as well as allow for corpus investigations of how such particles are used with nouns, verbs, and other particles.

3.3.2.2 Word order considerations

In the development of the previous iteration of the Plains Cree parser, one issue in the identification of participants concerned proximity of nominals to verbs and the relative ordering of constraints. For example, if two verbs occurred in the same CG clause, a noun between them might be associated with the incorrect verb, simply because the constraints were ordered in such a way that nominals were associated with preceding verbs before following verbs. This decision was motivated by previous research, as verb-initial word orders are generally identified as more basic in Plains Cree, regardless of their frequency (e.g., Mühlbauer, 2007). However, observations throughout the modelling process demonstrated once again that language use is more complex than is often described. Following the relative success of incremental constraints for particles, first looking immediately before and after the particle, then two places away, etc., a similar approach was introduced for verbs and arguments. Three main distinctions are used in this new approach:

the type of nominal (noun vs. pronoun), the type of verb (independent vs. *ê*-conjunct vs. other types of conjunct), and the proximity of the verb and nominal. The first set of constraints looks only at independent verbs and starts with nouns immediately adjacent to verbs, first following then preceding the verb; next is pronouns immediately adjacent to verbs; then, these constraints are repeated, but now with one word intervening between the noun and the verb; finally, the set concludes with nouns anywhere in a clause. This full set is then repeated for *ê*-conjunct verbs and then a reduced set is created for all other conjunct verbs, though the distance from the verb is not modelled in stages.

Several assumptions are made in choosing this approach for the argument constraints that result in at least some inaccurate application of function tags. Arguments following verbs are recognised before those preceding verbs, nouns are labelled before pronouns, arguments are assigned to independent verbs (e.g., matrix clauses) before conjunct verbs, and goals are labelled before actors, and the division between conjunct verbs attempts to address some differences in the use of different conjunct types. While these changes did not make great differences within the test corpus, improvements were seen in spot-checks of the full corpus. Further examination of larger portions of the corpus is needed to assess the effectiveness of this approach.

3.4 For future development

3.4.1 Remaining issues

Despite the considerable improvements to the CG-based parser for Plains Cree (see Chapter 4 for more details on coverage), several issues persist. Disambiguation of pronoun-particles, such as *aya*, *ôma*, and *wiya* remains elusive, as no clear syntactic context clues can be identified in many cases. Tests against the development corpus result incorrect disambiguation of *aya* and *ôma*, where disambiguation constraints apply, but remove the correct reading. These issues continue from earlier development (see Table 4.2 and Table 4.5 in Chapter 4). Ambiguous animate nouns, such as *môswa* ‘moose’, which can be either singular or obviative, generally cannot be adequately disambiguated without a demonstrative or verb. Similarly, verbs with further obviative marking with two overt obviative arguments defy correct labelling, as both can be either goals or actors, as in (8). In this sentence, *iskwêwa* ‘woman’ is the obviative goal of ‘to have’, which is inflected for

an obviative person acting on a further obviative person. However, the constraints are still too conservative to adequately assign the "goal" reading, and so an "actor" constraint later in the parser applies.

(8) Further obviative (Douquette, 1987, pp. 40–1)

… iskwêwa kik-âyâwâyit…

iskwêw	-a	kika-	ayâw-	-âyit
NA	3′	IPV	VTA	DIR
woman			have	3′>3″

'to have a wife'

The assignment of particle tags also presents some issues, as mentioned with respect to word order in §3.3.2.2 above. The best combination of incremental constraints for negative, temporal, locative, and quantifying particles remains to be determined. While looking first for following elements and then preceding ones has proven effective, incorrect function assignment still occurs, though to a lesser extent.

3.4.2 Looking forward

Higher-level relationships have again been left for future research, so that questions, matrix and embedded clauses, and coordination, among others, are not widely addressed in the present iteration of the Plains Cree parser. While the labelling of interrogative particles has been undertaken, both in the morphological gold standard and the parser, constraints to identify their relationship to other words have not. However, questions present more regular word-ordering principles in Plains Cree, which can be used in future parser development. The interrogative particle *cî*, which is used to form polar interrogatives, always occurs in the second position (Wolvengrey, 2011, p. 304). The word that occurs before *cî* is in focus, or that which is being questioned, and thus could be associated with the particle. In the examples given in (9), the element before *cî* could be stressed in the English gloss to indicate focus.

(9) Polar interrogatives (Wolvengrey, 2011, pp. 304–5)

a. *nôhtêhkatêw cî?*

nôhtêhkatê-	-w	cî
VAI	3SG	IPC
be.hungry		INTER

'Is s/he hungry?'

b. *otâkosîhk cî kî-takosin?*

otâkosîhk	cî	kî-	takosin	-Ø
IPC	IPC	PV	VAI	(3SG)
yesterday	INTER	PST	arrive	

'Did s/he arrive *yesterday*?'

Though *cî* is used only for polar interrogatives, content interrogatives have some set word orders as well. The question word (often formed in Cree with the element *tân-*) must occur in the initial, focused, position (Wolvengrey, 2011, p. 311). Some examples are given in (10). Note that unlike the polar interrogatives above, all content interrogatives occur with conjunct verbs, though either *ê-* or *kâ-* may be used (Cook, 2014, pp. 235–8; Blain, 1997, p. 69).

(10) Content interrogatives (Wolvengrey 2011, pp. 311–3)

a. *awîna ê-kî-pakamahwat?*

awîna	ê-	kî	pakamahw-	-at
PRON.A.SG	PV	PV	VTA	DIR
who	CNJ	PST	hit	2SG>3SG

'Who did you hit?'

b. *kîkwây ê-kî-pakamahaman?*

kîkwây	ê-	kî-	pakamah-	-aman
PRON.I.SG	PV	PV	VTI	2SG
what	CNJ	PST	hit	

'What did you hit?'

c. *tânispî ê-wî-sipwêhtêyan?*

tânispî	ê-	wî-	sipwêhtê-	-yan
IPC	PV	PV	VAI	2SG
when	CNJ	FUT	leave	

'When are you going to leave?'

d. *tânitahto (masinahikana) ê-kî-atâwêyan?*

tânitahto	(masinahikan	-a)	ê-	kî-	atâwê-	-yan
IPC	NI	PL	PV	PV	VAI	2SG
how.many	(book)		CNJ	PST	buy	

'How many (books) did you buy?'

Core components of questions can be identified, both linguistically (e.g., over 500 *tân*-word tokens[36] occur in the A-W subcorpus; over 200 instances of the interrogative particle *cî*), and textually (over 400 question marks). For both content and polar interrogatives, the parser could also make use of the clause boundary marker <?>, as the corpus makes use of standard English punctuation for the Cree texts. This can also allow for a distinction between interrogatives, like those above, and embedded *tân*-clauses. However, many of the question marks are also part of the annotations, such as indicating an uncertain word, rather than part of the texts themselves—how best to use textual clues remains to be seen.

At all levels of Plains Cree syntax, coordinators can also occur. The most common of these are *êkwa* 'and, then', *mâka* 'but', and *ahpô* 'or'. These can occur at the sentence level; *êkwa* is often used to link two sections of discourse together (e.g., "and then…") (e.g., Ogg, 1991). They can link clauses, nouns, noun phrases, verb phrases, etc. The identification of which type of unit a coordinator is linking may prove to be an interesting challenge for the parser, especially as *êkwa* and *mâka* are also commonly displaced to the second position of a phrase (Wolvengrey, 2011). Coordinating particles are not yet marked as such and so cannot be automatically identified by the parser.

[36] That is, Plains Cree *wh*-words, which overwhelmingly begin with *tân*-.

For the identification of arguments, one can rely primarily on the verb class and inflections to find relevant wordforms in the clause. While the verb class information is sufficient to determine which types of arguments are expected (animate, inanimate, actors, and goals), semantic information for subclasses may be useful for identifying ditransitive verbs, such as *asamêw* 's/he feeds s.t. to s.o.', exemplified in Chapter 5. For instance, many ditransitive verbs are benefactives that take both a recipient (that is, as the goal) plus an indirect object or THEME, which may be either animate or inanimate. As benefactives are often derived with the suffix *-amaw*, these may easily be found in the lexicon coded as such for the reference to argument structure for the parser (Wolvengrey, 2011, p. 43; Wolfart, 1973, p. 61). The VAI *minihkwêw*, 's/he drinks', mentioned in Chapter 2, could also be coded for an expected oblique, something that is drunk—this verb has the connotation of 'drinking alcohol' (not unlike English); without an overt object somewhere in the discourse, alcohol is assumed and so many speakers specify the object as though it were a VTI (Wolvengrey, 2001).[37] As the parsed corpus for Plains Cree is examined manually, various other verbs that seem to occur with specific thematic roles may also be noted and coded appropriately, though this remains for future work.

The current iteration of the parser also does not account for relationships between matrix and embedded clauses. Independent and conjunct clauses offer some distinction here: independent clauses can only be matrix clauses, and therefore cannot occur as relative clauses or clausal complements—these distinctions were used, for example, to incrementally apply the argument constraints, as noted in §3.3.2. Wolfart (1973) groups independent and imperative verbs together in contrast to conjunct and future conditional; the former can form clauses that can stand on their own while the latter require embedding.[38] In the most straightforward of sentences, wherein there is one independent and one conjunct clause, it may be straightforward to determine which is the matrix clause (independent), and which is embedded (conjunct). However, the distinction between independent and conjunct is far more complex than matrix vs. embedded clauses: conjunct clauses

[37] Otherwise, one might consider simply reclassifying this verb as a VTI.
[38] Independent and conjunct clauses have also been contrasted in terms of simple vs. progressive aspect. Okimâsis (2004, 2021), for example, has combined embedding and aspect to exemplify the differences in English glosses: e.g., *nimâton* 'I cry' vs. *ê-mâtoyân* 'as I am crying'.

frequently occur as matrix clauses, where they can instead be considered embedded within a non-linguistic context.

Cook (2014) conducted an in-depth investigation of independent and conjunct clauses, referring to them instead as indexical and anaphoric clauses respectively. Indexical clauses are evaluated against a speech situation and anaphoric clauses against a given context, which may be linguistic (e.g., a matrix clause) or extralinguistic. This distinction makes it possible for conjunct clauses to occur as matrix clauses: they are "embedded" in a real-world context. For example, conjunct verbs are used in the content interrogatives exemplified above: the fact that someone or something was hit, or that books were bought, can be considered known information, and instead *who*, *what*, or *how many* is what is being questioned. Unfortunately for a parser, this distinction does not necessarily simplify the situation or allow for easier recognition of interclausal relationships. However, different conjunct preverbs, which are marked in the morphological model, can be used to determine whether a verb is embedded. While the *ê-* preverb can occur as a matrix clause, most other instances of conjunct verbs must be syntactically embedded. For instance, *kâ-* (discussed in more detail below), must be embedded, as must any conjunct verb containing initial change[39]—but still, this embedding need not be linguistic. Conjunct verbs may also occur with other conjunct prefixes, or no prefix at all, and these must also be embedded. Similarly, future conditional verbs are always in embedded clauses (Cook, 2014, pp. 26–8). As such, matrix vs. embedded clause distinctions can be reliably determined from linguistic context only when both an independent and conjunct verb occur, though the preverb *kâ-* may offer a clue for relative clauses. Further means of recognising embedded clauses would require semantic classification of subordinating particles, though still these would not occur in every instance of an embedded clause. Any subordinating particle can only be used with conjunct verbs; these include *osâm* 'reason, because', *iyikohk* 'as far as', *kiyâm* 'although', *pâmwayês* 'before', *mayaw* 'as soon as', *âta* 'although', and *ayis* 'for, because', among others (Cook, 2014, pp. 60–1).

Semantic features such as those described above will allow for considerable improvement to a CG-based parser for Plains Cree. Some of these features may be best coded within the morphological

[39] Initial change is a change to the first vowel of a conjunct or future conditional verb (e.g., *i* > *ê*, *â* > *iya*) (Wolfart, 1973, pp. 82–3). For more on this rather archaic feature of Plains Cree, see Schmirler (in press).

analyser, such as certain subtypes of pronouns or particles (e.g., interrogatives), while others may be best indicated in a `SET` within the parser (e.g., benefactives, prepositional particles, etc.). As more features become available for reference in the parser, its capabilities will improve to the point that full syntactic analyses may be automatically produced for each sentence for inclusion within a corpus. These syntactic analyses can also be used for further large-scale investigations of Plains Cree syntax.

3.5 Conclusion

This chapter has presented the Constraint Grammar formalism, outlined its application to Plains Cree syntax, and discussed issues that remain for future development. In adding new constraints to the Plains Cree parser, morphological feature tags, lexicosemantic feature tags, and word order were taken into account. These include possession, zero-place predicates, and far more detailed actor and goal assignment constraints taking into account both distance between the verb and argument. The addition of negative, locative, temporal, and quantifying particles, though far from a complete classification of the large particle class in Plains Cree, allows the parser to label considerably more forms in a sentence, bringing us closer to a full syntactic analysis of a Plains Cree corpus.

Alongside adding new function tags and furthering disambiguation, the current Plains Cree parser improves the modelling of those relationships included in the earlier version (Schmirler et al., 2018). Noun phrases are more thoroughly modelled, which has also improved the assignment of actor and goal functions. Constraints for adpositions and obliques have been refined, reducing, though not entirely resolving, their overapplication in the development corpus. While these constraints are motivated by and tested against a development corpus (see Chapter 4 for coverage statistics), their efficacy in the full corpus has yet to be examined in great detail.

For every relationship modelled in the current parser, yet another remains. Interrogative particles, though now labelled, are not yet associated with other words, and thus questions are not identified. Relationships between clauses, such as embedding or coordination, are not implemented at all, though the morphological identification of independent, conjunct clauses, and future conditional clauses may be a step in the right direction.

Chapter 4

An expanded morphosyntactically tagged corpus of Plains Cree

4.1 Introduction

In this chapter, I describe the creation of a morphosyntactically tagged corpus of ~152,000 words of Plains Cree. This corpus includes Plains Cree texts collected at various times throughout the 20th century, which have been digitised for use in a digital, online corpus of Plains Cree. The texts have been divided into two subcorpora: the Bloomfield subcorpus and the Ahenakew-Wolfart subcorpus. These texts have been morphologically analysed by a Finite State Transducer (FST)-based morphological model (e.g., Snoek et al., 2014; Harrigan et al., 2017) for morphological features, such as those discussed in Chapter 2, which are then manually verified while retaining morphological ambiguity without reference to syntactic context. These verified morphological tags are then the input to the CG-based parser for Plains Cree described in Chapter 3, which provides disambiguation and syntactic function tags for the full corpus. The resulting morphosyntactically tagged corpus can then be used for various corpus linguistic studies, such as those that follow in Chapters 5 and 6, and for a tagged online searchable corpus available for researchers at https://korp.altlab.app/.

The texts included in the Plains Cree corpus are described in §4.2, along with details about the speakers and their lives. Section 4.3 describes the process of generating and validating the morphological tags in the two subcorpora and the testing and effectiveness of the parser. In §4.4, I give an overview of the morphosyntactic tags in the full corpus and describe differences between the two subcorpora. These differences are further discussed in §4.5 and the chapter concludes in §4.6.

4.2 The texts

The texts underlying the corpus are drawn from two sets of texts: those collected and edited by Leonard Bloomfield in the 1920s (BT "Bloomfield texts"), and those collected and edited by Freda Ahenakew and H. C. Wolfart in the last two decades of the 20th century (A-W "Ahenakew-Wolfart"). The volumes and their word counts are given in Table 4.1. When counting words, 'overall tokens/types' refers to the overall number of tokens/types in the corpus, regardless of which language they occur in or if they are numerals, punctuation, etc. 'Cree tokens/types' refers to the number of Plains Cree tokens/types remaining when all non-Cree words, numerals, metadata, and punctuation tokens are removed. Details about the speakers and the collection of the texts are given in §4.2.1 and §4.2.2; details on the content of the texts can be found in Chapter 6.

4.2.1 Bloomfield subcorpus

Leonard Bloomfield collected the texts of the BT subcorpus in the summer of 1925 at the Sweetgrass First Nation near Battleford, Saskatchewan. Both volumes include stories from the same group of monolingual Plains Cree speakers: Coming-Day (*kâ-kîsikâw-pîhtokêw*, 34 texts), Adam Sakewew (*sâkêwêw*, 13 texts), Mrs. Adam Sakewew (*kiyâkêskamikapiw*, 1 text), Maggie Achenam (*kâ-wîhkaskosahk*, 15 texts), Louis Moosomin (*nâh-nâmiskwêkâpaw*, 9 texts), Simon Mimikwas (*mimikwâs*, 6 texts), Chihtchikwayow (*cîhcîkwâyôw*, 3 texts), and Mrs. Coming-Day (*nakwêsis*, 1 text). Bloomfield's knowledge of Cree and his translations of the collected stories were facilitated by a number of younger bilingual Plains Cree-English speakers; he names Harry Watney, Norman Standinghorn (*mâyiskinîkiw*), and Baptiste Pooyak (Bloomfield, 1930, pp. 1–6; 1934, pp. iii–iv). Bloomfield described the Plains Cree spoken in Sweetgrass as representative of "an archaic type"; the volumes contain sacred stories and non-sacred stories (Bloomfield, 1930, p. 1).

Table 4.1: Volumes of the Plains Cree corpus

Subcorpus	Volume title & citation	All tokens (types)	Cree tokens (types)
BT	*Sacred Stories of the Sweetgrass Cree* (Bloomfield, 1930)	60,333 (10,015)	42,120 (10,000)
	Plains Cree Texts (Bloomfield, 1934)	42,550 (7,827)	30,354 (7,813)
A-W	*âh-âyîtaw isi ê-kî-kiskêyihtahkik maskihkiy / They Knew Both Sides of Medicine: Cree Tales of Curing and Cursing Told by Alice Ahenakew* (Ahenakew, 2000)	16,082 (3,458)	8,277 (2,689)
	kôhkominawak otâcimowiniwâwa / Our Grandmothers' Lives as Told in Their Own Words (Bear et al., 1998)	42,164 (6,968)	19,826 (5,583)
	ana kâ-pimwêwêhahk okakêskihkêmowina / The Counselling Speeches of Jim Kâ-Nîpitêhtêw (Kâ-Nîpitêhtêw, 1998)	10,520 (2,453)	6,629 (2,319)
	piko kîkway ê-nakacihtât: kêkêk otâcimowina ê-nêhiyawastêki / *She Can Do Anything: The Reminiscences of Cecilia Masuskapoe, Published in Cree* (Masuskapoe, 2010)	41,434 (7,495)	26,507 (6,975)
	kwayask ê-kî-pê-kiskinowâpahtihicik / Their Example Showed Me the Way: A Cree Woman's Life Shaped by Two Cultures (Minde, 1997)	18,061 (3,949)	11,733 (3,536)
	wâskahikaniwiyiniw-âcimowina / Stories of the House People, Told by Peter Vandall and Joe Douquette (Vandall & Douquette, 1987)	5,052 (1,222)	3,496 (1,202)
	kinêhiyâwiwininaw nêhiyawêwin / The Cree Language is Our Identity: The La Ronge Lectures of Sarah Whitecalf (Whitecalf, 1993)	5,333 (1,198)	3,542 (1,127)

4.2.2 Ahenakew-Wolfart subcorpus

The texts of the A-W subcorpus were collected between 1982 and 1996 by Freda Ahenakew and H. C. Wolfart, who then facilitated the transcription, editing, and, in most cases, translation of the texts. Both served as editors for the volumes, with the exception of Vandall & Douquette (1987). Freda Ahenakew, a Plains Cree woman and fluent speaker from *yêkawiskâwikamâhk* (Sandy Lake), Saskatchewan, was present for the recordings, and was often related to, or otherwise well acquainted with, the speakers.

The texts in Ahenakew (2000) were recorded by Freda Ahenakew at *yêkawiskâwikamâhk* (Sandy Lake), Saskatchewan, beginning in 1989. Alice Ahenakew (née Bush) was born in 1912 and raised in Sturgeon Lake, and was 77 at the time of recording. The stories are primarily personal narratives about her life, marriage, and other personal experiences, with a focus on the coexistence of Cree and Christian traditions (Ahenakew, 2000, pp. 1–33). Alice Ahenakew's husband, Andrew, was a cousin of Freda Ahenakew's grandfather.

The texts in Bear et al. (1998) were recorded by Freda Ahenakew over a number of years in different locations and include stories from several Cree women: Glecia Bear, Irene Calliou, Janet Feitz, Minnie Fraser, Alpha Lafond, Rosa Longneck, and Mary Wells. One of these, Janet Feitz (of La Ronge), was a speaker of Woods Cree and so her stories are not included in the Plains Cree Ahenakew-Wolfart subcorpus (Bear et al., 1998, p. 356). In the introduction to the texts, Wolfart notes that the stories are of various types, including personal stories, funny stories, and a few brief occurrences of counselling texts (p. 19). Glecia Bear's stories were recorded on Flying Dust First Nation, Meadow Lake, Saskatchewan, in November 1988. Wolfart notes that her Plains Cree seemed to have some influence of Woods Cree (Bear et al., 1998, p. 380), and would likely be considered northern Plains Cree (A. Wolvengrey, personal communication). Bear (née Laliberté), Freda Ahenakew's aunt, was born in *yêkawiskâwikamâhk* and was 75 at the time of recording (pp. 31–2). Minnie Fraser's stories were recorded in October 1986 in Prince Albert, Saskatchewan. Fraser (née Ahenakew, the paternal aunt of Freda Ahenakew's father) was born in 1896 at *yêkawiskâwikamâhk* and was 90 years at the time of recording (pp. 369–70). Irene Calliou's stories were recorded in Grouard, Alberta, in July 1988. Calliou was born in *pakicahwânisihk*, a Métis settlement in Alberta (p. 380). Mary Wells' stories were recorded in Grouard, Alberta, in July

1988. Wells was from Elizabeth Settlement, a Métis settlement north of *pakicahwânisihk*. While a fluent speaker of Plains Cree, Wells has referred to herself as *ocîpwayânîwiskwêw*, 'a Chipewyan [Dene] woman', born in a region with Cree-Dene bilingualism and intermarriage (p. 385). Calliou and Wells are friends of Freda Ahenakew from a Cree grammar and writing course (p. 32). A dialogue between Alpha Lafond (née Venne) and Rosa Longneck (née Lafond) was recorded in April 1988 on *opitihkwahâkêw*'s reserve at *maskêko-sâkahikanihk* (Muskeg Lake First Nation), Saskatchewan. Both speakers were born at *maskêko-sâkahikanihk*. Freda Ahenakew, who joins in with the dialogue during the recording, was born at *yêkawiskâwikamâhk*, but spent much of her life at *maskêko-sâkahikanihk* (p. 396) and was well acquainted with both women.

The texts in Kâ-Nîpitêhtêw (1998) were recorded from public speeches given between 1987 and 1989 and are comprised of counselling texts. Jim Kâ-Nîpitêhtêw (*kâ-pimwêwêhahk*), a monolingual Plains Cree speaker from *wîhcêkaskosîwi-sâkahikanihk* (Onion Lake First Nation), Saskatchewan, was in his early 80s at the time of the speeches. Unlike the texts in most of the other volumes in the A-W subcorpus, where the recordings were made in more private settings, these speeches were given publicly in several settings, including at the Saskatchewan Indian Cultural College (Saskatoon, Saskatchewan), the Thunderchild (*piyêsis-awâsis*) First Nation, and the Saskatchewan Indian Languages Institute (Kâ-Nîpitêhtêw, 1998, p. vii).

The texts in Masuskapoe (2010) were recorded between 1988 and 1996 in various locations. Cecilia Masuskapoe (*kêkêk*) was born Cecilia Rabbitskin in 1918 north of *mistatihkamêko-sâkahikanihk* (Whitefish Lake, or Big River First Nation), Saskatchewan, and moved to *yêkawiskâwikamâhk* upon her marriage. The texts include "reminiscences and historical accounts". Her Plains Cree demonstrates some variation with dialect features of Woods Cree (Masuskapoe, 2010, pp. v–vi) or northern Plains Cree.

The texts in Minde (1997) were recorded June 1988. Emma Minde (née Memnook) was born in 1907 at *onihcikiskwapiwinihk* (Saddle Lake), Alberta, and moved to *maskwacîsihk* (Maskwacîs, formerly Hobbema, Alberta) upon her marriage in 1927. The texts are Minde's reminiscences of her time in *maskwacîsihk*, interspersed with elements of counselling, wherein she shares her personal experiences as well as the lessons and values, both Cree and Catholic, that she gained throughout her life, though especially after her marriage (Minde, 1997, pp. ix–xv).

The texts in Vandall & Douquette (1987) were recorded in February 1982 at the Saskatchewan Indian Cultural College in Saskatoon, Saskatchewan. The texts were recorded in the span of an hour, with the two speakers alternating, and the stories to some extent flow from one to the next. Both speakers are, as the title of the volume suggests, *wâskahikaniwiyiniwak* (House People), which include the inhabitants of three nations: *atâhk-akohp*, *mistawâsis*, and *opitihkwahâkêw*, west of Prince Alberta, Saskatchewan. The texts include counselling stories, funny stories, and personal stories. The Plains Cree of Vandall and Douquette is described as "quite similar to that which Leonard Bloomfield heard at Sweet Grass Reserve near Battleford in 1925" (Vandall & Douquette, 1987, pp. x–xiii).

The texts in Whitecalf (1993) were recorded in January 1990 as a series of brief lectures given in response to student questions in a Cree class taught by Freda Ahenakew. Sarah Whitecalf was born in 1919 at *môsômininâhk* (Moosomin First Nation) near *kinosêwi-sâkahikanihk* (Jackfish Lake), Saskatchewan. Whitecalf spent much of her childhood between *môsômininâhk* and Sweetgrass, where Bloomfield collected his texts, and knew many of the speakers recorded by Bloomfield (Whitecalf & Whitecalf, 2021, p. xii). Her dialect is described as that of the *sîpîwiyiniwak* 'the River people'. The texts are lectures, illustrated with personal experiences (Whitecalf, 1993, pp. vii–xiii).

4.3 Morphosyntactic tags

4.3.1 Morphological tags: Morphological gold standards

For each subcorpus, a "morphological gold standard" (MGS) has been created to ensure that the morphological feature tags that feed into the parser and ultimately occur in the final morphosyntactically tagged corpus are as accurate as possible, through manual validation and correction of tags applied by the FST-based morphological model. The process for validating the features in the Ahenakew-Wolfart MGS began in 2016 with the application of the then-current morphological model for Plains Cree (e.g., Snoek et al., 2014; Harrigan et al., 2017) in "descriptive mode", attempting to cover as much of the subcorpus as possible with allowance for spelling variation. This model offered analyses for 10,036 of the 18,646 types (92,613 of 125,366 tokens) in this version of the A-W subcorpus (which at the time excluded Ahenakew, 2000 for testing

purposes). All tokens were then examined by the author and a colleague (A. Harrigan), and analyses were corrected and added where possible. The Ahenakew-Wolfart MGS has been under review since 2016, with the most recent pass of the original 18,646 types undertaken by the author in 2020; the unique types from Ahenakew (2000) were added in 2021, following the process used for the Bloomfield MGS (described below). This brings the total number of types in the A-W MGS (including punctuation, numerals, non-Cree words) to 20,502 (142,192 tokens), of which only 33 types (37 tokens) still defy analysis, both automatic and manual. Of the total types in the subcorpus, 17,976 types (81,565 tokens) are identified as Plains Cree, while the remainder include non-Cree words and punctuation. The A-W MGS is available at https://github.com/giellalt/lang-crk/blob/main/test/data/ahenakew_wolfart_MGS_tab-sep-anls_freq-sorted.txt.

The validation to create the Bloomfield MGS was undertaken in 2020, following a different plan of action from the original Ahenakew-Wolfart MGS. First, the updated morphological model was applied to the subcorpus in "normative" mode, which does not allow for any spelling variation. Thus, any forms successfully analysed by the model were deemed to have accurate analyses. This step analysed 11,506 of the 15,287 types (94,824 of 104,245 tokens). The remaining 3,620 types (9,421 tokens) were then analysed with the descriptive mode of the updated morphological model, which identified analyses for a further 1,210 types (4,098 tokens). All 3,620 types that were not analysed originally were then manually validated, correcting the descriptive mode analyses and adding missing analyses where needed. This was facilitated immensely by the digitised contents of the Bloomfield texts supplied to the Alberta Language Technology Lab from Kevin Russell and H.C. Wolfart, which they had transliterated to the Standard Roman Orthography (that used herein; Okimâsis & Wolvengrey, 2008) for Plains Cree and given morphological analyses. These were used to inform the validation, alongside the parallel sentence translation, to determine how each word's analysis transferred to the morphological model tags.[40] As a result of these detailed analyses and the morphological model, every wordform in the Bloomfield texts receives a full

[40] While it would have been perfectly possible to create a script that transferred the provided analyses to those in the style of the morphological model, there is not a one-to-one correspondence between the systems. Thus, the effort required to determine how to transfer one system to the other, and the fact that validation would be required to some degree regardless, the morphological tags were applied using the described method. In the end, the validation of the 3,620 forms was a matter of a few weeks' work and allowed for some errors in the original analyses to be corrected (though undoubtedly new errors were also introduced).

analysis in the MGS. Of these, 15,267 types (72,475 tokens) are identified as Plains Cree. The Bloomfield MGS is available at https://github.com/giellalt/lang-crk/blob/main/test/data/bloomfield_MGS_tab-sep-anls_freq-sorted.txt.

Between the two subcorpora, the full corpus contains 34,115 types (241,922 tokens), of which 31,616 types (152,405 tokens) are Plains Cree words. The MGS tags are applied using the script found at https://github.com/giellalt/lang-crk/blob/main/tools/shellscripts/AhenakewWolfart_add_MGS.sh.

4.3.2 Syntactic tags: CG test corpus & parser coverage

Following the morphological tags, syntactic tags can then be added; examples of the code used for this are found in the script at https://github.com/giellalt/lang-crk/blob/main/inc/crk-postprocess-cg.sh. However, it is important to note that the MGS analyses retain ambiguity: where a wordform has multiple morphological analyses, no attempt is made to disambiguate them. This is instead undertaken in the parser's disambiguation module, before syntactic tags are applied by the function assignment module. Additionally, while the morphological features of each wordform in the corpus have been validated, the application of the CG-based parser was tested using a small test corpus. One volume of the A-W subcorpus, Vandall & Douquette (1987), was manually annotated for disambiguation and syntactic functions, though the annotations are currently limited to only those tags which were included in the parser at this time. Disambiguation and function assignment were facilitated by the translations of Vandall & Douquette (1987), as the author of the present work does not speak Plains Cree. This development corpus, or SGS (syntactic gold standard), contains 4,734 tokens (1,221 types) of which 3,178 (1,201) are Plains Cree. Of the non-Cree tokens, 1,318 are clause boundary punctuation (7 types). For future testing, a separate portion of the corpus, or perhaps a new text altogether, could be similarly coded for disambiguation and syntactic functions to understand how well the parser applies to an unfamiliar Plains Cree text.

4.3.2.1 2017 parser coverage & disambiguation effects on A-W

The 2017 version of the Plains Cree parser, described in Schmirler et al. (2018) and Appendix A, gave promising results with minimal parser development. In this section, I present the results from this earlier parser, including first the efficacy and accuracy of disambiguation followed by the

recall and precision of syntactic function tags. Disambiguation coverage in the development corpus is presented in Table 4.2, which includes the number of wordforms, with their total readings before and after disambiguation, and a <+> to indicate that the reading manually identified as correct remained, or <-> if the correct reading was removed. These results also demonstrate that there is little ambiguity inherent in Plains Cree, as the vast majority of the words have only one analysis. Those words with multiple readings are therefore of more interest when exploring the efficacy of the disambiguation constraints, as these constraints cannot apply to wordforms with only one analysis. Overall, in 99.4% of wordforms the correct analysis is not removed, and in 93.0% of wordforms, the correct analysis is the only analysis after disambiguation. However, if the cases where only one analysis was given before disambiguation are excluded, in 96.6% of the remaining wordforms the correct analysis is not removed, but in only 56.7% of wordforms is the correct analysis the only analysis after disambiguation.

Table 4.2: Disambiguation results (Schmirler et al., 2018)

Number of wordforms	Analyses before disambiguation	Analyses after disambiguation	Accuracy	% of wordforms
2,704	1	1	+	83.8%
276	2	1	+	8.6%
199	2	2	+	6.2%
18	4	1	+	0.6%
13	2	1	-	0.4%
6	3	2	+	0.2%
5	4	1	-	0.2%
3	3	1	+	0.09%
2	4	2	+	0.06%
2	3	3	+	0.06%

While details for individual syntactic function tags are not available for the 2017 parser, the development corpus also allows for presentation of recall and precision statistics for syntactic function assignment. Schmirler et al. (2018), with actor/goal tags considered together, presents

92% for both recall and precision.[41] The effects of the 2017 parser on disambiguation for the A-W subcorpus, not just the development corpus, were not fully presented in Schmirler et al. (2018); though the underlying description of the 2017 parser and results are reproduced in Appendix A for reference. Disambiguation for types and tokens for this earlier version of the parser is presented in Table 4.3 and Table 4.4 respectively, though the accuracy cannot be determined as the subcorpus has not been manually annotated. The majority of forms with multiple analyses are demonstrative pronouns/particles and the pronoun/particle *aya*, which each have several functions.

Table 4.3: Type Ambiguity

	Before disambiguation		After disambiguation	
# of analyses	Number of types	% Types	Number of types	% Types
0[42]	2,255	11.7%	2,255	11.7%
1	14,563	75.8%	16,183	84.0%
2	2,205	11.5%	794	4.1%
3	109	0.6%	31	0.2%
4	88	0.46%	7	0.04%
5	3	0.02%	0	0.00%
6	1	0.005%	0	0.00%

Table 4.4: Token Ambiguity

	Before disambiguation		After disambiguation	
# of analyses	Number of tokens	% Tokens	Number of tokens	% Tokens
0	3,293	3.9%	3,293	3.9%
1	65,046	77.2%	75,134	89.2%
2	12,933	15.3%	5,739	6.8%
3	1,014	1.2%	103	0.1%
4	1,980	2.3%	7	0.008%
5	9	0.01%	0	0.00%
6	1	0.001%	0	0.00%

[41] *Recall*: the proportion of the manually identified actors and goals that the parser correctly identified (the number of forms manually identified and correctly tagged by the parser, divided by the total manually identified). *Precision*: the proportion of automatically identified actors and goals that were correct (the number of forms automatically and correctly assigned by the parser, divided by the total number of actor/goal tags assigned by the parser).

[42] Zero represents forms with no analysis, which include unrecognised Cree forms and other forms that are not yet explicitly tagged appropriately as non-Cree, such as English words or fragments.

4.3.2.2 Updated parser coverage

The first stage of the CG-based parser attempts to disambiguate the morphologically ambiguous forms offered by the MGS. To examine both the degree of ambiguity in the corpus and the efficacy of the disambiguation constraints, we can explore the number of tokens in the development corpus and see how the disambiguation constraints reduce these readings, and whether or not they apply correctly when compared to manual annotation. In Table 4.5, these results for the updated parser are presented as for the 2017 parser in Table 4.2.[43]

Table 4.5: Disambiguation update

Number of wordforms	Analyses before disambiguation	Analyses after disambiguation	Accuracy	% of wordforms
2663	1	1	+	83.53%
428	2	1	+	13.43%
45	2	2	+	1.41%
16	2	1	-	0.50%
12	4	1	+	0.38%
10	4	4	+	0.31%
8	3	1	+	0.25%
2	4	3	+	0.06%
2	3	3	+	0.06%
1	4	1	-	0.03%
1	3	2	+	0.03%

As Table 4.5 shows, the majority of words in the tagged development corpus drawn from Vandall & Douquette (1987) received only one reading from the morphological gold standard—though the number is now slightly lower than in Table 4.2, due to the inclusion of particle phrase constraints, which remove a number of wordforms by combining them with the preceding form.[44] Many of those that remain with more than one reading are those that remained in Schmirler et al. (2018), which could not be disambiguated with only context-based constraints and will require heuristic constraints. Overall, in 99.1% of wordforms the correct analysis is not removed, and in 97.6% of

[43] For these counts, punctuation, Roman numerals, and Arabic numerals were removed.
[44] A total of 3,568 particle phrases (205 types) are created using the constraints.

wordforms, the correct analysis is the only analysis after disambiguation. However, if the cases where only one analysis was given before disambiguation are excluded, in 96.8% of the remaining wordforms the correct analysis is not removed, but in 85.3% of wordforms is the correct analysis the only analysis after disambiguation. The first three values are comparable to those for the 2017 parser, though the last value is considerably higher now, indicating that disambiguation has indeed been improved.

Recall and precision for actor/goal tags has improved and new recall and precision numbers are given for functions not included in the 2017 version of either the parser or the development corpus. As CG development continued, so too did the inclusion of hand-coded syntactic functions in the development corpus, such as the particle function tags. With the exceptions of instruments and obliques, which have thus far undergone only cursory development, and adpositions, these numbers remain above 90% for all other functions (Table 4.6).

Table 4.6: Recall & precision update

Function	Recall (%)	Precision (%)
All	97.51%	97.46%
@ACTOR/GOAL	96.24%	95.27%
@N	98.85%	99.23%
@P	66.67%	100.00%
@Neg	96.00%	96.00%
@IPT -V	100.00%	98.88%
@IPL -V	94.57%	92.55%
@Quant	95.97%	91.67%
@Loc-V/IPL	100.00%	96.88%
@OBL	67.27%	87.80%
@INS	66.67%	66.67%
@PRED	99.64%	99.37%

4.3.2.3 Effects of disambiguation on full corpus

When the updated CG disambiguator is applied to the full Plains Cree corpus, the following results for types (Table 4.7) and tokens (Table 4.8) are achieved. Again, these cannot be checked for accuracy as only a small portion of the corpus is manually tagged, but instead offer a picture of

how ambiguity is reduced overall. When the updated CG disambiguator is applied to the A-W subcorpus, the following results for types (Table 4.9) and tokens (Table 4.10) are achieved. These demonstrate some small improvements from the last iteration of the parser (Table 4.3 and Table 4.4): after disambiguation, 97.7% of types (95.9% of tokens) have one reading, whereas previously this was 84.0% of types (89.2% of tokens). These results also demonstrate the improvements in the A-W morphological gold standard, as more forms with analyses are available—instead for 2,255 types (3,293 tokens) without an analysis, there are now only 7 types (327 tokens).

Table 4.7: Type Ambiguity (full corpus)

	Before disambiguation		After disambiguation	
# of analyses	Number of types	% Types	Number of types	% Types
0	7	0.02%	7	0.02%
1	28,186	87.00%	31,632	97.35%
2	3,711	11.45%	770	2.37%
3	332	1.02%	76	0.23%
4	148	0.46%	5	0.02%
5	10	0.03%	1	0.003%
6	5	0.02%	1	0.003%

Table 4.8: Token Ambiguity (full corpus)

	Before disambiguation		After disambiguation	
# of analyses	Number of tokens	% Tokens	Number of tokens	% Tokens
0	327	0.21%	327	0.21%
1	130,008	82.78%	148,946	96.86%
2	19,930	12.69%	2,855	1.86%
3	3,535	2.25%	433	0.28%
4	3,208	2.04%	1,211	0.79%
5	40	0.03%	7	0.005%
6	8	0.01%	1	0.001%

Table 4.9: Type Ambiguity (A-W)

	Before disambiguation		After disambiguation	
# of analyses	Number of types	% Types	Number of types	% Types
0	7	0.04%	7	0.04%
1	16,459	86.85%	18,711	97.69%
2	2,268	11.97%	414	2.16%
3	125	0.66%	16	0.08%
4	86	0.45%	4	0.02%
5	5	0.03%	1	0.005%
6	2	0.01%	0	0.00%

Table 4.10: Token Ambiguity (A-W)

	Before disambiguation		After disambiguation	
# of analyses	Number of tokens	% Tokens	Number of tokens	% Tokens
0	327	0.39%	327	0.40%
1	67,429	80.74%	78,187	95.90%
2	12,766	15.29%	1,480	1.82%
3	1,019	1.22%	318	0.39%
4	1,946	2.33%	1,207	1.48%
5	29	0.03%	7	0.01%
6	2	0.002%	0	0.000%

When the updated CG disambiguator is applied to the BT subcorpus, the following results for types (Table 4.11) and tokens (Table 4.12) are achieved. As little to no parser development occurred based on patterns found in the BT subcorpus, this can be taken as a decent approximation of a test corpus, though without manual coding against which the results can be checked. Thus, the parser can disambiguate effectively when confronted with entirely unfamiliar text, even if accuracy cannot be determined.

Table 4.11: Type Ambiguity (BT)

	Before disambiguation		After disambiguation	
# of analyses	Number of types	% Types	Number of types	% Types
0	0	0.00%	0	0.00%
1	13,258	88.13%	14,763	97.11%
2	1,506	10.01%	376	2.47%
3	209	1.39%	61	0.40%
4	62	0.41%	1	0.01%
5	5	0.03%	0	0.00%
6	3	0.02%	1	0.01%

Table 4.12: Token Ambiguity (BT)

	Before disambiguation		After disambiguation	
# of analyses	Number of tokens	% Tokens	Number of tokens	% Tokens
0	0	0.00%	0	0.00%
1	62,578	85.10%	70,759	97.93%
2	7,164	9.74%	1,375	1.90%
3	2,516	3.42%	115	0.16%
4	1,262	1.72%	4	0.01%
5	11	0.01%	0	0.00%
6	6	0.008%	1	0.001%

4.4 Corpus overview: morphosyntactic tags

In this section, I report on the numbers of various morphosyntactic tags in the full corpus. For more information, including more detailed tag information and numbers for each of the subcorpora, see Appendix B; for all tables here, they are repeated in at least some form in Appendix B for easier comparison to the subcorpora.

4.4.1 Morphological features

I begin with the morphological tags supplied by the morphological model and morphological gold standard.[45] First in Table 4.13, I give the overall types and tokens in the full corpus, alongside the Plains Cree tokens. Of the non-Cree tokens, 62,038 are clause boundary punctuation (8 types), which are referenced in the parser to disambiguate forms and assign functions. Type-token ratios (TTR) are also presented.

Table 4.13: Full corpus: overall tokens and types

Cree tokens	Cree types	All tokens	All types	TTR-Cree	TTR
152,405	31,616	241,922	34,115	0.21	0.14

4.4.1.1 Verbal features

Table 4.14 gives the counts for verbs and their subclasses, and Table 4.15 gives the same for quotative verbs. The totals for each verb class in the former table also include the quotative counts. VAIs are the most common transitivity class, followed by VTAs, though this difference is negligible within the quotative verbs. Quotative VTIs are quite rare, though the second most common overall class, with VIIs being the least common. For the VAI and VTA quotative verbs, the type-token ratios also demonstrate how frequently the same forms are used again and again, whereas verbs overall have much more variation in the types. Looking between the two subcorpora, for which tables are given in Appendix B (Table B.4 through Table B.9), BT has overall more verbs, more quotative verbs, and more VAIs and VTAs. However, A-W has more quotative VAIs and BT more quotative VTAs. The type-token ratios also differ between the two corpora, with higher ratios in the A-W subcorpus than the BT subcorpus. Here and throughout, one might notice that the numbers often total more than 100%; this is due to some overlap in classes resulting from ambiguity that is not yet removed by the parser.

[45] Many of the features presented herein are exemplified in Chapter 2.

Table 4.14: Full corpus: verbs and subclasses

	Cree tokens	Cree types	% Tokens	% Types	TTR		
Verbs	50,632	25,739	33.22%	81.41%	0.51		
					% Verb tokens	% Verb types	TTR
II	4,043	1,918	2.65%	6.07%	7.99%	7.45%	0.47
AI[46]	23,117	11,290	15.17%	35.71%	45.66%	43.86%	0.49
TI	8,318	4,192	5.46%	13.26%	16.43%	16.29%	0.50
TA	15,436	8,476	10.13%	26.81%	30.49%	32.93%	0.55

Table 4.15: Full corpus: quotative verbs and subclasses

	Cree tokens	Cree types	% Tokens	% Types	TTR		
Quot	5,574	361	11.01%	1.40%	0.06		
					% Quot tokens	% Quot types	TTR
AI	2,790	136	5.51%	0.53%	50.05%	37.67%	0.05
TI	33	13	0.07%	0.05%	0.59%	3.60%	0.39
TA	2,751	212	5.43%	0.82%	49.35%	58.73%	0.08

Table 4.16 gives VTA interaction features for the full corpus. Nonlocal verbs (non-SAP interactions) are the most common then mixed (SAPs and non-SAPs), then local (SAPs), which aligns with the overall more frequent occurrence of third persons compared to first and second. Direct verbs (more topical acting on less topical) are considerably more common than inverse verbs. Given the differences in person features (see the end of this subsection and Appendix B), it is unsurprising that A-W contains relatively more mixed verbs and Bloomfield contains more nonlocal, and, perhaps since BT contains overall more VTAs, also more local verbs (Table B.12 and Table B.14). However, though direct verbs are more common in both subcorpora, there is a greater proportion of inverse verbs in the BT subcorpus. Appendix B also contains tables that break down each of the mixed, local, and nonlocal verbs by direct and inverse (Table B.11, Table B.13,

[46] For the purposes of syntactic modelling, the VAI vs. VTI labels are decided by syntax (i.e., is it transitive or not) rather than morphology (i.e., does it have VAI-like morphology or not). Some ambiguity still occurs, such as for *ayâw*, which can be either VAI "s/he is there" or VTI "s/he has [it]".

and Table B.15); in general, mixed and nonlocal verbs occur more often with direct than inverse morphology, with nonlocal verbs containing an even greater proportion of direct verbs, while local verbs occur with slightly more inverse. These patterns highlight that more topical participants (proximate in nonlocal and SAPs in mixed) are more often actors than goals, and that first persons are often actors in local verbs, resulting in the prevalence of inverse verbs. The tables in the appendix also highlight the ambiguity particularly in mixed and local VTA morphology, and that this ambiguity occurs far more in the BT subcorpus than in the A-W subcorpus—this certainly suggests where future parser development may be focused.

Table 4.16: Full corpus: VTA features

Feature	Cree tokens	Cree types	% Tokens	% Types	% V tokens	% V types	TTR
Mixed	5,147	3,456	3.38%	10.93%	10.17%	13.43%	0.67
Local	1,185	775	0.78%	2.45%	2.34%	3.01%	0.65
Nonlocal	7,384	2,967	4.84%	9.38%	14.58%	11.53%	0.40
Direct	11,129	5,818	7.30%	18.40%	21.98%	22.60%	0.52
Inverse	4,378	2,699	2.87%	8.54%	8.65%	10.49%	0.62

Table 4.17 gives the counts for verbal orders in the full corpus. Verbs are considered all together here, though the tables in Appendix B also include counts for quotative verbs. Conjunct verbs are the most common order, followed by independent, and conjunct verbs show more variation with a higher type-token ratio. Future conditional and imperative verbs are considerably less frequent. Per Table B.16 through Table B.18 in Appendix B, the BT subcorpus contains generally more independent, future conditional, and imperative verbs, while the A-W subcorpus contains more conjunct verbs. The tables in the Appendix B also show that among quotative verbs, independent forms are overall more frequent.

Table 4.17: Full corpus: verbal orders

Order	Cree tokens	Cree types	% Tokens	% Types	% V tokens	% V types	TTR
IND	21,191	8,387	13.90%	26.53%	41.85%	32.58%	0.40
CNJ	26,863	15,717	17.63%	49.71%	53.06%	61.06%	0.59
COND	1,016	824	0.67%	2.61%	2.01%	3.20%	0.81
IMP	1,664	875	1.09%	2.77%	3.29%	3.40%	0.53

Tables for the frequency of verbal person features for the full corpus and both subcorpora, as they are rather lengthy, are left to Appendix B (Table B.19 through Table B.21). Overall, third person singular is the most common, followed by obviative, third person plural, and first person singular. The least common forms are inclusive and second person plural. Some different patterns are apparent between the two subcorpora: A-W generally contains more first and second persons, while BT contains more third persons. For some persons, the relative frequencies are very similar, such as for first person plural inclusive and for third person plural. Similar patterns hold within quotative verbs as well. For inanimate persons, A-W contains more proximate inanimate persons and Bloomfield more obviative; this aligns nicely with the relative occurrence of animate third persons. Person features are further discussed for the full corpus and both subcorpora in the text type analysis in Chapter 6.

4.4.1.2 Nominal features

Table 4.18 gives the noun features in the full corpus, beginning with the total number of nouns then the inanimate and animate nouns (including dependent nouns), followed by various morphological features such as number, obviation, locative, and possession. Note that some percentages, such as those for animate and inanimate, do not add up to 100%; this is due to ambiguity where nouns can be either animate or inanimate and the current parser is unable to disambiguate them from context. Animate nouns are more frequent than inanimate nouns and singular nouns are more common than plural nouns, which aligns with verbal person features as well, likely as these nouns are the actors and goals of these verbs. Possessed nouns are nearly all dependent/inalienable (as dependent nouns must be possessed, they must necessarily form a subset). The nominal type-token ratios are lower than those for verbs, as there is considerably less nominal inflection to result in different wordforms. The BT subcorpus contains overall more

nouns, as well as specifically more animate, obviative, locative, possessed, and dependent nouns. The A-W subcorpus contains more inanimate nouns and more plural nouns, while the two subcorpora contain singular nouns in similar frequencies (Table B.26 and Table B.27). Nominal types are also relevant to the text type analysis in Chapter 6 and further discussion is found therein.

Table 4.18: Full corpus: noun classes and features

	Cree tokens	Cree types	% Tokens	% Types	TTR		
Nouns	25,683	4,033	16.85%	12.76%	0.16		
					% N tokens	% N types	TTR
NI[47]	9,895	2,102	6.49%	6.65%	38.53%	52.12%	0.21
NA	17,759	1,994	11.65%	6.31%	69.15%	49.44%	0.11
NDI	1,192	318	0.78%	1.01%	4.64%	7.88%	0.27
NDA	5,660	558	3.71%	1.76%	22.04%	13.84%	0.10
SG	14,350	1,822	9.42%	5.76%	55.87%	45.18%	0.13
PL	6,094	1,094	4.00%	3.46%	23.73%	27.13%	0.18
OBV	6,597	714	4.33%	2.26%	25.69%	17.70%	0.11
LOC	1,540	443	1.01%	1.40%	6.00%	10.98%	0.29
PX	7,869	1,428	5.16%	4.52%	30.64%	35.41%	0.18

Table 4.19 gives pronominal features for the full corpus, starting with the total number of pronouns and then looking first at demonstrative pronouns and their animacy and features, then at personal pronouns and their persons. These are not the only types of pronouns, leaving about 10% of other pronouns accounted for only in the total. Unresolved ambiguity results in some uncertain counts; in particular, the third person pronoun *wiya* can also be a particle. Demonstratives are more frequent than personal pronouns and, like nouns, animate are more frequent than inanimate and singular more than plural. However, while there are more obviative than plural nouns, there are

[47] The dependent nominals (NDI and NDA) here are a subset of the overall nominals (NI and NA).

fewer obviative demonstratives. First person pronouns are the most frequent of the personal pronouns, then third person, with considerably fewer second person pronouns. The A-W subcorpus contains more pronouns overall as well as more pronouns of both subclasses. Additionally, the ratio between demonstrative and personal pronouns varies between the subcorpora, with A-W containing relatively fewer demonstrative pronouns out of the total pronouns, and BT fewer personal pronouns (Table B.29 and Table B.30). This is likely related to the different proportions of persons between the two subcorpora, as a greater proportion of demonstrative pronouns is not unexpected alongside a greater proportion of third persons.

Table 4.19: Full corpus: pronoun classes and features

	Cree tokens	Cree types	% Tokens	% Types	TTR		
PRON	13,926	172	9.14%	0.54%	0.012		
					% PRON tokens	% PRON types	TTR
DEM	10,507	62	6.89%	0.20%	75.45%	36.05%	0.0059
I	3,294	32	2.16%	0.10%	23.65%	18.60%	0.0097
A	7,459	36	4.89%	0.11%	53.56%	20.93%	0.0048
SG	7,378	28	4.84%	0.09%	52.98%	16.28%	0.0038
PL	2,005	25	1.32%	0.08%	14.40%	14.53%	0.012
OBV	1,351	14	0.89%	0.04%	9.70%	8.14%	0.010
Proximal	7,354	24	4.83%	0.08%	52.81%	13.95%	0.0033
Medial	3,048	28	2.00%	0.09%	21.89%	16.28%	0.0092
Distal	103	9	0.07%	0.03%	0.74%	5.23%	0.087
					% PRON tokens	% PRON types	TTR
PERS	1,940	38	1.27%	0.12%	13.93%	22.09%	0.020
1	849	13	0.56%	0.04%	6.10%	7.56%	0.015
2	259	8	0.17%	0.03%	1.86%	4.65%	0.031
3	717	11	0.47%	0.03%	5.15%	6.40%	0.015

4.4.1.3 Particles

Finally among the morphological tags are those for particles. Table 4.20 gives the particles and subclasses for the full corpus; particle phrases are also included in these counts. Quantifiers and locative particles are the most frequent of the labelled subclasses, followed by temporal and negative particles. Particles have the lowest type-token ratios of the word classes in the corpus, with next to no morphology possible and relatively few possible variants (e.g., reduplication, final vowel deletion, or variation between <t> and <c>). This table also demonstrates that only about a third of the particle types in the corpus have been semantically classified. The A-W subcorpus contains overall more particles than BT, though BT contains relatively more of each labelled subclass, though they remain in the same order of relative frequency (Table B.32 and Table B.33).

Table 4.20: Full corpus: particles

	Cree tokens	Cree types	% Tokens	% Types	TTR		
IPC	60,410	1,516	39.64%	4.80%	0.025		
					% IPC tokens	% IPC types	TTR
NEG	3,343	26	2.19%	0.08%	5.53%	1.72%	0.0078
IPL	5,777	93	3.79%	0.29%	9.56%	6.13%	0.016
IPT	4,838	131	3.17%	0.41%	8.01%	8.64%	0.027
QUANT	5,986	96	3.93%	0.30%	9.91%	6.33%	0.016
Other	40,466	1,170	26.55%	3.70%	66.99%	77.18%	0.029

4.4.2 Syntactic functions

As the tables detailing syntactic functions in Appendix B are generally large, they are not repeated here; instead, patterns and differences will be discussed in this section, and the tables may be referenced in Appendix B (Table B.34 through Table B.54). The full corpus contains 98,492 syntactic function tags as applied by the parser. Of these, 45,855 (46.6%) occur in the A-W subcorpus and 52,566 (53.4%) in the BT subcorpus. Of the tags, the largest proportion are those marking verbs—as either `@PRED` (all non-quotative verbs) or `@Quot`, plus transitivity class—then those marking actors and goals, then particles, and then noun phrase tags (`@N`) on demonstratives and numerals (other minor categories also occur, cf. the end of Appendix B). There are 45,049

`@PRED` tags and 5,574 `@Quot` tags in the full corpus (A-W: 20,223 (44.9% of `@PRED` tags) and 2,135 (38.3% of `@Quot` tags) respectively, BT: 24,826 (55.1%) and 3,439 (61.7%) respectively). As predicate and quotative function tags are very similar to the morphological tags for these verb classes, they are not further discussed here.

The category of tags with the most detail to be discussed is actor and goal tags, totalling 20,110 in the full corpus (A-W: 9,271, 46.1% of actor/goal tags; BT: 10,839, 53.9%). Throughout the corpus, actors are generally more common than goals, and nouns occur as actors and goals more often than pronouns, either demonstrative or personal. Singular actors are more common than plural, in line with the person morphological tags. Within the pronouns, demonstrative pronouns are more likely to occur as actors and goals than personal pronouns. Among personal pronouns, first persons occur more often as actors, and third persons more often as goals. Inanimate nouns and pronouns occur as goals more often than actors, while the reverse holds true for animate nouns.[48] Looking at differences between the subcorpora, a greater proportion of overall nominals and nouns in particular receive actor and goal tags in BT, though more pronouns receive tags in A-W. Though the overall pattern is to have more actors than goals, A-W contains slightly more goals for both demonstratives and nouns than BT. For personal pronouns, BT differs from A-W and the full corpus, with more third persons marked as actors and goals, though this is unsurprising given that there are generally more third person verbs in BT.

Particle tags, the next most common after actors and goals with 16,973 in the full corpus (A-W: 8,836, 52.1% of particle tags; BT: 8,137, 47.9%), are composed of, as discussed in Chapter 3, negative particles associated with either verbs, nouns or particles, temporal and locative tags associated with verbs, and quantifier tags associated with verbs or nouns. Tags on locative particles are the most common, followed by temporal particles, quantifiers, and negative particles. For the latter two, particles associated with verbs are most common, and for negative particles, they are next most likely to modify particles, with negated nouns being the rarest.[49] These patterns vary between the two subcorpora, for the A-W subcorpus, locatives are the most common, followed in order by quantifiers, negatives, and temporals, and for BT, temporals are the most common,

[48] The exception here is that obviative nominals are more likely to be goals than actors, this is discussed in far more detail in the following chapter.
[49] For more on negation in the Plains Cree corpus, see Schmirler & Arppe (2019a).

followed by locatives, quantifiers, and negatives. Thus, locatives being most common in both result in them being the most common in the full corpus, though there is a rather drastic difference in temporals that is obscured by the full corpus data.

Noun phrase tags, which can be assigned to demonstratives before or after nouns (a numeral may intervene), as well as to a number of other particles, pronouns, and numerals before a demonstrative or noun, occur 7,108 times in the full corpus (A-W: 3,557, 50.0% of noun phrase tags; BT: 3,551, 50.0%). Of these tags, the majority occur on demonstratives, though there are also considerably more non-demonstratives with `@N` tags in the A-W subcorpus than in the BT subcorpus. Demonstratives in noun phrases in the BT corpus also represent a far greater percentage of all demonstratives (~71%) compared to the A-W corpus (~37%). This can also be related back to the demonstrative actors and goals—a demonstrative can only be tagged as an actor or goal if it is not already tagged as part of a noun phrase, and there are more demonstrative actors and goals in A-W than in BT. The BT subcorpus thus must have considerably more fully specified noun phrases to occur as participants, while in A-W more pronouns occur as participants— these patterns are much obscured in the full corpus tags.

4.5 Discussion

The work undertaken to apply morphosyntactic tags to this corpus of Plains Cree has resulted in not only a tagged corpus, but morphological gold standards, which themselves allow for the identification of errors and missing phenomena in the morphological model, and several avenues for further development in the Plains Cree parser. The disambiguation results in §4.3.2 serve to demonstrate the considerable improvements to the A-W morphological gold standard, as where there were previously over 3,000 wordforms in the subcorpus without an analysis, there are now just over 300. Though many remaining issues with the parser were discussed previously in Chapter 3, the evaluation of the recall and precision of syntactic tags in §4.3.2 highlights the tag types that have not yet undergone substantial development and testing, especially the oblique and instrument function tags. Instrument tags, which are so far modelled with only three constraints with specific syntactic contexts (a non-locative noun, indeclinable nominal, or demonstrative pronoun preceding *ohci* ‘by means of’) are extremely limited in scope; full modelling of these is a consideration for future research and will also benefit from semantic classification of nouns, such as to make

reference to a label such as *tool* or *material*. Obliques are given slightly more consideration in the parser with eight constraints that essentially aim to label any non-locative nominal not already associated with a verb as an actor or goal. However, once again, research is required for future development. Surprisingly, the adposition tags (e.g., to associate locative nouns with locative particles) remain problematic, despite their apparently straightforward nature. As very few occur in the development corpus, the low recall can be directly related to a single sequence of locative noun-particle phrases. Indeed, for oblique, instrument, and preposition tags, their rarity in the development corpus is one of the primary reasons they are yet to be further developed and tested. Counts for these less-developed tags are given at the end of Appendix B for the sake of completeness, though their occurrence in the corpus must of course be taken with several grains of salt.

General differences between the subcorpora can be seen throughout both the morphological and syntactic tags. For example, there are more verbs in the BT subcorpus, as well as more third person tags, and then, unsurprisingly, more `@PRED` and `@Quot` tags and more overt actors and goals—the greater occurrence of actors and goals is likely caused by having both more verbs and more of those verbs having third person morphology, and thus more overt participants are possible (see also Chapters 5 and 6). However, outside of the verb and participant tags, the raw counts for syntactic function tags tend to be very close between the two subcorpora, though these represent higher relative frequencies in the BT subcorpus, as it contains slightly fewer words. I take this as representative of a difference in the editing and transcription styles of the two subcorpora: the BT subcorpus is more heavily edited and transcribed with less punctuation and in particular includes fewer commas (13,281 tokens, 43.7% of all clause boundary punctuation), while the A-W subcorpus contains considerably more commas (19,062 tokens, 60.2% of clause boundary punctuation), which are taken to offer at least some representation of how the texts were spoken.[50] If there is an intonational pause, the nominal in question can very likely be linked semantically and pragmatically to a nearby verb, but may not function syntactically as its actor or goal. As commas are used as `BARRIER`s in the parser to limit the search for actors and goals, fewer will be identified where more commas intervene. The occurrence of fewer commas likely results in at

[50] Compare to the full corpus, where commas represent 52.1% of clause boundary punctuation.

least some overapplication of actor and goal tags in the BT subcorpus; though the parser is performing as expected, if pauses are not represented by commas, more actor and goals will be identified. This may be confirmed in future research, through hand-coding a small section of the BT subcorpus; this can also serve to truly test the applicability of a parser developed on more modern Plains Cree to an older variety of the language.

Like the occurrence of commas, throughout this chapter differences between the subcorpora are often obscured by looking just at the full corpus counts, as previously noted (Schmirler & Arppe, 2019b). As more differences between and within the subcorpora come to light, a question arises of how to account for these differences when reporting on the corpus. As done in this chapter and its accompanying appendix, the full corpus is reported, but so too are the subcorpora included individually, at least to note major differences in the prose. This is also done in Chapter 5, where this same tagged corpus is used to explore word order patterns in the full corpus, and how they differ for the subcorpora. To report only the counts and frequencies in the full corpus is disingenuous: to claim that Plains Cree speakers use a given feature some percentage of the time is very likely simply untrue, if one uses the full corpus as the only benchmark. Therefore, in Chapter 6, a statistical study of the differences between and within the subcorpora is reported, exploring ways in which the corpus may be divided when considering patterns within.

Now that a morphosyntactically tagged corpus for Plains Cree exists, another important question is, quite simply, *what's next?* Beyond broad linguistic analyses, such as those undertaken using this corpus in the next two chapters, a digital, tagged corpus can be presented in an online corpus interface; for the Plains Cree corpus, this is within the Korp corpus interface (https://korp.altlab.app/). This interface, though it does not yet include the updated tagged corpus, allows for searches by whole words, parts of words, morphosyntactic features, and various detailed searches using regular expressions. As the Plains Cree corpus is drawn from published volumes that almost all include English translations, these may be aligned to the Plains Cree texts to present a parallel Cree-English corpus. Another question is that of public access; at present, access to this corpus can be requested by academics, though full public access requires further discussion regarding intellectual property. Access to Cree texts by Cree speakers is of course a priority for language data, alongside academic pursuits.

4.6 Conclusion

In this chapter, I described the texts that comprise a digital morphosyntactically tagged corpus of Plains Cree. These texts were collected at different times, from various speakers, and include different text types and editing styles, all of which may contribute to differences in the morphosyntactic tags. These morphosyntactic tags are drawn from two major elements: 1) the morphological gold standards for the subcorpora and 2) the parser that disambiguates and assigns syntactic function tags. The improvements to the A-W gold standard are evidenced when considering the disambiguation of the subcorpus, in addition to the Bloomfield morphological gold standard, the corpus now consists of ~152,000 words of Plains Cree with verified morphological analyses, and the verification process has also served to improve the morphological model. While the efficacy of the morphological model is beyond the scope of this work, the efficacy of the parser was reported on here, looking at the effectiveness of disambiguation, both with respect to the development corpus and the overall effects of disambiguation in the corpus, and the recall and precision of syntactic tags in the development corpus.

Much of the chapter has been dedicated to presenting the features that occur in the corpus and commenting on differences between the subcorpora. In this and following chapters, the differences between the corpora are of particular interest, as I aim to further explore the degree to which internal variation is obscured when the full corpus is considered without careful thought. The underlying differences in the style of transcription, the time periods in which the texts were created, and the content and text types run deep in the subcorpora and a thorough understanding of the corpus and its internal variation is key to accurately describing its morphosyntactic phenomena. Much like an analysis of English drawn from a corpus containing over a billion words requires consideration of whether the results are drawn from novels, newspaper articles, or tweets, the morphosyntactic features of Plains Cree cannot be well understood without their context. The differences between the subcorpora with respect to word order are explored in the Chapter 5, and a textual analysis of Plains Cree in Chapter 6 begins to tease apart the internal variation of the corpus.

Chapter 5

Argument realisation in a Plains Cree corpus

5.1 Introduction

In the previous chapter, I presented the Plains Cree corpus with respect to the number of various morphosyntactic tags in the full corpus and the ways in which the subcorpora differ from each other. In this chapter, I turn to the argument realisation patterns in the corpus using the morphological feature tags and actor/goal argument tags. Plains Cree word order is described as free, flexible, or non-configurational, with any permutation of arguments possible, and overt arguments are not required. Descriptive analyses of word order focus on the position of arguments with respect to the verb, with more topical or focused arguments expected preverbally. The morphosyntactically tagged corpus described in this dissertation is an ideal tool for exploring word order variation on a large scale; even without true pragmatic tags for topic or focus, many of the existing tags can be used to represent differences in relative topicality.

This chapter begins with a brief summary of what has been described for word order patterns in Plains Cree (§5.2) and how the corpus can be used to look into syntactic patterns (§5.3). The realisation and relative positions of actors and goals is presented through a series of tables in §5.4, starting first with different verb classes, then looking at different aspects of the morphosyntactic tags that can stand in for topicality, including direction, person, and nominal type. The results are briefly discussed in §5.5 and the chapter concludes in §5.6.

5.2 Word order patterns

With such a wealth of morphological information on nouns and verbs exemplified throughout this work, it is perhaps unsurprising that word order plays little to no role in establishing core relationships between nouns and verbs. The complex feature-marking on verbs as described above allows Plains Cree to be a language with flexible word order, in the sense that any order of the

actor, verb, and goal (A, V, and G respectively) is semantically (if not pragmatically) equivalent to any other order, as in (1). Word order has no bearing on the semantic roles of 'children' and 'ducks'; these are achieved through obviation on the nouns and direction morphology on the verb.

(1) Word order permutations: 'the children killed some ducks'
(adapted from Wolfart, 1996, p. 392)

a. *awâsisak* *nipahêwak* *sîsîpa*
children killed ducks
A V G

b. *awâsisak* *sîsîpa* *nipahêwak*
children ducks killed
A G V

c. *nipahêwak* *awâsisak* *sîsîpa*
killed children ducks
V A G

d. *nipahêwak* *sîsîpa* *awâsisak*
killed ducks children
V G A

e. *sîsîpa* *nipahêwak* *awâsisak*
ducks killed children
G V A

f. *sîsîpa* *awâsisak* *nipahêwak*
ducks children killed
G A V

While each of the above orders is possible, overt arguments are not required. There are also transitive clauses that contain AV or VA, VG or GV orders, or simply the verb without any overt arguments (Wolfart, 1996); overt arguments have also been described as adjuncts to the verb within generative formalisms (e.g., Blain, 1997).[51] Mühlbauer (2007) suggests that VSO (that is, VAG in the terminology used herein) is the basic or neutral word order in Plains Cree; while this order, with both arguments overtly realised, occurs in only 1.0% of the clauses examined in Wolvengrey (2011), verb-initial VA and VG orders are considerably more common (see §5.4.1 below).

Wolvengrey (2011) investigated the combinations of actors, verbs, and goals for monotransitive VTAs as found in one collection of Plains texts containing ~2,000 words, *Stories of the House People* (Vandall & Douquette, 1987),[52] and reported frequency of each word order permutation. These included clauses where one or both arguments were overtly realised as nominals, as well as the clauses with no overtly realised arguments, for a total of 286 clauses. In Vandall & Douquette (1987), the six possible word order permutations where both participants are overt make up only a small percentage of the total clauses, ranging from 0.7% for VGA (n = 2) to 3.8% for GVA (n = 11). Clauses with just an overt actor are more common: AV occurs in 3.8% of cases (n = 11) and VA in 5.2% (n = 15). It is even more common for only the object/goal to be overt: GV occurs in 18.9% of cases (n = 54) and VG in 29.7% (n = 85). Clauses in which neither participant is overtly realised (i.e. there is only a transitive verb with no nominals) occurred in 30.8% of cases (n = 88) (Wolvengrey, 2011, p. 202). As these numbers describe only one Cree text, they do not reflect counts that can be found in a larger corpus, as demonstrated in Schmirler et al. (2018) and Appendix A, where the overall occurrence of VTAs without arguments in the A-W subcorpus is considerably higher, at 46.8%, suggestive of differences between text types, speakers, or both.

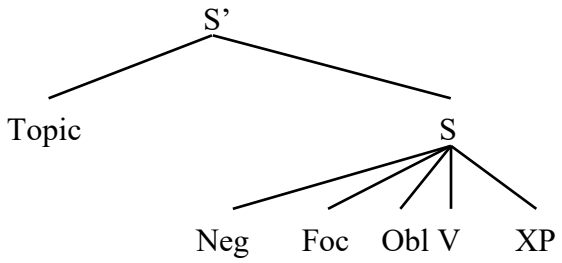
Though the various possible word orders are considered semantically equivalent, they are influenced by pragmatic factors (e.g., Wolvengrey, 2011, pp. 35–7). Across Algonquian languages, the first position of a clause frequently contains a topical argument, followed by

[51] Personal and demonstrative pronouns are also not required with verbs, and when they do occur as arguments, they serve emphatic functions.

[52] This collection has also been used as a development corpus in the creation of the Plains Cree parser, as well as being part of the A-W subcorpus (see Schmirler et al., 2018 or Appendix A, and Chapter 4 herein for more detail).

negation, focus, oblique elements, and then the verb, as in the template given in (2). The post-verbal material in a clause, however, does not present agreed-upon patterns, though some descriptions have been undertaken (e.g., Wolvengrey, 2011).

(2) The Algonquian clause (Dahlstrom, 1995, p. 3)



Wolvengrey (2011) explores the post-verbal portion of the clause, where arguments occur following the verb in verb-initial clauses, which are considered an unmarked order for many Algonquian languages (e.g., Mühlbauer, 2007). For transitive verbs, the highest-ranking argument in terms of semantic role (e.g., agent>recipient>patient, etc.) is more likely to be first, though this is only a tendency. Ditransitive VTAs display a more consistent pattern; when both the syntactic goal (the recipient in Cree) and the oblique (the theme) are animate, the grammatical information (animacy, obviation, direction) alone is not sufficient to determine their respective roles, as only one is marked morphologically on the verb. Wolvengrey (2011) presents the following examples and discusses comments made by Cree consultants.

(3) Ditransitive post-verbal restrictions (Wolvengrey, 2011, pp. 213–4)

a. *ana nâpêw kî-asamêw atimwa kinosêwa*.

ana	nâpêw	kî-	asam	-ê	-w	atimw-	-a	kinosêw	-a
DEM.A.SG	NA	IPV	VTA	DIR	3SG	NA	3′	NA	3′
that	man	PST	feed	3>3′		dog		fish	

'That man fed (a/the) dog(s) fish.'/'That man fed fish to (a/the) dog(s).'

b. *ana nâpêw kî-asamêw kinosêwa atimwa*.

ana	nâpêw	kî-	asam	-ê	-w	kinosêw	-a	atimw-	-a
DEM.A.SG	NA	IPV	VTA	DIR	3SG	NA	3′	NA	3′
that	man	PST	feed	3>3′		fish		dog	

'That man fed (a/the) fish dog.'/'That man fed dog to (a/the) fish.'

In the examples presented in (3), we see that word order in Plains Cree can be used to determine semantic roles in cases where direction and obviation are insufficient; the recipient occurs before the theme. However, pragmatic information can overrule such syntactic information. For example, in (4), an inalienably possessed form of 'dog' is used,[53] indicating a closer relationship between the possessor and the dog. When this occurs, ranking of semantic roles with respect to word order is demoted in favour of pragmatics; speakers would interpret both word orders the same way because it is pragmatically unlikely for a person to feed their own pet dog(s) to a fish. While (4)b. can be interpreted as a pet dog being fed to a fish, speakers find this unlikely (and rather humorous) and would avoid the word order (Wolvengrey, 2011, p. 215).

(4) Pragmatic post-verbal interpretations (Wolvengrey, 2011, p. 215)

a. *ana nâpêw kî-asamêw otêma kinosêwa*.

ana	nâpêw	kî-	asam	-ê	-w	o-	-têm-	-a	kinosêw	-a
DEM.A.SG	NA	IPV	VTA	DIR	3SG	3	NA	3′	NA	3′
that	man	PST	feed	3>3′			dog		fish	

'That man fed his dog(s) fish.'/'That man fed fish to his dog(s).'

b. *?ana nâpêw kî-asamêw kinosêwa otêma*.

ana	nâpêw	kî-	asam	-ê	-w	kinosêw	-a	o-	-têm-	-a
DEM.A.SG	NA	IPV	VTA	DIR	3SG	NA	3′	3	NA	3′
that	man	PST	feed	3>3′		fish			dog	

'That man fed his dog(s) fish.'/'That man fed fish to his dog(s).'

[53] When 'dog' is possessed, rather than using the free stem *atimw*-, the bound form *-têm-* which can mean 'dog' or 'horse' (more generally, 'animal companion') is used instead.

As these examples show, word order can have some bearing on determining syntactic and semantic roles of core arguments in Plains Cree. However, this rule cannot be categorically applied to Plains Cree sentences for two key pragmatic reasons. First, we see that pragmatics can overrule the syntax to allow for the most likely interpretation. Second, as many Cree arguments are not overtly realised as nominals, instances where both the recipient/goal and indirect object are present will be relatively rare. At present, the syntactic parser would identify both 'fish' and 'dog' as goals in the above sentences, leading to overapplication of the function tags; in future development, some pragmatically-based heuristics may be implemented, such as creating a ranking of animate participants (e.g., humans > companion animals > other animals, or inalienably possessed nouns > other nouns) to begin the process of automatically parsing such examples, such as the classifications in Dacanay et al. (2021). However, even with the current parser output, future research could scrutinise existing instances of multiple goals to identify and explore cases such as the above.

5.3 Exploring syntactic patterns in a corpus

5.3.1 Variation on a large scale

A morphosyntactically tagged corpus allows for the investigation of various patterns at a much larger scale than previously possible, and even without explicit tagging for pragmatic features, many of the above patterns described for Plains Cree arguments can be explored. This chapter primarily focuses on where overt participants are realised—for which nominal classes or verbal classes, whether overt nominals precede or follow the verb, and so on. The occurrence of arguments pre- and post-verbally can be examined with respect to discourse topicality, such as proximate and obviative features, or inherent topicality of SAPs vs. non-SAPs. In cases where both an actor and goal are overtly realised, their relative ordering can also be related to their topicality, as well as to their semantic role, with actor and goal standing in for agent and patient, for example.

5.3.2 Argument structure and overt participants

Argument structure describes the ways in which verbs and arguments combine morphosyntactically (word order, case, subjects or objects) and semantically (agents, patients, etc.), and, to some extent, pragmatically (in terms of discourse, topicality, etc.)—how syntactic

arguments are mapped to semantic roles, that is, alignment. Preferred Argument Structure (PAS) is a hypothesis that takes this a step further and explores not only which arguments are licensed by a predicate, but how they are most likely to be realised: overt or not, as nouns or pronouns, and in what order (for both flexible word order languages and for variation that occurs in languages like English, such as dative shift or passive constructions) (e.g., Du Bois, 1987, 2003). PAS focuses on spontaneous discourse (not unlike the spoken texts in the Plains Cree corpus) and explores why "certain configurations of arguments are systematically preferred over other grammatically possible alternatives"— it is "a model of argument realization" or "argument token selection" that finds that "certain argument realizations in certain argument positions" are preferred while others are dispreferred (Du Bois, 2003, p. 33).

These argument configurations can be described with respect to two constraints, quantity and role, along two dimensions, morphosyntax and pragmatics. For morphosyntax, the quantity constraint aims to "avoid more than one lexical core argument" and the role constraint aims to avoid a lexical agent, while for pragmatics, the constraints are the same but for new arguments/agents instead of lexical ones (Du Bois, 2003, pp. 33–4).[54] The focus on agents is key here, as subjects of intransitive clauses and objects of transitive clauses are more likely than agents to be realised as full noun phrases, and subjects and objects occur with similar frequencies. In Sacapultec Maya, an ergative language that began Du Bois's exploration of PAS, two overt arguments occur <3% of transitive clauses, while one or zero overt arguments occur in ~50% of clauses (Du Bois, 1987). For languages like English, which are 1) not ergative and 2) generally require an overt argument, Du Bois finds in spoken corpora that transitive clauses with zero overt lexical arguments are about as common as those with one (the others being pronominal), and those with two full noun phrase arguments occur less than 10% of the time (2003, pp. 33–4). Though of these languages, only Sacapultec Maya is morphosyntactically ergative, ergative-like frequencies for argument realisation are seen across languages.

DuBois (2003, p. 40), finds PAS useful on several levels: not only can it be used to statistically describe variation in argument realisation and point towards semantic, pragmatic, and

[54] As newness is not yet tagged for arguments in the Plains Cree corpus, this is set aside for the present work, but quantity can certainly be explored.

morphosyntactic motivations for certain preferences, but it also works toward a discourse-based explanation for the occurrence of ergativity in human languages: in terms of argument realisation, subjects of intransitive verbs and objects pattern together, differently from agents. Even for languages that are morphosyntactically accusative like English, the discourse information imparted through argument realisation (given vs. new, etc.) displays more ergative-like patterns. These types of frequencies have also been used to explore PAS in other types of alignment systems, including hierarchical. Matter (2020) finds for Mapudungun, a hierarchical language of South America, that subjects fall between agents and objects in terms of frequency of overt participants. Plains Cree, along with other Algonquian languages, is frequently cited as a canonical example of a language with direct-inverse or hierarchical alignment, though only certain members of the VTA paradigm truly represent such a system (e.g., Odribets & Oxford, in press). Algonquianists have also explored both ergative and accusative morphosyntactic patterns in Algonquian languages, when looking beyond the VTA paradigm and its direction morphology (e.g., Campana, 1989; Hewson, 1987; Odribets & Oxford, in press).

Ergative patterns in Plains Cree are identified in several ways. Transitive stems differ based on the animacy of the goal/patient rather than the actor/agent; some VTA goal marking is identical to VAI/VTI actor marking; and indefinite agents (unspecified actors) are marked by inflectional morphology, while indefinite goals are instead achieved by deriving VAIs (Hewson, 1987). In (5), two different forms of 'make' are used, the VTI stem with the inanimate 'shoes' and the VTA stem with the animate 'bread'; under Hewson's analysis, this is seen as some preferential treatment given to goals(/objects), and thus an ergative-like property.

(5) Transitive stems

a. VTI stem, required for an inanimate goal (Masuskapoe, 2010, p. 72)

osîhtâw maskisina

osîhtâ-	-w	maskisin	-a
VTI	3SG	NI	PL
make		shoe	

's/he makes shoes'

b. VTA stem, required for an animate goal (Lafond & Longneck, 1998, pp. 278–9)
nitosîhâw pahkwêsikan

ni(t)-	osîh	-â	-w	pahkwêsikan
1	VTA	DIR	3SG	NA
	make	1>3		bread

'I make bread'

In (6), the third person marker (and optional plural) is seen in both VTA and VAI independent morphology, though in the former, it marks the goal and in the latter it marks the actor(/subject). Like in ergative languages, this is a goal(/object) marked like an intransitive actor(/subject). These forms are drawn from paradigms of Plains Cree verbs, such as those found in Okimâsis (2004, 2021), Ratt (2016, 2022), Wolfart (1973), and Wolvengrey (2011).

(6) 3PL VTA goals and VAI actors

a. VTA with third person goal
niwâpamâw(ak)

ni-	wâpam	-â	**-w (-ak)**
1	VTA	DIR	3 PL
	see	1>3	

'I see him/her(/them)'

b. VAI with third person actor
nipâw(ak)

nipâ	**-w (-ak)**
VAI	3 PL
sleep	

's/he(/they) sleep'

In (7), the differences between an indefinite agent (unspecified actor) and an indefinite goal are exemplified. An indefinite agent is indicated solely through inflectional direction morphology, but for an indefinite goal, a new VAI stem is derived (*wîcih* + *-iwê*, a detransitiviser), which is

morphologically and syntactically intransitive. For Hewson (1987), this demonstrates how agents are less core to the predicate, as they can be removed with inflection, than objects, which require derivation to a new verb class.

(7) Unspecified actor and general object

a. *wîcihâw*

wîcih -â -w

VTA DIR 3SG

help X>3

'someone helps him/her'

b. *wîcihiwêw*

wîcihiwê -w

VAI 3SG

help.people

's/he helps people'

Accusative patterns are also observed in Plains Cree: while the morphology above demonstrates ways in which third person plural marking is the same for VTA goals and VAI actors, SAP morphology demonstrates an accusative pattern, as in (8), where the first person plural exclusive is represented by the same morphology as the actor of either a VTA or VAI. Odribets & Oxford (in press) argue that this type of pattern, as it occurs in direct forms, which are more frequent and may be considered less marked, is a default for Algonquian languages, supporting not only an accusative pattern, but suggesting the languages in question could be classified as accusative languages. However, I also include the inverse form of the VTA, as the same marking is used, only the direction changes, not the person morphology.

(8) SAP forms in VTAs and VAIs

a. *niwâpamânânak*

ni-	wâpam	-â	-nân	-ak
1	VTA	DIR	1PL	3PL
	see	1>3		

'we (excl.) see them'

b. *ninipânân*

ni-	nipâ	-nân
1	VAI	1PL
	sleep	

'we (excl.) sleep'

c. *niwâpamikonânak*

ni-	wâpam	-iko	-nân	-ak
1	VTA	INV	1PL	3PL
	see	1<3		

'they see us (excl.)'

These discussions have generally taken into account either the morphology of Plains Cree verbs or a small number of sentences with grammaticality judgements, and so are possible on a small scale. A tagged corpus, however, offers another opportunity to explore variation on a larger scale and add to the available data used in these typological arguments.

5.4 Occurrence of actors and goals

In this section, I present corpus results for the occurrence of overt actors and goals in the Plains Cree corpus. As in the previous chapter, I give tables for the full corpus here and discuss differences between subcorpora, with the tables for these given in Appendix C. I first explore the occurrences of actors and goals by verb class, which necessarily show some patterns for how animate and inanimate participants behave (§5.4.1). I then explore topicality in more detail, as it

can be inferred from the morphosyntactic tags available in the corpus so far (§5.4.2). Finally, I take a brief look at Preferred Argument Structure in the Plains Cree corpus (§5.4.3).

Throughout this section, cases where two of the same tag occur on either side of the verb (e.g., A V A) or with another argument tag intervening (e.g., V G A G) are excluded (<2% of clauses throughout). However, cases where two of the same tag occur without any other intervening material are included, considered just as e.g., "actor occurs before verb"; though erroneous tags can be expected to some extent here and throughout, these sequences still suggest that the actor occurred before the verb.

5.4.1 By verb class

First, I look at the occurrence of arguments by verb class, starting with just whether or not there is an overt actor or goal and whether that argument occurs before or after the verb (Table 5.1).[55] In reading this table, recall that actors for VIIs are inanimate and for all other verb classes are animate, and that for transitive verbs, the goals of VTIs are inanimate and the goals of VTAs are animate. This table demonstrates several patterns, though I focus on the differences between 1) transitive and intransitive verbs, 2) animate and inanimate participants, and 3) actors and goals. First, overt actors are seen with similar frequencies for intransitive verbs of both classes, which are somewhat higher than the frequencies for overt actors for transitive verbs, though the frequencies between the transitive classes are also very similar. Thus, transitivity plays a role in how frequently arguments are realised. Transitivity differences can also be seen looking beyond this table. Overall in the full corpus, 66% verbs occur without any overt arguments. For intransitive verbs (both VIIs and VAIs), 76% occur with no overt actor, while for transitive verbs, 51% of VTIs and 60% of VTAs occur with no overt actor or goal. This is unsurprising, as Table 5.1 also demonstrates that

[55] The tables in this chapter are generally quite dense, and often present several interconnected tables in one. Shading is used to demarcate rows while borders are used to demarcate headings of the subsections. The first row in Table 5.1 gives the total number of each verb transitivity class in the corpus, which are used to calculate the percentages within each class in the rows below. Then, actors and goals are looked at separately: the w/A row gives the number of verbs in that class with an overt actor, followed by the percentage of those verbs. Then, AV looks at the number of each class where an actor precedes a verb, then VA for where an actor follows a verb. The percentages in these rows, rather than being out of the total of each class, are instead out of the total with an overt actor. The same is done for goals for the transitive verbs. At the bottom of the table, actors and goals are totalled together to look at overt arguments in general. Tables throughout the chapter present information in similar ways, though are less complex than this first table.

goals are generally more likely to be overt than actors. Second, animate and inanimate participants pattern differently. For animate entities, both actors and goals are more likely to follow the verb than to precede it. Animate actors of VAIs occur postverbally more often than those for transitive verbs. However, inanimate actors and goals occur closer to 50% for each, with slightly more occurring preverbally than postverbally. Inanimate entities also occur as overt goals considerably more often than animate goals, and are more likely to occur postverbally than animate goals. Third, both animate and inanimate goals (of VTAs and VTIs respectively) occur more often than actors for either class.

Table 5.1: Actors and goals by verb class: full corpus

	VII	3,998	VAI	23,016	VTI	8,952	VTA	15,224	Total V	51,126
		% VII		% VAI		% VTI		% VTA		% V
w/A	925	23.5%	5,505	23.9%	1,302	14.5%	2,195	14.4%	9,927	19.4%
		% w/A		% w/A		% w/A		% w/A		% w/A
AV	479	51.8%	2,202	40.0%	565	43.4%	1,012	46.1%	4,258	42.9%
VA	446	48.2%	3,303	60.0%	737	56.6%	1,183	53.9%	5,669	57.1%
					VTI	8,952	VTA	15,224	Total VT	24,176
		% VII		% VAI		% VTI		% VTA		% VT
w/G	-	-	-	-	3,439	38.4%	4,409	29.0%	7,848	32.5%
		% w/G		% w/G		% w/G		% w/G		% w/G
GV	-	-	-	-	1,737	50.5%	1,460	33.1%	3,197	40.7%
VG	-	-	-	-	1,702	49.5%	2,949	66.9%	4,651	59.3%
		% w/A		% w/A		% w/A or G		% w/A or G		% w/A or G
all preV	479	51.8%	2,202	40.0%	2,302	48.6%	2,472	37.4%	7,455	41.9%
all postV	446	48.2%	3,303	60.0%	2,439	51.4%	4,132	62.6%	10,320	58.1%

The two subcorpora also demonstrate some differences (see Table C.1 and Table C.2 in Appendix C for comparison). The A-W subcorpus has more overt actors for VIIs and fewer for VAIs, while

the BT subcorpus is the opposite—thus these differences are obscured by the averages in the full corpus results. For the VAIs, more overt arguments in BT can likely be attributed to the more frequent occurrence of third persons and nominals overall, as the more frequent first and second person verbs in A-W are less likely to have overt actors. The different rates of overt actors for VIIs, conversely, cannot be readily explained. Transitive verbs occur with similar rates of actors and goals between the two subcorpora. The frequency with which actors and goals precede and follow verbs also differs between the subcorpora: for the BT subcorpus, all verb types are more likely to have arguments following the verb, while for A-W, VII, VTI, and VTA actors are all more likely to precede the verb than to follow, as are VTI goals. The overall occurrence of arguments differs considerably as well, while 42% of overt arguments precede the verb in the full corpus and 58% follow the verb, these values are 49% and 51% respectively for A-W, and 37% and 63% for BT. The occurrence of animate vs. inanimate overt participants do show some consistent patterns: in both subcorpora, inanimate actors and goals occur before and after verbs at similar rates, and in the BT subcorpus, the animate actors of both VAIs and VTIs also behave similarly to each other.

Table 5.2: Two overt arguments with transitive verbs: full corpus

	VTI	8,952	VTA	15,224	Total	24,176
		% VTI		% VTA		% VT
w/A & G	436	4.9%	517	3.4%	953	3.9%
		% w/A & G		% w/A & G		% w/A & G
VAG	52	11.9%	102	19.7%	154	16.2%
VGA	50	11.5%	56	10.8%	106	11.1%
AVG	133	30.5%	200	38.7%	333	34.9%
GVA	84	19.3%	55	10.6%	139	14.6%
AGV	38	8.7%	40	7.7%	78	8.2%
GAV	79	18.1%	64	12.4%	143	15.0%

When looking at transitive verbs, only about 4% of clauses have both arguments overtly realised, as seen in Table 5.2.[56] VTIs occur with both overt participants slightly more often; this aligns with the observation that goals are generally more common for VTIs than for VTAs, as shown above. For both VTIs and VTAs, AVG is the most common order, occurring for about one third of phrases where both are realised. For VTIs, GVA and GAV are the next most common, with the inanimate goal preceding both the verb and actor, then the verb-initial VAG and VGA. For VTAs, VAG is the next most common pattern, followed by GAV, VGA, and GVA. The least common pattern for both types is AGV, with both arguments preceding the verb. Again, variation is seen between the subcorpora. For the BT subcorpus, VTIs with both arguments realised are less frequent and VTAs more frequent than for the full corpus, while for A-W, VTIs with both arguments are more frequent and VTAs less frequent. AVG remains the most common for both, though less so for VTIs in the BT subcorpus. AGV remains by far the rarest in the BT subcorpus, though it is more common for A-W, where VGA is instead the least common. Again, more third person verbs may offer an explanation for a greater frequency of VTAs with two overt arguments in BT (Table C.3 and Table C.4).

5.4.2 Topicality

5.4.2.1 Direct and inverse

I next look at differences between direct and inverse VTAs, as direction combined with actor and goal tags offers a morphosyntactic means of exploring pragmatics and semantics: how is the occurrence of overt arguments related to their topicality? Table 5.3 includes 14,979 VTAs in the full corpus, including local, nonlocal, and mixed (cf. Chapter 2, §2.2.2).[57] First, this table again reinforces that direct verbs are more common than inverse (also seen in Chapter 4). Second, overt actors of inverse verbs actors are far more common than actors of direct verbs, and overt goals of direct verbs are far more common than goals of inverse verbs. Less topical entities (i.e. the obviative goals of direct verbs and the actors of inverse verbs) are generally realised overtly more

[56] Consider again Sacapultec Maya, mentioned above—though an ergative language, it does bear other similarities to Plains Cree in terms of verbal argument indexing. Du Bois (1987) found that two overt arguments occur ~3% of clauses, which is similar to the ~4% Here for Plains Cree.

[57] This number is slightly lower than the total for VTAs in Table 5.1 above; unspecified actor and inanimate actor VTAs are excluded here.

frequently than more topical entities (the proximate actors of direct and goals of inverse). Word order in relation to the verb can also be explored here: the more topical entities (actors of direct verbs, goals of inverse) precede the verb more often, while the less topical entities follow the verb more often. These tendencies also hold true in both subcorpora, though to differing degrees (see Table C.5 and Table C.6 in Appendix C).

Table 5.3: Actors and goals by direction: full corpus

	DIR	% VTA	INV	% VTA
	10,647	71.1%	4,332	28.9%
		% DIR		% INV
w/A	1,080	10.1%	1,098	25.3%
		% w/A		% w/A
AV	589	54.5%	411	37.4%
VA	491	45.5%	687	62.6%
		% DIR		% INV
w/G	4,164	39.1%	241	5.6%
		% w/G		% w/G
GV	1,323	31.8%	136	56.4%
VG	2,841	68.2%	105	43.6%

5.4.2.2 Persons

In this section, I look in more detail at how persons differ in terms of argument realisation, with particular focus on their relative topicality: SAPs are more topical than non-SAPs, proximate third persons are more topical than obviative third persons, and animate entities are more topical than inanimate entities. Singular and plural arguments are included together for proximate and inanimate. The inanimate values in Table 5.4 are the same as those in Table 5.1 above, arranged differently for comparison within and between persons rather than verb class. As above, inanimate participants, whether actors or goals, precede and follow the verb in roughly equal proportions. As actors, they are the least likely to be realised, but as goals, obviative and inanimate participants are

overtly realised at similar rates. Proximate participants are overtly realised more often as goals than as actors, though only to a small degree. However, both proximate actors and goals occur before or after the verb with almost identical proportions. Finally, obviative goals are more likely to be overt than obviative actors, and, though obviatives are in general the most likely to occur following the verb, obviative goals follow the verb more often than obviative actors. In the two subcorpora (Appendix C, Table C.7 and Table C.8) some differences emerge. Inanimate actors are far more common in the A-W subcorpus (also discussed above), on par with obviative actors, both more frequent than proximate actors. Arguments also occur more often preverbally in A-W than in BT.

Table 5.4: Actors and goals for non-SAPs: full corpus

	PROX actor	26,526	OBV actor	4,559	INAN actor	3,934
		% AI, TI, TA 3		% AI, TI, TA 3′		% II
w/A	7,341	27.7%	1,548	34.0%	925	23.5%
		% w/A		% w/A		% w/A
AV	2,909	39.6%	518	33.5%	479	51.8%
VA	4,432	60.4%	1,030	66.5%	446	48.2%
	PROX goal	5,918	OBV goal	5,828	INAN goal	8,952
		% TA 3O		% TA 3′O		% TI
w/G	1,837	31.0%	2,496	42.8%	3,439	38.4%
		% w/G		% w/G		% w/G
GV	726	39.5%	667	26.7%	1,737	50.5%
VG	1,111	60.5%	1,829	73.3%	1,702	49.5%
		% w/A or G		% w/A or G		% w/A or G
Total preV	3,635	39.6%	1,185	29.3%	2,216	50.8%
Total postV	5,543	60.4%	2,859	70.7%	2,148	49.2%

A similar exploration of SAPs is given in Table 5.5. In accordance with the Algonquian Person Hierarchy, which posits that second persons are more topical than first persons, I explore each of

these separately. First person here includes singular and plural exclusive, and second person includes singular and plural, while first person plural inclusive is separated from either: morphologically, the inclusive looks more like second person, but its syntactic behaviour, in terms of overt realisation, has not, to my knowledge, yet been explored.

Table 5.5: Actors and goals by SAPs: full corpus

	1 actor	9,497	21 actor	964	2 actor	3,915
		% AI, TI, TA 1		% AI, TI, TA 21		% AI, TI, TA 2
w/A	1,350	14.2%	121	12.6%	250	6.4%
		% w/A		% w/A		% w/A
AV	738	54.7%	68	56.2%	166	66.4%
VA	612	45.3%	53	43.8%	84	33.6%
	1 goal	2,101	21 goal	208	2 goal	939
		% TA 1		% TA 21		% TA 2
w/G	73	3.5%	6	2.9%	24	2.6%
		% w/G		% w/G		% w/G
GV	53	72.6%	6	100.0%	22	91.7%
VG	20	27.4%	0	0.0%	2	8.3%
		% w/A or G		% w/A or G		% w/A or G
Total preV	791	55.6%	74	58.3%	188	68.6%
Total postV	632	44.4%	53	41.7%	86	31.4%

Of course, compared to non-SAP animate participants above, SAPs are overall realised as overt arguments much less often. They are also far more likely to be realised as actors than goals, unlike non-SAPs, though these frequencies are skewed by the inclusion of VAIs and VTIs here; for just VTAs, SAPs occur at similar rates for both actors and goals. SAPs are also overall more likely to occur preverbally than postverbally, especially first and second person goals, though whether this is because they are inherently more topical or because they necessarily must be pronouns cannot be determined at this time (and, indeed, the difference may not be relevant). The occurrence of

inclusive first persons as overt actors is more similar to that for exclusive first persons, both in general and before and after verbs. Conversely, inclusive goals behave more like second person goals—though there are so few instances that these differences average out for the totals in the bottom two rows, where it seems the inclusive behaves syntactically more like the exclusive rather than second persons. However, a look at the individual subcorpora (see Table C.9 and Table C.10 in Appendix C) show that this is again a point where the full corpus obscures noticeable differences: in the A-W subcorpus inclusive arguments fall between first and second person arguments with respect to whether they precede or follow the verb, while in the BT subcorpus, inclusive actors are more likely to follow the verb, at the approximate 40/60 ratio seen for arguments in general in the above tables. For the BT subcorpus, all SAP goals occur preverbally over 90% of the time, which is also obscured when only the full corpus data is presented.

5.4.2.3 Nominal type

A third approach to topicality is the type of nominal that occurs as an overt actor or goal, with full nouns or noun phrases being considered less topical than pronouns. As already noted above, when SAPs are realised, they must occur as pronouns (perhaps not categorically true of the language, but true of the parser that assigns the tags at this stage of development), so for SAPs, their different aspects of topicality are conflated morphosyntactically. Here, therefore, I explore the occurrence of overt third person arguments before and after verbs, focusing on two main distinctions: whether the arguments are animate or inanimate and whether the arguments are personal pronouns, demonstrative pronouns, or nouns. Animate arguments will include actors of VAIs, VTIs, and VTAs, as well as VTA goals, and inanimate arguments will include VII actors and VTI goals. This approach will necessarily obscure differences between VAI, VTI, and VTA actors, though these have been commented on above (§5.4.1). For inanimate arguments, as there are no personal pronouns to explore, the distinction is only between demonstrative pronouns and nouns. Other pronoun classes are not explored here, nor are differences in person, number, or obviation.[58] The percentages in these tables aim only to explore the different proportions of overt demonstrative and personal pronominal arguments and the occurrence of pronominal arguments compared to

[58] These factors and the interactions between them are better served by multivariate analysis, a future goal for the argument realisation data that can be extracted from the Plains Cree corpus.

overt nominal arguments; the totals are not those of all actors and goals, but only those for these classes.

Table 5.6: Actors and goals by animate nominal type: full corpus

	DEM		PERS		PRON		N		Total
		% PRON		% PRON		% Total		% Total	
@A	1,289	57.4%	956	42.6%	2,245	25.5%	6,576	74.5%	8,821
		% @A		% @A		% @A		% @A	
AV	679	52.7%	770	80.5%	1,449	64.5%	2,126	32.3%	5,024
VA	610	47.3%	186	19.5%	796	35.5%	4,450	67.7%	6,042
	DEM		PERS		PRON		N		Total
		% PRON		% PRON		% Total		% Total	
@G	598	78.7%	162	21.3%	760	16.9%	3,738	83.1%	4,498
		% @G		% @G		% @G		% @G	
GV	255	42.6%	127	78.4%	382	50.3%	1,140	30.5%	1,904
VG	343	57.4%	35	21.6%	378	49.7%	2,598	69.5%	3,354
		% @A/G		% @A/G		% @A/G		% @A/G	
Total preV	934	49.5%	897	80.2%	1,831	60.9%	3,266	31.7%	
Total postV	953	50.5%	221	19.8%	1,174	39.1%	7,048	68.3%	

Table 5.6 offers a look at animate third person arguments. Overt nominal arguments are more common than pronominal arguments (% Total)—though recall that no overt arguments is the most common pattern—and demonstrative pronouns are more often assigned argument tags than personal pronouns (demonstrative pronouns are more frequent than personal overall; see Chapter 4). Greater differences arise in comparing whether these precede or follow verbs: as for overall arguments (e.g., Table 5.1), it is generally more common for actors to precede verbs than goals, and here we see that pronouns are more likely to precede verbs than nouns. However, personal pronouns are by far the most likely of these categories to occur preverbally. Pronouns are more often actors than goals, and nouns are more often goals than actors. Unsurprisingly, the BT

subcorpus contains more demonstrative and nominal arguments than A-W (Table C.11 and Table C.12), as BT also contains more third person verbs.

Table 5.7 looks at inanimate actors (of VIIs) and goals (of VTAs); similar patterns to those for animate nominals are evident, with more pronouns preceding the verb and more nouns following the verb, pronouns more often occurring as actors than goals and nouns occurring more often as goals than actors. However, like Table 5.1 above, the distribution of inanimate arguments before and after the verb is closer to 50/50 than for animates. Between the two subcorpora (Table C.13 and Table C.14), BT generally contains more nominal arguments and fewer pronominal compared to A-W, as well as fewer VII arguments overall. Additionally, inanimate pronominal actors precede the verb far more often in BT, while the nominal actors follow more often than in A-W.

Table 5.7: Actors and goals by inanimate nominal type: full corpus

	DEM		N		Total A
		% Total A		% Total A	
@A	311	31.2%	686	68.8%	997
		% @A		% @A	
AV	188	60.5%	313	45.6%	501
VA	123	39.5%	373	54.4%	496
	DEM		N		Total G
		% Total G		% Total G	
@G	914	25.5%	2,667	74.5%	3,581
		% @G		% @G	
GV	480	52.5%	1,291	48.4%	1,771
VG	434	47.5%	1,376	51.6%	1,810
		% @A/G		% @A/G	
Total preV	668	54.5%	1,604	47.8%	
Total postV	557	45.5%	1,749	52.2%	

5.4.3 PAS: A corpus approach to arguments

Finally, I briefly examine the occurrence of arguments as subjects (of VAIs), agents (actors of VTIs and VTAs), and objects (goals of VTAs). I focus here only on whether or not an overt nominal occurs, not the type of the nominal (noun/NP or pronoun). In contrast to the English findings discussed in §5.3.2 above, where nouns and pronouns are contrasted, for Plains Cree I instead contrast any overt participant with no overt participant—as Plains Cree demonstrates verbal person marking, which may also be considered bound pronouns, I consider Plains Cree verbs without overt argument analogous to verbs with only overt pronouns in a language like English (Du Bois, 2003). I also focus only on animate third persons, both proximate and obviative, to simplify the situation regarding animacy and topicality differences. Table 5.8 looks at proximate actors and goals and Table 5.9 at obviative actors and goals. As noted previously, VTI and VTA actors occur as overt arguments at similar rates, while VAI actors and VTA goals occur more frequently. This is especially clear in Table 5.8 for proximate third persons: they occur as overt transitive actors(/agents) for 22% of VTIs and VTAs, but as intransitive actors(/subjects) of VAIs much more frequently—and as overt transitive goals(/objects) of VTAs at a similar rate to VAIs actors. The obviative third persons in Table 5.9 show similar trends, though the frequencies span greater ranges. Thus, in terms of argument realisation, these results suggest an ergative-like pattern of argument realisation in Plains Cree.

However, once again, differences between the subcorpora are obscured here. The general patterns hold, in that VTI and VTA actors are more similar to each other than to VAI actors or VTA goals, but the differences between VTA goals and VAI actors is greater. In the A-W subcorpus, proximate VTA goals are realised for 38% of verbs and VAI actors for 32%, and obviative VTA goals are realised for 47% of clauses and VAI actors for 36%. In the BT subcorpus, proximate VTA goals are realised for only 25% of verbs (i.e., more like transitive actors) and VAI actors for 34%, but the occurrence of overt obviative VTA goals is nearly identical between A-W and BT, at 40% and 41% respectively (Table C.15 through Table C.18).

Table 5.8: Proximate actors and goals for VAIs, VTIs, VTAs: full corpus

	VAI	13,984		VTI	5,238	VTA	7,689
		% VAI			% VTI		% VTA
S	4,675	33.4%	A	1,146	21.9%	1,674	21.8%
						VTA	5,918
							% VTA
	-	-	G	-	-	1,837	31.0%

Table 5.9: Obviative actors and goals for VAIs, VTIs, VTAs: full corpus

	VAI	2,257		VTI	423	VTA	1,939
		% VAI 3′			% VTI 3′		% VTA
S	888	39.3%	A	100	23.6%	596	30.7%
						VTA	5,828
							% VTA
	-	-	G	-	-	2,496	42.8%

5.5 Discussion

Even using only morphosyntactic features, the results in this chapter demonstrate two main patterns. First, less topical arguments tend to be realised more often: inanimate arguments more than animates, non-SAPs more than SAPs, obviative more than proximate, patients more than agents, and nouns more than pronouns. Second, when arguments do occur, more topical arguments tend to occur more often before the verb: SAPs more than non-SAPs, proximate more than obviative, agents more than patients, and pronouns more than nouns. These patterns have been described on the basis of smaller samples of Plains Cree (e.g. Dahlstrom, 1995; Cook et al., 2003; Mühlbauer, 2007; Wolvengrey, 2011), but a morphosyntactically tagged corpus can be used to quantify these patterns and approach them from several angles more efficiently. Corpus data can also be used to add to typological discussions about the nature of a language, such as alignment patterns—for Plains Cree, hierarchical, ergative, and accusative patterns have been argued with respect to morphological patterns (e.g. Campana, 1989; Hewson, 1987; Odribets & Oxford, in

press; Wolfart, 1973), though the exploration of syntactic patterns, in terms of tendencies in large bodies of text, have not yet been feasible. The Plains Cree corpus demonstrates an ergative-like pattern in terms of overt argument realisation, where the objects of animate transitive verbs are realised with similar frequency as the subjects of intransitive verbs, while the agents of transitive verbs are realised less often.

These are just two examples of how a corpus, even with only basic morphosyntactic features automatically assigned, can add further data to impressionistic or small-scale statistical descriptions of Plains Cree word order and act as a starting point for future research questions. As already mentioned, more involved statistical methods could be used to explore how the features discussed herein interact in future research. Number was all but ignored in this chapter, though future corpus investigations may find it in fact plays a role. Further tagging of the corpus, especially tagging (manual or automatic) for pragmatic features, and semantic classification for nouns and pronouns (e.g. Dacanay et al., 2021), can be combined with morphosyntactic features to explore patterns in topicality and word order more accurately.

5.6 Conclusion

In this chapter, I have presented some aspects of variation in argument realisation in a Plains Cree corpus, exploring how actors and goals are realised with respect to verb class, nominal class, person, and more. This large-scale approach using only basic tags has allowed for an effective description of syntactic variation with respect to not only morphosyntax but also topicality—in accordance with previous descriptions, more topical entities (whether SAPs, pronouns, or agents) occur preverbally more often than their less topical counterparts. In the future, such results can be improved in a number of ways, not only through improvements to the parser underlying the tags and the inclusion of more involved statistical methods to better investigate the interactions between the many features under consideration, but also through further manual tagging of the corpus to improve our understanding of the pragmatic features. The inclusion of features such as whether animate entities are human or non-human, conceptually animate or inanimate (cf. Chapter 2, some grammatically animate nouns in Cree are not living beings, in at least the Western sense), whether an entity is newly introduced or previously mentioned, and whether the topical entity has shifted

can be combined with the morphosyntactic features already available to further explore what influences argument realisation and word order in Plains Cree.

The results in this chapter offer only a glimpse of the sort of data that can be extracted from a morphosyntactically tagged corpus, to explore variation in argument realisation patterns and contribute larger datasets to discussions that have previously smaller datasets. These results, in conjunction with the additional tables in Appendix C, also once again highlight the ways in which the two subcorpora can differ, sometimes drastically, and thus further research into their differences is required. Thus, in Chapter 6, I explore some of these differences in a text type analysis for Plains Cree: the variation seen in different text types, which differ considerably between the subcorpora, may further explain differences in argument realisation in future investigation.

Chapter 6

Plains Cree text type analysis of a morphosyntactically tagged corpus: A first look at registers with Principal Component Analysis

6.1 Introduction

In the previous chapters, I have detailed the ongoing construction of a Constraint Grammar-based parser for Plains Cree, applied this parser to a morphologically verified corpus of Plains Cree, and explored argument realisation within the Plains Cree corpus. Exploring the Plains Cree corpus highlights similarities and differences between the Bloomfield and Ahenakew-Wolfart subcorpora: for example, relative frequency of word classes and person features differ between them (cf. Chapter 4), as do word order patterns (cf. Chapter 5). Furthermore, an exploration of archaic morphology (Schmirler, in press) demonstrates that there can be significant differences between the corpora with respect to the frequency of some morphological features. When considering a corpus of Plains Cree, one must wonder to what extent these internal differences are obscured when considering the corpus as a whole, and in what ways, including and beyond the division of the subcorpora by editors, the corpus can be considered for a fuller picture of Plains Cree morphosyntax. Thus, in this chapter, I undertake a register analysis using this morphosyntactically tagged corpus of Plains Cree.[59]

Register analysis, as laid out by e.g., Biber (1991) and Biber & Conrad (2019), approaches texts as a combination of features and context and explores how these might be linked. One means of exploring corpora with many dozens or hundreds of features across many texts is to use a dimension reduction technique, such as Principal Component Analysis (PCA). While previous chapters herein have exemplified some differences between the subcorpora, these have generally

[59] A version of this chapter has been accepted for publication in the Papers of the 53rd Algonquian conference as Schmirler & Arppe (forthcoming), presented as Schmirler & Arppe (2021).

been small snapshots of individual patterns or phenomena that do not and, due to the scale of the question, cannot take into account the many individual texts and feature co-occurrences therein. With an approach such as PCA, co-occurrences and broad patterns within and between various texts within the Plains Cree corpus can be explored, allowing for a large-scale quantitative look at text types and morphosyntactic features in Plains Cree.

In §6.2, I describe the process of undertaking text type analysis in Plains Cree, beginning with a description of Plains Cree text types (§6.2.1), then an introduction to register analysis and PCA (§6.2.2 and §6.2.3 respectively), followed by the methods used in the current chapter (§6.2.4). Section 6.3 presents the results of text type analysis for the full Plains Cree corpus (§6.3.1), the BT subcorpus (§6.3.2), and the A-W subcorpus (§6.3.3). The analyses are followed by a general discussion (§6.4) and the chapter concludes in §6.5.

6.2 Undertaking text type analysis with a corpus of Plains Cree

In contrast with corpora for majority languages like English, the Plains Cree corpus, though analysed in written form, is entirely transcribed from spoken language. English corpora tend to favour written texts, thus most corpus investigations examine written texts: consider the Corpus of Contemporary American English, with approximately 1.27 million words of spoken English, out of over one billion total words (Davies, 2008-), or the British National Corpus, with a much higher proportion, approximately ten million spoken words out of one hundred million (*The British National Corpus*, 2001). Any investigation of the current Plains Cree corpus only examines transcriptions of spoken language, though written texts may be added in the future. A number of corpora for Indigenous languages were described in Chapter 1—the Plains Cree corpus, though not the largest (e.g., Inukitut, with several millions of words), is still larger than those for many other Indigenous languages.

Studies of texts in other Algonquian languages tend to look at a single text, and thus a much smaller number of words, than is possible with a tagged corpus of ~152,000 Plains Cree words. The works in Costa (2015) present linguistic and historical analyses of stories collected for various Algonquian languages, though the analyses are either restricted to the issues that arise in transliterating texts originally collected in earlier centuries, or descriptions of individual syntactic

phenomena using a small number of examples from the text at hand. The works in Costa (2015) do not present details such as the number of words in the texts—descriptive statistics are not the purpose of these studies; rather, the authors are exploring smaller bodies of text in great detail. One of the few instances where numbers are presented is in Brockie & Cowell (2015), where they report 80 unique verb stems in the Gros Ventre text under consideration. In comparison, the Plains Cree corpus has over 5,000 unique verb stems. While the works presented in Costa (2015) are undeniably valuable contributions to the field, they cannot be compared in scale with the studies that are possible using computational tools and many thousands of words. However, such smaller-scale studies as those in Costa (2015) often involve the contributions of speakers of the languages to the transcription, translation, and analysis. As a non-Cree linguist, my understanding of the purposes and features of various text types has only been gleaned from descriptions in the forewords of the volumes that comprise the current Plains Cree corpus; this is insufficient for in-depth textual analysis. In this vein, further expansion beyond the brief text analysis in this dissertation will require interviews with Cree speakers about the purposes of texts and features they include in their own storytelling.

6.2.1 Plains Cree text types

The texts in the A-W corpus represent a variety of text types and subtypes, as identified by the editors and described in prefacing material. The primary distinction between text types in Cree is between *âtayôhkêwina*, which are sacred stories, namely those that take place before the world was "in its present, definitive state" (Bloomfield, 1930, p. 6), and *âcimowina*, which are everything else, including discourses and narratives. Freda Ahenakew comments on some subcategories of *âcimowina* in the preface to Vandall & Douquette (1987)[60]: *kayâs-âcimowina* 'old-time stories' (the distant past, but not sacred stories), *kakêskihkêmowina* 'counselling texts' (especially commentary on the distinction between traditional Cree life and modern life), *wawiyatâcimowina* 'funny stories', and *âcimisowina* 'stories about oneself' or 'personal stories' (which can also be retellings of other people's life stories). Some further distinctions can be made on the basis of the type of text, though not necessarily ones presented by the speakers. For example, the *âtayôhkêwina* in the corpus include a Cree retelling of the story of Aladdin, which may demonstrate some

[60] Not all of these types are included in the A-W corpus.

different features from traditional legends; similarly, the corpus includes extended conversations which are identified only by the English term *dialogues*, which might be translated as *pîkiskwâtitowina*—this will be investigated as its own text type and compared with the others.

Table 6.1: Text types of the Plains Cree corpus, including rough volume/chapter divisions

Text type	Volumes/Chapters
âtayôhkêwina	*Sacred Stories of the Sweetgrass* Cree (Bloomfield, 1930) *Plains Cree Texts*, Part IV (Bloomfield, 1934)
âcimowina	*Plains Cree Texts*, Parts I, II, III (Bloomfield, 1934) *Stories of the House People*, Chapter 10 (Vandall & Douquette, 1987) *They Knew Both Sides of Medicine*, Chapters 11?, 12? (Ahenakew, 2000)
âcimisowina	*They Knew Both Sides of Medicine*, Chapters 1, 2, 3, 4, 5, 6?, 7?, 8, 9, 10, 11?, 12? (Ahenakew, 2000) *Stories of the House People*, Chapters 8, 9 (Vandall & Douquette, 1987) *Our Grandmothers' Lives as Told in Their Own Words*, Parts I & II (Bear et al., 1998) *piko kîkway ê-nakacihtât* (Masuskapoe, 2010) *Their Example Showed Me the Way* (Minde, 1997)
wawiyatâcimowina	*Stories of the House People*, Chapters 5, 6, 7 (Vandall & Douquette, 1987) *They Knew Both Sides of Medicine*, Chapters 6?, 7? (Ahenakew, 2000)
kaskêskihkêmowina	*The Counselling Speeches of Jim Kâ-Nîpitêhtêw* (Kâ-Nîpitêhtêw, 1998) *The Cree Language is Our Identity* (Whitecalf, 1993) *Stories of the House People*, Chapters 1, 2, 3, 4 (Vandall & Douquette, 1987) *Their Example Showed Me the Way* (Minde, 1997)
pîkiskwâtitowina	*Our Grandmothers' Lives as Told in Their Own Words*, Part III (Bear et al., 1998)

Ahenakew (2000) includes several narratives, generally of the speaker's life and events in the lives of people close to her. Thus, these might be classed as *âcimowina*, some of which are *âcimisowina*. Bear et al. (1998), presents a group of personal narratives followed by a set of dialogues; as the title *Our Grandmother's Lives as Told in Their Own Words* suggests, the narrative sections of this book can be classed as *âcimisowina*, while the dialogues might be called *pîkiskwâtitowina*. The editors note that while these texts present much more intimate, unstructured narrative than other collections in the A-W corpus, they do have some small sections that are more similar to *kakêskihkêmowina*, in that they comment on the differences between past and present ways of life for Cree people.[61] Kâ-Nîpitêhtêw (1998) is a collection of *kakêskihkêmowina*. Unlike many of the other volumes in the A-W corpus, these speeches were not recorded in small sessions, but were publicly delivered. Some portions of the text also have some aspects of *âcimowina*, in that they present a report of some aspect of Cree life or history. Masuskapoe (2010) is a collection of *âcimowina*, especially *âcimisowina*, about the speaker's life, with some aspects of *pîkiskwâtitowina* where the speaker is in dialogue with others present for the recordings. Minde (1997) is described by the editors as a collection of *âcimowina* that alternates between *âcimisowina* and *kakêskihkêmowina* (p. xv), though some sections of the text are recounting stories told to the speaker by others about their own lives. Vandall & Douquette (1987) contains some *kakêskihkêmowina*, followed by *wawiyatâcimowina*, *âcimisowina*, and one *âcimowin*; these are clearly labelled by the editor (p. xii-xiii). Whitecalf (1993) contains a collection of *kakêskihkêmowina*, though the editors note that these are sometimes illustrated using examples from the speaker's life. Bloomfield divides his texts into *âtayôhkêwina* and *âcimowina*. *Sacred Stories of the Sweetgrass Cree* (SSSC; Bloomfield, 1930) contains a collection of *âtayôhkêwina* and *Plains Cree Texts* (PCT; Bloomfield, 1934) contains both *âtayôhkêwina* and *âcimowina*. In PCT, some stories might be classed as *kayâs-âcimowina*, as they describe old ways of living and past events.

As this description of the volumes and their text types demonstrates, there are rarely clear-cut distinctions. This is not unusual: consider an anecdote (a narrative) told as part of a larger conversation—both text types can and do occur together. The types can be summarised as in Table

[61] One speaker in this collection, representing two chapters, is a speaker of Woods Cree; her portions are thus excluded from the present corpus of Plains Cree.

6.1; subtypes of *âcimowina* are not necessarily straightforward to identify, and thus some occur in different categories marked with a question mark. Each of these volumes, regardless of the identified text types within, are divided by chapter for the analysis.

6.2.2 Register analysis approach

Register analysis is one approach to textual analysis. Register analysis focuses on linguistic features in combination with the situational context in which the text is produced, and how these features function for the communicative purpose of the text. In contrast, genre analysis is concerned with conventions within a text type, such as the beginnings and ends of letters and stories (*Dear Sir/Madam*, *Sincerely*; *once upon a time*, *the end*). Style analysis, like register analysis, focuses more on linguistic features than conventions; however, where register is concerned with the function of a text, style more is concerned with aesthetic or other preferences, which can be associated with the author or speaker (e.g., Biber, 1991; Biber & Conrad, 2019). Any of these three approaches might be used to describe and analyse the text types in the Plains Cree corpus: the various Plains Cree text types described above can be associated with features, and these can be functionally motivated; the narratives in the BT subcorpus in particular demonstrate genre conventions, such as formulaic opening and closing lines in stories; individual speakers may have particular stylistic ways of speaking. However, as Plains Cree speakers already have a set of words that label the content and function of various texts, I have chosen a register analysis approach for this work.

Register analysis, as laid out by Biber and colleagues (e.g., Biber, 1991; Biber et al., 1998, 2002; Biber & Conrad, 2019), involves three main steps. The first step is a description of the situational context: who is speaking, and to whom; what is their relationship; what is the purpose of the interaction? The second step is to look at the relative frequencies of various linguistic features in each type of text: certain features will be more pervasive in certain types of texts. For example, a cursory exploration of texts in Plains Cree finds that personal stories demonstrate more first person exclusive verbs (‘we, but not the addressee’), while lectures demonstrate relatively more first person inclusive verbs (‘we, including the addressee’) (Schmirler & Arppe, 2020). The third step involves identifying functional relationships between the context and the linguistic features: why does a certain text type involve certain features? For the different first persons, speakers sharing

personal stories are talking about their own lives, their families, etc., which do not include the listeners, while speakers giving lectures are talking about shared Cree identity and actions taken by the community, increasing the use of the inclusive forms. Importantly, the notion of *text* includes not only those bodies of text that began as written words, but also those that are transcribed from spoken language.

The situational context provides a categorisation of the types of texts involved. These categories can be approached in different ways: the Plains Cree corpus has two broad sets of texts, those recorded in the earlier part of the twentieth century (the BT subcorpus) and those from the latter part (the A-W subcorpus)—a chronological distinction. The different types of texts as laid out above, *âtayôhkêwina* 'legends' and *âcimowina* 'stories', as well as the various subtypes of *âcimowina*, are another means of dividing the texts, and one used primarily for the analyses in this chapter. The above description also includes other aspects of the situational context: sometimes the speakers speak to (or with) the person recording, sometimes to other people present, and sometimes at public events.

Linguistic features in the present study fall into several categories: verbal (transitivity, person, tense, etc.), nominal (gender, number, possession, etc.), and adverbial (expressed in Cree through particles, including categories such as time, location, negation, etc.). The frequencies of various linguistic features are presented as normed frequencies; that is, the number of occurrences of a particular feature per a set number of words, e.g., 100 or 1000. These normed frequency can then be compared to determine the relative frequency of these features in different text samples. After relative frequencies are determined, the connection between these features and context of the text can then be investigated. For example, lectures contain more imperatives that direct listeners to act in a certain way, while narratives contain more past tense verbs that describe past events. In dialogue, second person singular verbs are more frequent than in narrative, while third person singular verbs, both animate and inanimate, are more frequent in narrative than in dialogue. Not all features may display significantly different frequencies between texts, and those that differ are not necessarily motivated by text type, but may be better attributed to e.g., speaker differences. Finally, Biber & Conrad (2019) stress the importance of the cyclical nature of this type of analysis: after the features of the text and their relevance to the text type are discussed, one then returns to

the situational context and purpose of the text, re-evaluating on the basis of the linguistic features and their proposed importance.

As a full register analysis is beyond the scope of the present chapter, two main elements of register analysis I use herein are 1) the notion of situational context and 2) the cyclical nature of the analysis. Thus, the importance of features to the text type, and how and why a speaker imparts information, are discussed throughout the analysis, and the features feed back into determining distinctions beyond the labels traditionally used in Plains Cree.

6.2.3 Principal Component Analysis

Principal Component Analysis (PCA) is a statistical technique that reduces the dimensions of large datasets—particularly sparse datasets. This describes the present data well: there are 140 chapters in the full corpus, and up to 87 morphosyntactic features included, not all of which occur in every text, and many of the features are necessarily correlated, as they can only apply to certain word classes, for example. The simplification of large datasets allows for the most important independent variables to be indentified by creating new statistically independent variables that aggregate the values of the original dataset, calling these new variables principal components (PCs). PCs are created as combinations of all independent variables in the data, so as many PCs are generated as original variables, created so that each PC is independent of any others. PCs are then ranked by how well they account for the variance in the data, with the first accounting for the largest proportion of variance, the second the next largest proportion, etc. The dependent variables, in this case the chapters of the texts, can then be plotted along the PCs, which allows for visualisation of similarities and differences among the texts. Though PCs can be less interpretable than the original variables, for large datasets, the simplification can instead help to visualise the relationships, especially for determining which features and interactions one might include in future statistical analyses, such as logistic regression.

Due to these qualities, dimension reduction techniques like PCA have been used to analyse natural language data, such as that found in a corpus, and, in particular, for register analysis (e.g., Biber, 1986; Biber et al., 2006). Unsurprisingly, datasets drawn from corpora can be very large, as there are many possible linguistic features, many words in which those features can occur in

combination, and many different features of texts that one may want to account for. Reducing the dimensions of such a dataset is invaluable, as hundreds or thousands of features and their interactions are not easily interpreted. The resulting PCA can then be used to see how texts differ, and how they group with respect to similar features, and then any groups can be linked to existing categories of Plains Cree texts; this is in contrast to Schmirler & Arppe (2020), which explored the differences only between a small number of personal narratives and counselling speeches.

Each PC consists of negative and positive weightings from -1 to 1, which are associated with the original variables (i.e. the morphosyntactic features) that characterise different portions of the data; in this case the data are the texts (the individual chapters), and the features are morphosyntactic features. The texts can then be plotted, with PC1 on the *x*-axis and PC2 on the *y*-axis, based on the degree to which they are characterised by the features, either towards the positive end or the negative end, giving a two-dimensional perspective of the morphosyntactic similarities and differences among the texts. I again return to the contrast between lectures and personals stories to exemplify this: in such an analysis, first person inclusive would be weighted towards one end of the PC and exclusive towards the other. Texts that are characterised more by inclusive, i.e., lectures, would then appear on that side of the plot, and texts characterised by exclusive, i.e., personal stories, on the other side. I will primarily focus on the first and second PCs in the following analyses. Throughout the analyses herein, the first two PCs together account for approximately 20% of the variation in the data. The third PC generally accounted for about the same proportion of variance as the second, however, as two PCs are more easily visualised and discussed, I do not explore the third PC here.

6.2.4 Method

The present work approaches textual analysis from a register analysis perspective, though with some concessions to complexity for the sake of brevity. Biber & Conrad (2019) provide a suggested list of features to be included for register analysis; I have adapted this to Plains Cree, suggesting similar features that could be used in future analyses (see Appendix D for this suggested list of Plains Cree features that may be used in a full register analysis, as well as an indication of the categories included in the datasets below). For the purposes of this small-scale investigation, only morphosyntactic features that are present in the corpus are considered. Many higher-level

features, such as the type-token ratio, the number of multifunction words, or the occurrence of hapax legomena, are excluded, as well as the co-occurrence of morphosyntactic features (e.g., verbs and actors are considered separately, while a fuller analysis would take into account how many verbs occur with actors, and whether the actor precedes or follows the verb, etc.). The analysis herein is undertaken in R (version 4.1.0; R Core Team, 2021) using the built-in prcomp function.

6.2.4.1 Data

The dataset used to examine text types in Plains Cree are drawn from the morphosyntactic tags present in the morphological gold standard and added by the CG-based parser. Here, the disambiguation and syntactic function assignment modules are augmented by an additional, ad hoc module that groups various features into higher-level tags: verbs, nouns, and pronouns with first person plural tags are marked as inclusive and exclusive as appropriate; verbs are marked as direct and inverse, as well as local, nonlocal, and mixed based on their combinations of person tags. After the application of the parser to the MGS corpus, non-Cree tokens are removed (English, French, punctuation, etc.). This choice is motivated by the previously noted differences between the two subcorpora, e.g., far more non-Cree words and punctuation are present in A-W than in BT: the present investigation focuses on how Plains Cree features differ between texts and text types, rather than how the editing style of the texts is reflected in PCA, which might otherwise obscure the differences in the Cree features.[62] Additionally, the future conditional, which is represented by two tags, `Fut+Cond`, is reduced to just `Cond` so that the future tag refers only to the use of future preverbs. After these forms were removed and adjusted, the tags were converted into a table where each row represented one chapter of the corpus, identified by speaker initials, chapter number, and in the case of BT the volume name, as speakers overlapped, and each column represented a morphological feature tag or syntactic function. Each cell then represents how many times a particular feature tag occurs in a text. This method does not consider how different morphological features interact (e.g., how many VAIs are first person singular, how many actors are nouns vs.

[62] Future investigations would benefit from considering how much of the English in A-W occurs in introductions and notes, as opposed to code-switching or other use of English within the spoken Plains Cree. While some of the A-W speakers were monolingual, most had some knowledge of English and so it is unsurprising that some English words, especially people and place names, appear in their Cree speech.

pronouns)—the individual features allow for somewhat more interpretable results, though interactions between features should be included in future analyses.

After the table was created, further individual tags (rather than whole words and all their features) were trimmed; these were primarily metalinguistic tags included for individual words, such as identifying morphological model errors (e.g., due to orthographical variation or morphological phenomena not fully incorporated into the model), marking rare features of interest for easier extraction later (e.g., diminutive VTAs), or indications that the analysis is uncertain and should be checked at a later time. Additionally, among the syntactic functions that are included in the parser, only actor and goal tags are retained for the purposes of this small-scale investigation. This choice removes negation of nouns, verbs, and particles, the occurrence of prepositional phrases, quantifiers modifying nouns and verbs, and temporal and locative particles modifying nouns and verbs. Instrument and oblique function tags are also removed, as they are not yet fully implemented in the parser. Initial modelling found several related features that ranked similarly within the same component, e.g., future would group with its subtypes, definite and intentional, indicating that the future tense overall, rather than a particular type of future, characterised certain portions of the texts, thus the subfeatures were trimmed. Similarly, immediate and delayed imperatives co-occurred with the overall imperative tag and so were trimmed, leaving only the imperative.

In applying PCA, the raw counts are transformed to relative frequencies, dividing the number of occurrences of each feature in a chapter by the total number of words, thus achieving normed frequency for dimension reduction techniques in register analysis (e.g., Biber, 1986; Biber et al., 2006). The analysis below is presented in three stages. The first of these includes the full corpus, the second just the BT subcorpus, and the third just the A-W subcorpus. This approach demonstrates similarities and differences between the subcorpora and as well as within them; the by-chapter division of the texts also demonstrates similarities and differences between different texts produced by the same speaker.

6.3 Exploring Plains Cree registers with PCA

6.3.1 The Plains Cree corpus

6.3.1.1 Results

The full Plains Cree corpus consists of 140 individual chapters from 20 speakers, drawn from nine volumes. Before preprocessing the morphosyntactically tagged corpus for PCA, the full Plains Cree corpus contains 241,152 tokens (37,941 types), including non-Cree words, punctuation, and metadata. After preprocessing, these are reduced to 152,405 Plains Cree word tokens (31,616 types). The dataset constructed from this preprocessed data includes 571 feature tags, which are trimmed to 87 for the purposes of PCA. Of the 484 removed, 378 (78.1%) were preverb or prenoun tags.

The first two PCs account for 17.72% and 7.65% of variance respectively, totaling 25.37%. These PCs are visualised in Figure 6.1, which plots the texts of the full corpus along the first and second principal components as reported in Table 6.2 below, with the distribution of each subcorpus shown by ellipses. As the two subcorpora group with the zero point of the *x*-axis nearly dividing them, it is then possible to say that each subcorpus is characterised by the positive (A-W) and negative (BT) features respectively along PC1. For PC2 along the *y*-axis, both subcorpora contain features from the positive and negative ranges. Even without reference to features, we can observe that the subcorpora, while different in some ways, are similar in others. For further comment on the distribution of the subcorpora and its relationship to the morphosyntactic features, see the interim discussion in §6.3.1.2.

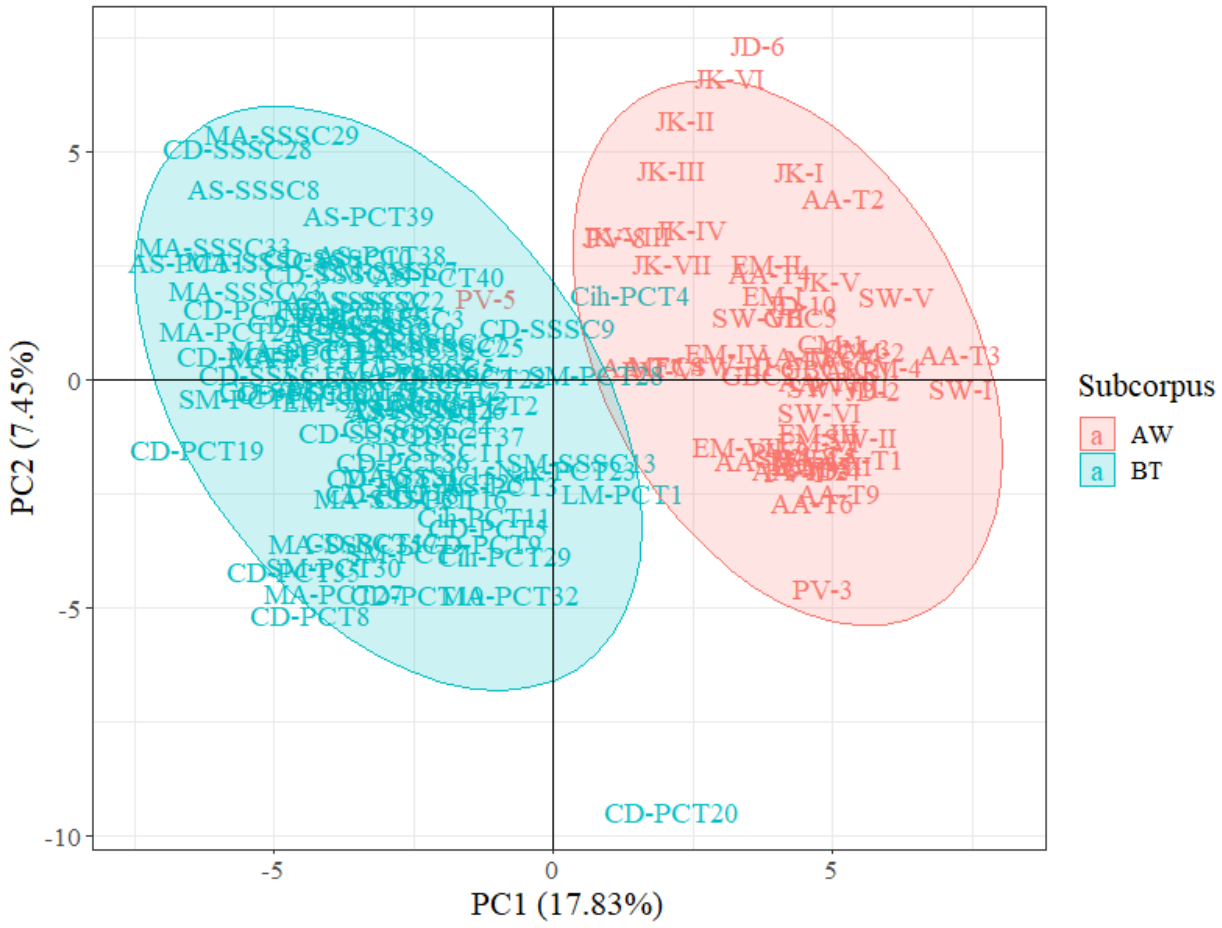


Figure 6.1: Full corpus PCA: Chapters along PC1 & PC2

Table 6.2 presents a selection of the top 25 positively and negatively weighted features for each of the first two PCs. These are not necessarily presented in order of the contribution to the PC weighting, but are instead listed, where possible, in pairs or sets of mutually exclusive or otherwise related features across the positive and negative sides of each PC (e.g., word class, tense, verbal order, person features, other verbal and nominal features). Not all features in the results are included here for the sake of brevity.[63]

[63] Full tables for the three analyses in this chapter, including the weights of features in each PC, are given in Appendix E.

Table 6.2: Full corpus PCA: Positive & negative features

	Positive	Negative
PC1	· Particles, pronouns · Past tense · Conjunct verbs · EXCL, 0SG, 1SG, PX21PL, X, INCL, PX1PL · Mixed VTAs · Inanimate, plural nominals	· Verbs · Present, future tense · Independent, imperative verbs · 3′, 3′-G, PX3SG, 3SG, PX3′, 0′SG, PX2SG, 2SG · Nonlocal, direct, local VTAs · Obviative, dependent, animate, vocative nominals · @<ACTOR
PC2	· VTAs, quotative verbs, personal pronouns · Future tense · Imperative, future conditional verbs · 1SG, 1SG-G, 3SG-G, 2SG-G, PX1SG, 2SG, PX21PL, INCL, PX1PL, 0SG, PX2SG · Mixed, local, inverse VTAs · Singular, dependent nominals · Interrogative, focus particles, interjections	· VAIs, particles · Past tense · 3′-G, PX3PL, PX3′, 0′SG, 3′-G, 3′, PX3SG · Nonlocal VTAs · Plural, obviative, animate nominals, locative, proper nouns · Numerals, quantifiers, temporal particles, indeclinable nominals · @ACTOR>, @GOAL>

In PC1, there is a contrast between more particles and pronouns on the positive side (i.e., A-W) and more verbs on the negative (i.e., BT). Within word classes, especially verbs, we see the occurrence of past tense and conjunct forms contributing to a positive score along PC1, and the occurrence of present, future, independent, and imperative verbs contributing to a negative score along PC1; mixed VTAs on the positive and nonlocal, direct, and local VTAs on the negative; first persons (both singular and plural) on the positive side, and third person (both proximate and obviative) and second person singular on the negative. There are inanimate and plural nominals on the positive side, and obviative, dependent, animate, and vocative nominals on the negative. Many of these features can be interpreted in groups on each side of PC1 as well, especially the negative: we see the occurrence of nonlocal verbs alongside third person proximate and obviative

tags, as well as more nominals with animate features, such as obviatives, dependent nouns, and vocatives. These features co-occur with the prevalence of overt actor tags. The negative side also shows a prevalence of imperative verbs and second person tags, which likely occur together, and, despite the relative lack of speech act participants, local VTAs as well—perhaps due to the overall prevalence of verbs here. Groups are less clear on the positive side, though a prevalence of first persons may be connected to the prevalence of mixed VTAs, and the prevalence of inanimate person tags to the prevalence of inanimate nominals.

While PC1 highlights differences between the subcorpora, PC2 highlights similarities in their internal variation. In PC2, there is a contrast between VTAs, quotative verbs, and personal pronouns on the positive side (the top half of the plot), with VAIs and particles on the negative. Future tense contrasts with past, as for PC1, but while imperative and future conditional verbs characterise the positive side, no verbal order characterises the negative. There is a contrast in persons as well, though now more strongly divided between SAPs and non-SAPs, though 3SG-G does occur alongside the SAPs. Again, the person contrasts align with VTA feature contrasts: where there are SAPs, mixed and local VTAs also occur, in contrast to third persons and nonlocal VTAs. Obviative and animate nominals again characterise the negative side, alongside third persons and overt actors and goals. However, dependent nouns now occur alongside the first and second persons: likely the possessors of these nouns. Despite these differences, a comparison of the feature lists for each PC also demonstrates that the same features appear in both.

6.3.1.2 Interim discussion

While the plot in Figure 6.1 demonstrates a clear division between the subcorpora, the distinction only occurs along PC1. PC1 accounts for a considerable degree of variance in the texts, thus suggesting that they do contain different features and feature co-occurrences, but PC2 demonstrates almost complete overlap of the two subcorpora. Indeed, similar features distinguish the texts along both PCs. Pronouns and particles appear in both PCs, as do future tense and imperative verbs. Similar persons appear repeatedly in their various roles, first person singular, second person singular, third person proximate singular and obviative: as actors, pronoun tags, goals, and possessors. Person features occur in similar combinations as well, contrasting more SAPs with more non-SAPs. In both PCs, the person features also co-occur with broader VTA

features: where SAPs occur, mixed and local VTAs occur, where non-SAPs occur, nonlocal VTAs occur. This pattern carries through the VIIs as well; 0SG tags occur with SAPs and 0′SG tags with non-SAPs. Finally, nominal features also occur in both PCs, especially animate, obviative, and plural nouns and pronouns—these also generally co-occur with third persons, unsurprisingly—as well as dependent nouns (e.g., kinship terms). Thus, many of the features and groups that separate the subcorpora also explain their internal variation—a closer look at the subcorpora is in order.

The similarities and differences between the subcorpora have been of ongoing interest in the development of the morphosyntactic models: it is wonderful to have a 152,000-word tagged corpus of an Indigenous language, but to what extent does considering this as one group of texts obscure the internal differences? Previously, in Chapters 4 and 5 respectively, these similarities and differences have been explored with respect to the frequency of morphosyntactic features and argument realisation patterns, finding in some cases consistency between the subcorpora, and in others rather noticeable differences. A PCA analysis appears to be no different: though major differences are found between the subcorpora, similarities are there as well.

Despite the similarities between the PCs, their differences bring to light two major distinctions: that between the content of subcorpora (PC1), and that between what I interpret as narrative with dialogue and that without (PC2). Recall that BT consists entirely of narrative, while A-W contains a combination of narratives, dialogues, and lectures. In PC1, the features that distinguish between the two subcorpora align with the content: on the negative, the non-SAPs and related features (nonlocal verbs, animate nominals) point towards narrative, especially more third person narrative, while on the positive, SAPs and related features point towards at the very least a different kind of narrative, with more first and second person features.

Many of these features also lend themselves to an interpretation of dialogue vs. non-dialogue narrative within PC2: where SAPs and related verbs (e.g., local, mixed) occur, so too do quotative verbs, imperative verbs, interrogative particles, and interjections. Quotative verbs, introducing quotation, are a clear clue here—one might expect that within that direct quotation, first and second persons address each other and interact, issue commands, ask questions, express surprise. Conversely, where non-SAPs occur, along with their related features of nonlocal verbs, overt actors and goals, etc., the accompanying features, such as numerals, temporal particles, nominal

features (e.g., nominals that accompany third person verbs), and proper nouns point towards a lack of quotation. As this distinction is more apparent along PC2, which explains more internal variation within the subcorpora rather than variation between them, features like this are also of interest in analysing each subcorpus below.

6.3.2 The Bloomfield subcorpus

6.3.2.1 Results

The BT subcorpus consists of 82 chapters from eight speakers, drawn from two volumes. Before preprocessing the morphosyntactically tagged corpus for PCA, BT contains 102,962 tokens (15,287 types), including non-Cree words, punctuation, and metadata. After preprocessing, these are reduced to 72,475 Plains Cree word tokens (15,267 types). The dataset constructed from this preprocessed data includes 354 feature tags, which are trimmed to 84 for the purposes of PCA. Of the 270 tags removed, 187 (69.26%) were preverb or prenoun tags.

For BT, the first two PCs account for 11.88% and 9.04% of variance respectively, totaling 20.92%. These PCs are visualised in Figure 6.2, which plots the texts of the BT subcorpus along the first and second principal components as reported in Table 6.3 below, with the distribution of each volume shown by ellipses. The volumes are used here as a rough approximation of text type, as *Sacred Stories of the Sweetgrass Cree* (SSSC, Bloomfield, 1930) is entirely sacred stories and *Plains Cree Texts* (PCT, Bloomfield, 1934) is mostly non-sacred stories with a small section of sacred stories. Unlike for the full corpus, the plot does not present any immediately apparent division between the volumes. However, the distribution of SSSC is almost entirely within that of PCT, suggesting that sacred stories behave as a subtype of narrative.

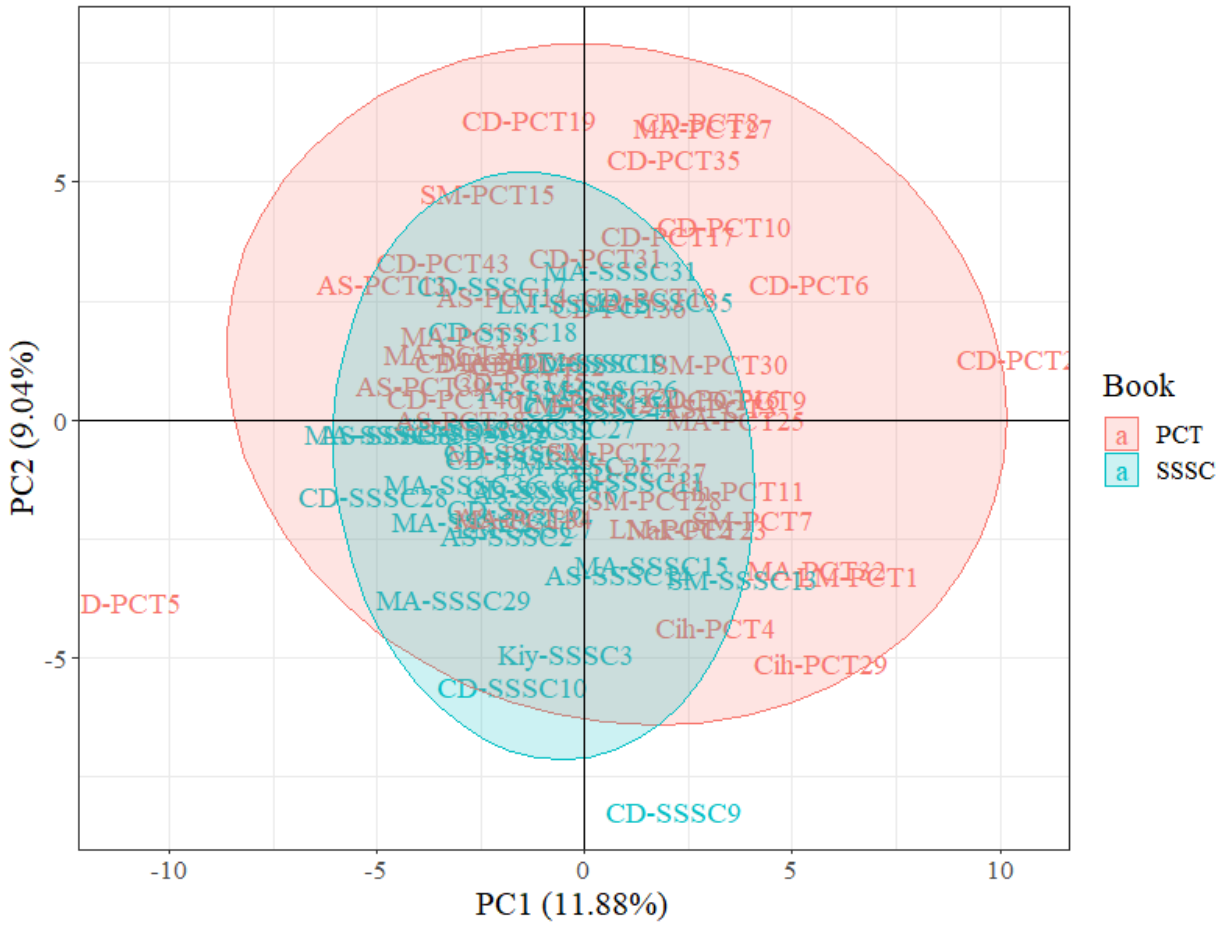


Figure 6.2: BT subcorpus PCA: Chapters along PC1 & PC2

Table 6.3 presents a selection of the top 25 positive and negative features for each of the first two PCs. Where possible, they are arranged in pairs or sets for reference, as for the full corpus results. In contrast to the full corpus, the BT subcorpus varies less clearly in terms of word class. For PC1, verbs in general, as well as VTAs and quotatives specifically, contrast with particles, though different classes of particles still characterise each side of the PC. Similarly, in PC2, VIIs and VTIs contrast with VTAs. However, in PC2, particles characterise the negative side, alongside a subset of particle classes, namely particle phrases, temporal and interrogative particles, and indeclinable nominals. In the positive range, pronouns in general and demonstrative pronouns in particular characterise the texts. As above, the persons listed often occurred in multiple roles within each PC, though not always.

Table 6.3: BT subcorpus PCA: Positive & negative features

	Positive	Negative
PC1	· Particles, VIIs · Past tense · Conjunct verbs · PX3PL, 0′SG, PX3′, EXCL · Plural, inanimate nominals, proper nouns · Quantifiers, numerals, locative, temporal, negative particles, indeclinable nominals · @ACTOR>, @GOAL>	· Verbs, VTAs, quotative verbs · Future tense · Future conditional, imperative, independent verbs · PX1SG, 2SG-G, 1SG-G, 3SG-G, 1SG, 3SG, 2SG, PX3SG, PX2SG, PX1PL · Mixed, local, inverse VTAs · Dependent, singular, vocative nominals · Interjections
PC2	· VTAs, pronouns, demonstratives · Conjunct verbs · 3′-G, PX3SG, 3′-G, PX3′, 3′, PX2SG, PXX · Nonlocal, direct VTAs · Obviative, animate, dependent, plural nominals · Numerals · @ACTOR>, @<GOAL, @GOAL>, @<ACTOR	· Particles, VIIs, quotative verbs, VTIs, verbs · Future tense · Independent · 1SG, 0SG, 0′SG, 1SG-G, PX1SG, PX21PL, 2SG-G, Excl · Local, mixed VTAs · Inanimate nominals · Particle phrases, interrogative, temporal particles, indeclinable nominals

The feature pairs in the table again serve to illustrate the differences along the axes of the plot in Figure 6.2. In PC1, there is a general contrast between particles and verbs, as well as different tenses, verbal orders, and persons. Unlike the full corpus, here there is no clear-cut distinction between SAPs and non-SAPs, though there tend to be more SAPs on the negative and more non-SAPs on the positive. The mix of persons is again highlighted in the negative range with especially mixed and local verbs; however, these features alongside the prevalence of direct and inverse is

likely also tied to the increased prevalence of VTAs in general (cf. Chapter 4). Differences in nominal features may also illuminate different features of the texts: on the negative side, both dependent and vocative nouns occur, suggesting these may be frequently used together. The occurrence of quotatives here suggests more direct quotation, within which dependent nouns, vocative forms, and SAPs likely often occur. In contrast, the positive range is characterised by 3PL and 3′, as well as overt actors, suggesting more narration.

In PC2, different word classes occur in both the positive and negative ranges, with particles and quotative verbs in the negative and VTAs and pronouns in the positive. Independent verbs contrast with conjunct verbs, as do, more clearly this time, SAPs and non-SAPs. However, in accordance with the overall prevalence of VIIs on the positive side, both proximate and obviative singular VIIs characterise the texts here. The person information is also supported by features of VTAs, where in the negative range, SAPs occur with local and mixed VTAs, indicating that SAPs are interacting with each other and third persons, while in the negative range, non-SAPs occur with nonlocal and direct VTAs, suggesting more proximate third persons acting on obviatives. This is further strengthened by the occurrence of overt actors and goals in the positive range, where they co-occur with non-SAPs and VTAs. Quotative verbs again co-occur with SAPs and local VTAs, suggesting direct quotation. The occurrence of numerals in the positive range may also indicate that these numerals are used to describe situations and referents in more third-person heavy narrative.

6.3.2.2 Interim discussion

As for the full corpus, some of the above features can be tied to a distinction between dialogue and non-dialogue narrative. Ranges that include quotative verbs occur with SAPs (and mixed/local verbs), interjections, and imperative verbs. The ranges of the PCs with third persons, nonlocal verbs, and overt participants, on the other hand, do not occur with features that imply direct quotation. The re-emergence of these patterns in the BT subcorpus draws attention to the fact that future tense also occurs alongside the SAP feature combinations. The occurrence of more future tense in directly quoted dialogue may be one way dialogue differs from narration, as stories are often set in the past, while those speaking within stories may talk about future plans. These features do not clearly align with the distributions of the two volumes; the overlap of the volumes, with the

distribution of SSSC almost entirely within that of PCT, may indicate that sacred stories behave as a subset of all stories; thus, future research into the internal variation of this subcorpus, especially with regard to sacred and non-sacred stories, might look to the occurrence of direct quotation in each subset.

6.3.3 The Ahenakew-Wolfart subcorpus

6.3.3.1 Results

The A-W subcorpus consists of 58 individual chapters from eleven speakers, drawn from seven volumes. Before preprocessing the morphosyntactically tagged corpus for PCA, the A-W subcorpus contains 136,036 tokens (22,872 types), including non-Cree words, punctuation, and metadata. After preprocessing, these are reduced to 79,930 Plains Cree word tokens (18,212 types). The dataset constructed from this preprocessed data includes 453 feature tags, which are trimmed to 85 for the purposes of PCA. Of the 368 removed, 282 (76.63%) were preverb or prenoun tags.

For A-W, the first two PCs account for 12.45% and 10.77% of variance respectively, totaling 23.22%. Figure 6.3 plots the texts, as individual chapters of the A-W subcorpus, along the first and second principal components as reported in Table 6.4 below. Ellipses are not presented here to show any groups, namely the speakers, though groups can still be identified.[64] Some clusters that emerge are those formed by the texts of Jim Kâ-Nîpitêhtêw (JK, in blue), Emma Minde (EM, in green), Sarah Whitecalf (SW, in pink), Alice Ahenakew (AA, in orange), Cecilia Maskuskapoe (CM, in yellow), and Glecia Bear (GB, in teal)—for these speakers, while there is internal variation within their texts, they are still relatively close to each other. In some cases, namely Alpha Lafond and Rosa Longneck's dialogue (ALRL), and Minnie Fraser's (MF), Irene Caillou's (IC), and Mary Wells' (MW) chapters of Bear et al. (1998), there is only one text per speaker (or speaker pair, in the case of the dialogue), so intra-speaker comparison is not possible. The greatest differences occur for Peter Vandall (PV, in pink) and Joe Douquette (JD, in turquoise), which occur on

[64] A plot with ellipses is given in Appendix E, as it more clearly demonstrates how some speakers' texts are internally similar, where others' vary considerably. While such a plot is interesting, the large ellipses required for some speakers change the scale of the plot, so the version here is somewhat easier to read.

opposite corners of the plot—these differences, as well as those among other speakers, are further discussed below.

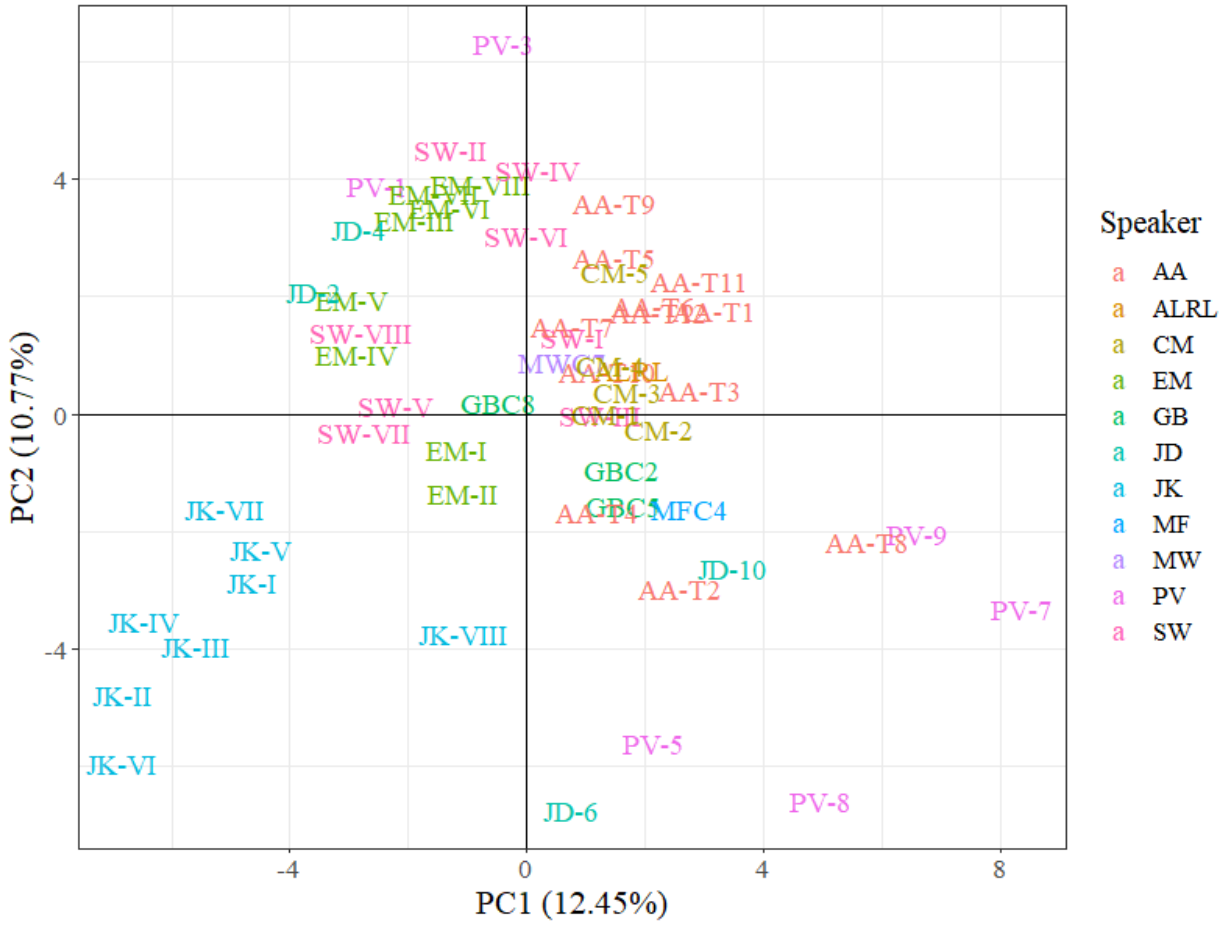


Figure 6.3: A-W subcorpus PCA: Chapters along PC1 & PC2

Table 6.4 presents a selection of the top 25 positive and negative features for each of the first two PCs. Where possible, they are arranged in pairs for reference, as for the previous results. Like the BT subcorpus, the A-W subcorpus varies less clearly in terms of word class compared to the full corpus analysis. The positive range of PC1 is characterised by verbs, especially VAIs and quotative verbs, yet the negative range by VTIs, VIIs, and VTAs. In PC2, particles and VTIs occur in the positive range, while verbs in general, VTAs, VIIs, and quotatives occur in the negative. In both PCs, the positive range is characterised by particles overall, though in both cases the negative range

is also characterised by at least some particle subclasses. As above, the same persons often occurred in multiple roles within each PCs, though not always.

Table 6.4: A-W subcorpus PCA: Positive & negative features

	Positive	Negative
PC1	· Verbs, VAIs, quotative verbs, particles, emphatic personal pronouns	· VTIs, VIIs, VTAs
		· Future tense
	· Independent verbs	· Conjunct, future conditional verbs
	· 3SG, EXCL, 1SG, 1SG-G, PX1SG	· PX21PL, INCL, 0SG, PX2SG, 2SG-G, 2SG
		· Local, inverse, mixed VTAs
	· Locative, proper, vocative, diminutive nouns	· Inanimate, singular nominals
	· Locative particles, numerals, quantifiers, interjections	· Indeclinable nominals, negative particles
		· @<GOAL, @GOAL>, @ACTOR>, @<ACTOR
PC2	· Particles, VTIs	· Verbs, quotative verbs, VTAs, VAIs
	· Past tense	· Future, present tense
	· Conjunct verbs	· Independent, imperative verbs
	· 0′SG, 3′-G, PX3PL, 3′, X, PX3SG	· 1SG, 1SG-G, 3SG-G, 3SG, 2SG-G, 2SG, PX2SG, PX21PL, PX3′
	· Nonlocal VTAs	· Local, mixed, inverse VTAs
	· Plural, obviative, animate nominals, proper nouns	· Singular nominals, vocative, diminutive nouns
	· Quantifiers, temporal particles, particle phrases, interjections	· Locative particles
	· @GOAL>, @ACTOR>	

The feature pairs in the table again serve to illustrate the differences along the axes of the plot Figure 6.3. In PC1, there is a general contrast between verb classes, verbal orders, and persons. Unlike the full corpus, here there is no clear-cut distinction between SAPs and non-SAPs, while first person singular, first person plural exclusive, and third person singular characterise the positive range, first person plural inclusive and second person singular characterise the positive range. The negative range also includes proximate VIIs (0SG), which is again in line with fewer

animate third persons (as obviative inanimate are the result of an animate third person in the discourse) and the overall prevalence of VIIs. While non-SAPs do not explicitly characterise the texts in the negative range, the overall occurrence of third persons is still supported by the occurrence of nonlocal VTAs and overt actors and goals. Inverse VTAs may indicate interactions between non-SAPs and SAPs as well. Differences in nominal features are again present: proper names, locative nouns, and vocatives characterise the negative range, suggesting a recounting of events where people and places feature, and where people address each other. In contrast, the positive range is characterised by more inanimate and singular nominals, which do not suggest any particular textual features, aside from the occurrence of more inanimate nouns, though it is worth noting that these do co-occur with VTIs and VIIs; the inanimate nominals may be contributing to the overt actors and goals seen here.

PC2 also offers a number of obvious feature pairings. Particles and VTIs contrast with verbs of various other types; past tense with present and future. For persons, 3′, 0'SG, and 3PL contrast with other persons, similar to BT PC1. Both sides contain prevalent possessors, which can be tied to dependent non-kinship nouns and to vocatives; as for the other analyses, vocatives also co-occur here with quotatives, suggesting their use within quotation. As for PC1, in PC2 only one side of the range is characterised by the occurrence of overt actors and goals.

I draw particular attention to the bottom left quadrant of the plot, that is, negative PC1 and negative PC2. There is considerable overlap of the features here: both contain VTAs, future tense, first person plural inclusive, second person singular, local, mixed, and inverse VTAs, more singular nominals, and inanimate actor. As this quadrant is dominated by the texts of Jim Kâ-Nîpitêhtêw, which are described as counselling texts, some of these features may be directly linked to this register: future tense is used to direct future action, and inclusive and second person is used to address the audience (see §6.3.3.2 for further discussion). The other features are less illuminating: there are likely more VTA-specific features because there are more VTAs overall, fewer obviative entities and thus fewer non-local VTAs, and more singular nominals alongside more singular persons.

6.3.3.2 Interim discussion

Unlike the previous results, where the distribution of the texts does not necessarily align with Plains Cree text type labels, the distribution for the A-W subcorpus in Figure 6.3 can broadly be attributed to text types, as noted in the prefaces and notes of the various volumes (e.g., §6.2.1 and Table 6.1 above). Nearly all the texts that fall on the lefthand side of the plot (PC1, negative values) fall broadly into the category of *kakêskihkêmowina*, counselling texts, while those on the righthand side fall into the category of some sort of *âcimowina*, or narrative. This is not to say that those on the left contain no narrative elements; the texts in Minde (1997, EM in the plot) are generally personal narratives that contain elements of *kakêskihkêmowina*, and those in Whitecalf (1993, SW) may be considered the opposite—*kakêskihkêmowina* that are illustrated with personal narrative. Though Vandall & Douquette (1987, PV and JD) contains texts of several different types, the first four, which all fall on the left side, are those designated as counselling texts, while the rest are some other form of narrative. Otherwise, the bottom left is dominated by the texts of Kâ-Nîpitêhtêw (1998, JK), which are all described as *kakêskihkêmowina*.

As previously discussed, *kakêskihkêmowina* are unsurprisingly characterised by more use of the first person plural inclusive: these are texts spoken by Plains Cree people, directed to Plains Cree people, with an aim to share what their way of life was like in the past, perhaps to mourn their way of life now, and how to live Plains Cree lives in the face of settler colonialism and imposed Western practices. The differences among the *kakêskihkêmowina* as represented in PC2 along the *y*-axis divides the texts along the format to some extent; Kâ-Nîpitêhtêw (1998) is addressing groups in a more structured style of public speeches, with frequent noticeable use of inclusive even when describing the past, while the other volumes perhaps use more third person singular and plural in similar contexts, which are more narrative-like features.

Personal narratives may be less of a clear category along PC1, though in PC2, they tend to group towards the top half of the plot: Ahenakew (2000, AA), the first section of Bear et al., (1998, GB, IC, MF, MW), much of Masuskapoe (2010, CM), Minde (1997), and Whitecalf (1993), which all contain considerable personal narrative, and some of the texts in Vandall & Douquette (1987), which are retellings with either first person dialogue, first person narration, or both. However, it is also worth noting that much of the righthand side of the plot, the positive values of PC1, also

include personal narratives, and within that half of the plot, the positive range of PC2 is characterised by first person plural exclusive, suggesting personal narratives that include family or other groups, while the negative range of PC2 focuses more on first person singular rather than plural, suggesting stories that instead focus on events experienced by an individual. A cursory glance at the texts seems to support the observation of a divide between groups and individuals in these first person stories, offering an excellent avenue for future investigations that focus more closely on the content of individual texts.

6.4 Discussion

Though the three different analyses of the corpus demonstrate a number of different patterns, many of the same features occur repeatedly: persons group in similar ways alongside related verbal features, as do tenses and particle types. In the full corpus analysis, the two subcorpora show strong differences along PC1, yet with similar internal variation along PC2, and the patterns of PC2 are reinforced in the individual analyses, especially with respect to dialogue and non-dialogue narrative. The BT volumes overlap considerably, thus suggesting considerable overlap of the features that occur in *atâyohkêwina* and *âcimowina*, though patterns of direct quotation vs. narration emerge instead. The A-W texts suggest a stronger divide between *kakêskihkêmowina* and *âcimowina* than might have been suggested by previous research (e.g. Schmirler & Arppe, 2020), but also the ways in which these types can overlap and how narrative can be used to counsel, often through more personal or historical narrative.

In all three analyses, a distinction arises between texts whose features suggest direct quotation—quotative verbs, SAP features, vocatives, interjections—and those that do not, and in general thus contain more non-SAP features and more animate nominal features. The direct quotation features also frequently occur with features that do not immediately suggest quotation, but nonetheless can be tied to it: more imperatives, more independent verbs, more future tense, more interrogatives. Groups of related features are apparent throughout the analyses: where SAPs occur, mixed and local VTAs also occur, where non-SAPs occur, nonlocal VTAs and overt participants are also present. Patterns like these, repeating themselves within and across the corpus, bring to light relationships that cannot be straightforwardly explored without the use of an approach like PCA.

Throughout the results, it is also worth noting the patterns that arise with respect to clusivity and obviation. The occurrence of first person plural features in general distinguishes the A-W subcorpus from the BT subcorpus, and yet within the A-W subcorpus, inclusive and exclusive characterise the texts along both PCs and can be directly related to known Plains Cree text types. There are also various points throughout the analyses where third person proximate singular occurs in opposition to obviative, though obviative frequently occurs with third person plural. When this occurs, third person singular instead occurs alongside SAPs; this aligns with the non-occurrence of SAPs with obviatives observed in the corpus (cf. Chapter 4). This of course does not indicate that proximate singular and obviative persons do not frequently overlap, but instead seems to indicate that singular third persons interact with SAPs at least as much as with obviative persons, while plural third persons occur less often with SAPs. These features, being relatively rare typologically, appear repeatedly among the distinguishing features in the analyses, highlighting their role in the grammar of, and communication in, Plains Cree.

6.5 Conclusion

This initial exploration of text types in Plains Cree has offered insight into a number of similarities and differences across the corpus. An analysis of the full corpus highlights strong differences between the subcorpora, which remain despite removing the non-Cree and editorial elements that had been hypothesised as key differences in Schmirler & Arppe (2020). However, the second PC in the full analysis also serves to illustrate how similar the two subcorpora are, with similar internal variation. Within the BT subcorpus, the similarities and differences between the volumes begin to suggest perhaps a more structured nature to *âtayôhkêwina*, as they fall within the range of *âcimowina*. In this analysis, however, one begins to see features that point towards the use of quotation in narrative, dividing texts along these lines rather than along those of Plains Cree text types. Conversely, in the A-W subcorpus, text types can be used to explain the distribution of texts, with *kakêskihkêmowina* contrasting against other *âcimowina*, though internal variation in these types is also apparent: first person narration in personal stories, whether or not they are used to counsel, third person narration, and quotation distinguish the narratives along the first two PCs in addition to the broad text type distinction.

This PCA approach, beyond the analyses and discussion in the present chapter, offers several paths for future exploration. A deeper look at the full corpus and the similarities between the two subcorpora is of interest: how does the overlap along PC2 in the full analysis align with textual features in each subcorpus, and how do these relate to the analyses for the individual subcorpora? In the BT subcorpus, text types were only examined using the volumes as a proxy; more careful consideration of the actual types will strengthen the hypothesis that *âtayôhkêwina* behave as a subset of narratives in general. Additionally, some degree of genre analysis would be of interest here; during the validation process for the BT subcorpus, formulaic opening and closing lines became apparent—however, beyond this observation, no detailed analysis of story structure is undertaken. Instead, the primary distinction seen for the BT subcorpus is between narrative type with respect to dialogue; this can be seen to some degree throughout the full analysis and the A-W analysis, and so is an excellent candidate for future research questions. Finally, within the clear text type distinction in the A-W subcorpus, the different ways in which stories can be used to counsel or simply share information comes to light. In this vein, manual text-by-text coding for Plains Cree text type and internal structure (e.g., type of narration or prevalence of dialogue) will allow for more options in visualising the results of the analyses and illuminate more similarities and differences among texts.

In addition to exploring these questions, the data can also be expanded. The present analysis, though undertaken using the corpus as disambiguated by the Plains Cree parser, includes only actor and goal tags from among the numerous syntactic functions assigned by the parser. Additional function tags, as well as interactions between functions and morphology, can be used in both PCA and other statistical analyses—not only including whether actors and goals occur, and before or after verbs, but other features of those participants: nouns, pronouns, full noun phrases, etc. A full register analysis approach should also make use of higher-level features of the texts, such as those presented in Appendix D, following Biber & Conrad (2019); the inclusion of these in future analyses is key to a thorough exploration of Plains Cree text types and registers. Finally, the results of these analyses may also lend themselves to future parser development: it is clear that quotation plays an important role in distinguishing texts in Plains Cree, and thus including quotation in the model is perhaps a priority for modelling interclausal relationships, alongside subordination, as discussed in Chapters 2 and 3. Data expansion can also go far beyond the

morphosyntactic tags included in the analysis and improvements to the underlying models, but to texts themselves. Expanding the Plains Cree corpus is an ongoing process and, as more texts are added, of types like and unlike those included in the present corpus, a fuller picture of Plains Cree text types will be possible.

Chapter 7

General discussion and conclusion

This dissertation presented the creation of a morphosyntactically tagged and morphologically validated corpus of ~152,000 words of Plains Cree and demonstrated ways in which this corpus can be used. As corpora are a key component of "digital language infrastructure" in language revitalisation (Arppe et al., 2016, p. 1), the development of a corpus, especially with at least some degree of automatic tagging, is a considerable step in the right direction. In this work, three studies were undertaken to demonstrate the use of the tagged corpus: the frequencies of morphosyntactic tags are presented, word order frequencies are explored, and a text type analysis is undertaken. Each of these is only a first step in the use of a tagged corpus for Plains Cree, offering many directions for future research.

The Plains Cree corpus described and used in this dissertation is built from three key components, beyond the digitised texts: 1) a morphological model that analyses wordforms to determine morphological features, 2) a hand-validated set of these morphological analyses (a morphological "gold standard" or MGS), and 3) a syntactic model that disambiguates words and identifies relationships between words. The focus of this work is primarily on the syntactic model, a Constraint Grammar (CG)-based parser for Plains Cree. The CG formalism is a practical choice for approaching Plains Cree syntax, as linear order plays only a small role in syntactic relationships, while the morphological features take centre stage. The ordered constraints in CG take morphological features and linear order together, allowing for morphologically rich languages with flexible word order to be modelled with relative efficiency, without using previously tagged data, as would be needed for machine learning. Though the development of the morphological model is not a focus of this work, I have been involved in the ongoing process, and the MGS described herein has been an important tool in furthering the morphological model. The underlying models that provide the tags, the manual validation, and the availability of a corpus are all applicable beyond just the existence of a tagged corpus. These tools can be used for other

applications, like spell checkers and grammar checkers, the manual validation can be used to identify weaker areas of modelling, and the tagged corpus can be used for a variety of other pedagogical and research efforts, from literary analysis to machine learning on a larger scale than previously possible for Plains Cree.

In §7.1, I summarise each of the previous chapters in this work: morphosyntactic features in §7.1.1 (Chapter 2), the CG-based parser for Plains Cree in §7.1.2 (Chapter 3), the application of this parser and a description of the tagged corpus in §7.1.3 (Chapter 4), an exploration of word order patterns in §7.1.4 (Chapter 5), and a text type analysis of Plains Cree §7.1.5 (Chapter 6). Word order patterns and text types are discussed further in §7.2 and §7.3 respectively, where I consider how the results presented in this work can be tied to broader typological and cross-linguistic research. I conclude the dissertation in §7.4.

7.1 Summary

7.1.1 Chapter 2

Chapter 2 described the morphosyntactic features of Plains Cree that underlie the current syntactic parser, as well as others that can be used in future parser development. These features included nominal features such as animacy (i.e., grammatical gender), person, and number, the verbal transitivity classes and verbal person marking, as well as the hierarchical alignment system, with direct and inverse marking and a topicality hierarchy.

The Algonquian Person Hierarchy and the direction morphology described in Chapter 2 are key components throughout the remainder of the dissertation, as the inherent topicality differences between speech act participants and third persons, and the discourse-based topicality differences between proximate and obviative third persons, are used to assign syntactic roles, explore how word order and relative topicality interact, and examine differences in Plains Cree narratives. The chapter concluded with a look at morphosyntactic features not yet greatly, if at all, implemented in the parser, but that will help in future development to more accurately label the relationships already implemented and to begin the process of implementing new ones.

7.1.2 Chapter 3

Chapter 3 introduced the Constraint Grammar formalism used to build the Plains Cree parser. This included a history of CG and the motivations for using CG for Plains Cree. As CG was initially applied to Finnish, a language with rich verbal and nominal morphology and flexible word order, features broadly shared by Plains Cree, the formalism was well suited to Algonquian morphosyntax.

Chapter 3 then laid out the previous development for the Plains Cree parser, which can be found in more detail in Appendix A. The new developments were then described, including the addition of more detailed morphological and lexical features, such as particle classes, and constraints to describe more syntactic functions, such as particles that modify nouns or verbs, obliques dependent on verbs, and more noun phrase options, including possession. Alongside these additions, improvements to the previously-developed constraints were made, such as those that assign the primary actor(/subject) and goal(/object) relationships. The chapter concluded with a callback to Chapter 2, exploring how those not-yet-implemented features and relationships might be added in the future.

7.1.3 Chapter 4

Chapter 4 discussed the application of the parser. This included a description of the texts within the corpus, an explanation of the morphological tagging process, an evaluation of the parser's effectiveness, and finally an overview of the frequencies of the various morphosyntactic tags in the corpus.

The corpus consists of two subcorpora, one drawn from the texts edited by Leonard Bloomfield (Bloomfield, 1930, 1934) and one from texts edited by Freda Ahenakew and H.C. Wolfart (Ahenakew, 2000; Bear et al., 1998; Kâ-Nîpitêhtêw, 1998; Masuskapoe, 2010; Minde, 1997; Vandall & Douquette, 1987; Whitecalf, 1993). These totalled 152,405 Plains Cree words and included several different text types, from 21 different speakers in various communities, in different time periods.

The morphological tags were initially supplied by an FST-based morphological model (Snoek et al., 2014; Harrigan et al., 2017), which were then manually validated—missing analyses were

added, and incorrect analyses were corrected. The parser evaluation took into consideration how well the parser disambiguated ambiguous forms and assigned syntactic functions, tested against a manually tagged development corpus. When compared to an earlier version of the parser (Schmirler et al., 2018; Appendix A), disambiguation improved in tests against the hand-coded development corpus, and overall improvements to the number of forms disambiguated in the A-W subcorpus also increased from those disambiguated by the earlier version of the parser. Improvements were also seen in the recall and precision of the syntactic function tags, and several new functions were added and performed reasonably well.

Finally, the corpus was described in terms of the frequencies of the various morphosyntactic features, with a focus on the full corpus and discussion of differences between the subcorpora. While there were many similarities, differences in verb types and persons were particularly apparent, which points to the need for more detailed analyses of the variation within the corpus, examples of which were undertaken in the following chapters.

7.1.4 Chapter 5

Chapter 5 examined the word order patterns, i.e., the relative order of verbs, actors, and goals in the Plains Cree corpus. As a language with verbal argument indexing and flexible word order, not only can the verb, actor, and goal occur in any order relative to each other, but actors and goals need not occur as overt nominals alongside the verb. Using a corpus to examine how often actors and goals are overtly realised, and when overt, how often they occur before or after the verb, allows for such variation to be considered on a larger scale than in previous studies (e.g., Cook et al., 2003; Mühlbauer, 2007; Wolvengrey, 2011).

As the corpus at this point is only tagged for morphosyntactic features and very few, if any, semantic or pragmatic ones, it was not as straightforward to explore the motivations behind word order choices as is possible with a smaller collection of clauses looked at in more detail, with specific attention paid to reference tracking, topicality, etc. Instead, I took into account those morphosyntactic features that can give clues to topicality: animate being more topical than inanimate, proximate more than obviative, first and second persons more than third persons. With these relationships in mind, it was found that less topical arguments were generally more likely to

occur as overt arguments, and, when overt, more topical arguments were more likely to occur preverbally than postverbally. These findings aligned well with smaller-scale observations (e.g., Wolvengrey, 2011); as the tagging is expanded in the corpus, more detailed statistical analyses could be undertaken to explore how semantic and pragmatic features interact with morphosyntax. Additionally, corpus results can be used to identify cases of interest (e.g., less frequent word orders) and allow them to be examined individually.

7.1.5 Chapter 6

Finally, Chapter 6 presented a text type analysis of the Plains Cree corpus, taking a register analysis approach. The text types of Plains Cree were initially identified as those traditionally described: sacred stories and non-sacred texts, the latter of which can be broken down into several subtypes, including personal stories, funny stories, stories about the past, and counselling texts. Many of the texts in the corpus are explicitly identified as containing aspects of these types in the forewords or other editors' comments in the published volumes, and these were taken as the starting point for the analysis. Broken down by chapter and by subcorpus, the texts were analysed using Principal Component Analysis (PCA) for dimension reduction, which simplified the large datasets into more visually interpretable results.

The results showed several main patterns. First, it was apparent that there were considerable differences between the subcorpora, though many similarities appeared as well. For the full corpus and BT analyses, there was little to indicate any strong morphosyntactic differences between traditional text types. The main similarity between the contents of the subcorpora is narrative—suggesting that narrative, with sacred or non-sacred content, is morphosyntactically consistent. However, a distinction that arose for the full corpus as well as for each subcorpus was a distinction between dialogue and non-dialogue narrative—stories with more first and second person, more quotative verbs, and more interrogatives or interjections, contrasted with stories with more third persons and more overt nominals. I propose that those first and second persons are more likely occurring within the speech introduced by the quotative verbs, alongside exclamations and questions.

Within A-W, where there was a greater variety of text types to consider, differences and similarities between counselling texts and narratives became apparent. Two patterns of counselling texts emerged: those that appear include more personal narrative, offering counsel through personal stories, and those that appear to counsel more directly. These then contrasted with narratives that were not intended to counsel, and this distinction can especially be found in the texts of Vandall & Douquette (1987), where the two speakers tell both narrative counselling texts and then a variety of dialogue and non-dialogue narratives, and these range across the resulting plot (Figure 6.3 in Chapter 6), alongside texts of similar content.

Like the word order results in Chapter 5, the text type results can be used to inform more detailed descriptive studies. These interpretations can be verified on a smaller scale, which in turn may find more features or patterns for consideration in a larger scale corpus analysis. It is the value of even straightforward corpus analysis to the exploration of Plains Cree morphosyntax that I hope I have demonstrated herein, and which I further attempt to link to broader typological and cross-linguistic research in the remainder of this chapter.

7.2 Corpora and typology

A typological exploration of Plains Cree syntax necessarily requires a closer look at individual features and small groups thereof; beyond the Algonquian language family, the combination of verbal argument indexing, hierarchical alignment, and flexible word order co-occur relatively rarely—for example, WALS includes only five languages that share these features: Plains Cree, Passamaquoddy-Maliseet (also Algonquian, spoken in New Brunswick and Maine), Karok (an isolate, spoken in California), Kawaiisu (Uto-Aztecan, also spoken in California), and Nunggubuyu (Gunwinyguan, spoken in Australia) (Dryer & Haspelmath, 2013). Of these, Plains Cree has by far the most speakers (Moseley, 2010), and, as described herein, there now exists a tagged corpus of Plains Cree which can be used to explore these features on a larger scale. Argument realisation in a corpus of another Algonquian language, Menomini (Shields, 2004), can be directly compared to the Plains Cree results herein, and thus verbal argument indexing, hierarchical alignment, and flexible word order are all relevant. However, I primarily discuss the occurrence of these features individually among the world's languages in comparison to Plains Cree. First, I look at Shields' small corpus study of Menomini word order patterns (2004) and

discuss similarities to Plains Cree. I then look beyond the Algonquian language family, first at null vs. overt arguments particularly in other languages with systems of hierarchical alignment and flexible word order, followed by more familiar null subject languages.

7.2.1 Hierarchical alignment and null arguments

7.2.1.1 Within Algonquian

I begin closest to home with a brief discussion of patterns in a corpus of Menomini, a highly endangered Algonquian language spoken in Wisconsin, that broadly shares the above-mentioned typological features with Plains Cree (as a less well-documented language than Plains Cree, its features are not fully included on WALS). Menomini word order has been described simply as free (e.g., Guile, 2001), though Bloomfield (1962, pp. 441–3) examined Menomini word order and listed tendencies for certain types of elements to precede or follow a verb. Pronouns, numerals, objects (goals), obviative subjects (actors) tended to precede verbs, and proximate subjects (actors) and goals of inverse transitive verbs (VTAs) tended to follow verbs—Shields (2004) summarises this as OVS(/GVA) for VTIs (transitive verbs with inanimate goals) and direct VTAs, SVO(/AVG) for inverse VTAs, and VS(/VA) for VAIs and VIIs (intransitive verbs with animate and inanimate actors respectively; p. 374). Shields (2004) then explored a corpus of Menomini (like the Plains Cree BT subcorpus examined in this dissertation, the Menomini corpus was also collected by Leonard Bloomfield in the 1920s), examining 297 clauses with at least one overt nominal. While this is a much smaller corpus than that for Plains Cree, and patterns were examined and reported differently, some comparisons can be made. In Table 7.1, I include a simplified version of Table 5.1 from Chapter 5, which gives the Plains Cree values, to be compared to those for Menomini in Table 7.2, taken from Shields (2004).

Table 7.1: Overt actors and goals in Plains Cree

	VII		VAI		VTI		VTA		Total V	
w/A	925		5,505		1,302		2,195		9,927	
		% w/A		% w/A		% w/A		% w/A		% w/A
AV	479	51.8%	2,202	40.0%	565	43.4%	1,012	46.1%	4,258	42.9%
VA	446	48.2%	3,303	60.0%	737	56.6%	1,183	53.9%	5,669	57.1%
					VTI		VTA		Total VT	
w/G	-	-	-	-	3,439		4,409		7,848	
		% w/G		% w/G		% w/G		% w/G		% w/G
GV	-	-	-	-	1,737	50.5%	1,460	33.1%	3,197	40.7%
VG	-	-	-	-	1,702	49.5%	2,949	66.9%	4,651	59.3%
		% w/A		% w/A		% w/A or G		% w/A or G		% w/A or G
all preV	479	51.8%	2,202	40.0%	2,302	48.6%	2,472	37.4%	7,455	41.9%
all postV	446	48.2%	3,303	60.0%	2,439	51.4%	4,132	62.6%	10,320	58.1%

While these are drawn from a much smaller corpus for Menomini, similar proportions are seen for Plains Cree in many cases. Actors across all verb classes occur pre- and postverbally in nearly identical proportions, with just under 60% occurring after the verb, and, while goals are less similar between the languages, they are in both languages more likely to occur following the verb.

Table 7.2: Overt actors and goals in Menomini (Shields, 2004)

	VII		VAI		VTI		VTA		Total V	
w/A	10		135		12		40		197	
		% w/A		% w/A		% w/A		% w/A		% w/A
AV	6	60.0%	55	40.7%	5	41.7%	17	42.5%	83	42.1%
VA	4	40.0%	80	59.3%	7	58.3%	23	57.5%	114	57.9%
					VTI		VTA		Total VT	
w/G	-	-	-	-	33		68		101	
		% w/G		% w/G		% w/G		% w/G		% w/G
GV	-	-	-	-	12	36.4%	21	30.9%	33	32.7%
VG	-	-	-	-	21	63.6%	47	69.1%	68	67.3%
		% w/A		% w/A		% w/A or G		% w/A or G		% w/A or G
all preV	6	60.0%	55	40.7%	17	37.8%	38	35.2%	116	38.9%
all postV	4	40.0%	80	59.3%	28	62.2%	70	64.8%	182	61.1%

For transitive verbs where both the actor and goal are overt, Plains Cree values are given in Table 7.3 (simplified from Table 5.2) and the Menomini values in Table 7.4. For VTIs and the overall transitive clauses, both languages use AVG quite often, and VGA is among the less common orders for both languages. However, AGV, which is the least common order in the Plains Cree corpus, is as frequent as AVG for Menomini—however, as only 20 clauses were reported by Shields (2004), a larger corpus of Menomini would be needed to better assess similarities to Plains Cree.

Table 7.3: Two overt arguments with transitive verbs, Plains Cree

	VTI		VTA		Total	
w/A & G	436		517		953	
		% w/A & G		% w/A & G		% w/A & G
VAG	52	11.9%	102	19.7%	154	16.2%
VGA	50	11.5%	56	10.8%	106	11.1%
AVG	133	30.5%	200	38.7%	333	34.9%
GVA	84	19.3%	55	10.6%	139	14.6%
AGV	38	8.7%	40	7.7%	78	8.2%
GAV	79	18.1%	64	12.4%	143	15.0%

Table 7.4: Two overt arguments with transitive verbs, Menomini

	VTI		VTA		Total	
w/A & G	9		11		20	
		% w/A & G		% w/A & G		% w/A & G
VAG	1	11.1%	2	18.2%	3	15.0%
VGA	0	0.0%	0	0.0%	0	0.0%
AVG	4	44.4%	2	18.2%	6	30.0%
GVA	2	22.2%	2	18.2%	4	20.0%
AGV	2	22.2%	4	36.4%	6	30.0%
GAV	0	0.0%	1	9.1%	1	5.0%

Shields (2004) also presents the occurrence of preverbal and postverbal arguments by argument type (proximate, obviative, inanimate), clause type (independent, conjunct, imperative), discourse factors (first mention of a referent, or a shift from obviative to proximate), and nominal type (pronoun, quantifier (phrase), and head of a relative clause). Only the first and last of these can be compared to Plains Cree using the current tags and results presented in Chapter 5. I begin with argument type—Table 7.5 is a simplified version of Table 5.4, and Table 7.6 gives comparable results for Menomini. The two languages are fairly similar with respect to animate third persons: both proximate and obviative arguments as either actors or goals are more likely to follow a verb, proximate actors occur pre- and postverbally at about the same rates as proximate goals and,

though the values differ between the languages, obviative goals occur postverbally more often than obviative actors. Inanimates behave somewhat differently: for Plains Cree, they occur pre- and postverbally rather equally, while in Menomini they behave like animate third persons and follow verbs more often.

Table 7.5: Actors and goals for third persons, Plains Cree

	PROX actor		OBV actor		INAN actor	
w/A	7,341		1,548		925	
		% w/A		% w/A		% w/A
AV	2,909	39.6%	518	33.5%	479	51.8%
VA	4,432	60.4%	1,030	66.5%	446	48.2%
	PROX goal		OBV goal		INAN goal	
w/G	1,837		2,496		3,439	
		% w/G		% w/G		% w/G
GV	726	39.5%	667	26.7%	1,737	50.5%
VG	1,111	60.5%	1,829	73.3%	1,702	49.5%
		% w/A or G		% w/A or G		% w/A or G
Total preV	3,635	39.6%	1,185	29.3%	2,216	50.8%
Total postV	5,543	60.4%	2,859	70.7%	2,148	49.2%

The data extracted for nominal types in Plains Cree is not so neatly comparable to Menomini to present in tables, so I address some general similarities and differences here. In both languages, pronominal actors are more likely to occur before verbs than after them. While this is also true of pronominal goals in Menomini, in Plains Cree this averages out at 50%, with considerable differences between demonstrative pronouns (more likely to be postverbal) and personal pronouns (more likely to be preverbal). In Plains Cree, these tendencies for pronouns contrast with those for nouns, which are more likely to occur following verbs (Table 5.6) regardless of their nominal type. In Menomini, new or more accessible referents occur more often preverbally (Shields, 2004). While these cannot be compared to Plains Cree at this time, newness and accessibility are of great

interest for future studies using the Plains Cree corpus, though these features will require hand coding of at least some portion of the corpus.

Table 7.6: Actors and goals for third persons, Menomini (Shields, 2004)

	PROX actor		OBV actor		INAN actor	
w/A	145		22		11	
		% w/A		% w/A		% w/A
AV	52	35.9%	11	50.0%	5	45.5%
VA	93	64.1%	11	50.0%	6	54.5%
	PROX goal		OBV goal		INAN goal	
w/G	26		43		33	
		% w/G		% w/G		% w/G
GV	10	38.5%	9	20.9%	12	36.4%
VG	16	61.5%	34	79.1%	21	63.6%
		% w/A or G		% w/A or G		% w/A or G
Total preV	62	36.3%	20	30.8%	17	38.6%
Total postV	109	63.7%	45	69.2%	27	61.4%

7.2.1.2 Beyond Algonquian

I next look further afield to Kutenai, spoken in British Columbia. Though an isolate, it shares a number of features with Algonquian languages, due to apparent areal effects when the Algic languages were spoken in the same geographical area (e.g., Dryer, 2007). Like Plains Cree, Kutenai animate third persons are either proximate or obviative in a transitive clause; however, with less overt person marking on Kutenai verbs, a bare verb form is direct and a marked verb is inverse. Dryer (1994) analyses a corpus containing 503 nonlocal (i.e. no first or second persons) transitive clauses, of which 138 are passive and are excluded here for comparison to Plains Cree. Table 7.7 compares the occurrence of direct and inverse nonlocal clauses in Kutenai (Dryer, 2004)

and Plains Cree, which are not dissimilar.[65] This is perhaps indicative of common relative proportions cross-linguistically—studies for other direct and inverse systems would further test such a hypothesis.

Table 7.7: Nonlocal direct and inverse clauses in Kutenai and Plains Cree

Language	Kutenai		Plains Cree	
# clauses	365		7,365	
	#	% Clauses	#	% Clauses
Direct	295	80.8%	5,701	77.4%
Inverse	70	19.2%	1,664	22.6%

Table 7.8: Overt actors and goals in Kutenai and Plains Cree

Language	Kutenai				Plains Cree			
	Direct		Inverse		Direct		Inverse	
# clauses	295		70		5,701		1664	
		% DIR		% INV		% DIR		% INV
w/A	46	15.6%	31	44.3%	867	15.2%	484	29.1%
No A	249	84.4%	39	55.7%	4834	84.8%	1180	70.9%
		% DIR		% INV		% DIR		% INV
w/G	172	58.3%	2	2.9%	2439	42.8%	152	9.1%
No G	123	41.7%	68	97.1 %	3262	57.2%	1512	90.9%

Dryer (1994) also reported the occurrence of overt and null arguments for nonlocal direct and inverse clauses in Kutenai, which are compared to Plains Cree in Table 7.8. Recall that proximate and obviative categories are key here: for direct verbs, actors are proximate and goals are obviative, while for inverse verbs, the actors are obviative and the goals are proximate. Curiously, here the similarities are strongest for proximate forms (actors of direct verbs and goals of inverse)—as

[65] The tables in Chapter 5 did not explore direct and inverse verbs specifically for nonlocal VTAs, so these values are newly extracted and thus are a subset of what is seen in Table 5.3.

actors they occur almost identically in the two languages (~15% overt), and as goals they are the least likely category to be overt (<10%). Inverse actors (obviative) are also similar between the two languages in that they are less often overtly realised, while direct goals (also obviative) are reversed between the languages. The variation especially between obviative arguments in the two languages is worth further exploration in 1) larger corpora and 2) other languages with similar typological features.

Next, I look at two South American isolates that display hierarchical alignment, Mapudungun and Movima. Like Plains Cree, Mapudungun, spoken in Chile and Argentina, uses direct and inverse verbs with respect to more or less topical third person referents. Matter (2020) explores the relative occurrence of subjects, agents, and patients in Mapudungun based on a number of features, including animacy features (human, non-human), salience (givenness and newness in the discourse), etc. Table 7.9 compares the occurrence of subjects, agents, and patients of animate and inanimate third person verbs in Mapudungun to the same categories in Plains Cree, where S will be equivalent to actors of intransitive verbs (VII and VAI), and A and P will be equivalent to actors and goals of transitive verbs (VTI and VTA). The two languages contrast considerably for S and P, with Mapudungun subjects and patients more likely to occur as overt nominals than Plains Cree subjects and patients—this difference is particularly dramatic for patients(/goals). The occurrence of overt agents is more similar, though in Plains Cree they are slightly more frequent. Matter (2020) demonstrates that intransitive subjects in Mapudungun do not pattern like either agents (i.e. an accusative-like pattern) or patients (i.e. an ergative-like pattern), but rather occur as an intermediate category with respect to how frequently they are overtly realised (and the factors that influence their realisation). The Plains Cree results here instead suggest more similarities between subjects and agents. However, I must note that these results combine animate and inanimate participants in Plains Cree, whereas in Chapter 5 (Table 5.8 and Table 5.9) looked separately at proximate and obviative animate third persons, and found neither the intermediate pattern seen for Mapudungun, nor the more accusative-like pattern seen for Cree here, but a more ergative-like pattern where subjects (intransitive actors) and patients (transitive goals) patterned similarly. These differences further demonstrate that inanimate and animate arguments do not pattern the same way in Plains Cree—a deeper consideration of the differences between proximate and

obviative third persons, animate and inanimate participants, and other discourse factors that influence the overt realisation of arguments is needed.

Table 7.9: Overt subjects, agents, and patients in Mapudungun and Plains Cree

Language	Mapudungun		Plains Cree	
S clauses	2,171		20,052	
	#	% Clauses	#	% Clauses
w/S	1,083	49.9%	6,285	31.3%
No S	1,088	50.1%	13,767	68.7%
A clauses	766		23,569	
	#	% Clauses	#	% Clauses
w/A	153	20.0%	6,530	27.7%
No A	613	80.0%	17,039	72.3%
P clauses	736		23,569	
	#	% Clauses	#	% Clauses
w/P	562	76.4%	2,741	11.6%
No P	174	23.6%	20,828	88.4%

Movima, spoken in Bolivia, also demonstrates hierarchical alignment, though with differences in verbal morphology and syntax that make comparisons to Plains Cree less straightforward than for Menomini or Mapudungun. Like Kutenai, third person transitive verbs do not have overt verbal person marking, but the proximate participant is always marked by means of an enclitic pronoun or an enclitic article specifying a noun. The obviative may also be marked by an enclitic pronoun or NP following the proximate, or it can be omitted (Haude, 2014). This would then be a stark difference to Plains Cree and the other languages mentioned here, where a proximate argument is less likely to be overtly realised than an obviative—the differences between bound pronouns and clitics and free forms across these different languages and descriptive traditions add a further layer of complexity to the comparison, which is beyond the present scope. However, some data drawn

from a small Movima corpus of 1,260 transitive clauses can be compared to the Plains Cree results. Haude (2014, p. 300) finds that these clauses are, overall, 94% direct and 6% inverse, a much higher proportion of direct clauses than for Plains Cree (77%) or Kutenai (80%), as reported above. As the proximate participant is always overt, I instead compare the occurrence of obviatives in transitive clauses with two animate arguments, that is, with direct goals and inverse actors, in Movima to those in Plains Cree; these are presented in Table 7.10. Yet another difference emerges here: for Movima, the occurrence of an overt obviative is essentially identical whether the clause is direct or inverse (i.e., whether they are the goal or the actor), while for Plains Cree, obviative participants are more likely to occur as overt goals than overt actors.

Table 7.10: Overt and null obviative participants in Movima and Plains Cree

Language	Movima				Plains Cree			
	Direct		Inverse		Direct		Inverse	
# clauses	1,177		81		5,701		1,664	
		% DIR		% INV		% DIR		% INV
Overt OBV	787	66.9%	53	65.4%	2439	42.8%	484	29.1%
Null OBV	390	33.1%	28	34.6%	3262	57.2%	1180	70.9%

Movima thus presents the most different case from Plains Cree of the languages considered thus far in this chapter, further demonstrating the value of corpus studies, of any size, especially for languages with rarer features like hierarchical alignment—to take any of these languages as representative of the category does not fully consider the possible variation. In these studies of argument realisation, the authors considered word order and the occurrence of overt arguments with respect to a number of other features, including the relative animacy of the participants (human, non-human, and inanimate), the given or new status of the referents, and the occurrence of pronouns vs. nouns as overt arguments, among others. These are certainly all categories that the Plains Cree corpus, with some additional tagging and postprocessing, will be well-suited to investigating at a larger scale than previously done for the language. Further comparisons among these hierarchical alignment languages will then be possible.

7.2.2 Beyond Algonquian: null subjects and flexible word order

7.2.2.1 Null subjects

Looking beyond hierarchical alignment, I look at Spanish and Portuguese, prototypical pro-drop or null subject languages. Different varieties of these languages, as well as their use in different time periods, have demonstrated different frequencies of null vs. overt subjects.[66] With less flexibility in word order in these languages, the focus is instead on the occurrence rather than position of arguments. In the Plains Cree corpus, 20% of verbal clauses occur with an overt actor. This is similar to Mexican Spanish, Madrid Spanish, and European Portuguese, where overt subjects occur in 19%, 24%, and 22% of clauses respectively (da Silva, 2006, p. 81; Otheguy et al., 2007, p. 785). Thus, overt actors in Plains Cree appear to occur at the same rate as overt actors in prototypical pro-drop languages. However, many other varieties of Spanish and Brazilian Portuguese demonstrate a much higher occurrence of overt subjects, from 27% in Ecuadoran Spanish, 41% in Dominican Spanish, and >70% in Brazilian Portuguese (Duarte, 2000; Otheguy et al., 2007; Smith, 2013). Smith (2013) also reports the occurrence of overt subjects in different registers, including oral texts. These provide perhaps a closer analogue to the Plains Cree corpus, as it is comprised entirely of oral language. These Brazilian Portuguese oral texts used fewer overt subjects than the written texts, with only 53% of clauses showing an overt subject; additionally, fiction used fewer overt subjects (62%) than academic text (78%) or newspaper articles (88%) (Smith, 2013, pp. 60–1). The importance of registers in corpus analysis is once again highlighted here; as the given figures for Spanish do not take into account different registers, it may be that oral and written texts in these languages behave differently and thus written texts are less comparable to the Plains Cree data—certainly a question for future consideration, which is touched on briefly in §7.3.

For Brazilian Portuguese, the use of more overt subjects is considered a change over time, as the occurrence of overt subjects has increased from 20% in 1845 to 74% in 1992, based on a corpus of plays (Camacho, 2008; Duarte, 2000). Dominican Spanish is similarly analysed (e.g., Toribio,

[66] Language change is also of interest in the Plains Cree corpus, as the Bloomfield subcorpus was collected several decades earlier than the Ahenakew-Wolfart subcorpus (Schmirler, in press), providing a possible parallel to the null subject languages discussed here.

2000): for example, in Dominican Spanish, older speakers use fewer overt subjects than younger speakers, though speakers who are also influenced by other varieties of Spanish use fewer overt subjects, such as university students who are exposed to academic language. However, there does not seem to be an influence of English on the use of overt subjects in Spanish (Cabrera-Puche, 2008, pp. 349–60; Flores-Ferrán, 2004). In this vein, Nagy et al. (2011) found that heritage speakers of three pro-drop languages (though not Spanish) in Toronto do not use significantly more overt pronouns due to English influence. In contrast, Otheguy et al. (2007, pp. 795–7) find a change in pronoun use among Spanish-speaking immigrants in New York City, with first generation speakers using fewer pronouns than second and third generation speakers (though they also note that Spanish as a heritage language is frequently lost within a few generations).

These findings then pose two questions for Plains Cree: is there a change in the use of overt actors? And does such a change indicate internal change over time or evidence of English influence, or some combination thereof? The two subcorpora demonstrate significant differences in the use of overt actors; 20% of clauses in the BT subcorpus and 18% in the A-W subcorpus, χ^2 (1, N = 50,042) = 32.019, $p < .001$. However, as per the results throughout Chapters 4, 5, and 6, the occurrence of more narratives containing third person verbs (that are more likely to take overt actors) in the BT subcorpus may account for these differences. Of course, a direct comparison of narratives, with differences in dialogue and non-dialogue narratives accounted for, will better indicate if there is a change over time. Additionally, newer texts may also show more influence of English than just the differences between the BT and A-W subcorpora. Such questions are excellent candidates for future corpus studies for Plains Cree.

7.2.2.2 Flexible word order

In this section, I briefly consider the relative frequency of argument realisation patterns in Plains Cree compared to other languages with flexible word order. I focus on the occurrence of subjects and objects in Uralic languages, which bear similarities to Plains Cree in terms of not only word order flexibility, but also in their rich morphology (though not direction morphology). I look at three Uralic language corpora of varying sizes and content, annotated using the Universal Dependencies scheme: Moksha, Erzya, and Hungarian. Due to the annotation scheme, I compare only whether the overt arguments occur before or after the verb and I do not consider overt subjects

and objects together.[67] The frequencies are summarised in Table 7.11. (Rueter, 2018 for Moksha; Rueter & Tyers, 2018 for Erzya; Vincze & Csirik, 2010 for Hungarian). For Plains Cree, the corpus results show that overt actors and goals are more likely to occur after a verb than before it. Conversely, in all three Uralic languages under consideration, overt subjects are more likely to occur preverbally rather than postverbally, and in Hungarian, this is also the case for overt objects.

Table 7.11: Overt subjects and objects in Uralic languages and Plains Cree

Language	Moksha		Erzya		Hungarian		Plains Cree	
# of Tokens	3,175		17,336		42,032		152,405	
Composition	Fiction		Fiction		News articles		Narrative, lectures	
	#		#		#		#	
Total overt S(/A)	269		1,593		2,618		9,927	
	#	% Overt	#	% Overt	#	% Overt	#	% Overt
Preverbal S(/A)	189	70.3%	1,040	65.3%	1,984	75.8%	4,258	42.9%
Postverbal S(/A)	80	29.7%	553	34.7%	634	24.2%	5,669	57.1%
	#		#		#		#	
Total overt O(/G)	146		849		1,763			
	#	% Overt	#	% Overt	#	% Overt	#	% Overt
Preverbal O(/G)	34	23.3%	356	41.9%	993	56.3%	3,197	40.7%
Postverbal O(/G)	112	76.7%	493	58.1%	770	43.7%	4,651	59.3%

There is considerable variation among the Uralic data for both subjects and objects, though objects more noticeably so. The closest figures to the Plains Cree results are for objects in Moksha, with

[67] A brief comparison can be made to Erzya, where less than 20% of transitive clauses occur with two overt arguments (J. Rueter, personal communication). In Plains Cree (Table 5.2), this is ~4%.

42% occurring preverbally—otherwise, there are few similarities between these Uralic languages and Plains Cree (or Menomini, as discussed in §7.2.1, which is far more similar to Plains Cree than the Uralic languages are to each other). This variation or lack thereof within and between families could be a simple fact of the different families and languages, or the different sizes or content of the corpora used could be obscuring similarities or highlighting differences. Regardless, even just five languages from two families can demonstrate the extent to which argument realisation and flexible word order can vary, and future corpus investigations of this type will further illuminate the possible variation.

7.3 Text types across languages

As for typology, a comparison to a similar situation—an entirely spoken, morphosyntactically-tagged corpus of another under-resourced language for which a similar register analysis has been undertaken—is difficult, if not impossible, to provide in this work. Instead, I compare register analyses of majority languages to the Plains Cree results presented in Chapter 6, with reference to spoken portions of corpora where possible. Many such studies explore both written and spoken (or written-to-be-spoken) texts within a corpus, analysing various text types as more "literate" and more "oral". Analyses of written and spoken works in majority Indo-European languages, like English and Spanish, demonstrate similar patterns to dialogue vs. non-dialogue narrative in spoken Plains Cree, especially in prose fiction and plays; speeches, which might be likened to counselling texts, fall somewhere between spoken and written texts. Thus, in this section, I explore narratives and speeches.

When register analyses look at written vs. spoken or written-to-be-spoken text, it tends to be the first dimension (analogous to PC1 in Chapter 6) which has the strongest correlation to whether the texts are more literate (written, edited) or oral (spoken, spontaneous), and further dimensions tend to point towards narrative vs. non-narrative elements, or dialogue vs. non-dialogue elements, of the text types. One study, Guinovart & Guerra (2000), finds that English spoken texts have more features that suggest interaction, such as interrogatives, discourse particles, and interjections, while their written-to-be-spoken counterparts have features more reminiscent of written texts, with more informational features—the authors analyse this distinction as "dynamic deictic reference" in the spoken text and "notional richness" in the written text (pp. 57–9). Another study, Biber (1986)

explores spoken and written texts in English, finding a primary distinction between "Interactive vs. Edited Text", with more first and second persons, contractions, and questions on the interactive (spoken) side and longer words and a more varied vocabulary on the edited (written) side (pp. 393–5). Conversation and interviews are the most interactive and academic prose, press reports, and official documents are the most edited (p. 398). As the Plains Cree corpus is all spoken text, PC1 could not show a difference between spoken/oral or written/literate modes, but instead shows a difference between the two subcorpora. As explored in Chapter 6, this difference may arise to some extent due the content differences, but it may be also explained by the editing—the BT subcorpus displays more carefully edited text while the A-W subcorpus more closely represents the spoken text—which may be a similar to the distinction between literate and oral texts, though further research is required.

7.3.1 Dialogue vs. non-dialogue narrative

Many of the features that distinguished dialogue-heavy narrative in Plains Cree from non-dialogue narrative can also be seen in other languages. For English, Biber & Finegan (2011) find that non-dialogue narratives occur with more nouns, longer words, and a more varied vocabulary, while drama and dialogue narratives occur with more first and second person pronouns, verbs, discourse particles, and time and place adverbials. These distinctions are labelled as "informational vs. involved production" and "elaborated vs. situation-dependent reference" (p. 690). Thus, non-dialogue narrative is more likely to occur with more detailed, elaborated information, while dialogue occurs with more involvement between the speakers. Similarly, Assenova (2010) gives an "informational vs. interactive" label to describe the differences between non-dialogue and dialogue narrative in Bulgarian texts, much like the informational vs. involved distinction found by Biber & Finegan (2011). For Spanish texts, Biber et al. (2006) also find a distinction between "addressee-focused interaction" (drama, conversation), with more second person forms, questions, and exclamatives, and "informational reports of past events" (non-dialogue fiction), with more past tense. They also find patterns in irrealis marking in spoken discourse, where drama tends to be more irrealis and fiction much less so. Similar patterns are seen in the Plains Cree results presented in Chapter 6, where similar features group: the dialogue-heavy narrative, with more direct quotation, occurs with first and second person features, interjections (e.g., discourse

particles, exclamatives), questions, future tense (e.g., irrealis), and often with locative particles (e.g., place adverbials). However, locative and temporal particles are also often found with the non-dialogue narrative, alongside third person features and past tense where they are likely presenting more detailed information about the setting (e.g., informational reporting). Thus, the dialogue vs. non-dialogue distinction in Plains Cree narratives bears many similarities to not only dialogue and non-dialogue fiction in other languages, but also to spoken vs. written texts more generally.

7.3.2 Narrative vs. speeches

Though the BT subcorpus is entirely narrative, the A-W subcorpus allows for the exploration of the differences between other text types, namely narratives and counselling texts. In this section, I look at how speeches differ from other text types, especially narratives. Biber et al. (2006) found for Spanish that speeches tend to behave more like oral text types than literate types, between conversation/drama on the more oral end of the spectrum and fiction on the more literate. Features that characterise the more oral types include first and second person features, interrogatives, present and future tense, and third person pronouns. A starker difference is seen, however, in another of the dimensions described by Biber et al. (2006), where fiction and drama group together with respect to features labelled as "narrative discourse", where features such as past tense and third persons cooccur, and absence of such features characterise other text types, including speeches. Biber (1986) finds similar patterns for English text types, where conversations are among the most interactive text types, (written) narrative among the most edited, and speeches fall between them. Interactive features are similar to those for oral text types in Spanish, with interrogatives and first and second person pronouns. Speeches then contrast with narrative and conversation as more abstract, though not as abstract as official documents, and, similar to Spanish, spontaneous speeches are less abstract than planned speeches. Speeches contrast again with narrative as more immediate, rather than reported, with spontaneous speeches more immediate than planned ones.

With the exception of interrogatives, which are found more in the dialogue narratives for Plains Cree, the oral, interactive text type features also characterise the negative ranges of PC1 and PC2 of the A-W analysis in Chapter 6, which are identified as counselling texts, while the narrative

features are seen especially in the non-dialogue narrative throughout the analyses. Less abstract and less immediate texts, like narratives, tend to occur with third persons, past tense, and adverbs describing place and time, as already noted for Plains Cree narratives; these are again points where counselling texts differ. While Plains Cree, English, and Spanish narratives and dialogue also share some features with speeches, this highlights the variable nature of counselling texts commented upon in Chapter 6, where they can contain directives with some narrative to support or explain, or they can primarily use narrative to explain good ways of acting and living.

7.4 Conclusion

In this dissertation, I introduced morphosyntactic features of Plains Cree and applied them using the Constraint Grammar formalism to create an automatic syntactic parser for Plains Cree. I then applied this parser to a corpus of ~152,000 words of Plains Cree and presented results in three stages: 1) a general description of the frequencies of morphosyntactic features in the corpus, 2) an exploration of argument realisation in the corpus, focusing on how often actors and goals are overtly realised, and, when realised, how often they precede or follow the verb, and 3) a text type analysis of the content of the corpus, examining both existing text type labels and describing other distinctions that arose. Argument realisation in Plains Cree was compared to that in languages with similar typological features, especially hierarchical alignment and flexible word order, where it was observed that more topical participants are less likely to occur as overt nominals than less topical participants, and, when they are overt, more topical participants (whether pronouns, SAPs, or proximate third persons) are more likely to occur before a verb. Plains Cree text types were compared to those in majority languages, with the distinction between dialogue and non-dialogue oral narratives in Plains Cree found to be similar to the distinctions between written and spoken texts and between dialogue and non-dialogue written fiction in majority languages, with their much larger corpora containing both written and spoken material. The Plains Cree distinction between narratives and counselling texts also bears many similarities to differences between speeches and stories in other languages. Even with a relatively small corpus of limited scope and morphosyntactic tags, the results are informative not only regarding Plains Cree, but across languages.

Morphological and syntactic models, and corpora to which they are applied, are just a small step in language maintenance and linguistic research. Studies undertaken using such corpora, even if the tags are far from exhaustive, can still add to pre-existing descriptions and analyses based on smaller datasets, and indicate new questions for future research, for Plains Cree or other Algonquian languages, and for languages with rich morphology or flexible word order worldwide. In a time where technology in everyday life becomes increasingly ubiquitous, 21st century tools for Indigenous languages can play a valuable role in language maintenance and revitalisation. A majority language user can open a computer or phone and be confident they will be able to find keyboards, spell checkers, autocomplete, speech recognition, and automatic translation for their language; they can search for multiple dictionaries and hear pronunciations of words across various regions, and find large bodies of text for reading, research, or educational purposes. A morphosyntactically tagged corpus, its underlying models, and its searchable, online presence are a considerable step towards making these tools more widely available. A corpus can be linked to a dictionary, such as https://itwewina.altlab.app/, where users can see how words are used in context and which words are attested. A parallel corpus, though not yet ready for Plains Cree, can be used more effectively by learners and researchers, and can serve to further machine translation development. A parser can form the basis of a grammar checker and be used in language learning applications to build exercises and check answers. These tools, and others under development, must always be created with and for the language communities, to be best suited to their needs. Despite the arcane, impenetrable academic studies in works such as this dissertation, my foremost aim is always to contribute to tools that can and will be used by the Indigenous language speakers I work alongside.

References

Aboriginal Language Resources. (n.d.). South Slave Divisional Education Council. Retrieved September 22, 2022, from http://www.ssdec.nt.ca/ablang/

About the Innu Language. (n.d.). Retrieved August 29, 2022, from https://www.innu-aimun.ca/english/about/the-innu-language/

Acatlán Mixtec Dictionary. (2014). SIL International. https://www.sil.org/resources/archives/56285

Ahenakew, A. (2000). *âh-âyîtaw isi ê-kî-kiskêyihtahkik maskihkiy / They Knew Both Sides of Medicine: Cree Tales of Curing and Cursing Told by Alice Ahenakew* (H. C. Wolfart, Ed.). University of Manitoba Press.

Ahenakew, E. (1995). *Voice of the Plains Cree* (R. Buck, Ed.; Illustrated edition). University of Regina Press.

Ahenakew, F. (1987). *Cree language structures: A Cree approach*. Pemmican Publications.

Algonquian Dictionaries and Language Resources Project. (n.d.). Retrieved September 22, 2022, from https://www.algonquianlanguages.ca/

Algonquian Linguistics Atlas. (n.d.). Retrieved August 29, 2022, from https://www.atlas-ling.ca/

Antonsen, L., & Trosterud, T. (2011). Next to nothing—A cheap South Saami disambiguator. In B. S. Pederson, G. Nešpore, & I. Skadiņa (Eds.), *Proceedings of the 18th Nordic Conference of Computational Linguistics (NoDaLiDa 2011) Workshop Constraint Grammar Applications* (Vol. 14, pp. 61–69).

Arapaho Lexical Dictionary. (n.d.). Retrieved September 22, 2022, from https://verbs.colorado.edu/arapaho/public/view_search

Arppe, A. (2000). Developing a grammar checker for Swedish. *Proceedings of the 12th Nordic Conference of Computational Linguistics (NODALIDA 1999)*, 13–27. https://aclanthology.org/W99-1002

Arppe, A., Cox, C., Hulden, M., Lachler, J., Moshagen, S. N., Silfverberg, M., & Trosterud, T. (2017). *Computational Modeling of Verbs in Dene Languages: The Case of Tsuut'ina* (Working Papers in Athabaskan (Dene) Languages, pp. 51–69). Alaska Native Language Center.

Arppe, A., Junker, M.-O., & Torkornoo, D. (2017). Converting a comprehensive lexical database into a computational model: The case of East Cree verb inflection. *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, 52–56. https://doi.org/10.18653/v1/W17-0108

Arppe, A., Lachler, J., Trosterud, T., Antonsen, L., & Moshagen, S. N. (2016). Basic language resource kits for endangered languages: A case study of Plains Cree. *Proceedings of the 2nd Workshop on Collaboration and Computing for Under-Resourced Languages Workshop (CCURL 2016), Portorož, Slovenia*, 1–8.

Arppe, A., Wolvengrey, A., Poulin, J., Santos, E. A., Johnson, R., Harrigan, A, Thunder, D., Trosterud, T., & Antonsen, L. (under development). *itwêwina: nêhiyawêwin – âkayâsîmowin / English – Plains Cree intelligent on-line dictionary*. http://itwewina.altlab.app/

Assenova, D. (2010). Spoken vs. Written or Dialogue vs. Non-Dialogue?: Frequency Analysis of Verbs, Nouns and Prepositional Phrases in Bulgarian. *Slovo: Journal of Slavic Languages and Literatures*, *51*, 115–127.

Assini, A. A. (2013). *Natural language processing and the Mohawk language: Creating a finite state morphological parser of Mohawk formal nouns* [Master's thesis]. University of Limerick.

Bear, G. (1998). Lost and Found. In Bear et al., *kôhkominawak otâcimowiniwâwa // Our Grandmothers' Lives: As Told in Their Own Words* (F. Ahenakew & H. C. Wolfart, Eds.; Bilingual edition, pp. 123–144). University of Regina Press.

Bear, G., Calliou, I., Feitz, J., Fraser, M., Lafond, A., Longneck, R., & Wells, M. (1998). *kôhkominawak otâcimowiniwâwa // Our Grandmothers' Lives: As Told in Their Own Words* (F. Ahenakew & H. C. Wolfart, Eds.; Bilingual edition). University of Regina Press.

Biber, D. (1986). Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language*, *62*(2), 384–414.

Biber, D. (1991). *Variation across Speech and Writing*. Cambridge University Press.

Biber, D., & Conrad, S. (2019). *Register, Genre, and Style* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/9781108686136

Biber, D., Davies, M., Jones, J. K., & Tracy-Ventura, N. (2006). Spoken and written register variation in Spanish: A multi-dimensional analysis. *Corpora*, *1*(1), 1–37.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press.

Biber, D., & Finegan, E. (2011). The linguistic evolution of five written and speech-based English genres from the 17th to the 20th centuries. In *The linguistic evolution of five written and speech-based English genres from the 17th to the 20th centuries* (pp. 688–704). De Gruyter Mouton. https://doi.org/10.1515/9783110877007.688

Biber, D., Fitzmaurice, S. M., & Reppen, R. (2002). *Using Corpora to Explore Linguistic Variation*. John Benjamins.

Bick, E., & Didriksen, T. (2015). Cg-3—Beyond classical constraint grammar. *Proceedings of the 20th Nordic Conference of Computational Linguistics, NODALIDA 2015, May 11-13, 2015, Vilnius, Lithuania*, 31–39. https://aclanthology.org/W15-1807

Birn, J. (2000). Detecting grammar errors with Lingsoft's Swedish grammar checker. *Proceedings of the 12th Nordic Conference of Computational Linguistics, NODALIDA 1999, December 9-10, 1999, Trondheim, Norway*, 28–40. https://aclanthology.org/W99-1003

Blackfoot Online Dictionary. (n.d.). Retrieved September 22, 2022, from https://dictionary.blackfoot.atlas-ling.ca/

Blain, E. M. (1997). *Wh-constructions in nêhiyawêwin (Plains Cree).* [Doctoral dissertation, University of British Columbia]. https://dx.doi.org/10.14288/1.0087953

Bloomfield, L. (Ed.). (1930). *Sacred Stories of the Sweet Grass Cree*. F.A. Acland.

Bloomfield, L. (Ed.). (1934). *Plains Cree Texts* (Vol. 13). G.E. Stechert & Co.

Bloomfield, L. (1946). Algonquian. In *Linguistic structures of Native America* (Vol. 6, pp. 85–129). Viking Fund Publications in Anthropology.

Bloomfield, L. (1962). *The Menomini Language* (C. Hockett, Ed.). Yale University Press.

Bontogon, M., Arppe, A., Antonsen, L., Thunder, D., & Lachler, J. (2018). Intelligent computer assisted language learning (ICALL) for nêhiyawêwin: An in-depth user-experience evaluation. *Canadian Modern Language Review*, *74*(3), 337–362.

Bowers, D., Arppe, A., Lachler, J., Moshagen, S., & Trosterud, T. (2017). A morphological parser for Odawa. *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, 1–9.

Brixey, J., & Artstein, R. (2020). ChoCo: A multimodal corpus of the Choctaw language. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-020-09494-5

Brockie, T., & Cowell, A. (2015). Editing a Gros Ventre (White Clay) Text. In D. J. Costa (Ed.), *New Voices for Old Words: Algonquian Oral Languages* (pp. 9–20). University of Nebraska Press.

Bungarten, T. (1979). Das Korpus als empirische Grundlage in der Linguistik und Literaturwissenschaft. In Bergenholtz, H. and Schaeder, B. (Eds.), *Empirische Textwissenschaft. Aufbau und Auswertung von Text-Corpora*. Monografien Linguistik und Kommunikationswissenschaft 39. Scriptor.

Cabrera-Puche, M. J. (2008). *Null subject patterns in language contact: The case of Dominican Spanish* [PhD dissertation]. Rutgers - New Brunswick.

Camacho, J. (2008). Syntactic variation: The case of Spanish and Portuguese subjects. *Studies in Hispanic and Lusophone Linguistics*, *1*(2), 415–434.

Campana, M. (1989). Algonquian and the Absolutive Case Hypothesis. In W. Cowan (Ed.), *Actes du vingtième congrès des algonquianistes*. Carleton University.

Chamacoco Talking Dictionary. (n.d.). Retrieved September 22, 2022, from https://talkingdictionary.swarthmore.edu/chamacoco/

Chiruzzo, L., Amarilla, P., Ríos, A., & Giménez Lugo, G. (2020). Development of a Guarani—Spanish Parallel Corpus. *Proceedings of the 12th Language Resources and Evaluation Conference*, 2629–2633. https://aclanthology.org/2020.lrec-1.320

Coler, M., & Homola, P. (2014). Rule-based machine translation for Aymara. In M. C. Jones (Ed.), *Endangered Languages and New Technologies* (pp. 67–80). Cambridge University Press.

Cook, C. (2014). *The Clause-Typing System of Plains Cree: Indexicality, Anaphoricity, and Contrast*. Oxford University Press.

Cook, C., Rose-Marie Déchaine, & Mühlbauer, J. (2003). *Rhetorical Structure of a Plains Cree Counselling Speech*. 35th Algonquian Conference, Winnipeg, MB.

Cordova, J., & Nouvel, D. (2021). Toward Creation of Ancash Quechua Lexical Resources from OCR. In *Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP). Proceedings of the First Workshop* (pp. 163–167). The Association for Computational Linguistics. https://hal.archives-ouvertes.fr/hal-03610330

Costa, D. J. (2015). *New Voices for Old Words: Algonquian Oral Literatures*. University of Nebraska Press.

Cowell, A. (Ed.). (2012). *Dictionary of the Arapaho Language*. https://homewitharapaho.files.wordpress.com/2015/03/arapaho-dictionary1.pdf

Cree – English Online Dictionary. (n.d.). Retrieved September 22, 2022, from https://moosecree.ca/

da Silva, H. S. (2006). *O Parâmetro Do Sujeito Nulo: Confronto Entre O Português E O Espanhol* [Master's thesis]. Universidade Federal do Rio de Janeiro.

Dacanay, D., Harrigan, A., & Arppe, A. (2021). Computational Analysis versus Human Intuition: A Critical Comparison of Vector Semantics with Manual Semantic Classification in the Context of Plains Cree. *Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, 33–43. https://aclanthology.org/2021.computel-1.5

Dahlstrom, A. (1991). *Plains Cree Morphosyntax*. Taylor and Francis.

Dahlstrom, A. (1995). *Topic, focus and other word order problems in Algonquian*. Voices of Rupert's Land.

Dakota Dictionary Online. (n.d.). Retrieved September 22, 2022, from https://fmp.cla.umn.edu/dakota/

Dakota-English Dictionary. (n.d.). Retrieved September 22, 2022, from https://dictionary.swodli.com/

Davies, M. (2008-). *The Corpus of Contemporary American English (COCA): One billion words, 1990-2019.* https://www.english-corpora.org/coca/

Davis, F., Santos, E. A., & Souter, H. (2021). On the Computational Modelling of Michif Verbal Morphology. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2631–2636.

Demers, P., McIlwraith, N. L., & Thunder, D. (2010). *The Beginning of Print Culture in Athabasca Country: A Facsimile Edition & Translation of a Prayer Book in Cree Syllabics by Father Émile Grouard, OMI, Prepared and Printed at Lac La Biche in 1883 with an Introduction by Patricia Demers*. The University of Alberta Press.

Domeij, R., Karlsson, O., Moshagen, S., & Trosterud, T. (2019). Enhancing information accessibility and digital literacy for minorities using language technology—The example of Sami and other national minority languages in Sweden. In C. Cocq & K. Sullivan (Eds.), *Perspectives on Indigenous writing and literacies* (Vol. 37, pp. 113–137). Brill.

Douquette, J. (1987). Wishful Thinking. In J. Vandall & P. Douquette, *wâskahikaniwiyiniw-âcimowina / Stories of the House People, Told by Peter Vandall and Joe Douquette* (F. Ahenakew, Ed.; pp. 70–75). University of Manitoba Press.

Dryer, M. S. (1994). The discourse function of the Kutenai inverse. In T. Givón (Ed.), *Voice and inversion* (pp. 65–99). John Benjamins.

Dryer, M. S. (2007). Kutenai, Algonquian, and the Pacific Northwest from an areal perspective. In H. C. Wolfart (Ed.), *Papers of the 38th Algonquian Conference* (pp. 155–206). University of Manitoba Press.

Dryer, M. S. (2013). Order of Subject, Object and Verb. In M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology. https://wals.info/chapter/81

Dryer, M. S., & Haspelmath, M. (Eds.). (2013). *WALS Online*. Max Planck Institute for Evolutionary Anthropology. https://wals.info/

Du Bois, J. W. (1987). The discourse basis of ergativity. *Language*, *63*(4), 805–855.

Du Bois, J. W. (2003). Argument structure: Grammar in use. In J. W. Du Bois, L. E. Kumpf, & W. J. Ashby, *Preferred argument structure: Grammar as architecture for function* (Vol. 14, pp. 11–60). John Benjamins.

Duarte, M. E. L. (2000). The loss of the "avoid pronoun" principle in Brazilian Portuguese. In M. A. Kato & E. Negrão (Eds.), *Brazilian Portuguese and the Null Subject Parameter* (pp. 17–36). Vervuert.

East Cree dialects. (n.d.). Retrieved August 29, 2022, from https://www.eastcree.org/cree/en/grammar/east-cree-dialects/

Eastern James Bay Cree Dictionary on the Web: Home. (n.d.). Retrieved July 17, 2022, from https://dictionary.eastcree.org/

English-Inuttut Dictionary. (n.d.). Retrieved September 22, 2022, from http://www.labradorvirtualmuseum.ca/english-inuttut.htm

Florence, M. (2019). *kimotinâniwiw itwêwina / Stolen Words* (D. Sand & G. Weenie, Trans.; Bilingual edition). Second Story Press.

Flores-Ferrán, N. (2004). Spanish subject personal pronoun use in New York City Puerto Ricans: Can we rest the case of English contact? *Language Variation and Change*, *16*(1), 49–73.

Fort Severn Online Dictionary: Home. (n.d.). Retrieved July 17, 2022, from https://fortsevern.atlas-ling.ca/

Frey, B. (2018). *"Data is Nice:" Theoretical and pedagogical implications of an Eastern Cherokee corpus*. University of Hawai'i Press. http://hdl.handle.net/10125/24931

Gasser, M. (2018). *Mainumby: Un Ayudante para la Traduccíon Castellano-Guaraní*. https://doi.org/10.48550/arXiv.1810.08603

Grammar. (n.d.). Retrieved September 22, 2022, from https://www.eastcree.org/cree/en/grammar/

Greenlandic Dictionaries. (n.d.). Retrieved September 22, 2022, from https://ordbog.gl/

Guile, T. (2001). Sketch of Menominee grammar. In *An Anthology of Menominee sayings, with translations, annotations, and grammatical sketch* (pp. 452–501).

Guinovart, J. G., & Guerra, J. P. (2000). A multidimensional corpus-based analysis of English spoken and written-to-be-spoken discourse. *Cuadernos de Filología Inglesa*, *9*(1). https://revistas.um.es/cfi/article/view/66431

Gúusaaw Northern Haida dictionary. (n.d.). Retrieved December 16, 2022, from https://guusaaw.altlab.dev/

Harrigan, A. (forthcoming). *A quantitative account of* nêhiyawêwin *order: Using mixed effects modeling to uncover syntactic, semantic, and morphological motivations in* nêhiyawêwin. [Doctoral dissertation]. University of Alberta.

Harrigan, A., Schmirler, K., Arppe, A., Antonsen, L., Trosterud, T., & Wolvengrey, A. (2017). Learning from the computational modelling of Plains Cree verbs. *Morphology*, *27*(4), 565–598. https://doi.org/10.1007/s11525-017-9315-x

Harrigan, A., Mills, T., & Arppe, A. (2019). A Preliminary Plains Cree Speech Synthesizer. *Proceedings of the 3rd Workshop on Computational Methods for Endangered Languages Volume 1 (Papers)*, 64–73.

Harris, E. K. (2018). *tipiskâwi-kêhkêhk êkwa apisci wâwâskêsiw.* (M. Kyplain, Trans.). TaleFeather Publishing.

Haude, K. (2014). Animacy and inverse in Movima: A corpus study. *Anthropological Linguistics*, *56*(3), 294–314.

Heritage Michif Dictionary. (n.d.). Retrieved September 22, 2022, from http://www.metismuseum.ca/michif_dictionary.php

Hewson, J. (1987). Are Algonquian Languages Ergative? In W. Cowan (Ed.), *Papers of the 18th Algonquian Conference*. Carleton University.

Holden, J., Cox, C., & Arppe, A. (2022). An Expanded Finite-State Transducer for Tsuut'ina Verbs. *Proceedings of the Language Resources and Evaluation Conference*, 5143–5152. https://aclanthology.org/2022.lrec-1.551

Hubert, I., Arppe, A., Lachler, J., & Santos, E. A. (2016). Training & quality assessment of an optical character recognition model for Northern Haida. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 3227–3234. https://aclanthology.org/L16-1514/

Innu Dictionary. (n.d.). Retrieved July 17, 2022, from https://dictionary.innu-aimun.ca/Words

Inuktitut Transcoder. (n.d.). Retrieved September 25, 2022, from https://www.inuktitutcomputing.ca/Transcoder/index.php

Inuktut Tusaalanga. (n.d.). Retrieved September 22, 2022, from https://tusaalanga.ca/dialect

IOM Dictionary. (n.d.). Retrieved September 22, 2022, from http://www.iowayotoelang.nativeweb.org/dictionary.htm

Joanis, E., Knowles, R., Kuhn, R., Larkin, S., Littell, P., Lo, C., Stewart, D., & Micher, J. (2020). The Nunavut Hansard Inuktitut–English Parallel Corpus 3.0 with Preliminary Machine

Translation Results. *Proceedings of the 12th Language Resources and Evaluation Conference*, 2562–2572. https://aclanthology.org/2020.lrec-1.312

Jones, A. (2022). *Finetuning a Kalaallisut-English machine translation system using web-crawled data*. http://arxiv.org/abs/2206.02230

Kâ-Nîpitêhtêw, J. (1998). *ana kâ-pimwêwêhahk okakêskihkêmowina / The Counselling Speeches of Jim Kâ-Nîpitêhtêw* (F. Ahenakew & H. C. Wolfart, Eds.). University of Manitoba Press.

Karlsson, F. (1990). Constraint grammar as a framework for parsing running text. In *COLING 1990 Volume 3: Papers presented to the 13th International Conference on Computational Linguistics*.

Karlsson, F. (2011a). Designing a parser for unrestricted text. In F. Karlsson, A. Voutilainen, J. Heikkilae, & A. Anttila (Eds.), *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text* (pp. 1–40). Walter de Gruyter.

Karlsson, F. (2011b). The formalism and environment of Constraint Grammar Parsing. In F. Karlsson, A. Voutilainen, J. Heikkilae, & A. Anttila (Eds.), *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text* (pp. 41–88). Walter de Gruyter.

Karlsson, F., Voutilainen, A., Heikkilae, J., & Anttila, A. (2011). *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Walter de Gruyter.

Kazantseva, A., Maracle, O. B., Maracle, R. J., & Pine, A. (2018). Kawennón:nis: The Wordmaker for Kanyen'kéha. *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, 53–64. https://aclanthology.org/W18-4806

Kazeminejad, G., Cowell, A., & Hulden, M. (2017). Creating lexical resources for polysynthetic languages—The case of Arapaho. *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, 10–18. https://doi.org/10.18653/v1/W17-0102

Knowles, R., Stewart, D., Larkin, S., & Littell, P. (2020). NRC Systems for the 2020 Inuktitut-English News Translation Task. *Proceedings of the Fifth Conference on Machine Translation*, 156–170. https://aclanthology.org/2020.wmt-1.13

Kuznetsova, A., & Tyers, F. (2021). A finite-state morphological analyser for Paraguayan Guaraní. *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, 81–89. https://doi.org/10.18653/v1/2021.americasnlp-1.9

Lafond, A., & Longneck, R. (1998). Reminiscences of Muskeg Lake. In F. Ahenakew & H. C. Wolfart (Eds.), *kôhkominawak otâcimowiniwâwa // Our Grandmothers' Lives: As Told in Their Own Words* (Bilingual edition). University of Regina Press.

Language Resources. (n.d.). Sealaska Heritage. Retrieved September 22, 2022, from https://www.sealaskaheritage.org/institute/language/resources

Lavallee, R., & Silverthorne, J. (2014). *Honouring the Buffalo: A Plains Cree Legend* (1st edition). Your Nickel's Worth Publishing.

Le, T. N., & Sadat, F. (2020). Low-Resource NMT: An Empirical Study on the Effect of Rich Morphological Word Segmentation on Inuktitut. *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, 165–172. https://aclanthology.org/2020.amta-research.15

LeClaire, N., Cardinal, G., & Chalifoux, T. J. (1998). *Alberta Elders' Cree Dictionary / alperta ohci kehtehayak nehiyaw otwestamâkewasinahikan* (E. H. Waugh, Ed.; Illustrated edition). University of Alberta Press.

Leech, G. (2007). New resources, or just better old ones? The Holy Grail of representativeness. In M. Hundt, N. Nesselhauf, & C. Biewer (Eds.), *Corpus linguistics and the web* (Vol. 59, pp. 133–149). Brill. https://brill.com/view/book/edcoll/9789401203791/B9789401203791-s009.xml

Littell, P. (2018). Finite-state morphology for Kwak'wala: A phonological approach. *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, 21–30. https://aclanthology.org/W18-4803

Liu, Z., Spence, J., & Tucker Prud'hommeaux, E. (2022). Enhancing Documentation of Hupa with Automatic Speech Recognition. *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, 187–192. https://doi.org/10.18653/v1/2022.computel-1.23

Lovick, O., Cox, C., Silfverberg, M., Arppe, A., & Hulden, M. (2018, May). A Computational Architecture for the Morphology of Upper Tanana. *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*. LREC 2018, Miyazaki, Japan. https://aclanthology.org/L18-1294

MacKenzie, M. E. (1980). *Towards a dialectology of Cree-Montagnais-Naskapi* [Doctoral dissertation]. University of Toronto.

Martin, J., Johnson, H., Farley, B., & Maclachlan, A. (2003). Aligning and Using an English-Inuktitut Parallel Corpus. *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, 115–118. https://aclanthology.org/W03-0320

Martínez-Gil, C., Zempoalteca-Pérez, A., Soancatl Aguilar, V., Estudillo-Ayala, M., Lara, J., & Alcántara-Santiago, S. (2012). Computer Systems for Analysis of Nahuatl. *Research in Computing Science*, *47*, 11–16. https://doi.org/10.13053/rcs-47-1-1

Masuskapoe, C. (2010). *piko kîkway ê-nakacihtât: kêkêk otâcimowina ê-nêhiyawastêki* (H. C. Wolfart & F. Ahenakew, Eds.). Algonquian and Iroquoian Linguistics.

Matter, F. (2020). An inspection of Preferred Argument Structure in Mapudungun narratives. *International Journal of American Linguistics*, *86*(1), 59–93.

McLeod, N., & Wolvengrey, A. (2016). *100 Days of Cree* (Bilingual edition). University of Regina Press.

Micher, J. (2017). Improving coverage of an Inuktitut morphological analyzer using a segmental recurrent neural network. *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, 101–106.

Michif Online Dictionary. (n.d.). Retrieved September 22, 2022, from https://dictionary.michif.atlas-ling.ca/

Mi'gmaq Mi'kmaq Micmac Online Talking Dictionary. (n.d.). Retrieved September 22, 2022, from https://www.mikmaqonline.org/

Minde, E. (1997). *kwayask ê-kî-pê-kiskinowâpatihicik / Their Example Showed Me the Way: A Cree Woman's Life Shaped by Two Cultures* (F. Ahenakew & H. C. Wolfart, Eds.). University of Alberta Press.

Miso, J. (2020). *ᐲᓯᒣᐧᓵᐱᐩ*. Jean Miso.

Moeller, S., Kazeminejad, G., Cowell, A., & Hulden, M. (2018). A Neural Morphological Analyzer for Arapaho Verbs Learned from a Finite State Transducer. *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, 12–20. https://aclanthology.org/W18-4802

Moose & Eastern Swampy Cree Online Dictionary. (n.d.). Retrieved September 22, 2022, from https://dictionary.moosecree.atlas-ling.ca/

Moseley, Christopher (Ed.). (2010). *Atlas of the World's Languages in Danger* (3rd ed.). Paris, UNESCO Publishing. Online version: http://www.unesco.org/culture/en/endangeredlanguages/atlas

Mother tongue by geography, 2021 Census. (n.d.). Retrieved September 22, 2022, from https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/dv-vd/language-langue/index-en.html

Mühlbauer, J. (2007). *Word order and the interpretation of nominals* [Manuscript].

Myaamia-Peewaalia Dictionary. (n.d.). Retrieved September 22, 2022, from https://mc.miamioh.edu/ilda-myaamia/dictionary

Nagy, N. G., Aghdasi, N., Denis, D., & Motut, A. (2011). Null subjects in heritage languages: Contact effects in a cross-linguistic context. *University of Pennsylvania Working Papers in Linguistics*, *17*(2), 135–144.

Nasa yuwe Talking Dictionary. (n.d.). Retrieved September 22, 2022, from https://talkingdictionary.swarthmore.edu/paez/

Naskapi Online Dictionary. (n.d.). Retrieved September 22, 2022, from https://dictionary.naskapi.atlas-ling.ca/

Nishnaabemwin Online Dictionary. (n.d.). Retrieved September 22, 2022, from https://dictionary.nishnaabemwin.atlas-ling.ca/

NLD Online v.5. (n.d.). Retrieved September 22, 2022, from https://nldo.lakotadictionary.org/

Northern Michif Dictionary. (n.d.). Retrieved September 22, 2022, from http://www.metismuseum.ca/northern_michif_dictionary.php

Odribets, Z., & Oxford, W. (in press). Algonquian languages are not ergative. In M. Macaulay & M. Noodin (Eds.), *Papers of the 52nd Algonquian Conference* (18 pages). MSU Press.

Ogg, A. C. (1991). *Connective particles and temporal cohesion in Plains Cree narrative* [Master's thesis]. University of Manitoba.

Okimâsis, J. L. (2004). *Cree: Language of the Plains / nêhiyawêwin: paskwâwi-pîkiskwêwin*. University of Regina Press.

Okimâsis, J. L. (2021). *Cree: Language of the Plains / nêhiyawêwin: paskwâwi-pîkiskwêwin* (Second Edition). University of Regina Press.

Okimâsis, J. L., & Wolvengrey, A. (2008). *How to spell it in Cree: The standard Roman orthography*. Miywâsin Ink. http://dare.uva.nl/personal/pure/en/publications/how-to-spell-it-in-cree–the-standard-roman-orthography(02ae6224-fbd6-4516-94cd-a3ad6dde0e64).html

Online Atikamekw Dictionary. (n.d.). Retrieved September 22, 2022, from https://dictionnaire.atikamekw.atlas-ling.ca/

Online Cree Dictionary, Cree Language, Cree: Words, Alberta Elders' Dictionary, Maskwacis. (n.d.). Retrieved September 22, 2022, from https://www.creedictionary.com/

Oppenneer, M. (2013, August 13). *Indigenous Language Apps & Online Indigenous Language Dictionaries*. The Ethnos Project. https://www.ethnosproject.org/indigenous-language-apps-online-indigenous-language-dictionaries/

Oqaasileriffik har indsamlet grønlandske ord [Oqaasileriffik has collected Greenlandic words]. (2021, January 24). *Sermitsiaq*. https://sermitsiaq.ag/node/226847

Ortega, J. E., Castro Mamani, R., & Cho, K. (2020). Neural machine translation with a polysynthetic low resource language. *Machine Translation*, *34*(4), 325–346. https://doi.org/10.1007/s10590-020-09255-9

Otheguy, R., Zentella, A. C., & Livert, D. (2007). Language and dialect contact in Spanish in New York: Toward the formation of a speech community. *Language*, *83*(4), 770–802.

Oxford, W. (2008). *A grammatical study of Innu-aimun particles*. Algonquian and Iroquoian Linguistics.

Pankratz, E., Arppe, A., & Lachler, J. (2022). Low hanging fruit and the Boasian trilogy in digital lexicography of morphologically rich languages: Lessons from a survey of Indigenous language resources in Canada. *Nordlyd*, *46*(1), Article 1. https://doi.org/10.7557/12.6441

Park, H. H., Schwartz, L., & Tyers, F. (2021). Expanding Universal Dependencies for Polysynthetic Languages: A Case of St. Lawrence Island Yupik. *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, 131–142. https://doi.org/10.18653/v1/2021.americasnlp-1.14

Pine, A., & Turin, M. (2018). Seeing the Heiltsuk orthography from font encoding through to Unicode: A case study using convertextract. *Proceedings of the LREC 2018 Workshop "CCURL 2018–Sustaining Knowledge Diversity in the Digital Age,"* 27–30.

Plains Cree: Online dictionary. (n.d.). Retrieved September 22, 2022, from https://dictionary.plainscree.atlas-ling.ca/

Proto-Algonquian Online Dictionary. (n.d.). Retrieved September 22, 2022, from https://protoalgonquian.atlas-ling.ca/

Proulx, P. (1988). The demonstrative pronouns of Proto-Algonquian. *International Journal of American Linguistics*, *54*(3), 309–330.

Pugh, R., Huerta Mendez, M., Sasaki, M., & Tyers, F. (2022). Universal Dependencies for Western Sierra Puebla Nahuatl. *Proceedings of the Language Resources and Evaluation Conference*, 5011–5020. https://aclanthology.org/2022.lrec-1.535

Pugh, R., & Tyers, F. (2021). Towards an Open Source Finite-State Morphological Analyzer for Zacatlán-Ahuacatlán-Tepetzintla Nahuatl. *Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, 80–85. https://aclanthology.org/2021.computel-1.10

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org

Ratt, S. (2016a). *mâci-nêhiyawêwin / Beginning Cree*. University of Regina Press.

Ratt, S. (2022). *âhkami-nêhiyawêtân / Let's Keep Speaking Cree*. University of Regina Press.

Rehrig, S. (2017). *A Guide to Building a Diphone Speech Synthesis System for Kalaallisut* [Honours Thesis, Bryn Mawr College]. https://scholarship.tricolib.brynmawr.edu/handle/10066/19065

Resources – Language Secretariat. (n.d.). Retrieved September 17, 2022, from https://oqaasileriffik.gl/en/resources/

Rhodes, R., & Todd, E. (1981). Subarctic Algonquian languages. In *Subarctic* (Vol. 6, pp. 52–66). Smithsonian Institution.

Rios, A. (2015). *A Basic Language Technology Toolkit for Quechua* [Doctoral dissertation]. University of Zurich. https://www.zora.uzh.ch/id/eprint/119943/

Rockthunder, M. L. (2021). *kayas nohcin: I Come from a Long Time Back* (J. L. Okimâsis & A. Wolvengrey, Eds.). University of Regina Press.

Roest, C., Edman, L., Minnema, G., Kelly, K., Spenader, J., & Toral, A. (2020). Machine Translation for English–Inuktitut with Segmentation, Data Acquisition and Pre-Training.

Proceedings of the Fifth Conference on Machine Translation, 274–281. https://aclanthology.org/2020.wmt-1.29

Rueter, J. (2018). *Open Erme Moksha* (v1.0). https://github.com/rueter/erme-ud-moksha/tree/v1.0

Rueter, J., Fernanda Pereira de Freitas, M., Da Silva Facundes, S., Hämäläinen, M., & Partanen, N. (2021). Apurinã Universal Dependencies Treebank. *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, 28–33. https://doi.org/10.18653/v1/2021.americasnlp-1.4

Rueter, J., & Tyers, F. (2018). Towards an open-source universal-dependency treebank for Erzya. *Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages*, 106–118.

Sainte-Marie, B. (2022). *tâpwê êkwa mamâhtâwastotin* (S. Ratt, Trans.). Greystone Kids.

Santos, E., & Harrigan, A. (2020). Design and evaluation of a smartphone keyboard for Plains Cree syllabics. *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, 88–96. https://aclanthology.org/2020.sltu-1.12

Schmirler, K. (in press). Infrequent Morphosyntactic Phenomena in Plains Cree: Bloomfield's Text Collections and the Ahenakew-Wolfart Corpus. In M. Macaulay & M. Noodin (Eds.), *Papers of the 52nd Algonquian Conference* (pp. 285–302). MSU Press.

Schmirler, K., Arppe, A., Trond Trosterud, & Lene Antonsen. (under development). *A Constraint Grammar syntactic parser for Plains Cree*. https://github.com/giellalt/lang-crk/tree/main/src/cg3

Schmirler, K., & Arppe, A. (2019a). Modelling Plains Cree Negation with Constraint Grammar. *Proceedings of the NoDaLiDa 2019 Workshop on Constraint Grammar-Methods, Tools and Applications, 30 September 2019, Turku, Finland*, 27–34.

Schmirler, K., & Arppe, A. (2019b). *Plains Cree actors and goals: Across time periods and genres*. Paper presented at the 51st Algonquian Conference at McGill University in Montreal, Quebec, October 25-27, 2019.

Schmirler, K., & Arppe, A. (2020). *A quantitative look at Plains Cree text types: âtayôhkêwina vs. â*cimowina *in Bloomfield's texts and âcimisowina vs. kakêskihkêmowina in the Ahenakew-Wolfart corpus*. 52nd Algonquian Conference at UW Madision [hosted online], October 23-25, 2020.

Schmirler, K., & Arppe, A. (2021). Plains Cree textual analysis with PCA: Across the Bloomfield and Ahenakew-Wolfart subcorpora. Paper presented at the 53rd Algonquian Conference at Carleton University [hosted online], October 14-17, 2021.

Schmirler, K., & Arppe, A. (forthcoming). Plains Cree textual analysis with PCA: Across the Bloomfield and Ahenakew-Wolfart subcorpora. In M. Macaulay & M. Noodin (Eds.), *Papers of the 53rd Algonquian Conference*. MSU Press.

Schmirler, K., Arppe, A., Trosterud, T., & Antonsen, L. (2018). Building a Constraint Grammar Parser for Plains Cree Verbs and Arguments. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Shields, R. (2004). Word Order and Discourse in Menominee. In H. C. Wolfart (Ed.), *Papers of the 35th Algonquian Conference* (pp. 373–388). University of Manitoba Press.

Siewierska, A. (2013a). Alignment of Verbal Person Marking. In M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology. https://wals.info/chapter/100

Siewierska, A. (2013b). Verbal Person Marking. In M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology. https://wals.info/chapter/102

Siletz Talking Dictionary. (n.d.). Retrieved September 22, 2022, from https://siletz.swarthmore.edu/

Smith, S. D. (2013). *Pro-drop and word-order variation in Brazilian Portuguese: A corpus study* [Master's Thesis]. Brigham Young University.

Snoek, C., Thunder, D., Lõo, K., Arppe, A., Lachler, J., Moshagen, S., & Trosterud, T. (2014). Modeling the Noun Morphology of Plains Cree. *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, 34–42. http://www.aclweb.org/anthology/W14-2205

Spoken Cree , Cree Legends and Narratives: Glossary. (n.d.). Retrieved September 22, 2022, from https://www.spokencree.org/glossary

Stoney Nakoda Dictionary. (n.d.). Retrieved September 22, 2022, from https://dictionary.stoneynakoda.org/

Tapanainen, P. (1996). The Constraint Grammar Parser CG-2. *Publications of the Department of General Linguistics*, *27*.

Teodorescu, D., Matalski, J., Lothian, D., Barbosa, D., & Demmans Epp, C. (2022). Cree Corpus: A Collection of nêhiyawêwin Resources. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6354–6364. https://doi.org/10.18653/v1/2022.acl-long.440

The British National Corpus, version 2 (BNC World). 2001. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

The Lenape Talking Dictionary | Home. (n.d.). Retrieved September 22, 2022, from https://www.talk-lenape.org/

The Ojibwe People's Dictionary. (n.d.). Retrieved August 24, 2016, from http://ojibwe.lib.umn.edu/en

The Passamaquoddy-Maliseet Dictionary. (n.d.). Passamaquoddy-Maliseet Language Portal. Retrieved September 22, 2022, from https://pmportal.org/browse-dictionary

Thomas, G. (2019). Universal Dependencies for Mbyá Guaraní. *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, 70–77. https://doi.org/10.18653/v1/W19-8008

Tłı̨chǫ Yatıì Multimedia Dictionary. (n.d.). Retrieved September 22, 2022, from http://tlicho.ling.uvic.ca/users/mainview.aspx?AspxAutoDetectCookieSupport=1

Toribio, A. J. (2000). Setting parametric limits on dialectal variation in Spanish. *Lingua*, *110*(5), 315–341.

Torres, A. I., Miller, J., Oncevay, A., & Zariquiey Biondi, R. (2021). Representation of Yine [Arawak] Morphology by Finite State Transducer Formalism. *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, 102–112. https://doi.org/10.18653/v1/2021.americasnlp-1.11

Trosterud, T., & Moshagen, S. (2021). Soft on errors? The correcting mechanism of a Skolt Sami speller. In University of Helsinki, M. Hämäläinen, N. Partanen, & K. Alnajjar (Eds.), *Multilingual Facilitation* (pp. 197–207). University of Helsinki. https://doi.org/10.31885/9789515150257.19

Truth & reconciliation: Calls to action. (2015). National Centre for Truth and Reconciliation. https://nctr.ca/map.php

Tu Idau. (n.d.). Retrieved September 22, 2022, from http://mixtec.nativeweb.org/

Tyers, F., & Henderson, R. (2021). A corpus of K'iche' annotated for morphosyntactic structure. *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, 10–20. https://doi.org/10.18653/v1/2021.americasnlp-1.2

Valentine, J. R. (2001). *Nishnaabemwin reference grammar*. University of Toronto Press. http://myaccess.library.utoronto.ca/login?url=http://books.scholarsportal.info/viewdoc.html?id=/ebooks/ebooks0/gibson_crkn/2009-12-01/6/418193

van Eijk, J. (2013). *Lillooet-English Dictionary* (Vol. 2). https://lingpapers.sites.olt.ubc.ca/files/2018/01/Van_Eijk_Lillooet-English-Dictionary1-1.pdf

Vandall, P., & Douquette, J. (1987). *wâskahikaniwiyiniw-âcimowina / Stories of the House People, Told by Peter Vandall and Joe Douquette* (F. Ahenakew, Ed.). University of Manitoba Press.

Vincze, V., & Csirik, J. (2010). Hungarian corpus of light verb constructions. *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, 1110–1118.

Wagner, I., Cowell, A., & Hwang, J. D. (2016). Applying Universal Dependency to the Arapaho Language. *Proceedings of the 10th Linguistic Annotation Workshop Held in Conjunction with ACL 2016 (LAW-X 2016)*, 171–179. https://doi.org/10.18653/v1/W16-1719

Western Abenaki Dictionary. (n.d.). Retrieved September 22, 2022, from http://westernabenaki.com/

Whitecalf, S. (1993). *kinêhiyawiwiniwaw nêhiyawêwin / The Cree Language is Our Identity: The La Ronge Lectures of Sarah Whitecalf* (H. C. Wolfart & F. Ahenakew, Eds.). University of Manitoba Press.

Whitecalf, S., & Whitecalf, T. (2021). *mitoni niya nêhiyaw / Cree is Who I Truly Am: nêhiyaw-iskwêw mitoni niya / Me, I am Truly a Cree Woman* (H. C. Wolfart & F. Ahenakew, Trans.). University of Manitoba Press.

Wiechetek, L. (2018). *When grammar can't be trusted—Valency and semantic categories in North Sámi syntactic analysis and error detection*. https://munin.uit.no/handle/10037/12726

Wiechetek, L., & Arriola, J. M. (2011). An experiment of use and reuse of verb valency in morphosyntactic disambiguation and machine translation for Euskara and North Sámi. *Proceedings of the 18th Nordic Conference of Computational Linguistics (NoDaLiDa 2011) Workshop Constraint Grammar Applications*, *14*, 61–69.

Wolfart, H. C. (1973). *Plains Cree: A Grammatical Study: Vol. 63.5*. American Philosophical Society.

Wolfart, H. C. (1996). Sketch of Cree, an Algonquian Language. In *Handbook of American Indians. Volume 17: Languages* (Vol. 17, pp. 390–439). Smithsonian Institute.

Wolvengrey, A. (2001). *nêhiyawêwin: itwêwina / Cree: Words* (Bilingual edition). University of Regina Press.

Wolvengrey, A. (Ed.). (2007). *wawiyatâcimowinisa / Funny Little Stories* (Illustrated edition). University of Regina Press.

Wolvengrey, A. (2011). *Semantic and pragmatic functions in Plains Cree syntax*. LOT. http://dare.uva.nl/record/1/342704

Zacarías, D., & Meza, I. V. (2021). Ayuuk-Spanish Neural Machine Translator. *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, 168–172. https://doi.org/10.18653/v1/2021.americasnlp-1.19

Zapotec Talking Dictionaries. (n.d.). Retrieved September 22, 2022, from https://talkingdictionary.swarthmore.edu/zapotecs/

Zhang, S., Frey, B., & Bansal, M. (2020). *ChrEn: Cherokee-English Machine Translation for Endangered Language Revitalization*. https://doi.org/10.48550/arXiv.2010.04791

Appendix A

Computational Modelling of Plains Cree Syntax

The following link directs to a PDF of Schmirler (2017), the unpublished generals paper describing the first iteration of the Plains Cree parser expanded upon herein. This manuscript offers more details than the published version, Schmirler et al. (2018).

Schmirler, K. (2017). *Computational modelling of Plains Cree syntax: A Constraint Grammar approach to verbs and arguments in a Plains Cree corpus*. [Unpublished PhD qualifying paper]. University of Alberta.

In the case that the above link no longer functions, see https://www.kschmirler.com/publications.

Appendix B

Morphosyntactic Feature Frequency in a Plains Cree Corpus

B.1 Morphological tags

The tags presented in this section are those drawn from the gold standard, after the disambiguator module of the parser has been applied.

B.1.1 Tokens/types

All tokens/type include non-Cree words, metadata, and punctuation. Cree tokens/types have filtered these out. Following these first three tables, only Cree words are considered. Of the non-Cree tokens in the full corpus, 62,038 are clause boundary punctuation (8 types), which are referenced in the parser to disambiguate forms and assign functions. These represent 31,643 clause boundary tokens in the A-W subcorpus (7 types) and 30,395 tokens (6 types) in the BT subcorpus.

Table B.1: Overall tokens and types in the full corpus

Cree tokens	Cree types	All tokens	All types	TTR-Cree	TTR
152,405	31,616	241,922	34,115	0.21	0.14

Table B.2: Overall tokens and types in the A-W subcorpus

Cree tokens	Cree types	All tokens	All types	TTR-Cree	TTR
79,930	18,212	138,960	20,699	0.23	0.15

Table B.3: Overall tokens and types in the BT subcorpus

Cree tokens	Cree types	All tokens	All types	TTR-Cree	TTR
72,475	15,267	102,962	15,287	0.21	0.15

B.1.2 Verbs

The total for verbs includes all verbs, while the quotatives pull out a subset of the total verbs.

Table B.4: Verbs and subclasses in the full corpus

	Cree tokens	Cree types	% Tokens	% Types	TTR		
Verbs	50,632	25,739	33.22%	81.41%	0.51		
					% V tokens	% V types	TTR
II	4,043	1,918	2.65%	6.07%	7.99%	7.45%	0.47
AI	23,117	11,290	15.17%	35.71%	45.66%	43.86%	0.49
TI	8,318	4,192	5.46%	13.26%	16.43%	16.29%	0.50
TA	15,436	8,476	10.13%	26.81%	30.49%	32.93%	0.55

Table B.5: Verbs and subclasses in the A-W subcorpus

	Cree tokens	Cree types	% Tokens	% Types	TTR		
Verbs	22,366	14,393	27.98%	79.03%	0.51		
					% V tokens	% V types	TTR
II	2,051	1,135	2.57%	6.23%	9.17%	7.89%	0.55
AI	9,938	6,192	12.43%	34.00%	44.43%	43.02%	0.62
TI	4,255	2,712	5.32%	14.89%	19.02%	18.84%	0.64
TA	6,192	4,390	7.75%	24.10%	27.68%	30.50%	0.71

Table B.6: Verbs and subclasses in the BT subcorpus

	Cree tokens	Cree types	% Tokens	% Types	TTR		
Verbs	28,266	12,400	39.00%	81.22%	0.51		
					% V tokens	% V types	TTR
II	1,992	877	2.75%	5.74%	7.05%	7.07%	0.44
AI	13,179	5,549	18.18%	36.35%	46.62%	44.75%	0.42
TI	4,063	1,709	5.61%	11.19%	14.37%	13.78%	0.42
TA	9,244	4,374	12.75%	28.65%	32.70%	35.27%	0.47

Table B.7: Quotative verbs and subclasses in the full corpus

	Cree tokens	Cree types	% V tokens	% V types	TTR		
Quot	5,574	361	11.01%	1.40%	0.06		
					% Quot tokens	% Quot types	TTR
AI	2,790	136	5.51%	0.53%	50.05%	37.67%	0.05
TI	33	13	0.07%	0.05%	0.59%	3.60%	0.39
TA	2,751	212	5.43%	0.82%	49.35%	58.73%	0.08

Table B.8: Quotative verbs and subclasses in the A-W subcorpus

	Cree tokens	Cree types	% V tokens	% V types	TTR		
Quot	2,135	227	9.55%	1.58%	0.11		
					% Quot tokens	% Quot types	TTR
AI	1,352	92	6.04%	0.64%	63.33%	40.53%	0.07
TI	25	9	0.11%	0.06%	1.17%	3.96%	0.36
TA	758	126	3.39%	0.88%	35.50%	55.51%	0.17

Table B.9: Quotative verbs and subclasses in the BT subcorpus

	Cree tokens	Cree types	% V tokens	% V types	TTR		
Quot	3,439	196	12.17%	1.58%	0.06		
					% Quot tokens	% Quot types	TTR
AI	1,438	73	5.09%	0.59%	41.81%	37.24%	0.05
TI	8	4	0.03%	0.03%	0.23%	2.04%	0.50
TA	1,993	119	7.05%	0.96%	57.95%	60.71%	0.06

Table B.10: VTA features in the full corpus

Feature	Cree tokens	Cree types	% Tokens	% Types	% V tokens	% V types	TTR
Mixed	5,147	3,456	3.38%	10.93%	10.17%	13.43%	0.67
Local	1,185	775	0.78%	2.45%	2.34%	3.01%	0.65
Nonlocal	7,384	2,967	4.84%	9.38%	14.58%	11.53%	0.40
Direct	11,129	5,818	7.30%	18.40%	21.98%	22.60%	0.52
Inverse	4,378	2,699	2.87%	8.54%	8.65%	10.49%	0.62

Table B.11: Direct and inverse for mixed, local, and nonlocal verbs in the full corpus

Feature	Cree tokens	Cree types		
Mixed	5,147	3,456		
			% Mixed tokens	% Mixed types
Direct	3,381	2,248	65.69%	65.05%
Inverse	1,857	1,273	36.08%	36.83%
Local	1,185	775		
			% Local tokens	% Local types
Direct	599	387	50.55%	49.94%
Inverse	660	438	55.70%	56.52%
Nonlocal	7,384	2,967		
			% Nonlocal tokens	% Nonlocal types
Direct	5,806	2,244	78.63%	75.63%
Inverse	1,673	792	22.66%	26.69%

Table B.12: VTA features in the A-W subcorpus

Feature	Cree tokens	Cree types	% Tokens	% Types	% V tokens	% V types	TTR
Mixed	3,249	2,197	4.06%	12.06%	14.53%	15.26%	0.68
Local	305	215	0.38%	1.18%	1.36%	1.49%	0.70
Nonlocal	1,702	1,216	2.13%	6.68%	7.61%	8.45%	0.71
Direct	4,316	3,012	5.40%	16.54%	19.30%	20.93%	0.70
Inverse	1,913	1,403	2.39%	7.70%	8.55%	9.75%	0.73

Table B.13: Direct and inverse for mixed, local, and nonlocal verbs in the A-W subcorpus

Feature	Cree tokens	Cree types		
Mixed	3,249	2,197		
			% Mixed tokens	% Mixed types
Direct	2,074	1,383	63.84%	62.95%
Inverse	1,176	815	36.20%	37.10%
Local	305	215		
			% Local tokens	% Local types
Direct	129	92	42.30%	42.79%
Inverse	177	124	58.03%	57.67%
Nonlocal	1,702	1,216		
			% Nonlocal tokens	% Nonlocal types
Direct	1,401	978	82.31%	80.43%
Inverse	304	241	17.86%	19.82%

Table B.14: VTA features in the BT subcorpus

Feature	Cree tokens	Cree types	% Tokens	% Types	% V tokens	% V types	TTR
Mixed	1,898	1,343	2.62%	8.80%	6.71%	10.83%	0.71
Local	880	577	1.21%	3.78%	3.11%	4.65%	0.66
Nonlocal	5,682	1,904	7.84%	12.47%	20.10%	15.35%	0.34
Direct	6,813	3,025	9.40%	19.81%	24.10%	24.40%	0.44
Inverse	2,465	1,365	3.40%	8.94%	8.72%	11.01%	0.55

Table B.15: Direct and inverse for mixed, local, and nonlocal verbs in the BT subcorpus

Feature	Cree tokens	Cree types		
Mixed	1,898	1,343		
			% Mixed tokens	% Mixed types
Direct	1,307	919	68.86%	68.43%
Inverse	681	488	35.88%	36.34%
Local	880	577		
			% Local tokens	% Local types
Direct	470	301	53.41%	52.17%
Inverse	483	325	54.89%	56.33%
Nonlocal	5,682	1,904		
			% Nonlocal tokens	% Nonlocal types
Direct	4,405	1,393	77.53%	73.16%
Inverse	1,369	577	24.09%	30.30%

Table B.16: Verbal orders in the full corpus

Order	Cree tokens	Cree types	% Tokens	% Types	% V tokens	% V types	TTR
IND	21,191	8,387	13.90%	26.53%	41.85%	32.58%	0.40
CNJ	26,863	15,717	17.63%	49.71%	53.06%	61.06%	0.59
COND	1,016	824	0.67%	2.61%	2.01%	3.20%	0.81
IMP	1,664	875	1.09%	2.77%	3.29%	3.40%	0.53
Order	Cree tokens	Cree types	% Tokens	% Types			TTR
QUOT	5,574	361	11.01%	1.40%			0.06
					% Quot tokens	% Quot types	TTR
IND	4,423	120	2.90%	0.38%	79.35%	33.24%	0.03
CNJ	1,080	210	0.71%	0.66%	19.38%	58.17%	0.19
COND	36	22	0.02%	0.07%	0.65%	6.09%	0.61
IMP	39	11	0.03%	0.03%	0.70%	3.05%	0.28

Table B.17: Verbal orders in the A-W subcorpus

Order	Cree tokens	Cree types	% Tokens	% Types	% V tokens	% V types	TTR
IND	6,686	3,654	8.36%	20.06%	29.89%	25.39%	0.55
CNJ	15,056	10,230	18.84%	56.17%	67.32%	71.08%	0.68
COND	299	277	0.37%	1.52%	1.34%	1.92%	0.93
IMP	316	222	0.40%	1.22%	1.41%	1.54%	0.70
Order	Cree tokens	Cree types	% Tokens	% Types			TTR
QUOT	2,135	227	9.55%	1.58%			0.11
					% Quot tokens	% Quot types	TTR
IND	1,479	70	1.85%	0.38%	6.61%	0.49%	0.05
CNJ	640	145	0.80%	0.80%	2.86%	1.01%	0.23
COND	7	7	0.01%	0.04%	0.03%	0.05%	1.00
IMP	10	6	0.01%	0.03%	0.04%	0.04%	0.60

Table B.18: Verbal orders in the BT subcorpus

Order	Cree tokens	Cree types	% Tokens	% Types	% V tokens	% V types	TTR
IND	14,505	5,064	20.01%	33.17%	51.32%	40.84%	0.35
CNJ	11,807	6,145	16.29%	40.25%	41.77%	49.56%	0.52
COND	717	557	0.99%	3.65%	2.54%	4.49%	0.78
IMP	1,348	709	1.86%	4.64%	4.77%	5.72%	0.53
Order	Cree tokens	Cree types	% Tokens	% Types			TTR
QUOT	3,439	196	12.17%	1.58%			0.06
					% Quot tokens	% Quot types	TTR
IND	2,944	73	4.06%	0.48%	10.42%	0.59%	0.02
CNJ	440	100	0.61%	0.66%	1.56%	0.81%	0.23
COND	29	16	0.04%	0.10%	0.10%	0.13%	0.55
IMP	29	8	0.04%	0.05%	0.10%	0.06%	0.28

Table B.19: Verbal person features in the full corpus

Person	Cree tokens	Cree types	% Tokens	% Types	% V tokens	% V types	TTR
1SG	8,303	4,988	5.45%	15.78%	16.40%	19.38%	0.60
2SG	3,928	2,547	2.58%	8.06%	7.76%	9.90%	0.65
3SG	24,075	9,799	15.80%	30.99%	47.55%	38.07%	0.65
3′	10,517	4,759	6.90%	15.05%	20.77%	18.49%	0.41
1PL	2,400	1,777	1.57%	5.62%	4.74%	6.90%	0.74
21PL	1,025	805	0.67%	2.55%	2.02%	3.13%	0.79
2PL	1,070	811	0.70%	2.57%	2.11%	3.15%	0.76
3PL	8,481	5,139	5.56%	16.25%	16.75%	19.97%	0.61
X	3,031	2,192	1.99%	6.93%	5.99%	8.52%	0.72
0SG	2,292	1,012	1.50%	3.20%	4.53%	3.93%	0.44
0PL	411	301	0.27%	0.95%	0.81%	1.17%	0.73
0′SG	1,185	505	0.78%	1.60%	0.78%	1.60%	0.43
0′PL	230	139	0.15%	0.44%	0.45%	0.54%	0.60

Table B.20: Verbal person features in the A-W subcorpus

Person	Cree tokens	Cree types	% Tokens	% Types	% V tokens	% V types	TTR
1SG	5,455	3,270	6.82%	17.96%	24.39%	22.72%	0.60
2SG	1,163	895	1.46%	4.91%	5.20%	6.22%	0.77
3SG	8,650	4,999	10.82%	27.45%	38.67%	34.73%	0.58
3′	2,180	1,635	2.73%	8.98%	9.75%	11.36%	0.75
1PL	2,127	1,588	2.66%	8.72%	9.51%	11.03%	0.75
21PL	573	466	0.72%	2.56%	2.56%	3.24%	0.81
2Pl	352	285	0.44%	1.56%	1.57%	1.98%	0.81
3PL	4,565	3,362	5.71%	18.46%	20.41%	23.36%	0.74
X	1,876	1,468	2.35%	8.06%	8.39%	10.20%	0.78
0SG	1,519	734	1.90%	4.03%	6.79%	5.10%	0.48
0PL	331	250	0.41%	1.37%	1.48%	1.74%	0.76
0′SG	164	118	0.21%	0.65%	0.21%	0.65%	0.72
0′PL	46	40	0.06%	0.22%	0.21%	0.28%	0.87

Table B.21: Verbal person features in the BT subcorpus

Person	Cree tokens	Cree types	% Tokens	% Types	% V tokens	% V types	TTR
1SG	2,848	1,882	3.93%	12.33%	10.08%	15.18%	0.66
2SG	2,765	1,730	3.82%	11.33%	9.78%	13.95%	0.63
3SG	16,198	5,623	22.35%	36.83%	57.31%	45.35%	0.35
3′	9,293	3,693	12.82%	24.19%	32.88%	29.78%	0.40
1PL	273	224	0.38%	1.47%	0.97%	1.81%	0.82
21PL	452	350	0.62%	2.29%	1.60%	2.82%	0.77
2PL	718	546	0.99%	3.58%	2.54%	4.40%	0.76
3PL	3,916	1,997	5.40%	13.08%	13.85%	16.10%	0.51
X	1,155	775	1.59%	5.08%	4.09%	6.25%	0.67
0SG	773	343	1.07%	2.25%	2.73%	2.77%	0.44
0PL	80	58	0.11%	0.38%	0.28%	0.47%	0.73
0′SG	1,021	406	1.41%	2.66%	1.41%	2.66%	0.40
0′PL	184	104	0.25%	0.68%	0.65%	0.84%	0.57

Table B.22: Quotative verbal person features in the full corpus

Person	Cree tokens	Cree types	% Tokens	% Types	% V tokens	% V types	TTR
1SG	745	97	0.49%	0.31%	1.47%	0.38%	0.13
2SG	175	60	0.11%	0.19%	0.35%	0.23%	0.34
3SG	4,730	160	3.10%	0.51%	9.34%	0.62%	0.03
3′	2,003	67	1.31%	0.21%	3.96%	0.26%	0.03
1PL	56	33	0.04%	0.10%	0.11%	0.13%	0.59
21PL	15	10	0.01%	0.03%	0.03%	0.04%	0.67
2PL	32	25	0.02%	0.08%	0.06%	0.10%	0.78
3PL	332	63	4.00%	0.20%	0.66%	0.24%	0.19
X	254	68	0.17%	0.22%	0.50%	0.26%	0.27

Table B.23: Quotative verbal person features in the A-W subcorpus

Person	Cree tokens	Cree types	% Tokens	% Types	% V tokens	% V types	TTR
1SG	647	70	0.81%	0.38%	2.89%	0.49%	0.11
2SG	30	21	0.04%	0.12%	0.13%	0.15%	0.70
3SG	1,720	102	2.15%	0.56%	7.69%	0.71%	0.06
3′	140	35	0.18%	0.19%	0.63%	0.24%	0.25
1PL	48	30	0.06%	0.16%	0.21%	0.21%	0.63
21PL	6	4	0.01%	0.02%	0.03%	0.03%	0.67
2PL	9	8	0.01%	0.04%	0.04%	0.06%	0.89
3PL	182	47	3.34%	0.26%	0.81%	0.33%	0.26
X	128	45	0.16%	0.25%	0.57%	0.31%	0.35

Table B.24: Quotative verbal person features in the BT subcorpus

Person	Cree tokens	Cree types	% Tokens	% Types	% V tokens	% V types	TTR
1SG	98	44	0.14%	0.29%	0.35%	0.35%	0.45
2SG	145	45	0.20%	0.29%	0.51%	0.36%	0.31
3SG	3,010	92	4.15%	0.60%	10.65%	0.74%	0.03
3′	1,863	45	2.57%	0.29%	6.59%	0.36%	0.02
1PL	8	7	0.01%	0.05%	0.03%	0.06%	0.88
21PL	9	6	0.01%	0.04%	0.03%	0.05%	0.67
2PL	23	19	0.03%	0.12%	0.08%	0.15%	0.83
3PL	150	23	5.27%	0.15%	0.53%	0.19%	0.15
X	126	35	0.17%	0.23%	0.45%	0.28%	0.28

B.1.3 Nouns

Table B.25: Noun features in the full corpus

	Cree tokens	Cree types	% Tokens	% Types	TTR		
Nouns	25,683	4,033	16.85%	12.76%	0.16		
					% N tokens	% N types	TTR
NI[68]	9,895	2,102	6.49%	6.65%	38.53%	52.12%	0.21
NA	17,759	1,994	11.65%	6.31%	69.15%	49.44%	0.11
NDI	1,192	318	0.78%	1.01%	4.64%	7.88%	0.27
NDA	5,660	558	3.71%	1.76%	22.04%	13.84%	0.10
SG	14,350	1,822	9.42%	5.76%	55.87%	45.18%	0.13
PL	6,094	1,094	4.00%	3.46%	23.73%	27.13%	0.18
OBV	6,597	714	4.33%	2.26%	25.69%	17.70%	0.11
LOC	1,540	443	1.01%	1.40%	6.00%	10.98%	0.29
PX	7,869	1,428	5.16%	4.52%	30.64%	35.41%	0.18

[68] The dependent nominals (NDI and NDA) here are a subset of the overall nominals (NI and NA).

Table B.26: Noun features in the A-W subcorpus

	Cree tokens	Cree types	% Tokens	% Types	TTR		
Nouns	11,750	2,438	14.70%	13.39%	0.21		
					% N tokens	% N types	TTR
NI	5,958	1,318	7.45%	7.24%	50.71%	54.06%	0.22
NA	7,478	1,148	9.36%	6.30%	63.64%	47.09%	0.15
NDI	293	151	0.37%	0.83%	2.49%	6.19%	0.52
NDA	2,080	324	2.60%	1.78%	17.70%	13.29%	0.16
SG	7,299	1,123	9.13%	6.17%	62.12%	46.06%	0.15
PL	3,854	693	4.82%	3.81%	32.80%	28.42%	0.18
OBV	2,636	390	3.30%	2.14%	22.43%	16.00%	0.15
LOC	700	261	0.88%	1.43%	5.96%	10.71%	0.37
PX	2,717	709	3.40%	3.89%	23.12%	29.08%	0.26

Table B.27: Noun features in the BT subcorpus

	Cree tokens	Cree types	% Tokens	% Types	TTR		
Nouns	13,933	2,031	19.22%	13.30%	0.15		
					% N tokens	% N types	TTR
NI	3,937	970	5.43%	6.35%	28.26%	47.76%	0.25
NA	10,281	1,100	14.19%	7.21%	73.79%	54.16%	0.11
NDI	899	211	1.24%	1.38%	6.45%	10.39%	0.23
NDA	3,580	329	4.94%	2.15%	25.69%	16.20%	0.09
SG	7,051	893	9.73%	5.85%	50.61%	43.97%	0.13
PL	2,240	493	3.09%	3.23%	16.08%	24.27%	0.22
OBV	3,961	428	5.47%	2.80%	28.43%	21.07%	0.11
LOC	840	231	1.16%	1.51%	6.03%	11.37%	0.28
PX	5,152	868	7.11%	5.69%	36.98%	42.74%	0.17

B.1.4 Pronouns

Table B.28: Pronoun features in the full corpus

	Cree tokens	Cree types	% Tokens	% Types	TTR		
PRON	13,926	172	9.14%	0.54%	0.012		
					% PRON tokens	% PRON types	TTR
DEM	10,507	62	6.89%	0.20%	75.45%	36.05%	0.0059
I	3,294	32	2.16%	0.10%	23.65%	18.60%	0.0097
A	7,459	36	4.89%	0.11%	53.56%	20.93%	0.0048
SG	7,378	28	4.84%	0.09%	52.98%	16.28%	0.0038
PL	2,005	25	1.32%	0.08%	14.40%	14.53%	0.012
OBV	1,351	14	0.89%	0.04%	9.70%	8.14%	0.010
Proximal	7,354	24	4.83%	0.08%	52.81%	13.95%	0.0033
Medial	3,048	28	2.00%	0.09%	21.89%	16.28%	0.0092
Distal	103	9	0.07%	0.03%	0.74%	5.23%	0.087
					% PRON tokens	% PRON types	TTR
PERS	1,940	38	1.27%	0.12%	13.93%	22.09%	0.020
1	849	13	0.56%	0.04%	6.10%	7.56%	0.015
2	259	8	0.17%	0.03%	1.86%	4.65%	0.031
3	717	11	0.47%	0.03%	5.15%	6.40%	0.015

Table B.29: Pronoun features in the A-W subcorpus

	Cree tokens	Cree types	% Tokens	% Types	TTR		
PRON	8,423	136	10.54%	0.75%	0.016		
					% PRON tokens	% PRON types	TTR
DEM	6,053	54	7.57%	0.30%	71.86%	0.64%	0.0089
I	2,596	28	3.25%	0.15%	30.82%	0.33%	0.0108
A	3,631	29	4.54%	0.16%	43.11%	0.34%	0.0080
SG	4,325	26	5.41%	0.14%	51.35%	0.31%	0.0060
PL	1,384	22	1.73%	0.12%	16.43%	0.26%	0.0159
OBV	499	8	0.62%	0.04%	5.92%	0.09%	0.0160
Proximal	3,248	20	4.06%	0.11%	38.56%	0.24%	0.0062
Medial	2,773	26	3.47%	0.14%	32.92%	0.31%	0.0094
Distal	30	7	0.04%	0.04%	0.36%	0.08%	0.2333
					% PRON tokens	% PRON types	TTR
PERS	1,373	28	1.72%	0.15%	16.30%	20.59%	0.020
1	643	13	0.80%	0.07%	7.63%	9.56%	0.020
2	99	8	0.12%	0.04%	1.18%	5.88%	0.081
3	516	11	0.65%	0.06%	6.13%	8.09%	0.021

Table B.30: Pronoun features in the BT subcorpus

	Cree tokens	Cree types	% Tokens	% Types	TTR		
PRON	5,503	61	7.59%	0.40%	0.011		
					% PRON tokens	% PRON types	TTR
DEM	4,454	19	6.15%	0.12%	80.94%	31.15%	0.0043
I	698	9	0.96%	0.06%	12.68%	14.75%	0.013
A	3,828	13	5.28%	0.09%	69.56%	21.31%	0.0034
SG	3,053	8	4.21%	0.05%	55.48%	13.11%	0.0026
PL	621	8	0.86%	0.05%	11.28%	13.11%	0.013
OBV	852	6	1.18%	0.04%	15.48%	9.84%	0.0070
Proximal	4,106	8	5.67%	0.05%	74.61%	13.11%	0.0019
Medial	275	6	0.38%	0.04%	5.00%	9.84%	0.022
Distal	73	5	0.10%	0.03%	1.33%	8.20%	0.069
					% PRON tokens	% PRON types	TTR
PERS	567	12	0.78%	0.08%	10.30%	19.67%	0.021
1	206	4	0.28%	0.03%	3.74%	6.56%	0.019
2	160	4	0.22%	0.03%	2.91%	6.56%	0.025
3	201	4	0.28%	0.03%	3.65%	6.56%	0.020

B.1.5 Particles

Here, the particle count also includes particle phrases (IPH).

Table B.31: Particle classes in the full corpus

	Cree tokens	Cree types	% Tokens	% Types	TTR		
IPC	60,410	1,516	39.64%	4.80%	0.025		
					% IPC tokens	% IPC types	TTR
NEG	3,343	26	2.19%	0.08%	5.53%	1.72%	0.0078
IPL	5,777	93	3.79%	0.29%	9.56%	6.13%	0.016
IPT	4,838	131	3.17%	0.41%	8.01%	8.64%	0.027
QUANT	5,986	96	3.93%	0.30%	9.91%	6.33%	0.016
Other	40,466	1,170	26.55%	3.70%	66.99%	77.18%	0.029

Table B.32: Particle classes in the A-W subcorpus

	Cree tokens	Cree types	% Tokens	% Types	TTR		
IPC	38,761	1,222	48.49%	6.71%	0.032		
					% IPC tokens	% IPC types	TTR
NEG	1,846	22	4.76%	1.80%	4.76%	1.80%	0.012
IPL	3,273	76	8.44%	6.22%	8.44%	6.22%	0.023
IPT	2,480	112	6.40%	9.17%	6.40%	9.17%	0.045
QUANT	3,337	76	8.61%	6.22%	8.61%	6.22%	0.023
Other	27,825	936	34.81%	5.14%	71.79%	76.60%	0.034

Table B.33: Particle classes in the BT subcorpus

	Cree tokens	Cree types	% Tokens	% Types	TTR		
IPC	23,641	715	32.62%	4.68%	0.030		
					% IPC tokens	% IPC types	TTR
NEG	1,497	12	6.33%	1.68%	6.33%	1.68%	0.008
IPL	2,504	43	10.59%	6.01%	10.59%	6.01%	0.017
IPT	2,358	66	9.97%	9.23%	9.97%	9.23%	0.028
QUANT	2,649	51	11.21%	7.13%	11.21%	7.13%	0.019
Other	14,633	543	20.19%	3.56%	61.90%	75.94%	0.037

B.2 Syntactic tags

The full corpus contains 98,421 syntactic function tags as applied by the parser. Of these, 45,855 occur in the A-W subcorpus and 52,566 in the BT subcorpus.

B.2.1 Predicate tags

The following tables give the total number of predicate tags assigned in the corpus, followed by those for each verbal subclass. Note that the percentages will not add up to the expected 100%, as many common verbs are ambiguous and cannot yet be automatically disambiguated and thus receive multiple tags. Unlike above, where quotative morphological tags are a subset of verb tags, here `@PRED` and `@Quot` tags are counted separately.

Table B.34: Predicate tags in the full corpus

	Tags	% All tags (98,421)
`@PRED`	39,916	40.56%
		% `@PRED` tags
`@PRED-II`	3,892	9.75%
`@PRED-AI`	19,245	48.21%
`@PRED-TI`	8,020	20.09%
`@PRED-TA`	12,197	30.56%
		% All tags (98,421)
`@Quot`	5,570	5.66%
		% `@Quot` tags
`@Quot-AI`	2,790	50.09%
`@Quot-TI`	33	0.59%
`@Quot-TA`	2,749	49.35%

Table B.35: Predicate tags in the A-W subcorpus

	Tags	% All tags (45,855)
@PRED	17,564	20.63%
		% @PRED tags
@PRED-II	1,968	11.20%
@PRED-AI	8,057	45.87%
@PRED-TI	4,073	23.19%
@PRED-TA	5,208	29.65%
		% All tags (45,855)
@Quot	2,132	4.65%
		% @Quot tags
@Quot-AI	1,352	63.41%
@Quot-TI	25	1.17%
@Quot-TA	756	35.46%

Table B.36: Predicate tags in the BT subcorpus

	Tags	% All tags (52,566)
@PRED	22,352	42.52%
		% @PRED tags
@PRED-II	1,924	8.61%
@PRED-AI	11,188	50.05%
@PRED-TI	3,947	17.66%
@PRED-TA	6,989	31.27%
		% All tags (52,566)
@Quot	3,438	6.54%
		% @Quot tags
@Quot-AI	1,438	41.83%
@Quot-TI	8	0.23%
@Quot-TA	1,993	57.97%

B.2.2 Particle tags

Tagged particles here include both particles and particle phrases, labelled just as IPC for brevity. For negative and quantifier tags, the total number is given as well as each subtype.

Table B.37: Particle tags in the full corpus

	Tags	% All tags	% All IPC
IPC w/tag	16,973	17.25%	28.10%
		% IPC w/tags	
`@Neg`	3,645	21.48%	
			% `@Neg`
`@Neg-V`	3,342	19.69%	91.69%
`@Neg-N`	74	0.44%	2.03%
`@Neg-IPC`	229	1.35%	6.28%
`@IPL-V`	4,938	29.09%	
`@IPT-V`	4,363	25.71%	
`@Quant`	4,027	23.73%	
			% `@Quant`
`@Quant-V`	3,427	20.19%	85.10%
`@Quant-N`	600	3.54%	14.90%

Table B.38: Particle tags in the A-W subcorpus

	Tags	% All tags	% All IPC
IPC w/tag	8,836	19.27%	22.80%
		% IPC w/tags	
@Neg	2,115	23.94%	
			% @Neg
@Neg-V	1,870	21.16%	88.42%
@Neg-N	62	0.70%	2.93%
@Neg-IPC	183	2.07%	8.65%
@IPL-V	2,583	29.23%	
@IPT-V	1,926	21.80%	
@Quant	2,212	25.03%	
			% @Quant
@Quant-V	1,852	20.96%	83.73%
@Quant-N	360	4.07%	16.27%

Table B.39: Particle tags in the BT subcorpus

	Tags	% All tags	% All IPC
IPC w/tag	8,137	15.48%	34.42%
		% IPC w/tags	
@Neg	1,530	18.80%	
			% @Neg
@Neg-V	1,472	18.09%	96.21%
@Neg-N	12	0.15%	0.78%
@Neg-IPC	46	0.57%	3.01%
@IPL-V	2,355	28.94%	
@IPT-V	2,437	29.95%	
@Quant	1,815	22.31%	
			% @Quant
@Quant-V	1,575	19.36%	86.78%
@Quant-N	240	2.95%	13.22%

B.2.3 Noun phrases

The following tables give the counts noun phrase tags, @N> or @<N, which can be assigned to demonstratives before or after nouns (a numeral may intervene), as well as to a number of other particles, pronouns, and numerals before a demonstrative or noun. For the demonstratives, which can only be marked to specify a noun, the rightmost columns indicate the percentage of demonstratives that receive an @N tag, the percentage of nouns these numbers indicate occur in noun phrases.

Table B.40: Noun phrase tags in the full corpus

	Tags	% All tags			
All @N	7,136	7.25%			
			% All @N		
Non-DEM w/@N	1,749	1.78%	24.51%		
				% All DEM	% All N
DEM w/@N	5,418	5.50%	75.92%	51.57%	21.10%
		% DEM w/@N			
I	1,047	19.32%	14.67%	9.96%	4.08%
A	4,371	80.68%	61.25%	41.60%	17.02%
SG	3,822	70.54%	53.56%	36.38%	14.88%
PL	845	15.60%	11.84%	8.04%	3.29%
OBV	751	13.86%	10.52%	7.15%	2.92%
Proximal	4,380	80.84%	61.38%	41.69%	17.05%
Medial	1,009	18.62%	14.14%	9.60%	3.93%
Distal	29	0.54%	0.41%	0.28%	0.11%

Table B.41: Noun phrase tags in the A-W subcorpus

	Tags	% All tags			
All @N	3,557	7.76%			
			% All @N		
Non-DEM w/@N	1,327	2.89%	37.31%		
				% All DEM	% All N
DEM w/@N	2,247	4.90%	63.17%	37.12%	19.12%
		% DEM w/@N			
I	711	31.64%	19.99%	11.75%	6.05%
A	1,536	68.36%	43.18%	25.38%	13.07%
SG	1,620	72.10%	45.54%	26.76%	13.79%
PL	460	20.47%	12.93%	7.60%	3.91%
OBV	167	7.43%	4.69%	2.76%	1.42%
Proximal	1,343	59.77%	37.76%	22.19%	11.43%
Medial	902	40.14%	25.36%	14.90%	7.68%
Distal	2	0.09%	0.06%	0.03%	0.02%

Table B.42: Noun phrase tags in the BT subcorpus

	Tags	% All tags			
All @N	3,551	6.76%			
			% All @N		
Non-DEM w/@N	394	0.75%	11.10%		
				% All DEM	% All N
DEM w/@N	3,171	6.03%	89.30%	71.19%	22.76%
		% DEM w/@N			
I	336	10.60%	9.46%	7.54%	2.41%
A	2,835	89.40%	79.84%	63.65%	20.35%
SG	2,202	69.44%	62.01%	49.44%	15.80%
PL	385	12.14%	10.84%	8.64%	2.76%
OBV	584	18.42%	16.45%	13.11%	4.19%
Proximal	3,037	95.77%	85.53%	68.19%	21.80%
Medial	107	3.37%	3.01%	2.40%	0.77%
Distal	27	0.85%	0.76%	0.61%	0.19%

B.2.4 Actors & Goals

Table B.43: Nouns as actors and goals in the full corpus

	Tags	% All tags	% All Nom				
@A/G	20,110	20.43%	58.82%				
		% All tags	% @A/G		Tags	% All tags	% @A/G
@A	11,123	11.30%	55.31%	@G	8,987	9.13%	44.69%
		% @A	% All N			% @G	% All N
N w/@A	7,716	69.37%	30.04%	N w/@G	6,904	76.82%	26.88%
			% N w/@A				% N w/@G
NI	735	6.61%	9.53%	NI	2,881	32.06%	41.73%
NA	6,981	62.76%	90.47%	NA	4,023	44.76%	58.27%
NDI	91	0.82%	1.18%	NDI	402	4.47%	5.82%
NDA	1,808	16.25%	23.43%	NDA	1,585	17.64%	22.96%
SG	4,722	42.45%	61.20%	SG	3,023	33.64%	43.79%
PL	1,412	12.69%	18.30%	PL	1,354	15.07%	19.61%
OBV	1,582	14.22%	20.50%	OBV	2,511	27.94%	36.37%
Total PX	2,028	18.23%	26.28%	Total PX	2,404	26.75%	34.82%
PXI	171	1.54%	2.22%	PXI	755	8.40%	10.94%
PXA	1,857	16.70%	24.07%	PXA	1,649	18.35%	23.88%

Table B.44: Pronouns as actors and goals in the full corpus

	Tags	% @A	% All Pron		Tags	% @G	% All Pron
Pron w/@A	3,099	27.86%	22.25%	Pron w/@G	1,921	21.38%	13.79%
			% Pron w/@A				% Pron w/@G
Dem w/@A	1,768	15.89%	57.05%	Dem w/@G	1,649	18.35%	85.84%
			% Dem w/@A				% Dem w/@G
I	339	3.05%	19.17%	I	998	11.10%	60.52%
A	1,429	12.85%	80.83%	A	651	7.24%	39.48%
SG	1,278	11.49%	72.29%	SG	1,165	12.96%	70.65%
PL	359	3.23%	20.31%	PL	288	3.20%	17.47%
OBV	131	1.18%	7.41%	OBV	196	2.18%	11.89%
Proximal	1,143	10.28%	64.65%	Proximal	940	10.46%	57.00%
Medial	593	5.33%	33.54%	Medial	698	7.77%	42.33%
Distal	32	0.29%	1.81%	Distal	11	0.12%	0.67%
			% Pron w/@A				% Pron w/@G
Pers w/@A	1,097	9.86%	35.40%	Pers w/@G	185	2.06%	9.63%
			% Pers w/@A				% Pers w/@G
1	523	4.70%	47.68%	1	70	0.78%	37.84%
2	119	1.07%	10.85%	2	20	0.22%	10.81%
3	455	4.09%	41.48%	3	94	1.05%	50.81%
SG	845	7.60%	77.03%	SG	152	1.69%	82.16%
PL	252	2.27%	22.97%	PL	32	0.36%	17.30%

Table B.45: Particles as actors and goals in the full corpus[69]

		% @A			% @G
IPC/IPN w/@A	308	2.77%	IPC/IPN w/@G	162	1.80%

[69] Particles that can occur as actors and goals include quantifiers, such as *kahkiyaw* 'all' (§3.3.2.1).

Table B.46: Nouns as actors and goals in the A-W subcorpus

	Tags	% All tags	% All Nom				
@A/G	9,271	20.22%	51.72%				
		% All tags	% @A/G		Tags	% All tags	% @A/G
@A	4,860	10.60%	52.42%	@G	4,411	9.62%	47.58%
		% @A	% All N			% @G	% All N
N w/@A	2,694	55.43%	22.93%	N w/@G	2,907	65.90%	24.74%
			% N w/@A				% N w/@G
NI	492	10.12%	18.26%	NI	1,370	31.06%	47.13%
NA	2,202	45.31%	81.74%	NA	1,537	34.84%	52.87%
NDI	22	0.45%	0.82%	NDI	63	1.43%	2.17%
NDA	795	16.36%	29.51%	NDA	496	11.24%	17.06%
SG	1,745	35.91%	64.77%	SG	1,515	34.35%	52.12%
PL	712	14.65%	26.43%	PL	692	15.69%	23.80%
OBV	237	4.88%	8.80%	OBV	688	15.60%	23.67%
Total PX	860	17.70%	31.92%	Total PX	672	15.23%	23.12%
PXI	59	1.21%	2.19%	PXI	164	3.72%	5.64%
PXA	801	16.48%	29.73%	PXA	508	11.52%	17.48%

Table B.47: Pronouns as actors and goals in the A-W subcorpus

	Tags	% @A	% All Pron		Tags	% @G	% All Pron
Pron w/@A	2,040	41.98%	24.22%	Pron w/@G	1,466	33.24%	17.40%
			% Pron w/@A				% Pron w/@G
Dem w/@A	1,147	23.60%	56.23%	Dem w/@G	1,290	29.25%	87.99%
			% Dem w/@A				% Dem w/@G
I	278	5.72%	24.24%	I	844	19.13%	65.43%
A	869	17.88%	75.76%	A	446	10.11%	34.57%
SG	821	16.89%	71.58%	SG	963	21.83%	74.65%
PL	284	5.84%	24.76%	PL	234	5.30%	18.14%
OBV	42	0.86%	3.66%	OBV	93	2.11%	7.21%
Proximal	621	12.78%	54.14%	Proximal	637	14.44%	49.38%
Medial	520	10.70%	45.34%	Medial	647	14.67%	50.16%
Distal	6	0.12%	0.52%	Distal	6	0.14%	0.47%
			% Pron w/@A				% Pron w/@G
Pers w/@A	760	15.64%	37.25%	Pers w/@G	107	2.43%	7.30%
			% Pers w/@A				% Pers w/@G
1	405	8.33%	53.29%	1	45	1.02%	42.06%
2	48	0.99%	6.32%	2	8	0.18%	7.48%
3	307	6.32%	40.39%	3	53	1.20%	49.53%
SG	561	11.54%	73.82%	SG	83	1.88%	77.57%
PL	199	4.09%	26.18%	PL	23	0.52%	21.50%

Table B.48: Particles as actors and goals in the A-W subcorpus

		% @A			% @G
IPC/IPN w/@A	126	2.59%	IPC/IPN w/@G	38	0.86%

Table B.49: Nouns as actors and goals in the BT subcorpus

	Tags	% All tags	% All Nom				
@A/G	10,839	20.62%	66.64%				
		% All tags	% @A/G		Tags	% All tags	% @A/G
@A	6,263	11.91%	57.78%	@G	4,576	8.71%	42.22%
		% @A	% All N			% @G	% All N
N w/@A	5,022	80.19%	36.04%	N w/@G	3,997	87.35%	28.69%
			% N w/@A				% N w/@G
NI	243	3.88%	4.84%	NI	1,511	33.02%	37.80%
NA	4,779	76.31%	95.16%	NA	2,486	54.33%	62.20%
NDI	69	1.10%	1.37%	NDI	339	7.41%	8.48%
NDA	1,013	16.17%	20.17%	NDA	1,089	23.80%	27.25%
SG	2,977	47.53%	59.28%	SG	1,508	32.95%	37.73%
PL	700	11.18%	13.94%	PL	662	14.47%	16.56%
OBV	1,345	21.48%	26.78%	OBV	1,823	39.84%	45.61%
Total PX	1,168	18.65%	23.26%	Total PX	1,732	37.85%	43.33%
PXI	112	1.79%	2.23%	PXI	591	12.92%	14.79%
PXA	1,056	16.86%	21.03%	PXA	1,141	24.93%	28.55%

Table B.50: Pronouns as actors and goals in the BT subcorpus

	Tags	% @A	% All Pron		Tags	% @G	% All Pron
Pron w/@A	1,059	16.91%	19.24%	Pron w/@G	455	9.94%	8.27%
			% Pron w/@A				% Pron w/@G
Dem w/@A	621	9.92%	58.64%	Dem w/@G	359	7.85%	78.90%
			% Dem w/@A				% Dem w/@G
I	61	0.97%	9.82%	I	154	3.37%	42.90%
A	560	8.94%	90.18%	A	205	4.48%	57.10%
SG	457	7.30%	73.59%	SG	202	4.41%	56.27%
PL	75	1.20%	12.08%	PL	54	1.18%	15.04%
OBV	89	1.42%	14.33%	OBV	103	2.25%	28.69%
Proximal	522	8.33%	84.06%	Proximal	303	6.62%	84.40%
Medial	73	1.17%	11.76%	Medial	51	1.11%	14.21%
Distal	26	0.42%	4.19%	Distal	5	0.11%	1.39%
			% Pron w/@A				% Pron w/@G
Pers w/@A	337	5.38%	31.82%	Pers w/@G	78	1.70%	17.14%
			% Pers w/@A				% Pers w/@G
1	118	1.88%	35.01%	1	25	0.55%	32.05%
2	71	1.13%	21.07%	2	12	0.26%	15.38%
3	148	2.36%	43.92%	3	41	0.90%	52.56%
SG	284	4.53%	84.27%	SG	69	1.51%	88.46%
PL	53	0.85%	15.73%	PL	9	0.20%	11.54%

Table B.51: Particles as actors and goals in the BT subcorpus

		% @A			% @G
IPC/IPN w/@A	182	2.91%	IPC/IPN w/@G	124	2.71%

B.2.5 Other tags

Table B.52: Other tags in the full corpus

	Tags	% All N
@OBL	1,824	7.10%
@INS	119	0.46%
		% Possessed N
N w/@Pos	124	1.58%
		% Loc N
IPL w/@P	58	3.77%
		% Loc N + IPL
Loc/IPL w/@Loc-V	697	9.53%

Table B.53: Other tags in the A-W subcorpus

	Tags	% All N
@OBL	1,048	8.92%
@INS	72	0.61%
		% Possessed N
N w/@Pos	33	1.21%
		% Loc N
IPL w/@P	32	4.57%
		% Loc N + IPL
Loc/IPL w/@Loc-V	321	8.08%

Table B.54: Other tags in the BT subcorpus

	Tags	% All N
@OBL	776	5.57%
@INS	47	0.34%
		% Possessed N
N w/@Pos	91	1.77%
		% Loc N
IPL w/@P	26	3.10%
		% Loc N + IPL
Loc/IPL w/@Loc-V	376	11.24%

Appendix C

Argument realisation in a Plains Cree corpus: The subcorpora

C.1 By verb class

Table C.1: Actors and goals by verb class: A-W

	VII	1,985	VAI	9,674	VTI	4,580	VTA	6,092	Total	22,331
		% VII		% VAI		% VTI		% VTA		% V
w/A	638	32.1%	2,014	20.8%	647	14.1%	834	13.7%	4,133	18.5%
		% w/A		% w/A		% w/A		% w/A		% w/A
AV	357	56.0%	979	48.6%	379	58.6%	432	51.8%	2,147	51.9%
VA	281	44.0%	1,035	51.4%	268	41.4%	402	48.2%	1,986	48.1%
					VTI	4,580	VTA	6,092	Total VT	10,672
		% VII		% VAI		% VTI		% VTA		% VT
w/G	-	-	-	-	1,910	41.7%	1,824	29.9%	3,734	35.0%
		% w/G		% w/G		% w/G		% w/G		% w/G
GV	-	-	-	-	1,035	54.2%	683	37.4%	1,718	46.0%
VG	-	-	-	-	875	45.8%	1,141	62.6%	2,016	54.0%
		% w/A		% w/A		% w/A or G		% w/A or G		% w/A or G
all preV	357	56.0%	979	48.6%	1,414	55.3%	1,115	41.9%	3,865	49.1%
all postV	281	44.0%	1,035	51.4%	1,143	44.7%	1,543	58.1%	4,002	50.9%

Table C.2: Actors and goals by verb class: BT

	VII	1,949	VAI	12,873	VTI	4,363	VTA	9,132	Total	28,317
		% VII		% VAI		% VTI		% VTA		% V
w/A	286	14.7%	3,489	27.1%	728	16.7%	1,361	14.9%	5,864	20.7%
		% w/A		% w/A		% w/A		% w/A		% w/A
AV	122	42.7%	1,245	35.7%	268	36.8%	580	42.6%	2,215	37.8%
VA	164	57.3%	2,244	64.3%	460	63.2%	781	57.4%	3,649	62.2%
					VTI	4,363	VTA	9,132	Total VT	13,495
		% VII		% VAI		% VTI		% VTA		% VT
w/G	-	-	-	-	1,567	35.9%	2,585	28.3%	4,152	30.8%
		% w/G		% w/G		% w/G		% w/G		% w/G
GV	-	-	-	-	741	47.3%	777	30.1%	1,518	36.6%
VG	-	-	-	-	826	52.7%	1,808	69.9%	2,634	63.4%
		% w/A		% w/A		% w/A or G		% w/A or G		% w/A or G
all preV	122	42.7%	1,245	35.7%	1,009	44.0%	1,357	34.4%	3,733	37.3%
all postV	164	57.3%	2,244	64.3%	1,286	56.0%	2,589	65.6%	6,283	62.7%

Table C.3: Two overt arguments with transitive verbs: A-W

	VTI	4,580	VTA	6,092	Total	10,672
		% VTI		% VTA		% VT
w/A & G	244	5.3%	181	3.0%	425	4.0%
		% w/A & G		% w/A & G		% w/A & G
VAG	25	10.2%	25	13.8%	50	11.8%
VGA	17	7.0%	12	6.6%	29	6.8%
AVG	85	34.8%	71	39.2%	156	36.7%
GVA	44	18.0%	28	15.5%	72	16.9%
AGV	31	12.7%	19	10.5%	50	11.8%
GAV	42	17.2%	26	14.4%	68	16.0%

Table C.4: Two overt arguments with transitive verbs: BT

	VTI	4,363	VTA	9,132	Total	13,495
		% VTI		% VTA		% VT
w/A & G	191	4.4%	336	3.7%	527	3.9%
		% w/A & G		% w/A & G		% w/A & G
VAG	27	14.1%	77	22.9%	104	19.7%
VGA	32	16.8%	44	13.1%	76	14.4%
AVG	48	25.1%	129	38.4%	177	33.6%
GVA	40	20.9%	27	8.0%	67	12.7%
AGV	7	3.7%	21	6.3%	28	5.3%
GAV	37	19.4%	38	11.3%	75	14.2%

C.2 Topicality

C.2.1 Direct and inverse

Table C.5: Actors and goals by direction: A-W

	DIR	% VTA	INV	% VTA
	4,109	68.5%	1,887	31.5%
		% DIR		% INV
w/A	334	8.1%	488	25.9%
		% w/A		% w/A
AV	206	61.7%	217	44.5%
VA	128	38.3%	271	55.5%
		% DIR		% INV
w/G	1,714	41.7%	102	5.4%
		% w/G		% w/G
GV	622	36.3%	56	54.9%
VG	1,092	63.7%	46	45.1%

Table C.6: Actors and goals by direction: BT

	DIR	% VTA	INV	% VTA
	6,619	73.0%	2,445	27.0%
		% DIR		% INV
w/A	746	11.3%	610	24.9%
		% w/A		% w/A
AV	383	51.3%	194	31.8%
VA	363	48.7%	416	68.2%
		% DIR		% INV
w/G	2,450	37.0%	135	5.5%
		% w/G		% w/G
GV	701	28.6%	76	56.3%
VG	1,749	71.4%	59	43.7%

C.2.2 Persons

Table C.7: Actors and goals for non-SAPs: A-W

	PROX actor	26,526	OBV actor	703	INAN actor	1,985
		% AI, TI, TA 3		% AI, TI, TA 3′		% II
w/A	2,809	28.5%	235	33.4%	638	32.1%
		% w/A		% w/A		% w/A
AV	1,326	47.2%	99	42.1%	357	56.0%
VA	1,483	52.8%	136	57.9%	281	44.0%
	PROX goal	2,893	OBV goal	1,408	INAN goal	4,580
		% TA 3O		% TA 3′O		% TI
w/G	1,090	37.7%	660	46.9%	1,910	41.7%
		% w/G		% w/G		% w/G
GV	424	38.9%	238	36.1%	1,035	54.2%
VG	666	61.1%	422	63.9%	875	45.8%
		% w/A or G		% w/A or G		% w/A or G
Total preV	1,750	44.9%	337	37.7%	1,392	54.6%
Total postV	2,149	55.1%	558	62.3%	1,156	45.4%

Table C.8: Actors and goals for non-SAPs: BT

	PROX actor	16,682	OBV actor	3,856	INAN actor	1,949
		% AI, TI, TA 3		% AI, TI, TA 3′		% II
w/A	4,532	27.2%	1,313	34.1%	286	14.7%
		% w/A		% w/A		% w/A
AV	1,583	34.9%	419	31.9%	122	42.7%
VA	2,949	65.1%	894	68.1%	164	57.3%
	PROX goal	3,025	OBV goal	4,420	INAN goal	4,365
		% TA 3O		% TA 3′O		% TI
w/G	747	24.7%	1,794	40.6%	1,567	43.0%
		% w/G		% w/G		% w/G
GV	302	40.4%	433	24.1%	741	47.3%
VG	445	59.6%	1,361	75.9%	826	52.7%
		% w/A or G		% w/A or G		% w/A or G
Total preV	1,885	35.7%	852	27.4%	863	46.6%
Total postV	3,394	64.3%	2,255	72.6%	990	53.4%

Table C.9: Actors and goals by SAPs: A-W

	1 actor	6,877	21 actor	501	2 actor	1,196
		% AI, TI, TA 1		% AI, TI, TA 21		% AI, TI, TA 2
w/A	955	13.9%	75	15.0%	65	5.4%
		% w/A		% w/A		% w/A
AV	520	54.5%	49	65.3%	47	72.3%
VA	435	45.5%	26	34.7%	18	27.7%
	1 goal	1,310	21 goal	151	2 goal	295
		% TA 1O		% TA 21O		% TA 2O
w/G	46	3.5%	4	2.6%	8	2.7%
		% w/G		% w/G		% w/G
GV	27	58.7%	4	100.0%	7	87.5%
VG	19	41.3%	0	0.0%	1	12.5%
		% w/A or G		% w/A or G		% w/A or G
Total preV	547	54.6%	53	67.1%	54	74.0%
Total postV	454	45.4%	26	32.9%	19	26.0%

Table C.10: Actors and goals by SAPs: BT

	1 actor	2,583	21PL actor	563	2 actor	2,719
		% AI, TI, TA 1		% AI, TI, TA 21		% AI, TI, TA 2
w/A	362	14.0%	46	8.2%	185	6.8%
		% w/A		% w/A		% w/A
AV	206	56.9%	19	41.3%	119	64.3%
VA	156	43.1%	27	58.7%	66	35.7%
	1 goal	791	21PL goal	57	2 goal	644
		% TA 1O		% TA 21O		% TA 2O
w/G	27	3.4%	2	3.5%	16	2.5%
		% w/G		% w/G		% w/G
GV	26	96.3%	2	100.0%	15	93.8%
VG	1	3.7%	0	0.0%	1	6.3%
		% w/A or G		% w/A or G		% w/A or G
Total preV	232	59.6%	21	43.8%	134	66.7%
Total postV	157	40.4%	27	56.3%	67	33.3%

C.2.3 Nominal type

Table C.11: Actors and goals by animate nominal type: A-W

	DEM		PERS		PRON		N		Total
		% PRON		% PRON		% Total		% Total	
@A	785	53.9%	672	46.1%	1,457	41.5%	2,058	58.5%	3,515
		% @A		% @A		% @A		% @A	
AV	394	50.2%	509	75.7%	903	62.0%	803	39.0%	2,609
VA	391	49.8%	163	24.3%	554	38.0%	1,255	61.0%	2,363
	DEM		PERS		PRON		N		Total
		% PRON		% PRON		% Total		% Total	
@G	408	81.9%	90	18.1%	498	25.8%	1,432	74.2%	1,930
		% @G		% @G		% @G		% @G	
GV	163	40.0%	60	66.7%	223	44.8%	510	35.6%	163
VG	245	60.0%	30	33.3%	275	55.2%	922	64.4%	245
		% @A/G		% @A/G		% @A/G		% @A/G	
Total preV	557	46.7%	569	74.7%	1,126	57.6%	1,313	37.6%	
Total postV	636	53.3%	193	25.3%	829	42.4%	2,177	62.4%	

Table C.12: Actors and goals by animate nominal type: BT

	DEM		PERS		PRON		N		Total
		% PRON		% PRON		% Total		% Total	
@A	504	64.0%	284	36.0%	788	14.9%	4,518	85.1%	5,306
		% @A		% @A		% @A		% @A	
AV	285	56.5%	261	91.9%	546	69.3%	1,323	29.3%	2,415
VA	219	43.5%	23	8.1%	242	30.7%	3,195	70.7%	3,679
	DEM		PERS		PRON		N		Total
		% PRON		% PRON		% Total		% Total	
@G	190	72.5%	72	27.5%	262	10.2%	2,306	89.8%	2,568
		% @G		% @G		% @G		% @G	
GV	92	48.4%	67	93.1%	159	60.7%	630	27.3%	948
VG	98	51.6%	5	6.9%	103	39.3%	1,676	72.7%	1,882
		% @A/G		% @A/G		% @A/G		% @A/G	
Total preV	377	54.3%	328	92.1%	705	67.1%	1,953	28.6%	
Total postV	317	45.7%	28	7.9%	345	32.9%	4,871	71.4%	

Table C.13: Actors and goals by inanimate nominal type: A-W

	DEM		N		Total
		% Total		% Total	
@A	252	35.7%	454	64.3%	706
		% @A		% @A	
AV	140	55.6%	238	52.4%	378
VA	112	44.4%	216	47.6%	328
	DEM		N		Total
		% Total		% Total	
@G	767	37.8%	1,263	62.2%	2,030
		% @G		% @G	
GV	404	52.7%	670	53.0%	1,074
VG	363	47.3%	593	47.0%	956
		% @A/G		% @A/G	
Total preV	544	53.4%	908	52.9%	
Total postV	475	46.6%	809	47.1%	

Table C.14: Actors and goals by inanimate nominal type: BT

	DEM		N		Total
		% Total		% Total	
@A	59	20.3%	232	79.7%	291
		% @A		% @A	
AV	48	81.4%	75	32.3%	123
VA	11	18.6%	157	67.7%	168
	DEM		N		Total
		% Total		% Total	
@G	147	9.5%	1,404	90.5%	1,551
		% @G		% @G	
GV	76	51.7%	621	44.2%	697
VG	71	48.3%	783	55.8%	854
		% @A/G		% @A/G	
Total preV	124	60.2%	696	42.5%	
Total postV	82	39.8%	940	57.5%	

C.3 PAS

Table C.15: Proximate actors and goals for VAIs, VTIs, VTAs: A-W

	VAI	5,337		VTI	2,085	VTA	2,560
		% VAI			% VTI		% VTA
S	1,704	31.9%	A	496	23.8%	672	26.3%
						VTA	2,893
							% VTA
	-	-	G	-	-	1,090	37.7%

Table C.16: Obviative actors and goals for VAIs, VTIs, VTAs: A-W

	VAI	226		VTI	57	VTA	392
		% VAI 3′			% VTI 3′		% VTA
S	96	36.1%	A	17	29.8%	127	32.4%
						VTA	1,408
							% VTA
	-	-	G	-	-	660	46.9%

Table C.17: Proximate actors and goals for VAIs, VTIs, VTAs: BT

	VAI	8,647		VTI	3,153	VTA	5,129
		% VAI			% VTI		% VTA
S	2,971	34.4%	A	650	20.6%	1,002	19.5%
						VTA	3,025
							% VTA
	-	-	G	-	-	747	24.7%

Table C.18: Obviative actors and goals for VAIs, VTIs, VTAs: BT

	VAI	1,991		VTI	366	VTA	1,547
		% VAI 3′			% VTI 3′		% VTA
S	792	39.8%	A	83	22.7%	469	30.3%
						VTA	4,420
							% VTA
	-	-	G	-	-	1,794	40.6%

Appendix D

Features for register analysis in Plains Cree

The following list of features has been adapted from Biber & Conrad (2019, pp. 65–8) to indicate which Plains Cree features would a) serve a similar function to those listed in Biber & Conrad's list (2019) and b) to indicate, in boldface, the features that are represented in the final dataset used in Chapter 6.

1. Vocabulary features
    - multifunction words: *aya*, *ôma*, *awa*, *êkwa*
    - type/token ratio
    - average word length
    - number of hapax legomena (perhaps with a focus on stems for Plains Cree, as the morphology makes unique types quite frequent)

2. Content word classes
    - **nouns**
    - **verbs**
    - **particles**

3. Function word classes
    - **pronouns**
        - **Dem**
        - **Pers**
        - other
    - **particles**

4. Derived words
    - nominalizations (*-win*, *-ikan*, *-ihkân*)
    - **diminutives** (Der/Dim)
    - compounds (e.g., prenouns, preverbs)

5. Verb features
    - valency
        - **II, AI, TI, TA**
    - PV
        - tense: **Pst**, **Fut**
        - grammatical: *ê-*, *kâ-*, *ka-*, initial change, etc.
        - lexical
    - voice
        - **direct/inverse**
    - person
        - **1SG, 2SG, 3SG, 3′, 1PL, 21PL, 2PL, 3PL, 0SG, 0PL, 0′SG, 0′PL**
        - **1SG-G, 2SG-G, 3SG-G, 3′O, 1PL, 21PL-G, 2PL-G, 3PL-G**

6. Pronoun features
    - personal
        - **number**, **person**
    - demonstrative
        - **animacy**, **number**, proximity

7. Reduced forms
    - elided second person prefix
    - sandhi
    - fragments, hesitations

8. Prepositional phrases

- adposition with locative noun

9. Coordination
    - *êkwa*, *mîna*, others

10. Clause type
    - order
        - **Ind**, **Cnj**, Fut **Cond**, **Imp**
    - interrogative
        - **Qst**
        - *cî* vs. *tân-* (polarity vs. content)
    - average sentence length

11. Noun phrases
    - **animacy**
    - **number**
    - **possession**
        - possessed w/overt possessor
    - occurs with demonstrative

12. Adverbials
    - Particles
        - **locative, temporal, negative, quantifiers, interjections**
        - other categories

13. Complement clauses
    - *kâ-* preverb, others

14. Word order choices
    - verbs and arguments

        - occurs with **actor/goal before or after verb**: noun, pronoun, other (e.g., kahkiyaw, pêyak)
        - occurs with no overt actor/goal
    - Dem N vs. N Dem

15. Special features of conversation
    - For future consideration

Appendix E

PCA results: Further details

Table E.1: PCA feature weights: full corpus

PC1				PC2			
Positive		Negative		Positive		Negative	
Feature	Weight	Feature	Weight	Feature	Weight	Feature	Weight
IPC	0.20313	V	-0.21039	Mixed	0.27895	Num	-0.18549
Prt	0.15101	Nonlocal	-0.20621	1Sg	0.25193	Quant	-0.18009
Pers	0.14311	Prs	-0.19938	1SgO	0.24539	Pl	-0.17865
Pron	0.14263	Ind	-0.19471	3SgO	0.22756	4Sg/PlO	-0.15339
Excl	0.12811	4Sg/Pl	-0.19315	2SgO	0.20696	Obv	-0.14406
IPC	0.12798	4Sg/PlO	-0.18970	Local	0.20278	@ACTOR>	-0.12614
Qst	0.12346	Obv	-0.18259	Px1Sg	0.20210	Nonlocal	-0.12514
Foc	0.12168	Px3Sg	-0.17390	Inverse	0.20082	Prop	-0.11356
Emph	0.11150	3Sg	-0.17380	Sg	0.16029	Px3Pl	-0.11219
Dem	0.10828	TA	-0.17172	Quot	0.15743	Px4Sg/Pl	-0.11035
0Sg	0.09810	Direct	-0.17098	2Sg	0.15328	4Sg	-0.10195
Cnj	0.08532	D	-0.16600	Fut	0.14901	IPT	-0.07299
Pl	0.08404	Fut	-0.16253	Cond	0.14528	A	-0.07052
IPH	0.08074	A	-0.15937	Inan	0.13747	5Sg/PlO	-0.06615
1Sg	0.07817	Imp	-0.15087	Qst	0.13408	@GOAL>	-0.06596
Neg	0.07756	@<ACTOR	-0.12987	Foc	0.11454	4Sg/Pl	-0.05776
IPL	0.07086	Voc	-0.12672	Px21Pl	0.11298	RdplS	-0.04673
X	0.06510	AI	-0.12402	TA	0.10829	IPN	-0.04463
Quant	0.06460	IC	-0.11923	Imp	0.10679	Rel	-0.03995
Px21Pl	0.05586	Px4Sg/Pl	-0.11658	INCL	0.10540	Loc	-0.03824
RdplS	0.05428	4Sg	-0.11400	Px1Pl	0.09682	AI	-0.03693
Mixed	0.05415	Px2Sg	-0.11323	Pers	0.09212	Prt	-0.03468
Rflx	0.04389	2Sg	-0.11252	D	0.08617	Dub	-0.03462
Incl	0.03994	Local	-0.10702	0Sg	0.07991	IPC	-0.03316
Px1Pl	0.03515	Cond	-0.10170	Px2Sg	0.07896	Px3Sg	-0.03286

Table E.2: PCA feature weights: BT subcorpus

PC1				PC2			
Positive		Negative		Positive		Negative	
Feature	Weight	Feature	Weight	Feature	Weight	Feature	Weight
Quant	0.18993	Px1Sg	-0.22728	Obv	0.29740	II	-0.19959
IPC	0.18581	D	-0.20442	A	0.27517	1Sg	-0.17052
Num	0.17903	Quot	-0.20399	4Sg/PlO	0.26565	0Sg	-0.14765
Pl	0.17309	2SgO	-0.19954	Nonlocal	0.23977	IPC	-0.14512
IPL	0.15692	Mixed	-0.19224	Pron	0.22072	Local	-0.14415
Prt	0.12251	1SgO	-0.19131	Dem	0.22034	Fut	-0.13212
II	0.10708	3SgO	-0.17872	Direct	0.17916	IPH	-0.12235
Zero	0.10675	1Sg	-0.17649	Px3Sg	0.15259	4Sg	-0.12022
Prop	0.09564	Fut	-0.17093	5Sg/PlO	0.13634	PrtHT	-0.11363
IPN	0.09251	Cond	-0.16651	@ACTOR>	0.13478	1SgO	-0.11359
I	0.08879	V	-0.16337	TA	0.13355	IPT	-0.10395
@ACTOR>	0.08115	3Sg	-0.16167	@<GOAL	0.12856	Px1Sg	-0.09647
Px3Pl	0.07795	Inverse	-0.16024	@GOAL>	0.11898	Quot	-0.09641
4Sg	0.07503	2Sg	-0.15271	Px4Sg/Pl	0.10296	Ind	-0.09450
IPT	0.07346	TA	-0.15263	@<ACTOR	0.10218	IPN	-0.09225
Neg	0.05395	Local	-0.14935	X4Sg.Pl	0.09204	TI	-0.08892
0Sg	0.05335	Imp	-0.14907	D	0.07863	Px21Pl	-0.08746
PrtHT	0.04926	Sg	-0.13626	Pl	0.07736	I	-0.08387
RdplS	0.04575	Px3Sg	-0.12057	Num	0.06406	Qst	-0.08245
Indef	0.04130	Ind	-0.11850	Px2Sg	0.06288	Mixed	-0.07080
Cnj	0.04072	IPJ	-0.11389	Cnj	0.06163	EXCL	-0.07068
Px4Sg/Pl	0.04039	Voc	-0.11189	PxX	0.05696	IC	-0.06862
@GOAL>	0.03080	Inan	-0.10673	Rel	0.04975	2SgO	-0.06682
EXCL	0.00960	Px2Sg	-0.10664	RdplW	0.04550	2Sg/PlO	-0.05870
4Sg/PlO	0.00802	Px1Pl	-0.09354	RdplS	0.04326	V	-0.05798

Table E.3: PCA feature weights: A-W subcorpus

PC1				PC2			
Positive		Negative		Positive		Negative	
Feature	Weight	Feature	Weight	Feature	Weight	Feature	Weight
AI	0.23346	Cnj	-0.22677	Pl	0.16265	V	-0.25610
RdplS	0.20231	@<GOAL	-0.21977	Prt	0.15798	Prs	-0.21479
Ind	0.19343	Px21Pl	-0.21667	Nonlocal	0.14786	1Sg	-0.21212
Quot	0.17259	INCL	-0.21627	IPC	0.14406	Fut	-0.21162
Der/Dim	0.16187	TI	-0.19065	0Sg	0.14222	Ind	-0.20960
RdplW	0.14354	Cond	-0.18703	Obv	0.13027	Local	-0.19214
EXCL	0.11354	I	-0.18517	4Sg/PlO	0.12660	Mixed	-0.18302
IPC	0.10130	II	-0.16785	Px3Pl	0.12430	3Sg	-0.18045
Quant	0.08493	@GOAL>	-0.16227	Quant	0.11786	3SgO	-0.17791
3Sg	0.08331	Fut	-0.15850	@GOAL>	0.10884	1SgO	-0.17663
IPL	0.07746	Inverse	-0.14113	PxX	0.08721	TA	-0.16791
Prop	0.07216	Px2Sg	-0.13755	4Sg/Pl	0.08613	Quot	-0.15863
Loc	0.07020	Sg	-0.13428	X	0.07997	Imp	-0.15259
PrtHT	0.06706	Inan	-0.13028	Rel	0.07327	IPL	-0.15231
1Sg	0.06531	2SgO	-0.12641	A	0.06209	Inverse	-0.15048
IPT	0.06218	@<ACTOR	-0.12349	IPH	0.05859	Voc	-0.14784
Emph	0.05884	TA	-0.12032	IPT	0.05209	2SgO	-0.14780
V	0.05604	@ACTOR>	-0.11700	Cnj	0.04413	Px2Sg	-0.13461
Num	0.04994	Local	-0.11660	@ACTOR>	0.04390	2Sg	-0.13364
Dub	0.04099	2Sg	-0.11596	IPJ	0.03982	Px4Sg/Pl	-0.12043
1SgO	0.03232	Direct	-0.10508	Prop	0.03780	AI	-0.11677
Rflx	0.02856	Mixed	-0.09931	Px3Sg	0.03694	Inan	-0.11376
IPJ	0.02723	IPN	-0.09910	TI	0.03555	Px21Pl	-0.10487
Px1Sg	0.02708	Neg	-0.09886	Dub	0.03494	Px1Pl	-0.09799
Pers	0.02341	Imp	-0.09254	IPN	0.03491	Sg	-0.09612

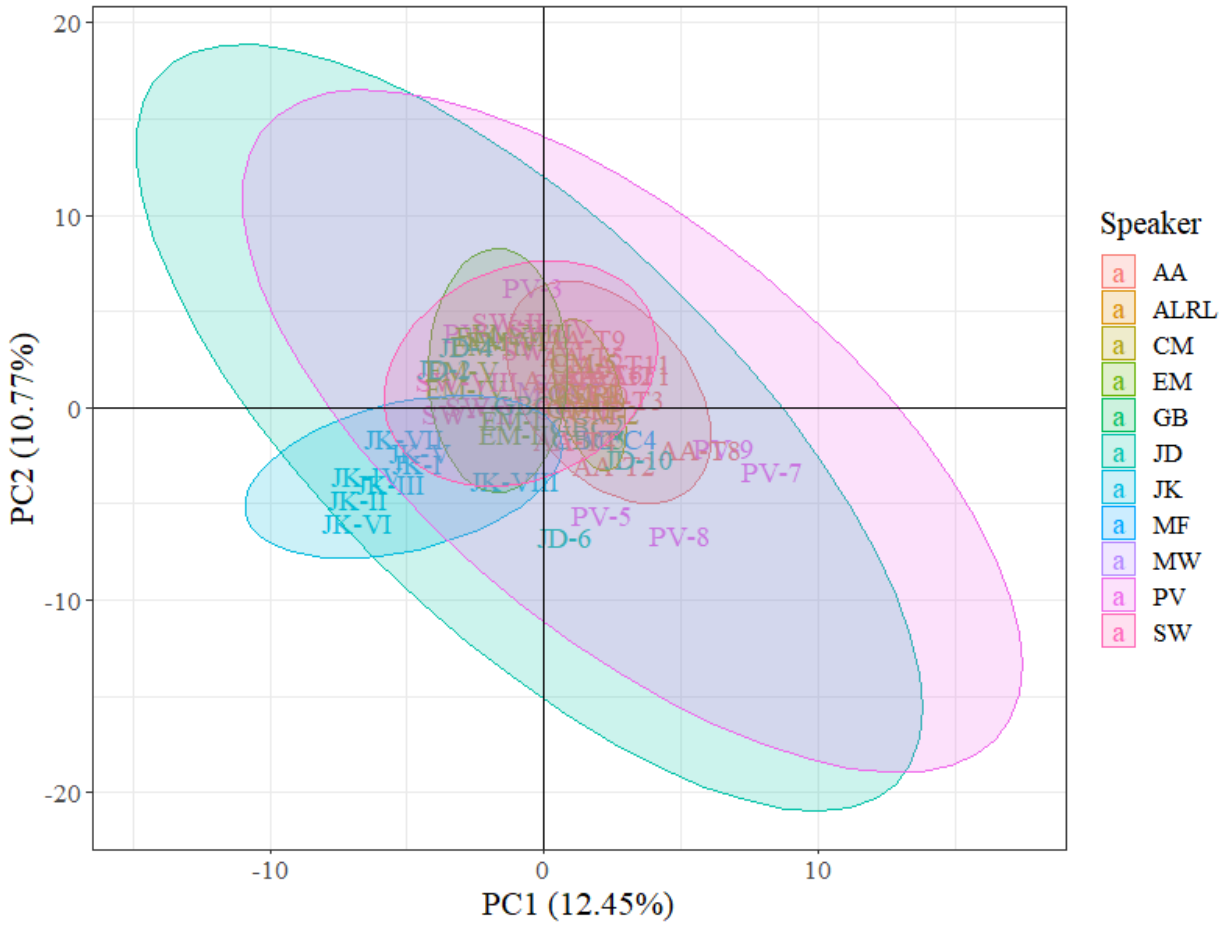


Figure E.1: A-W subcorpus PCA: Chapters along PC1 & PC2 (with ellipses)