Building a Constraint Grammar Parser for Plains Cree Verbs and Arguments

Katherine Schmirler[1], Antti Arppe[1], Trond Trosterud[2], Lene Antonsen[2]
[1]University of Alberta, [2]UIT Arctic University of Norway
{schmirle, arppe}@ualberta.ca, {trond.trosterud, lene.antonsen}@uit.no

Abstract

This paper discusses the development and application of a Constraint Grammar parser for the Plains Cree language. The focus of this parser is the identification of relationships between verbs and arguments. The rich morphology and non-configurational syntax of Plains Cree make it an excellent candidate for the application of a Constraint Grammar parser, which is comprised of sets of constraints with two aims: 1) the disambiguation of ambiguous word forms, and 2) the mapping of syntactic relationships between word forms on the basis of morphological features and sentential context. Syntactic modelling of verb and argument relationships in Plains Cree is demonstrated to be a straightforward process, though various semantic and pragmatic features should improve the current parser considerably. When applied to even a relatively small corpus of Plains Cree, the Constraint Grammar parser allows for the identification of common word order patterns and for relationships between word order and information structure to become apparent.

Keywords: Plains Cree, Constraint Grammar, automatic parser

1. Introduction

This paper presents the first developmental stage of a Constraint Grammar (CG) parser for Plains Cree, a North American Indigenous language spoken by several thousands of people primarily in the Canadian provinces of Alberta and Saskatchewan. It is a member of the Algonquian language family and the westernmost language of the Cree-Montagnais-Naskapi continuum spoken across much of Canada. A number of texts have been published in Plains Cree, totalling several hundreds of thousands of words; while these comprise a small corpus compared to many available for languages such as English, they do allow for the development and testing of computational tools such as morphological and syntactic models. Once such tools have been developed, they can be used not only for corpus investigations, but can be implemented in various language technological applications for speakers, teachers, and students of Plains Cree. These include morphologically “intelligent” online dictionaries, morphosyntactically tagged corpora, spell checkers, grammar checkers, and intelligent computer-assisted language learning (ICALL) applications.

1.1 Research Questions

The present research aims to answer three questions. First, to what extent can basic syntactic relationships in Plains Cree be adequately modelled using only morphological information? Second, of those syntactic relationships that cannot be identified, how can they be accounted for in future development? Third, when applied to a corpus of Plains Cree, what word order patterns can be identified?

2. Plains Cree Morphosyntax

Like all Algonquian languages, Plains Cree is highly polysynthetic with a complex morphological system. The morphosyntactic features and inflectional system of Plains Cree are modelled by means of a morphological analyser (Snoek et al, 2014; Harrigan et al., 2017), the output of which includes morphosyntactic feature tags and all possible analyses of ambiguous forms. These analyses become the input for the Constraint Grammar parser. The features discussed in this section are tagged in the output of the morphological analyser and are referenced in the constraints specified in the current CG parser.

2.1 Nominal and Verbal Features

2.1.1 Nominal Features

Plains Cree nouns are classified in terms of animacy, with two genders, animate and inanimate. All words for people, animals, and trees are animate, while most other forms are inanimate. However, there are many words such as *asikan* ‘sock’ or *kôna* ‘snow’[1] that are semantically inanimate but behave as animate nouns in Plains Cree. In this way, it is truly a grammatical distinction rather than a purely semantic one (e.g. Ahenakew, 1987). Sex-based gender distinctions are not made in Cree (e.g. *wiya* ‘he, she, it [animate]’). Animate nouns are also inflected for obviation, a pragmatic category that distinguishes between more topical (proximate) and less topical (obviative) animate third persons. Plains Cree obviative third persons are marked with the suffix *-a* and do not distinguish number (Wolfart, 1973). Obviation occurs, and overt marking is required, whenever more than one animate third person participant is present in a discourse. This includes one animate third person acting on another, but also when one animate third person possesses another.

Obviation can be considered part of a larger person hierarchy, which is particularly relevant in the discussion of transitive verbs below. In this hierarchy, animates are ranked over inanimates, proximate over obviative, and speech act participants (local) over non-speech act participants (non-local). Additionally, second person is ranked over first person. The person hierarchy is visualized in (1).

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

$$2 > 1 >> 3 > 3' >> 0$$

Both nouns and demonstrative pronouns may stand alone as the arguments of verbs, though nouns and demonstrative pronouns may also co-occur. However, due to extensive

[1] Circumflexes or macrons are used to mark long vowels in written Plains Cree; all Cree forms throughout this manuscript are written using the Standard Roman Orthography.

verbal morphology (discussed below), it is more common that no overt arguments are present.

2.1.2 Verbal Features

Plains Cree verbs are classified by their transitivity and by the animacy of their participants. In the context of parsing, these classes allow the CG to determine how many participants (transitivity) of which noun classes (animacy) can potentially be associated with a given verb. Following Algonquianist tradition, verbal arguments are referred to as *actors* and *goals*, rather than as *subjects* and *objects*.[2] There are four classes, inanimate intransitive (VII), animate intransitive (VAI), transitive inanimate (VTI), and transitive animate (VTA). The intransitive classes are each marked for one participant, VIIs may take an inanimate actor and VAIs an animate actor. VTIs are also only marked for one animate actor, but an inanimate goal may be syntactically present. VTAs present the most complex person morphology, as two animate participants are marked on the verb. Furthermore, Plains Cree makes use of a direct-inverse system, where, rather than coding for actor/subject or goal/object (e.g. as done in case-marking languages), the direction of an action is marked on the verb. The person and number marking does not mark the role of a participant; instead, the direction theme sign indicates which is the actor and which is the goal (Wolvengrey, 2011, p. 173-6). Direction is determined by means of the person hierarchy (above): direct morphology occurs when a more topical participant acts on a less topical one and inverse morphology occurs in the opposite situation. Examples are given in (2).

(2) Plains Cree (Wolvengrey, 2011, p. 175)

a. *câniy kî-wîcihêw mêrîwa*

câniy	kî-wîcih-**ê**-w	mêrî-wa
John.PROX	PST-help.VTA-**DIR**-3SG	Mary-OBV

'Johnny helped Mary.'

b. *câniy kî-wîcihik mêrîwa*

câniy	kî-wîcih-**ik(w)**-(w)	mêrî-wa
John.PROX	PST-help.VTA-**INV**-3SG	Mary-OBV

'Mary helped Johnny.'

Sentences like these also demonstrate the key way in which proximate and obviative marking are unlike grammatical cases. In Plains Cree, regardless of the semantic or syntactic role of an argument, the nominal marking remains unchanged while the roles are indicated by direction morphology and the relative topicality of the arguments.

2.2 Syntactic relationships

As Plains Cree is a non-configurational language, word order is not used to determine syntactic relationships such as actor and goal, as demonstrated in the above examples. When determining relationships between nouns and verbs, one can rely almost entirely on the morphology of the verb and any lexicalised arguments. However, there are still some linear relationships that can be used in a CG parser. For example, though both nouns and demonstratives may occur on their own, when they occur adjacent to each other with agreeing features, they can be described as noun phrases (NPs), specifying something akin to "this/that N". Though linear order is not used to determine syntactic relationships, pragmatic positions have been identified, e.g. topical and focused nouns generally occur before a verb (e.g. Dahlstrom, 1995). Some patterns can be seen in a corpus investigation, suggesting avenues for future research.

3. The Plains Cree Corpus

The current corpus for Plains Cree consists of narratives, dialogues, speeches, and lectures recorded by F. Ahenakew in the 1980s and 1990s. These were then transcribed, translated, and edited by F. Ahenakew and H.C. Wolfart and are available in several published volumes (Ahenakew, 2000; Bear et al., 1992; Kâ-Nîpitêhtêw, 1998; Masuskapoe, 2010; Minde, 1997; Vandall & Douquette, 1987; Whitecalf, 1993). Together, these total 108,413 tokens (18,649 types), of these, 73,189 tokens (15,994 types) are identifiable Plains Cree words.

This corpus has been evaluated using a morphological analyser for Plains Cree, the first versions of which are described in Snoek et al. (2014) and Harrigan et al. (2017). After the initial analysis was performed, the corpus was hand-verified, correcting erroneous analyses and adding analyses that the analyser was unable to produce. This has served to both identify areas for improvement in the morphological model and, more importantly for the present purpose, to include as many correct analyses as possible for use as input to the syntactic parser. For further information on the annotation process, see Harrigan et al. (2017). The hand-verified corpus is referred to as the morphological Gold Standard. Using the morphological Gold Standard, the CG parser described below was tested using a smaller portion of this corpus, hand-coded for basic syntactic relationships. This is referred to as the syntactic Gold Standard. The online version of the corpus uses the Korp interface, based on the open-source tools in the IMS Open Corpus Workbench (Evert & Hardie, 2011; Borin et al., 2012) and can be found at http://altlab.ualberta.ca/korp. The corpus resources, including the texts, the model coverage compared to the hand-verified Gold Standards, and the online corpus interface and its search capabilities are discussed in detail in Arppe et al. (2017).

Additional Plains Cree texts, both historical (e.g. Bloomfield, 1930, 1934; Demers et al., 2010) and modern (e.g. Wolvengrey 2007, a number of children's books, etc.) are available for future inclusion in the corpus. Further development of the morphological analyser to model archaic morphological features is also underway to more thoroughly analyse historical texts.

4. Constraint Grammar

A Constraint Grammar (or Constraint Grammar Parser) is a list of descriptive, context-based constraints designed to parse natural language. The constraints disambiguate forms using morphological and lexical information to output a single surface morphosyntactic reading for each utterance. Unlike a generative approach to grammar, no input is considered incorrect or ungrammatical; every input is instead analysed as best as the parser allows, regardless of

[2] *Actor* and *goal* do not correspond to semantic roles, however, as they can be agents, experiencers, etc.

its grammaticality (Karlsson, 1990). The fundamental underpinnings of Constraint Grammar are the rules of a language that a linguist can easily identify as categorical, and do not need to be learned as tendencies; the constraints can then be tested against a corpus of natural language to determine how accurate the rules identified by linguists are (F. Karlsson, pers. comm. to A. Arppe).

The input for the CG is morphological analysis, such as that returned by a morphological analyser. The morphological analyses offered for each word form are the *readings* in its *cohort*. The CG then has two main goals: 1) the disambiguation of forms with a cohort of more than one morphological analysis and 2) the assignment of syntactic functions (e.g. dependencies) using the sentential context of each word form. CG can delineate the range in which to look for dependencies by referencing clause boundary punctuation in the text, such as periods, question marks, exclamation points, commas, colons, quotation marks, etc.

Constraints are used to narrow down (disambiguate) the relationships between words in a sentence to return an analysis (Karlsson, 1995a). The constraints in CG are of several different types: 1) constraints that disambiguate based on the context, 2) constraints that map the clause boundaries using punctuation and capitalization, and 3) constraints that map the syntactic functions of word forms. Alongside context-based constraints, heuristic-based parsing can also be used to improve the analysis. Heuristic constraints may be used to disambiguate where context cannot, such as choosing a reading because it was contextually selected previously, or to simplify issues by ignoring constraints and enforcing analyses. They can also be purely probabilistic and choose the analysis that is the most likely based on prior quantitative analysis (Karlsson, 1995b). Though not yet widely implemented for Plains Cree, heuristic constraints would prove useful for a number of frequent ambiguous forms for which context is not sufficient for disambiguation (see below).

The CG formalism for the current Plains Cree parser uses the VISLCG-3 compiler (e.g. Bick & Didrikson, 2015; documentation can be found at http://visl.sdu.dk/cg3.html). This newer compiler includes various capabilities not available in earlier versions, such as the use of regular expressions in constraints, the ability to specify relationships between constraints, easier control over the scope of parameters, the chunking of heads and modifiers, and the identification of dependencies between objects and their complements or between anaphoric or discourse relationships. Of these, only regular expressions are currently used in the Plains Cree CG parser, though other capabilities will be used as development of the parser progresses. Additionally, the current formalism also allows for the implementation of lexicosemantic information that can serve to refine the constraints further and better represent syntactic relationships; adding lexicosemantic information is planned for the immediate future.

Though an admittedly rather archaic technology, the CG formalism offers several advantages for modelling the syntax of languages such as Plains Cree. One, the rich agreement morphology of Plains Cree lends itself readily to the identification of arguments within the non-configurational syntax, a well-known capability of the formalism. Two, the implementation of categorical rules identified by linguists is straightforward within the formalism, requiring very little training to compose basic constraints and begin testing. Three, these categorical rules are often sufficient to model much of the syntax, which is particularly advantageous for languages with only several tens of thousands of words available in a corpus, where stochastic modelling would be impractical; straightforward categorical rule identification also speeds the model development. Four, for understudied or endangered languages such as Plains Cree, speed of development is crucial for tools such as syntactic parsers for inclusion in applications for use by speakers such as grammar checkers and ICALL applications. It is for these reasons that Constraint Grammar has been selected for modelling Plains Cree syntax.

5. Building a Parser for Plains Cree

The current iteration of the Plains Cree parser implements 67 disambiguation constraints and 105 function constraints. The basic patterns found in the constraints, as well as their coverage of a hand-coded text of ~3,200[3] words (Vandall & Douquette, 1987) are laid out below.

5.1 Disambiguation

Twenty-seven of the 67 of the constraints used for disambiguation in the Plains Cree parser are required not because of ambiguity inherent to the language, but because of ambiguity introduced by a descriptive version of the morphological analyser, which ignores vowel length distinctions in favour of recognising as many forms as possible. For example, the verbal suffix *-yân* marks first person singular in certain verb forms, and constitutes a minimal pair with *-yan*, which marks second person singular. The descriptive analyser offers both analyses regardless of how the word is spelled, so we have written constraints that choose the one that matches the spelling, which for the current texts we deem to be accurate.[4] While the ambiguity introduced by the analyser presents a number of challenges, here we instead discuss how the constraints handle the inherent ambiguities.

Diminutive nouns present one such ambiguity. The morphological analyser may give up two analyses, one where the diminutive noun is its own lexical entry, if one is available, and one where it is derived from another lexical entry. For these situations, we have opted to choose the reading where the diminutive is its own lexical entry, as it often has slightly different semantics than simply “little X”. In the following example, *nêhiyâsis* does not simply mean “small Plains Cree person” (from *nêhiyaw* ‘Plains Cree person’), but ‘young Plains Cree person’. We use the

[3] These ~3,200 words comprise approximately 4.5% of the current Plains Cree corpus; additional texts in a hand-coded reference corpus would be beneficial to future development of the CG parser.

[4] This solution, while workable for the present corpus, will not be sufficient in the eventual implementation of a generally applicable CG parser, as many speakers do not mark vowel length.

constraint given in (3) to disambiguate the cohort given in (4); a semicolon <;> marks the removed reading.[5]

(3) REMOVE:DerNo   Der/Dim (0C N) ;

(4) "<nêhiyâsisak>"   'young Plains Cree person'
;   "nêhiyaw" N AN Der/Dim N AN Pl
    "nêhiyâsis" N AN Pl

Similarly, there are a number of forms that can be either proximate or obviative. These may be homophonous animate and inanimate nouns, but more frequently these are demonstrative pronouns that may be either inanimate plural or animative obviative. These are disambiguated based on their context, such as an adjacent noun with agreeing features. An exemplary constraint is given in (5).[6]

(5) REMOVE:DemANObvnotIN IN + Pl (1 N + AN + Obv) (0 Dem + AN + Obv) ;

This type of constraint applies in a case such as the following example in (6), where an ambiguous pronoun is identified as obviative when adjacent to an obviative noun.

(6) "<ôhi>"   'this/those'
;   "ôma" Pron Dem Prox IN Pl
    "awa" Pron Dem Prox AN Obv
"<otêma>"   'his dog/dogs (obviative)'
    "atim" N AN Sg Pl Px3Sg Obv

5.1.1 Coverage

The effectiveness of the disambiguation constraints can be examined with a text that has been manually coded for disambiguation and function assignment. In this text of 3,226 Plains Cree words, of which 524 have more than one possible reading before disambiguation, 544 readings were manually marked to be removed. For this same text, the disambiguation constraints remove a total of 374 readings. Of these, 335 are removed in both the manually-coded text and by the constraints. Therefore, the recall rate for removal of an ambiguous reading is 62% while the precision rate is 90%. Many of the problematic cases are those which cannot be determined by sentential context alone, and so lower rates are to be expected. For readings that are not removed (i.e., treated as correct or preferred), 3,241 are marked as correct in the manually-coded text and 3,202 remain after the disambiguation constraints have been applied; therefore, the recall rate for correctly preferred readings is 99%. In Table 1 below, the number of Plains Cree word forms and the numbers of readings both before and after disambiguation are given. When a correct analysis remains, this is indicated with a plus sign <+>; when a correct analysis is removed, this is indicated with a minus sign <->.

n	Before	After	Accuracy	%
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

Table 1: Pre- and post-disambiguation results

Of these word forms, those that have only one reading both before and after disambiguation are assumed to be correct as they are drawn from the hand-verified morphological Gold Standard corpus.[7] There are 529 word forms with two or more readings before disambiguation; 297 of these have one correct reading after disambiguation. This gives a recall rate of 56% for the selection of the one correct reading. After disambiguation, 315 forms are reduced to one reading; the 297 with the correct analysis remaining give a precision rate of 94%. There are 209 word forms with one or more reading remaining after disambiguation, 6% of the 3,226 total word forms presented here. However, in all cases where more than one reading remains, the correct reading has not been removed.

5.2 Function Assignment

5.2.1 Nouns and Demonstratives

Relationships between nouns and demonstratives are determined by their linear and morphological relationships. To identify a demonstrative as dependent on a noun, it must be immediately adjacent to the noun and agree for animacy, number, and obviation. An exemplary constraint for modelling such a relationship is given in (7).[8]

(7) MAP:DemNANSgR @<N TARGET Dem + AN + Sg IF (NOT -1 Obv)(-1 N + AN + Sg BARRIER CLB) ;

Clause boundaries (CLB) are used to limit the scope of constraints. Clause boundaries include a set of punctuation, including periods and other sentence-level punctuation, as well as semicolons, commas, etc. These do not result in clauses in the traditional sense, i.e. containing a verb, but simply serve to divide the text into more manageable sections within which the constraints attempt to identify relationships. The assumption made here is that punctuation identifies natural pauses, and therefore to some degree the intonational contours, that would occur in speech and ideally correspond to syntactic units of some kind. For the Plains Cree corpus, which has been transcribed from recorded speech, we have assumed that

[5] This constraint is interpreted as follows: *remove* the reading containing the tag *Der/Dim*, in the context that the word itself (*0*) must be (*C*) a noun (*N*). The name of the constraint for reference purposes is given after the direction REMOVE, here saying we do not want the derived reading (*DerNo*).

[6] This constraint also *removes* an unwanted reading; an inanimate plural (*IN+Pl*) demonstrative reading is removed when there is an animate obviative noun (*N+AN+Obv*) immediately to the right (position indicated by *1*). The second context indicates that the demonstrative itself (*0*) must also have an *AN+Obv* reading.

[7] Though single analyses here are understood as correct, a single analysis will never be removed by a CG parser. In texts where analyses have not been verified, a word form with only one analysis cannot be guaranteed to be correct.

[8] Function constraints *map* syntactic function tags. Here, the tag is *@<N*, which marks a demonstrative that is dependent on a noun (*N*) to the left (<). The position immediately to the left is indicated with *-1* in the context conditions. Function constraints also make use of *barriers*, which direct the constraint not to look past certain elements, here a clause boundary (*CLB*).

punctuation adequately approximates pauses in the recorded speech. While these boundaries may occasionally separate verbs and arguments, they are still used for the purposes of syntactic modelling.

Multiple constraints of this format are included in the CG parser for Plains Cree; they apply for animate and inanimate nouns, both plural and singular, and animate obviative, as well as looking both to the left and right of demonstrative. A total of 10 constraints are used for these relationships. However, further modification of these constraints is required in future development, as they do not yet account for intervening modifiers such as numerals.

5.2.2 Arguments of Verbs

The arguments of verbs, when lexicalised, can be either nouns, demonstrative pronouns, and personal pronouns. Personal pronouns rarely occur overtly with verbs, but constraints are also included for these relationships. These constraints are written to assign @ACTOR and @GOAL tags to the arguments of verbs, where they are present. As above, constraints are specified for animacy, number, and obviation combinations, as well as whether the verb upon which a nominal is dependent is to its left or right. Examples for assigning @ACTOR and @GOAL functions are given in (8) and (9) respectively.[9]

(8) MAP:AITIACTSgR @<ACTOR TARGET N + AN + Sg IF (NOT 0 Loc)(NOT 0 Obv)(*-1 AI + 3Sg OR TI + 3Sg BARRIER V OR CLB) ;

(9) MAP:TAGOAL3R @<GOAL TARGET N + AN + Sg IF (NOT 0 Loc)(NOT 0 Obv)(*-1 TA + 3SgO BARRIER V OR CLB) ;

A total of 72 constraints are required to map the actor and goal functions to nouns and demonstrative and personal pronouns in the current CG parser. Just as for the disambiguation constraints, some function constraints are also required due to limitations introduced by the morphological analyser. Ten constraints are required to mark the pronoun *êkoni* 'those (ones)' as dependent on an adjacent noun or demonstrative pronoun, rather than marking both as an actor or goal. The word class and agreement features of *êkoni* are not fully specified in the morphological analyser, so constraints targeting the form itself using regular expressions are implemented.

Though not yet tested and refined, broad constraints also assign @{<}OBL{>} (oblique) to nouns that are not morphologically associated with a nearby verb. Many current instances of @OBL are due to overapplication of these constraints and so they are not included in the present results, though further development of these constraints is underway. These nouns generally include roles that are not specified by the features of the verb (e.g. VAIs that may take goals, indirect objects of benefactive VTAs), as well as instruments that indicate the means by which an action is performed.[10] Such constraints will be implemented in the future with reference to lexicosemantic features. Some oblique nouns will, however, always be incorrectly marked to some extent. Chief among these are the use of inanimate nouns as animate nouns for pragmatic reasons—if a narrative requires an inanimate entity to act with some degree of agency, it will occur with animate verbs and demonstratives. However, as the morphological analyser will still identify these as inanimate nouns, such pragmatically animate nouns will likely never be parsed automatically.

5.2.3 Coverage

The morphological feature tags on nouns and verbs are generally sufficient for assigning syntactic roles; the CG has both a recall and precision rate of 92% for @ACTOR and @GOAL assignment when compared to the manually-coded text. Where mismatches of argument assignment occur (n = 15), the nouns have features that allow them to agree with verbs (as actor or goal) on either side of them and the parser has selected an option different from the manual coding. In situations where the CG has not identified a manually identified argument (n = 9), further refinement of constraints, particularly where pronouns and obviative nouns are concerned, is required. Incorrectly identified obliques, discussed above, also fall into this category. Where the CG has assigned an incorrect argument tag (n = 15), these are all instances of *ôma* 'this, it is this' being misidentified as a pronominal VII actor or VTI goal rather than a focus particle. As *ôma* is generally a problematic case for disambiguation, these situations cannot be solved only with morphosyntactic features and syntactic context and will instead require the addition of lexicosemantic information.

6. Results and Discussion

6.1 Overall Phrase Order Patterns

When applied to a corpus of ~73,000 Plains Cree words, the CG parser can be used to investigate word order patterns on a larger scale than previously possible for most indigenous languages. First and foremost, we can see to what extent overt arguments occur in the language: 47% of all clauses containing verbs contain no overt arguments. This is the most common pattern for all four verb classes, as seen in Table 2. This table contains the 22 most common phrase order patterns, excluding oblique elements and particles, out of the 19,734 phrases containing verbs in the corpus. Other general patterns can also be identified, for example, goals also occur more often than actors; phrase order patterns where VTAs occur with actors appear in less than 1% of the total verbal clauses.

[9] Unlike the above constraints, these allow for the verb with agreeing features to occur anywhere to the left of the target nominal; this is indicated by the asterisk <*> in the context conditions. Otherwise, these constraints are not unlike those for nouns and demonstratives: they assign functions when agreement conditions are met, and do not look beyond clause boundaries or, in these cases, other verbs (*V*).

[10] Fortunately for the syntactic modelling of Plains Cree, many oblique functions that are served by nouns in languages such as English (e.g. temporal or spatial functions) are instead achieved by verbal constructions (e.g. *kâ-nîso-kîsikâk* '(when) it is Tuesday' ~ 'on Tuesday') and particles. Spatial functions are performed by particles or by nouns with locative marking, which can never be the arguments of verbs; these are ruled out by the context (*NOT 0 Loc*), as in the constraints given in (8) and (9).

n	Verb class and arguments	%
4865	@PRED-AI	24.7
2128	@PRED-TA	10.8
1432	@PRED-TI	7.3
801	@PRED-II	4.1
539	@PRED-TA @<GOAL	2.7
521	@ACTOR> @PRED-AI	2.6
503	@PRED-AI @<ACTOR	2.5
450	@GOAL> @PRED-TI	2.3
386	@GOAL> @PRED-TA	2.0
324	@PRED-TI @<GOAL	1.6
242	@PRED-AI @PRED-AI	1.2
211	@ACTOR> @PRED-II	1.1
173	@ACTOR> @PRED-TA	0.9
162	@PRED-AI @N> @<ACTOR	0.8
159	@PRED-II @<ACTOR	0.8
148	@PRED-TA @<ACTOR	0.7
132	@PRED-TA @N> @<GOAL	0.7
116	@PRED-TA @PRED-AI	0.6
114	@PRED-TI @PRED-AI	0.6
111	@ACTOR> @PRED-TI	0.6
103	@PRED-TI @<ACTOR	0.5
76	@PRED-TI @N> @<GOAL	0.4

Table 2: Overall phrase order patterns

Following verbs with no overt arguments, the next most common patterns are transitive verbs (VTA and VTI) with goals, both following and preceding the verbs. VAIs with arguments are also seen with comparable frequencies, but as VAIs are the most common verb subtype in the corpus, their frequency with or without overt arguments is unsurprising. As frequency decreases, the patterns begin to demonstrate actors with transitive verbs;[11] it is not until phrase patterns make up less than 1% of those in the corpus that both actors and goals are lexicalised.

6.2 VTA Phrase Order Patterns

VTAs present opportunities for deeper investigations, as they allow for two arguments to be lexicalised and contain the greatest amount of morphological information. Clauses containing VTAs represent 4,551 of the overall 19,734 clauses. The 24 most common word order patterns for these 4,551 clauses are given in Table 3. The more frequent patterns from Table 2 are repeated here, though as larger proportions of VTAs. For example, where a given word order pattern with both a VTA and an actor occurs in less than 1% of overall clauses, similar clauses occur with a rate of less than 4% of all VTA clauses. Though a larger percentage than found in the overall corpus, it is still small, confirming that VTAs with actors are indeed rare. Still, as patterns decrease in frequency, similar patterns to those above are generally present: goals are lexicalised more often than actors, actors are more likely to precede the verb, and arguments without demonstrative pronouns associated with them are more common. Note that in some cases, ACTOR or GOAL tags occur multiple times in the same clause; this is due to nouns and demonstrative pronouns occurring with intervening material, so they are not correctly associated with each other. Further refinement of constraints is required to solve these issues.

n	VTA phrase orders	%
2128	@PRED-TA	46.8
539	@PRED-TA @<GOAL	11.8
386	@GOAL> @PRED-TA	8.5
173	@ACTOR> @PRED-TA	3.8
148	@PRED-TA @<ACTOR	3.3
132	@PRED-TA @N> @<GOAL	2.9
54	@PRED-TA @N> @<ACTOR	1.2
35	@ACTOR> @PRED-TA @<GOAL	0.8
28	@GOAL> @PRED-TA @<GOAL	0.6
23	@N> @GOAL> @PRED-TA	0.5
22	@PRED-TA @<GOAL @<GOAL	0.5
21	@ACTOR> @PRED-TA @<ACTOR	0.5
20	@GOAL> @GOAL> @PRED-TA	0.4
17	@ACTOR> @GOAL> @PRED-TA	0.4
17	@N> @ACTOR> @PRED-TA	0.4
13	@GOAL> @<N @PRED-TA	0.3
12	@PRED-TA @<ACTOR @<GOAL	0.3
12	@GOAL> @PRED-TA @<ACTOR	0.3
11	@GOAL> @PRED-TA @N> @<GOAL	0.2
11	@GOAL> @ACTOR> @PRED-TA	0.2
9	@ACTOR> @ACTOR> @PRED-TA	0.2
9	@PRED-TA @<ACTOR @<ACTOR	0.2
8	@ACTOR> @<N @PRED-TA	0.2
6	@ACTOR> @PRED-TA @N> @<GOAL	0.1

Table 3: Overall VTA phrase order patterns

Similar to all verb types combined, 47% of VTA clauses occur without overt arguments. Therefore, verbs with third person proximate and obviative participants are the better candidates to investigate how direct and inverse morphology combined with phrase order information can offer a surface-syntactic insight into how word order reflects information structure, namely topicality and focus. Phrase order patterns for direct and inverse third-person VTAs are presented in Tables 4 and 5 respectively. Note that for direct verbs, the actor is proximate and the goal is obviative, while the reverse is true for inverse verbs. Table 4 gives the 15 most common of 985 third person direct phrases, while Table 5 represents the nine most common of 217 third person inverse phrases in the corpus.

n	Phrase order	%
303	@PRED-TA	30.8
189	@PRED-TA @<GOAL	19.2
133	@GOAL> @PRED-TA	13.5
47	@ACTOR> @PRED-TA	4.8
46	@PRED-TA @N> @<GOAL	4.7
37	@PRED-TA @<ACTOR	3.8
29	@ACTOR> @PRED-TA @<GOAL	2.9
15	@PRED-TA @N> @<ACTOR	1.5
10	@ACTOR> @GOAL> @PRED-TA	1.0
8	@GOAL> @PRED-TA @<GOAL	0.8
8	@GOAL> @PRED-TA @<ACTOR	0.8
8	@GOAL> @ACTOR> @PRED-TA	0.8
8	@ACTOR> @PRED-TA @<ACTOR	0.8
7	@PRED-TA @<ACTOR @<GOAL	0.7
6	@PRED-TA @<GOAL @<GOAL	0.6
5	@ACTOR> @ACTOR> @PRED-TA	0.5

Table 4: Direct third person phrase order patterns

[11] VIIs are ignored in the present paper, as it is known that a number of errors are present in the assignment of actors; lexicosemantic information will be required in the parser to solve these issues.

n	Phrase order	%
87	@PRED-TA	40.1
21	@PRED-TA @<ACTOR	9.7
15	@PRED-TA @<GOAL	6.9
15	@ACTOR> @PRED-TA	6.9
7	@GOAL> @PRED-TA	3.2
4	@PRED-TA @N> @<ACTOR	1.8
2	@GOAL> @PRED-TA @N> @<ACTOR	0.9
2	@GOAL> @PRED-TA @<ACTOR	0.9
2	@ACTOR> @PRED-TA @<ACTOR	0.9

Table 5: Inverse third person phrase order patterns

In these patterns, we see that regardless of whether the actor or goal is obviative, the less topical argument is always more likely to be lexicalised, as qualitative descriptions would suggest (e.g. Dahlstrom, 1995; Wolvengrey, 2011). Additionally, the more topical participant is more likely to occur earlier in the phrase, particularly when both arguments are lexicalised.

This cursory investigation of VTAs is only one way in which the phrase order patterns can be examined. In the future, phrase order patterns may be investigated in a number of other ways. One such avenue is the internal structure of arguments: whether they are nouns, pronouns, or nouns plus demonstratives can be further indications of topicality. Additionally, phrase orders in each text can be compared; this is already to some degree possible with the VTA phrase order patterns in Vandall and Douquette (1987), as these were previously summarised in Wolvengrey (2011). Between just Vandall and Douquette (1987) and the entire corpus, the percentages of phrase orders demonstrate similar patterns: a single verb is most common and V GOAL ACTOR is the least frequent. In between, the phrase order frequencies descend in nearly identical orders. However, where VTA clauses without any lexicalised arguments make up nearly half (47%) of the VTA clauses in the overall corpus, they make up only 31% of the VTA clauses in Vandall and Douquette (1987).

These discrepancies may be due to inherent differences in the styles of texts in the collection; Vandall and Douquette (1987) is a collection of shorter narratives, while other texts such as Bear et al. (1992), Kâ-Nîpitêhtêw (1998), and Masuskapoe (2010) include longer narratives and dialogues, where there may be more opportunity for non-lexicalised arguments as more verbs refer to the same topical participant. Additionally, Vandall and Douquette (1987) speak mostly about history and relate others' stories, so there are many third person verbs, which can potentially occur with multiple overt arguments. On the other hand, texts like Bear et al. (1992) involve more stories about the speaker's own lives, and dialogues where speakers directly address each other, so there are likely to be a greater proportion of first and second person verbs, which rarely occur with overt arguments, such as personal pronouns.

7. Conclusions

The Constraint Grammar formalism has proven to be an excellent tool for the computational modelling of verbs and arguments in the non-configurational syntax of Plains Cree, making use of its rich agreement morphology. Furthermore, Plains Cree syntactic roles can be adequately modelled using only morphological feature information. Moreover, the type of semantic information concerning certain words and sets of words that would substantially improve the model coverage are readily apparent when the CG parsing rules are tested.

When applied to a Plains Cree corpus, the parser can be used to investigate the phrase order patterns in the language, which display generally expected patterns even without access to higher-level discourse information. A large-scale analysis of clauses reveals that nearly 50% of verb clauses in this corpus occur without overt lexicalised arguments; even for VTAs with two third person arguments, over one third occur without overt arguments. Among those that do occur, the interplay between semantic role and obviation also presents itself: the less topical participant is more likely to be lexicalised, but the more topical is more likely to occur earlier in a clause.

These investigations also suggest that different genres may occur with different lexicalisation patterns. Further development of the CG parser will allow not only for more accurate descriptions of Plains Cree syntax, but also for implementation within tools and resources that can be used by students and speakers of the language.

8. Acknowledgements

This work has been undertaken in the Alberta Language Technology Lab (URL: http://altlab.artsrn.ualberta.ca), University of Alberta, and it has been funded by the Social Sciences and Humanities Research Council of Canada (SSHRC) through a Connections Outreach Grant (611-2016-0207), a Partnership Development Grant (890-2013-0047), and a Doctoral Fellowship (752-2017-2105), as well as by a KIAS Research Cluster Grant (University of Alberta).

9. References

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

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.). Winnipeg: University of Manitoba Press.

Arrpe, A., Schmirler, K., Harrigan, A. G., & Wolvengrey, A. (2017). A morpho-syntactically tagged corpus for Plains Cree. Paper presented at the 49th Algonquian Conference, Montreal, QC, 27-29 October, 2017.

Bear, G., Fraser, M., Calliou, I., Wells, M., Lafond, A., & Longneck, R. (1992). *kôhkominawak otâcimowiniwâwa / Our Grandmothers' Lives as Told in Their Own Words*. F. Ahenakew & H.C. Wolfart (Eds.). Regina: Canadian Plains Research Center.

Bick, E., & Didriksen, T. (2015, May). CG-3—Beyond Classical Constraint Grammar. In *Proceedings of the 20th Nordic Conference of Computational Linguistics, NODALIDA 2015, May 11-13, 2015, Vilnius, Lithuania* (No. 109, pp. 31-39). Linköping University Electronic Press.

Borin, L, Forsberg, M., & Roxendal, J. (2012). Korp-the corpus infrastructure of Språkbanken. In *LREC*, pp. 474-78.

Bloomfield, L. (1930). *Sacred stories of the Sweet Grass Cree.* National Museum of Canada Bulletin, 60

(Anthropological Series 11). Reprinted 1993, Saskatoon: Fifth House.

Bloomfield, L. (1934). *Plains Cree texts*. American Ethnological Society Publications 16. New York. Reprinted 1974, New York: AMS Press.

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

Demers, P., McIlwraith, N.L., Thunder, D., & Wolvengrey, A. (Eds.). (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*. Edmonton: University of Alberta Press.

Evert, S., & Hardie, A. (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 conference*, University of Birmingham, UK.

Harrigan, A. G., Schmirler, K., Arppe, A., Antonsen, L., Trosterud, T., & Wolvengrey, A. (2017). Learning from the Computational Modelling of Plains Cree Verbs: Analysis and Generation Using Finite State Transduction. *Morphology*, *27*(4), 565-598.

Karlsson, F. (1990). Constraint grammar as a framework for parsing running text. In *Proceedings of the 13th Conference on Computational Linguistics, Vol. 3* (pp. 168-173). Association for Computational Linguistics.

Karlsson, F. (1995a). 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*. Walter de Gruyter, pp. 1–40.

Karlsson, F. (1995b). 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*. Walter de Gruyter, pp. 41–88.

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

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

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 and H.C. Wolfart (Eds.). Edmonton: University of Alberta Press.

Snoek, C., Thunder, D., Lõo, K., Arppe, A., Lachler, J., Moshagen, S., & Trosterud, T. (2014). Modeling the Noun Morphology of Plains Cree. *ComputEL: Workshop on the use of computational methods in the study of endangered languages*, 52nd Annual Meeting of the ACL.

Vandall, P. & Douquette, J. (1987). *wâskahikaniwiyiniw-âcimowina / Stories of the House People, Told by Peter Vandall and Joe Douquette*. F. Ahenakew (Ed.). Winnipeg: University of Manitoba Press.

Whitecalf, S. (1993). *kinêhiyawiwininaw nêhiyawêwin / The Cree Language is Our Identity: The La Ronge Lectures of Sarah Whitecalf*. H.C. Wolfart and F. Ahenakew (Eds.). Winnipeg: University of Manitoba Press.

Wolfart, H. C. (1973). *Plains Cree: A grammatical study* (Vol. 63.5). Philadelphia: American Philosophical Society.

Wolvengrey, A. (Ed.). (2007). *wawiyatācimowinisa / Funny Little Stories*. Regina: Canadian Plains Research Center.

Wolvengrey, A. E. (2011). *Semantic and pragmatic functions in Plains Cree syntax* (Unpublished doctoral dissertation). LOT. Retrieved from http://dare.uva.nl/record/1/342704.