Morphology
DOI 10.1007/s11525-017-9315-x

Learning from the computational modelling of Plains Cree verbs

Atticus G. Harrigan[1] · Katherine Schmirler[1] · Antti Arppe[1] · Lene Antonsen[2] · Trond Trosterud[3] · Arok Wolvengrey[4]

Received: 12 September 2016 / Accepted: 17 October 2017

Abstract This paper describes the ongoing process of creating a computational morphological model of Plains Cree, a language native to North America, making use of finite-state machines, and with a focus on verbs. We cover prior linguistic theoretical and descriptive models of Plains Cree, moving on to the computational implementation of (chiefly) inflectional phenomena, followed by relevant morphophonological processes. We evaluate the performance of our computational implementation with a hand-verified corpus of Plains Cree, and present a discussion of the morphological complexity found in the corpus, as compared to that of our model and its theoretical underpinnings. The results of this evaluation and research into natural language use inform us about the practical extent of morphological complexity for a polysynthetic language, and allow us to identify avenues for improvement of the model. Finally, this computational model for Plains Cree offers the opportunity to create various digital tools and applications for language users for the maintenance and revitalization of this language in the 21st century.

Keywords Plains Cree · Computational modelling · Finite state transducer · Morphological modelling

✉ A.G. Harrigan
galvin@ualberta.ca

1 4-32 Assiniboia Hall, University of Alberta, Edmonton, Alberta T6G 2E7, Canada

2 SVfak/HUMfak bygget, SVHUM A 3020A, Universitetet i Tromsø—Norges arktiske universitet, Tromso, Norway

3 ISK A 3006, Universitetet i Tromsø—Norges arktiske universitet, Tromso, Norway

4 1 First Nations Way, First Nations University of Canada, Regina, Saskatchewan S4S 7K2, Canada

Published online: 30 October 2017

1 Introduction

Finite State Transducer (FST) models have been used for the computational modelling of morphology for at least three decades, with one of the seminal works being Beesley and Karttunen (2003). This formalism allows for simple but powerful computational machines that can provide both the morphological analysis and generation of word forms. The advantage of these FST models is the ease with which they can be created from linguistic grammatical descriptions. Following the previous, provisional account of the computational modelling of Plains Cree nouns by Snoek et al. (2014), this paper details an endeavour in modelling the morphology of verbs in Plains Cree. This paper will first summarize the relevant grammatical aspects of Plains Cree as well as the basics of finite state morphology, before detailing attempts to theoretically describe and model Plains Cree verbs. Following this, we will describe how we have computationally modelled the Plains Cree verb. Finally, we will make use of a small corpus as well as theoretical descriptions of the language to assess the extent of actual morphological complexity evident in natural language use of Plains Cree, and to evaluate the utility and appropriateness of our computational model.

2 Background

Plains Cree (ISO 639-3: CRK) is a member of the Algonquian language family, which is one of the major language families of North America, spoken across much of Canada and the eastern United States. The Algonquian family includes two major dialect continua spoken in Canada: the Ojibwe and Cree-Montagnais-Naskapi languages, the latter of which stretches mainly from Alberta to Quebec and Labrador.[1]

Plains Cree is the westernmost member of the Cree continuum and is mostly spoken in Alberta, Saskatchewan, and northern Montana. There is said to be approximately 34,000 speakers of Plains Cree (Ethnologue 2016), most over the age of 30. This number is likely overestimated, though a previous account by Ethnologue was equally dubious, with a number of roughly 150. Statistics Canada (2015) reports over 83,000 native Cree speakers, but this is not divided by dialect and so the number who speak the Plains Cree variety is likely much smaller. Wolfart (1973) estimated 20,000 speakers, though the number has likely dropped since then. Although any of these numbers is dwarfed by the number of speakers of majority languages in Canada,

[1]The languages/dialects of Cree-Montagnais-Naskapi continuum share many similarities, as can be seen in the following words for 'person', where the reflexes of Proto-Algonquian **l* are given in boldface. The dialects are given in roughly west to east geographical distribution:

(1) Plains: *iyiniw*
Woods: *iðiniw*
Swampy: *ininiw*
Moose: *ililiw*
Atikamekw: *iriniw*
East: *iyiyiw/iyiyû/iyinû*
Naskapi: *iyiyû*
Innu: *ilnu/innu*

Plains Cree retains a strong presence, particularly for an Indigenous North American language, holding a classification of 5 (Developing) on the Extended Graded Intergenerational Disruption Scale (EGIDS) (Ethnologue 2016), a system for assessing language vitality based on domains of use, intergenerational transmission, and other sociolinguistic factors (Lewis and Simons 2012). With its comparatively large speaker base, Plains Cree has garnered attention from a variety of Americanists, in the form of grammars (e.g. Wolfart 1973; Dahlstrom 2014; Wolvengrey 2011), textbooks (e.g. Okimâsis 2004; Ratt 2016) and an online electronic dictionary (itwêwina[2]).

2.1 Nouns

Plains Cree exhibits a number of morphosyntactic features that differ considerably from the well-known characteristics of more familiar Indo-European languages. Unlike sex-based gender systems such as those found in many contemporary Indo-European languages, Algonquian languages have a two-way gender or noun classification system contrasting inanimate with animate nouns; this grammatical animacy has some basis in semantic animacy: all humans, animals, and trees are animate.

This distinction is not clear-cut though, as *êmihkwân*, 'spoon', is animate,[3] and thus the system is considered one of grammatical classification. Animacy is relevant to nominal and verbal morphology in Cree in various ways. Among nouns, this animacy distinction is manifested in two distinct plural markers, *-ak* for animate and *-a* for inanimate nouns; archaic singular marking is seen for monosyllabic roots, for example *maskw-a* 'bear (an.)' and *wâw-i* 'egg (inan.)'. Plains Cree has no grammatical case system, but it does have locative marking, generally *-ihk* for inanimate nouns and *-inâhk* for animate nouns (Wolfart 1973, 1996).

Plains Cree is a head-marking language, and so the person and number of the possessor is marked on the possessum. Singular possessors are marked only with prefixes: *ni-* for first person, *ki-* for second person, and *o-* for third person. For plural possessors, circumfixes are used: the prefixes are the same as for singular persons, which are matched with a set of suffixes: *ni- -(i)nân* for first person plural exclusive ('ours but not yours'), *ki- -(i)naw* for first person plural inclusive ('ours and yours'), *ki- -(i)wâw* for second person plural ('yours but not ours'), and *o- -(i)wâw* for third person plural. Cree also distinguishes between alienable and inalienable nouns; the latter category must occur with possession and includes kinship terms and body parts as well as some other intimate possessions or relationships, such as *nitâs* 'my pair of pants' versus *mitâs* '(someone's) pair of pants' and *nitêm* 'my dog or horse (animal companion)' versus *atim* '(a) dog' (Wolfart 1973, 1996; Wolvengrey 2011). This nominal possession marking is very similar to person marking seen on verbs, with *ni-* corresponding to first person, *ki-* to second person, and *ø-* to third person.

Within animate nouns, a pragmatic distinction is made between the proximate third person, or more topical entity in a discourse, and the obviative third person, or less topical entity or entities in the discourse. This distinction occurs any time more

[2] http://altlab.ualberta.ca/itwewina.

[3] It is worth noting that animacy is not consistent across dialects of Cree, or even communities of Plains Cree. Some words, such as *sîwinikan* 'sugar', are animate in some dialects and inanimate in others.

than one animate third person occurs in a discourse, such as when one third person animate entity acts on another or when a third person animate entity possesses another. An obviative noun is marked with the obviative suffix *-a* and no number distinction is made; this is conventionally marked with 3′ (or as the '4th person', with no number distinction). The further obviative, which occurs when two obviative entities occur in one discourse necessitating the demotion of one of them, is by convention marked with 3″ (or as the '5th person', also with no number distinction). As obviation is based in topicality rather than syntactic roles, it is generally not considered a marker of case. This is further exemplified with respect to verbal constructions below.

2.2 Verbs

Cree verbs are traditionally classified according to both their transitivity and the animacy of their arguments/participants. There are two classes of intransitive verbs, which can occur with one inanimate participant (VII—verb inanimate intransitive) or one animate participant (VAI—verb animate intransitive). The former includes impersonal verbs such as weather terms and attributive verbs used to describe inanimate objects, and the latter includes intransitive actions and attributive verbs used to describe animate objects (Bloomfield 1946; Wolfart 1973, 1996; Okimâsis 2004). The VII and VAI classes are exemplified in (1) and (2) respectively.[4,5]

(1) VII

a. mispo-n
be.snowing.VII-3SG
'it is snowing'

b. astotin wâpiskâ-w
hat.NI be.white.VII-3SG
'the hat is white'

(2) VAI
mîciso-w
eat.VAI-3SG
's/he eats, has a meal'

[4]Note that in our analyzer we have used '3′ to refer to both inanimate and animate participants. However, '0′ has also been used for inanimate entities (e.g. Wolvengrey 2011).

[5]Abbreviations: CNJ: Conjunct order; COM: comitative; DIR: Direction; IND12: First Person (incl.) in Independent Order; IND3: Third Person in Independent Order; INDP: Independent order; INV: Inverse; NA: Animate Noun; NI: Inanimate Noun; POSS: Possesseum; PV-GxN: Number of Grammatical Preverbs; PV-LxN: Number of Hesitation Marker Preverbs; PV-LxN: Number of Lexical Preverbs; PV: Preverb; RDPLH: Heavy Reduplication; RDPLL: Light Reduplication; THM: Theme sign; VAI-N: <n> final Intransitive Animate Verb stem; VAI-V: Vowel final Intransitive Animate Verb stem; VAI: Verb Animate Intransitive; VII-N: *n* final Intransitive Inanimate Verb stem; VII-V: Vowel final Intransitive Inanimate Verb stem; VII: Verb Inanimate Intransitive; VTA-1: Transitive Animate Verbs not otherwise specified; VTA-2: Vowel final Transitive Animate Verb stem; VTA-3: *c* final Transitive Animate Verb stem; VTA-4: *t* final Transitive Animate Verb stem where *t* is changed to <s> in some forms; VTA-5: Catch-all class for oddly behaved Transitive Animate Verbs; VTA: Verb Animate Transitive; VTI-1: Transitive Inanimate verbs not otherwise specified; VTI-2: Semantically transitive inanimate verbs which follow the VAI-v conjugation; VTI-3: Select group of odd semantically transitive inanimate verbs which follow the VAI-v conjugation; VTI: Verb Inanimate Transitive; 3′: Third person Obviative; 3″: Third Person Further Obviative.

Similarly, there are two classes of transitive verbs, though these are classified by the animacy of their second participant, often considered the object: transitive inanimate verbs (VTI) with an animate subject and an inanimate object, and transitive animate verbs (VTA) with two animate arguments.[6,7] Examples are given in (3) and (4); note that there are three different verbs for 'eat' depending on the transitivity and the animacy of participants.[8]

(3) VTI

mîci-w
eat.VTI-3SG

's/he eats something (inanimate)'

(4) VTA

mow-ê-w
eat.VTA-THM-3SG.SBJ.3′OBJ

's/he eats something (animate)'

Verbal inflections include marking for the subject (and object for VTAs). There are four marking strategies on verbs, used in different semantic and syntactic contexts, characterized by different systems of affixation: the Independent, Conjunct, Imperative, and Future Conditional. These strategies are traditionally called 'verbal orders.'[9] For each, person is marked differently. For the Independent order, which is often used for matrix clauses, similar prefixes and circumfixes to those used for nominal possession are used, though the third person prefixes are not used for verbs. For the Conjunct order, which is used for both matrix and embedded clauses, person information is encoded using only suffixes (which are formally dissimilar to those for the Independent order).[10] The Future Conditional order, which indicates a future or potential action

[6] *Subjects* and *objects* are conventionally called *actors* and *goals* in Algonquian literature (Bloomfield 1946; Wolvengrey 2011). For this paper, however, we make use of the general linguistic terminology of *subject*/*object*.

[7] As an alternative interpretation, Wolvengrey (2011) proposes a three-way distinction between verbs, based solely on the number of animate participants: *V0* containing any verb forms with no animate participants (corresponding to VII), *V1* containing verbs with only one animate participant (corresponding to VAI and VTI), and *V2* containing verbs with two animate participants (corresponding to VTA); this alternative interpretation is also morphologically motivated as VAI and VTI verbs share some of the affixes marking the person and number of the subject (Wolvengrey 2011).

[8] One may notice the similarities between the VTI stem and *mîciso-* in (2). Despite the similarity, any morphological relationship between these two stems can only be posited for Proto-Algonquian, and then only tentatively, and there is no synchronic relationship. However, there are other triplets of VAI, VTI and VTA verb stems which would appear to be based on the same root, such as *wâpi-* 'to (be able to) see (generally)', *wâpaht-* 'to see something (inan.)', and *wâpam-* 'to see someone (an.)' in examples (5) through (8).

[9] The Future Conditional is more commonly considered a subclass of Conjunct and is marked with similar morphology (Wolfart 1973), though we discuss them separately here as we code them separately in our model.

[10] While the Independent and Conjunct orders do have some differences in their syntactic behaviour (Cook 2014), their semantic differences remain poorly understood; however, there appear to be significant tendencies for individual lexemes to be used in either Independent or Conjunct order forms (cf. Harrigan and Arppe 2015; Arppe et al. 2016a).

(often translated as 'when' or 'if'), is generally marked using Conjunct suffixes plus a suffix *-i* or *-o*, though some suffixes differ from the Conjunct order. Finally, the Imperative order, which marks subject for only second person singular and plural and first person plural inclusive, makes use of its own set of suffixes for each class (excluding VII, for which imperatives cannot occur), with different suffixes for both immediate imperatives ('act now') and delayed imperatives ('act later') (Wolfart 1973; Wolvengrey 2011). Examples are given in (5) through (8).

(5) Independent Order

a. ni-wâpi-n
1SG.SBJ.INDP-see.VAI-1SG.SBJ
'I (am able to) see'

b. ni-wâpaht-ê-n
1SG.SBJ.INDP-see.VTI-THM-1SG.SBJ
'I see it (inanimate)'

c. ni-wâpam-â-w
1SG.SBJ.INDP-see.VTA-THM-1SG.SBJ.3SG.OBJ
'I see him/her (animate)'

(6) Conjunct Order

a. ê-wâpi-yân
CNJ-see.VAI-1SG.SBJ
'I (am able to) see'

b. ê-wâpaht-am-ân
CNJ-see.VTI-THM-1SG.SBJ
'I see it (inanimate)'

c. ê-wâpam-ak
CNJ-see.VTA-1SG.SBJ.3SG.OBJ
'I see him/her (animate)'

(7) Future Conditional Order

a. wâpi-yâni
see.VAI-1SG.SBJ
'if/when I (am able to) see'

b. wâpaht-am-âni
see.VTI-THM-1SG.SBJ
'if/when I see it (inanimate)'

c. wâpam-aki
see.VTA-1SG.SBJ.3SG.OBJ
'if/when I see him/her (animate)'

(8) Imperative Order

a. wâpi-tân
see.VAI-1PL.INCL.SBJ.HORT
'Let's see!'

b. wâpaht-a
see.VTI-2SG.SBJ.IMP
'See it (inanimate)!'

c. wâpam-â-tân
see.VTI-THM-1PL.INCL.SBJ.HORT
'Let's see him/her (animate)!'

As noted above, Cree does not have a case system to determine syntactic roles. There is no morphology that indicates the role of a noun and transitivity is determined by the class to which a verb stem belongs. Arguments agree with the verb according to animacy: inanimate subjects for VII and animate subjects for VAI, VTI, and VTA. The inanimate participant in a clause containing a VTI is the object of the verb, or some other oblique argument, but not the subject. The person marking on VII, VAI, and VTI verbs corresponds to the person and number of the subject. However, in VTAs, both arguments are animate and realized in the verbal morphology, with their respective roles determined by obviation and direction morphology, discussed below. Essentially, verbs and their arguments can be thought of as constructions where certain verb stems license a certain number of arguments of particular animacy and vice versa.

To determine the roles of participants in VTA clauses, Algonquian languages make use of a direct-inverse system. VTAs occur with two animate participants and there is no grammatical case or fixed word order by which to determine the semantic roles. Instead, direction is used as a method of determining which argument is the subject and which is the object. In Plains Cree, direction is determined by the relative topicality of participants, extended beyond the proximate-obviative distinction into a full hierarchy known as the Algonquian person hierarchy, given in (9). Direction is indicated by a theme morpheme (a *theme sign* according Algonquianist terminology), which indicates that the action is either direct or inverse. When a more topical participant acts on a less topical participant, the morphology or theme sign is direct (*-â-*, *-ê-*, *-i-*). When the opposite occurs, the morphology or theme sign is inverse (*-ik(w/o)-*, *-iti-*). As visualized in (9), second person is ranked topically above first person, and both of these speech act participants are ranked above all third persons, wherein obviation applies. Due to this hierarchy, first person acting on second necessarily always occurs with inverse morphology. In this way, it is not a passive form, but simply the only way of indicating first person acting on second. For this and a variety of other reasons not discussed herein, Cree inverse forms are not considered equivalent to passive voice in languages such as English (Wolfart 1973; Wolvengrey 2011).

(9) $2 > 1 >> 3 > 3' > 3''$

Table 1 gives a subset of a VTA paradigm, exemplifying direct and inverse forms for different pairs of participants for the VTA *wâpamêw* 's/he (animate) sees someone (animate)'. The person prefixes, and often the suffixes, remain the same while the direction morphology changes. While for VTA Independent forms the direction morphology and person morphology are distinct, this is not true of many VTA Conjunct forms, where portmanteau morphemes convey both direction and subject/object per-

Table 1 Examples of direct and inverse morphology

subject → object	Direct	Translation	object → subject	Inverse	Translation
1s → 3s	ni-wâpam-â-w	I see him/her	1s → 3s	ni-wâpam-ik	S/he sees me
2s → 3s	ki-wâpam-â-w	You see him/her	2s → 3s	ki-wâpam-ik	S/he sees you
1p EXCL → 3s	ni-wâpam-â-nân	We see him/her	1p EXCL → 3s	ni-wâpam-iko-nân	S/he sees us
1p INCL → 3s	ki-wâpam-â-naw	We all see him/her	1p INCL → 3s	ki-wâpam-iko-naw	S/he sees us all
2p → 3s	ki-wâpam-â-wâw	Y'all see him/her	2p → 3s	ki-wâpam-iko-wâw	S/he sees y'all
3s → 3	wâpam-ê-w	S/he (3SG) sees him/her (3′)	3s → 3′	wâpam-ik	S/he (3′) sees him/her (3SG)
3p → 3′	wâpam-ê-wak	They see him/her (3′)	3p → 3′	wâpam-ik-wak	S/he (3′) sees them
3′ → 3″	wâpam-ê-yiwa	S/he (3′) sees him/her (3″)	3′ → 3″	wâpam-iko-yiwa	S/he (3″) sees him/her (3′)
1s → 2s	ki-wâpam-i-n	You see me	1s → 2s	ki-wâpam-iti-n	I see you

son information.[11] With obviation marked on both nouns and verbs, sentences such as those in (10)a. are possible in Plains Cree. Additionally, both obviative and further obviative marking may be needed, depending on the number of third persons lexically specified, as in (10)b. However, when a VTI is involved, and so there is an inanimate object rather than an animate one, no object or obviative marking occurs on either the verb or the inanimate noun, as in (11) (Wolfart 1973; Wolvengrey 2011).

(10) VTA

a. cân pahkwêsikan-a mow-ê-w
John.3SG bread.NA-3′ eat.VTA-THM.DIR-3SG.SBJ.3′.OBJ
'John eats bread (animate).'

b. cân o-têm-a oskâtâskw-a
John.3SG 3.POSS-dog.NA-3′ carrot.NA-3″
mow-ê-yiwa
eat.VTA-THM.DIR-3′.SBJ.3″.OBJ
'John's (3SG) dog (3′) eats the carrot (animate, 3″).'[12]

(11) VTI

cân wiyâs mîci-w
John.3SG meat.NI eat.VTI-3SG
'John eats meat (inanimate).'

Alongside extensive person and direction morphology, several other categories may also be expressed on verbs.[13] Closely related to person is the relational suffix *-w-*, an inflectional category which indicates that an action is done in relation to another animate person but which does not increase the valency of the verb;[14] we have not yet completed the modelling of this relatively rare phenomenon. Preverbs (prefixes which add both lexical and grammatical information to the verb, but do not constitute person affixes) attach to the verb between person and the verb stem and serve several purposes expanded on below. There are two types of preverbs: grammatical and lexical. The outermost of grammatical preverbs include those such as the *ê-* form seen in Conjunct verbs above, as well as other Conjunct preverbs including *kâ-* and *ta-*. These serve as complementizers and may have further functions, such as marking future or relative clauses. Closer to the verbal stem, one can observe another type of

[11] While in many of the inverse forms (e.g. *niwâpamik*, 's/he sees me', the suffix appears to be a portmanteau morpheme, and could be analyzed as such, this suffix may be analyzed as *-ikw-w*, where the cluster is then simplified by regular phonological rules. Indeed, this form may also occur as *-ikow*, demonstrating another possible resolution of a *C-w-w* sequence (A. Wolvengrey p.c. 2016).

[12] As the marking for obviative and further obviative is formally the same, they must instead be distinguished on the basis of semantics and pragmatics.

[13] For a large (though not yet complete) overview of Plains Cree morphemes (including common preverbs) see Cook and Muehlbauer (2010).

[14] When in the relational form, VAIs and VTIs are not derived to VTAs, and the second animate participant is not marked as for VTAs. The relational only indicates that the VAI or VTI is performed while another animate person is present, or otherwise acknowledges the presence of an animate person that is not directly involved in the action (Cenerini 2014).

grammatical preverb for tense and aspect: *kî-* for past, *wî-* for intended future, and *ka-/ta-* for definite future. Closer still to the verb are lexical preverbs, e.g., *kakwê-* 'try (to)', *nihtâ-* 'be good at', *nitawi-* 'go and (do something)', *âpihtâ-* 'half (of)/halfway', *kihci-* 'large', etc. (Wolfart 1973, 1996; Wolvengrey 2001).

Furthermore, Cree also exhibits reduplication in the pre-stem position, which is not considered part of the class of preverbs, even though in some cases the resulting morphemes are identical to preverbs, especially grammatical ones (e.g. *ka-* and *kâh-* can be both reduplicative or grammatical preverbs). There are two reduplication templates which copy the initial consonant from the following morpheme: *Ca-* and *Câh-*. The former, known as light (or weak) reduplication, indicates an ongoing action, and the latter, heavy (or strong) reduplication, indicates a repeated action. Before vowels, the vowel is not copied but the morphemes surface as *ay-* and *âh-* respectively (though there appear to be some rare exceptions to this rule). Preverbs can be preceded by reduplication as well according to the same template, and both forms of reduplication can occur sequentially e.g. *Ca-Câh-* (Ahenakew and Wolfart 1983).

Though inflection is the primary focus of our computational model, there are two derivational processes that we also consider. The first of these is the associative or comitative category, which is expressed by means of a circumfix, the first component of which is the preverbal element *wîci-*, and the second component a suffix *-m* immediately after the verb stem. This expresses a joint action in VAIs, resulting in the derivation of a new VTA stem (e.g. *pîkiskwêw* 's/he speaks (VAI)' and *wîci-pîkiskwêmêw* 's/he speaks with someone (VTA)'). The second derivational process is verbal diminutivization *-si*, which can occur on verbs to denote a "smaller" or less intense version of the action. The diminutive forms can also be lexicalized with a conventionalized meaning, for example, *kimiwan* 'it is raining' is derived to *kimiwasin* 'it is drizzling', rather than 'it is a small rain' or 'it is raining a little'. For a word such as *kimiwasin*, two analyses would be offered by the analyzer, one which presents the derivation as a productive morphological process, and another which treats *kimiwasin* as lexicalized as its own lexical entry.

Moreover, diminutivization triggers a sound-symbolic phonological change on verbs, where /t/ (<t>) is changed to /tʃ/ (<c>). There is some uncertainty as to how far from the diminutive suffix palatalization will occur (e.g. when the VTI stem *itâtot-* 's/he tells s.t. thus' would be hypothetically diminutivized, there is variation or uncertainty among speakers as to what extent the palatalization will apply, with all three of the following as possibilities: *itâtoc-*, *itâcoc-*, or *icâcoc-* (A. Wolvengrey & J. Okimâsis, p.c. 2014, 2015)). Other common derivational processes, yet to be modelled, include the reflexive *-iso* and reciprocal *-ito* which derive VAIs from VTAs, and the benefactive *-stamaw*, which creates VTAs which take as their objects animate entities on whose behalf an action is performed. Nominalization, which commonly occurs with the suffix *-win* (though other nominalizing suffixes with various meanings, such as abstract concepts or tools, exist) is similarly not yet modelled, though to our knowledge, nearly every verb can undergo these derivations. Compound nouns, such as forms with a stem or root with a particle suffix *-i* followed by a free noun, are also possible, though not yet fully modelled (e.g. *kisiskâciwani-sîpiy* 'Saskatchewan River', from *kisiskâciwan* 'it flows swiftly' and *sîpiy* 'river') (Wolfart 1973, 1996; Wolvengrey 2011).

3 Theoretical modelling

As a polysynthetic language, Plains Cree exhibits considerable agglutinative verbal morphology in both derivation and inflection, which has been described in works such as Bloomfield (1946) and Wolfart (1973, 1996). Unsurprisingly then, there have been a number of attempts to theoretically model the morphology by means of verbal templates. For example, Bakker (2006) has proposed a verbal template based on a synthesis of previous attempts in the literature. His template proposes a verb comprised of 18 distinct cells. This treatment is maximally concatenative. As such, even the verbal stem is broken up in distinct units. As is common in the Algonquian literature, these units are described as *initials* or *roots*, *medials*, and *finals*. According to Goddard (1990), Algonquian initials represent actions or states of being, medials represent classificatory elements, and finals represent the manner by which an action proceeded. Goddard also placed valency-determining affixes as members of the final class (i.e. morphemes that determine the class (transitivity, animacy) of the verb) (1990).

3.1 Verbal template

According to Bakker, preceding a verb stem in the verbal complex is a series of seven prefixal cells. First in this series is the person marking (in the Independent mode) or Conjunct marking (in the Conjunct mode) slot (2006). This cell is followed by tense preverbs, followed by mood preverbs, and up to three aspectual preverb slots (with the final two aspect slots representing light and heavy reduplication, respectively). The final preverb slot represents *Aktionsart*, or lexical aspect.

Following the verb stem, Bakker (2006) proposes eight suffixal slots. Immediately after the stem final is a suffix marking obviation on the object of the transitive verb, followed by a suffixal slot for a theme marking direction (direction of the verbal action on the Algonquian hierarchy). After theme marking are the suffixal slots for valency and voice. Following voice, the Plains Cree verbal template contains a slot for marking obviation for the subject of an intransitive verb. This slot is followed by a slot for person marking, a slot for plurality of the absolutive,[15] and finally a slot for the conditional morpheme.

Similarly, Wolvengrey (2012) proposes a verbal template from a Functional Discourse Grammar perspective. An important difference between Bakker's and Wolvengrey's treatments of the verbal template is that Wolvengrey addresses the differences between the Independent and the Conjunct orders for preverbal ordering. These differences are essentially reflected in the Independent order's lack of subjective modality and relative tense preverbal slots (Wolvengrey 2012).[16]

[15]The use of *absolutive* in Algonquian languages is non-standard, and it is not immediately clear what Bakker's motivation in using this term is. As with other uses of the term, *absolutive* here refers to the object marking on a transitive verb or the subject marking on an intransitive verb.

[16]The subjective modality and relative tense slots are posited only for the Conjunct order. Subjective modality includes preverbs such as *ta-kî-* 'should, can, ought to' and relative tense is a separate slot for *kî-*, as tense marking on Conjunct verbs can be relative to previous tense marking on an Independent verb.

Bakker (2006)									
Person Marking (Independent)/ Conjunct Markers (Conjunct)	Tense	Mood	Aspect 1	Aspect 2	Aspect 3	Aktionsart	STEM (Initial)	STEM (Medial)	STEM (Final)
Wolvengrey (2012)									
Person Marking (Independent)/ Conjunct Markers (Conjunct)	Subjective Modality (Conjunct Only)	Absolute Tense	Relative Tense (Conjunct Only)	Perspectival Aspect	Participant-Oriented Modality	Phasal Aspect	Manner/Direction	STEM	

Fig. 1 Visualization of verbal prefix model (adapted from Bakker 2006; Wolvengrey 2012)

For Independent prefixes, Wolvengrey begins with person marking, followed by a slot for absolute tense, and next by a slot for perspectival aspect (such as the future intentional, which may be translated as 'going to', allowing for constructions such as *ê-kî-wî-mîcisoyân* meaning 'I was (in the past) going to (in the future relative to the past of the action, but not necessarily in the future of the utterance time) eat.'). This slot is subsequently followed by slots for participant-oriented modality (including preverbs that can be translated as 'want to'), then phasal aspect, and finally preverbs for manner/direction (including preverbs translatable as 'going to'). At this point, the verb stem and its component parts are given (a set of) slots. As discussed before, the Conjunct prefix order is essentially the same except that the person prefix is replaced by a Conjunct preverb; as well, subjective modality occurs after the Conjunct prefix and before absolute tense and a slot for relative tense occurs after that (and indicates that the time of a Conjunct verb action occurs relative to when the action of a preceding verb has taken place, rather than absolute to the time of speaking).

Following the stem slot(s), Wolvengrey (2012) proposes a set of three slots to represent valency: "thematic disjoint", represented by the suffix *-im*, which indicates that the persons involved in the action are further removed from each other on the person hierarchy (i.e., $1 > 3$ or $3 > 3'$ do not require disjoint suffixes as they are only one "level" apart in the hierarchy, while $1 > 3'$ or $3 > 3''$ are further removed from each other and so require a disjoint suffix); "theme signs", markers of direction across the Algonquian hierarchy; and "thematic obviation", a marker of obviation for one of the arguments as designated by the theme sign and person agreement. Following these valency slots, Wolvengrey designates slots for perfective aspect, followed by a slot to mark first and/or second person arguments, followed by two mood and tense slots, a slot for third person marking, a slot for marking third person plurality and obviation, and finally a slot for clausal and irrealis marking (roughly equivalent to Bakker's final slot of conditional marking).

These models for Plains Cree stand in contrast to models of related languages such as Ojibwe (Valentine 2001) which need to deal with further verbal complexity by marking for negation, prohibition, absence, and doubt. While Plains Cree previously marked verbs for such mood and/or aspect (A. Wolvengrey, p.c. 2015; Lacombe 1872), these categories are now obsolete, and their functions can instead be expressed by syntactic particles. Functionally, Bakker (2006)'s and Wolvengrey (2012)'s treatments of the verbal template are similar to each other, though Wolvengrey importantly specifies additional suffixal aspect, tense, and mood slots. A comparison of these models can be found in Figs. 1 and 2.

Bakker (2006)												
STEM (Initial)	STEM (Medial)	STEM (Final)	Possessed O	Theme	Valency	Voice	Posessed	Person	Plural Absolutive	Conditional		
Wolvengrey (2012)												
STEM			Thematic Disjoint	Theme	Thematic Obviative	Imperfective	Speech Act Participant	Tense/Mood	Tense/Mood	3rd Person	3PL / Obviative	Irrealis

Fig. 2 Visualization of verbal suffix model (adapted from Bakker 2006; Wolvengrey 2012)

3.2 Morpho-phonology

Following the morphological description above, Plains Cree morphophonology can be described. With ten consonants and seven vowels, the phonology of Plains Cree is relatively straightforward and there are very few exceptions to a small set of morphophonemic rules. For example, the person prefixes used in possession are as follows: *ni-*, *ki-*, and *o-* before consonants and *nit-*, *kit-*, and *ot-* before vowels. Other rules with quite general application include the contraction of both *Vyi* and *Vwi* sequences to *V:* across morpheme boundaries and the contraction of *wi* to *o* following a consonant. While the broad application of such rules is straightforward, there are a small number of exceptions that must be dealt with in the modelling. Alongside such general patterns, there are also rules specific to certain classes or subclasses of verbs. For example, in VIIs and some classes of VAIs, stem-final *an*, *in*, or *on* followed by suffixes beginning with *k* or *hk* results in the deletion of *n*.[17] For vowel-final VAIs, when the final vowel is *ê*, this becomes *â* before first and second person (singular or plural) Independent suffixes. All of these suffixes begin with *n*, allowing us to specify a clear phonological context for the change, but only within the VAI verbs. Conversely, for one VTI subclass, a verb-final *a* becomes *ê* in the same morphological and phonological context, before first and second person *n*-initial suffixes (Wolfart 1973, 1996).

Long-distance morphophonemic rules must also be modelled for Cree. For instance, the reduplicative prefixes take their initial consonants from that of the morpheme they precede; if this morpheme begins with a vowel, the reduplicative prefixes generally occur without a consonant and *Ca-* becomes *ay-* (*Câh-* simply becomes *âh-*).[18] Another rule that applies over a greater distance is the palatalization in diminutives, discussed above. Additionally, the palatalization can even occur to indicate a diminutive verb or noun without an overt suffix being present, which can be used for pragmatic reasons, such as a hypocoristic function (Wolfart 1973, p. 80).

However, not all morphophonemic rules apply as regularly as the above examples. In a subclass of VTAs, we can see stem-final /t/ becoming /s/ in a number of contexts, mainly when second person acts on a first person object, though the change occurs in a few instances when second person acts on a third person object. This change also occurs in other morphophonemic contexts. In indistinguishable contexts, we also see

[17] The change we have written here as *n-hk* > *hk* is in fact an example of a historical sound change rule where a nasal becomes /h/ before a stop. We present the change here in this format to indicate that this is how our model has dealt with the changes in the simplest set of rules. Similarly, many of our rules may appear odd or non-standard, but this is because they are representing how we have written them for the TWOLC formalism.

[18] Reduplication before *o/ô* may also occur with *w*.

t become /tʃ/ (represented with <c>). Historically, these changes were regular and predictable, though in modern Cree, the environments have been obscured by three mergers of sounds: /*θ/ and /*t/ merged to /t/, /*ʃ/ and /*s/ merged to /s/, and /*i/ and /*e/ merged to /i/. Historically, /*e/ caused no change to a preceding /*θ/ or /*t/ while /*i/ caused them to palatalize to /*ʃ/ and /*tʃ/ respectively. The mergers then resulted in (synchronically) unpredictable palatalization: sometimes *t* becomes *s* (as /*ʃ/ merged with /s/) when followed by *i*, and sometimes *t* becomes *c* but in other cases remains *t*, depending on the origins of both the *t* and the *i*. To account for these sorts of morphophonemic changes without treating each relevant lexical item as irregular, we could add further detail to our model to explicitly denote the historical phonological contexts in Plains Cree through special marking (e.g. /i/ derived from /*e/ might be coded as a special kind of /i/, e.g. <i2>, which is then taken into account in the contexts of morphophonological rules in our model and then realized as *i* once the relevant changes have taken place).

Example (12) shows the phonological rules we have found necessary for modelling Plains Cree (verbs). Classes of verbs which the rules apply to are given in parentheses (see Table 2 and Sect. 5.2 for more information on these verbal subclasses). However, many of these rules also have broader applications in Plains Cree and may be found in derivational contexts or nominal inflection as well.

(12)

a. $n \rightarrow \varnothing$ / V_+ (h)k (VII)

b. $\hat{e} \rightarrow \hat{a}$ / _+ n (VAIv)

c. $a \rightarrow \hat{e}$ / _+ n (VTI)

d. i. $V\begin{bmatrix} -\text{long} \end{bmatrix} \rightarrow V\begin{bmatrix} +\text{long} \end{bmatrix} / \_\begin{bmatrix} -\text{consonantal} \\ -\text{syllabic} \\ +\text{sonorant} \end{bmatrix} + \begin{bmatrix} \text{i}(<*\text{e}) \end{bmatrix}$ (VTAv)

ii. $\begin{bmatrix} -\text{consonantal} \\ -\text{syllabic} \\ +\text{sonorant} \end{bmatrix} + \begin{bmatrix} \text{i}<*\text{e} \end{bmatrix} \rightarrow \varnothing / V\begin{bmatrix} +\text{long} \end{bmatrix}\_$ (VAIv)

e. i. w → o / C_+ i (VTAc)

ii. i → ∅ / Co_+ (VTAc)

f. w → ∅ / C_# (VTAc)

g. t → s / _[i(<*i)] # (VTAt)

h. t → s / _# (VTAt)

i. Reduplication

i. $\varnothing \rightarrow C\begin{bmatrix} \alpha\ \text{place} \\ \alpha\ \text{manner} \\ \alpha\ \text{voice} \end{bmatrix} a / \#\_C\begin{bmatrix} \alpha\ \text{place} \\ \alpha\ \text{manner} \\ \alpha\ \text{voice} \end{bmatrix}$

ii. $\varnothing \rightarrow C\begin{bmatrix} \alpha\ \text{place} \\ \alpha\ \text{manner} \\ \alpha\ \text{voice} \end{bmatrix} \hat{a}h / \#\_C\begin{bmatrix} \alpha\ \text{place} \\ \alpha\ \text{manner} \\ \alpha\ \text{voice} \end{bmatrix}$

iii. ∅ → ay / #_V

iv. ∅ → âh / #_V

j. Comitative
   i. ∅ → i / C_m
   ii. ∅ → i / m_C
k. ∅ → h / # ê_+ V

Note that in (12d) and (12g), the notation *i(<*e)* or *i(<*i)* refers to situations where the following phoneme is historically derived from /*e/ or /*i/, respectively. As well, in typical descriptions of the sound change in (12e) this phenomenon would be achieved by a single rule e.g. *Cw-i > Co*. However, within the TWOLC formalism we employ in our computational model it is necessary to implement this as two interdependent rules, which is subsequently indicated here. For reduplication we give four rules: one for light reduplication in consonant initial stems (12i$_i$), one for heavy reduplication in consonant initial stems (12i$_{ii}$), one for light reduplication in vowel initial stems (12i$_{iii}$), and finally one for heavy reduplication in vowel initial stems (12i$_{iv}$). The rules in (12j) refer to rules for epenthesis required due to the comitative suffix possibly creating an illegal consonant cluster. Our final rule in (12j) is used for the alternative orthographical convention of placing an <h> after word initial <ê> in the Conjunct order, if not using a hyphen.

4 Computational modelling

While our computational model is based on the theoretical models described above, they differ in a few significant ways. For the purposes of creating our model of Plains Cree, we make use of a set of finite state transducers (FSTs) as described in e.g. Beesley and Karttunen (2003) and Karttunen (2003).

Interestingly, this latter source equates the two-level morphology used in FSTs to realizational morphology, as described in Stump (2001), providing a theoretical backdrop to the formalism.

Our development of a finite state model for Plains Cree makes use of the Xerox (XFST: Beesley and Karttunen 2003), Helsinki (HFST: Lindén et al. 2011) and FOMA (Hulden 2009) finite state compilers, as incorporated within the Giella infrastructure (cf. Trosterud 2006),[19] which allows for the rapid transformation and integration of a finite state computational model into linguistic modules providing (1) a spell-checking functionality as part of a word processor, (2) word-form recognition and analysis, and (3) word-form or paradigm generation as part of an Intelligent Electronic Dictionary or Intelligent Computer-Assisted Language Learning Application (cf. Arppe et al. 2015, 2016b).

All of the three above-mentioned finite state compilers (XFST, HFST, FOMA) make use of two formalisms: LEXC, which deals with morphological concatenation, and (with the exception of FOMA) TWOLC (TWO-Level Compiler), which deals with the morphophonological alternations. Within the LEXC formalism, we make use of flag diacritics to implement co-occurrence constraints for (typically discontinuous) morpheme combinations, as well as occasionally outputting a morphological tag at

[19] http://giellatekno.uit.no/index.eng.html.

some position other than where the corresponding morpheme/feature is actually observed or determined, in order to present morphological features in a standardized order.

Source files for our Plains Cree model can be found under the language code CRK in the Giella infrastructure.[20] This infrastructure is designed to keep plain text source code descriptions of morphological affixation and the lexicon/stems, as well as morphophonological processes, in separate files and directories for each major part of speech, which are then concatenated or composed together in the compilation of the finite state model.

4.1 Morphotactic modelling

Our model begins with verbal prefixes, having first a slot for person marking for Independent forms or the preverbs indicating any of the several Conjunct forms or Imperatives. Flag diacritics are used to signal which person prefix has been observed (including a null prefix in the Conjunct and third person forms). For Independent forms, these flag diacritics constrain the permissible corresponding suffixes, by requiring unification with a matching flag diacritic associated with an appropriate corresponding person/number suffix. Thus, we treat person marking as circumfixation. For instance, in VAIs the first person prefix *ni-* is only compatible with the first person singular suffix *-n* and first person plural exclusive suffix *-nân*, exemplified in ***ni****nipâ****n*** 'I sleep' and ***ni****nipâ****nân*** 'We (but not you) sleep', respectively. The aforementioned restrictions also preclude the grammatical Conjunct preverbs (e.g. *ê-*) from occurring if Independent prefixes/suffixes are present. In the second position, we have a single slot for marking both absolute and relative tense.[21]

After tense marking, we have a third slot for non-tense preverbs. Rather than separating preverbs into a sequence of various slots according to differences in modality and aspect, we simply divide preverbs into grammatical and lexical categories. Currently, lexical preverbs are allowed to be repeated without restriction, resulting in a cyclical slot. This allows, in principle, for accepting forms such as *ê-nôhtê-nôhtê-pê-nôhtê-nipayân*, 'I want to want to go and want to sleep' (a highly unlikely if not an entirely ungrammatical construction). This characteristic is not especially problematic for our purposes at this time. Our model is designed for maximal descriptive analysis, and lexical preverbs are generally not so short as to result in incorrect analyses based in this cyclicity (i.e., *nôhtê-nôhtê-* is not, in terms of edit distance, close to any more likely word form, whose misspelling would thus get incorrectly accepted). Furthermore, while theoretical models have been proposed for Plains Cree preverb combinatorics,[22] it remains an empirical question as to which preverbal sequences and combinations are likely and acceptable. This question may, in fact, be

[20] https://victorio.uit.no/langtech/trunk/langs/crk/src/.

[21] Our model does not yet allow relative tense to co-occur with absolute tense, though we intend to implement this in the near future.

[22] Wolvengrey (2015) has presented a *preliminary* corpus-based analysis of the Plains Cree preverbal template, though simply scrutinizing preverbal string sequences identifiable with the use of a separating hyphen, without the use of a grammatical analyzer as we will do later below.

partially answered through the use of a morphological analyzer based on a computational model such as ours, which allows for any combinations that may actually occur in naturally produced texts (cf. Schmirler et al. (to appear), for a preliminary analysis).

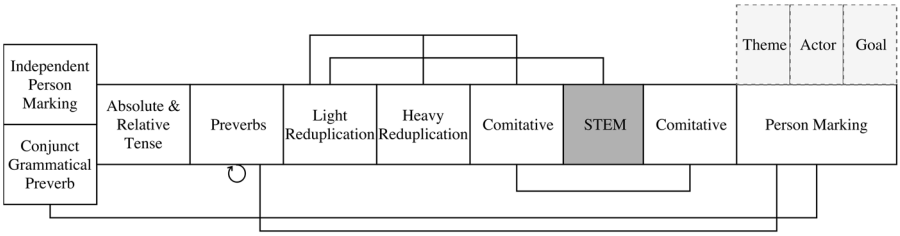
Following preverbs, we assign a fourth and a fifth slot for each of light and heavy reduplication (in that order). Currently, our model allows only for at most one light reduplicated morpheme, followed by one heavy reduplicated morpheme. Further, reduplication is based on the verb stem (or the comitative prefix), and if both light and heavy reduplication co-occur, they must occur adjacently. Thus, we would allow for light reduplication in *ê-na-nôcihikocik*, 'they are hunting for a while'; heavy reduplication in *ê-nâh-nôcihikocik*, 'they hunt habitually or repeatedly'; light reduplication followed by heavy reduplication in *ê-na-nâh-nôcihikocik*, 'they habitually or repeatedly hunt for a while'; but not heavy reduplication followed by light reduplication as in **ê-nâh-na-nôcihikocik* or with heavy reduplication on the *misi-* preverb as well as the verbal stem as in *ê-mâh-misi-nâh-nôcihikocik*, 'they repeatedly/habitually have very large hunts'. While these rules are appropriate for the first four examples (as **ê-nâh-na-nôcihikocik* is not in fact an acceptable construction), the final example was found naturally occurring in a corpus. As such, we should adapt our model to allow for reduplication on any lexical (but not grammatical) preverbs, besides the verb stem or the comitative prefix.

After reduplication, the sixth and final prefixal slot in our model allows for the prefixal component of the comitative circumfix, *wîci- -m*. As with person/Conjunct marking prefixes, using flag diacritics we require the *-m* suffix when the *wîci-* prefix is present. This process derives a VTA from a VAI or VTI, and the verb is then redirected to the relevant VTA suffixes in our affixation LEXC file. After the comitative, we have completed prefixation and move through verb stems toward suffixation. As with preverbs, the comitative can also be targeted for reduplication.[23]

Immediately following the verb stem we have a slot for the suffix component of the comitative circumfix, controlled by a flag diacritic. Unlike the previous theoretical/linguistic models discussed above, our computational model treats the subsequent suffixal morphology of Plains Cree as single chunks, associated with several morphological features, rather than maximally decomposed sequences of morphemes, such as treating each verbal suffix (e.g. the various valency, obviative, and person morphemes) separately with each their individual morphological features, and then combining them together, which would require the application of morphophonological rules at each morpheme juncture. For instance, in generating the word *mowikoyiwa* 's/he ($3''$) eats him/her (animate noun, $3'$)', our model adds the suffix chunk *-ikoyiwa* to the VTA stem *mow* rather than affixing the directional theme (*-iko-*), the obviative marker (*-yi-*), the third person marker (*-w*), and a final obviation marker (*-a*) serially.

A diagram exemplifying our model can be found in Fig. 3, while Fig. 4 details the computational flow of our FST (note that in Fig. 4 person, order, and tense morphemes are recognized in the prefix portion; however, the corresponding features are outputted only after the verb stem, resulting in an analysis such

[23]This comitative preverb *wîci-* is currently the only one for which we have allowed reduplication in our current model.



ni-	-kî-wî-	kakwê- ati- nitawi-	wa-	wâh-	wîci-	pîkiskwê	-m-	-imânâna
ê-	-kî-wî-	kakwê- ati- nitawi-	wa-	wâh-	wîci-	pîkiskwê	-m-	-imâyâhk

Independent nikî-wî-kakwê-ati-nâh-nitawi-wa-wîci-pîkiskwêmimânâna

Conjunct ê-kî-wî-kakwê-ati-nâh-nitawi-wa-wîci-pîkiskwêmimâyâhk

'We had intended to try to start repeatedly going to continuously speak with them (3’).'

Fig. 3 Our working Plains Cree verbal model with example complex verbs. (*Lines* connecting units represent dependency; *dashed boxes* above ‘Person Marking’ represent separate morphemes that we have chunked together)

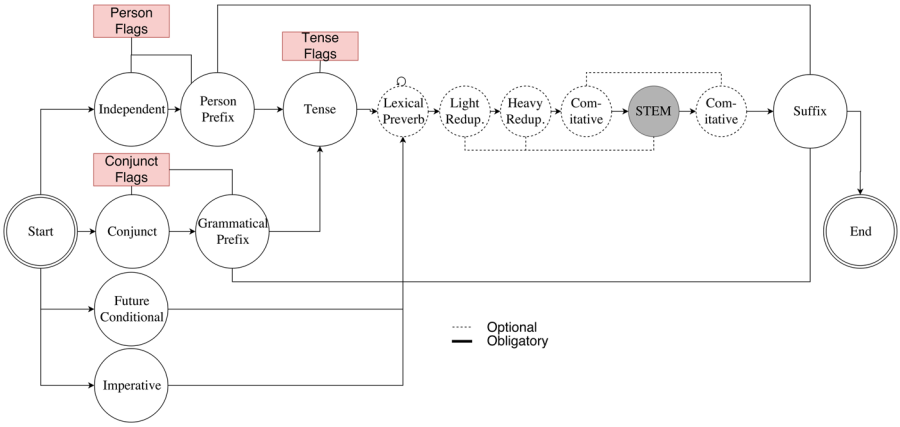


Fig. 4 Computational representation of our Plains Cree verbal model (*Lines* connecting units represent dependency; *rectangles* represent flag diacritics, *solids* represent obligatory units and *dashes* represent optional elements)

as *nipâw+Verb+Animate Intransitive+Independent+Past+1*SG *subject* for the string *nikî-nipan* ‘I slept’).

Our motivation for treating suffix sequences as chunks rather than splitting them into their individual constituent morphemes is due mainly in the interest of simplicity and having a ‘flat’, less hierarchical, model. The cost for this is that we need to repeat morpheme/feature pairings in all the chunks where they occur, even when the morphemes in question are affixed in an entirely agglutinative manner. However, in the case that such affixation is not entirely agglutinative, the reward for chunking is

that we have fewer morpheme junctures for which to consider morphophonological processes, since these are already 'preprocessed' within the chunks. Additionally, we do not necessarily lose information about complexity, as the number of morphemes in each suffix chunk can be determined, and the entire set of morphological features associated with such suffix chunks can be outputted when analyzing a word form. Although this computational approach is in principle un-parsimonious, this allows us to devise simpler morphological and phonological rule sets which are easier to write, understand, and debug. This is an extremely important aspect for the faster development of computational tools for endangered languages, as we have in practice observed this strategy to result in a substantially development time.

4.2 Inflectional subclasses

Our computational model of Plains Cree verbs currently incorporates all 10,365 verb-stems in the lexical database underlying *nêhiyawêwin: itwêwina / Cree: Words* by Wolvengrey (2001).[24] However, where Wolvengrey (2001) presents 12 total classes of verbs (two VII subclasses, two VAI subclasses, three VTI subclasses, and five VTA subclasses), our model makes use of 19 verb classes (six VII, four VAI, two VTI, seven VTA). The differences between our model and Wolvengrey's (2001) account are due mainly to issues of verbs that occur only in the singular or plural, and how each system treats them. In our model, plural-only and singular-only verbs require their own lexicon/class to restrict morphological number. In the lexical database underlying Wolvengrey (2001), these verbs simply contain a note in their entry to indicate such restrictions. Ignoring these subclasses, we have 11 total subclasses, though our breakdown still differs from Wolvengrey (2001) (with our model having two VII subclasses, two VAI subclasses, one VTI subclass,[25] and six VTA subclasses).

This difference in the number of verb classes is due to two main phenomena: first, Wolvengrey's (2001) final VTA subclass is essentially a class of miscellaneous irregular verbs which cannot be placed into a single, uniform "irregular" inflectional paradigm; second, of Wolvengrey's (2001) three VTI subclasses, two pattern identically the first class of VAI (which end in *-w* in the 3SG form), and so have been redirected to this lexicon/class in our model. Our model contains the same lexical items as Wolvengrey (2001); only our organization is different. Table 2 demonstrates the translation from Wolvengrey's verb classes to our own.

4.3 Morphophonological modelling

The TWOLC formalism is used to model morphophonemic changes once concatenation has occurred. For example, the person prefix variants with and without an epenthetic /t/ are achieved by designating a special character <t2> which is used in

[24] We intend to expand these with content from the *Maskwacîs Cree Dictionary* (Maskwachees Cultural College 1997) as well as the *Alberta Elders' Cree Dictionary* (LeClaire et al. 1998), but based on preliminary scrutiny we need to carefully review these resources in terms of the dialectal and areal variation we know them to contain and the consequences this may have on their orthographical consistency.

[25] Technically, we have three VTI subclasses, though two of these exist only to mark a stem as VTI before redirecting to VAI suffix morphology.

Table 2 Mapping of Wolvengrey's verb-stem categories to our verb categories

Wolvengrey's categories (# of stems)	Computational model categories
VII-v (826)	VIIw, VIIw-Sg, VIIw-Pl
VII-n (626)	VIIn, VIIn-Sg, VIIn-Pl
VAI-v (4474)	VAIv, VAIv-Pl
VAI-n (191)	VAIn, VAIn-Pl
VTI-1 (1468)	VTI, VTI-Pl
VTI-2 (360)	VAIv
VTI-3 (6)	VAIv
VTA-1 (1176)	VTA
VTA-2 (570)	VTAv, VTAv-Pl
VTA-3 (340)	VTAc
VTA-4 (324)	VTAt
VTA-5 (4)	VTAi, VTAti

the lexical entries for these prefixes, which may then do one of two things according to the TWOLC rules: before consonants, <t2> is deleted, and before vowels, it surfaces as <t>. Deletion is a common rule; for example, where *n*-final verb stems are followed by suffixes beginning with *-k* or *-hk* the <n> is deleted. This rule can be demonstrated through standard SPE phonological notation in (13) below. TWOLC rules work similarly enough, and can be seen in (14). These two examples show the similarity of the two formalisms.[26] For a full account of these rules as implemented according to the TWOLC formalism,[27] one can refer to our openly available source code accessible through the Giella infrastructure.[28]

(13) n → ∅ / V_+ (h)k

(14)
```
n:0 <=> [ a | o | i ] _ %>:0 (h) k
```

Above, we discussed issues regarding predictable historical morphophonemic changes that have resulted in apparently irregular forms in synchronic analyses. If we do not include the historical environments, this would require large scale coding to account for the irregularities and would result in a far more complex system. Instead, we have opted to include a 'trigger' (which always surfaces as nothing) on those suffixes we

[26] As a point of reference, the TWOLC statement in (14) below can be parsed as follows:

(a) `n:0` represents the deletion of an underlying *n* in the surface form.
(b) <=> represents that the environment always causes the change, while the change cannot happen without the environment to the right.
(c) [`a | o | i`] indicates that the *n* change occurs only when following *a, o,* or *i*
(d) _ `%>:0 (h) k` represents the context for *n* deletion: a morpheme boundary to the right (denoted by '>', which surfaces as nothing/is deleted; since '>' is a special character in the FST formalism (marking the end of a regular expression), it has to be "escaped" by the '%' character), followed by an optional *h*, followed by *k*.

[27] It may be the case that our rules can be simplified, which is a ongoing objective of ours.

[28] https://victorio.uit.no/langtech/trunk/langs/crk/src/phonology/crk-phon.twolc.

know begin with an *i* that is a reflex of /*i/ rather than /*e/; the trigger indicates that the suffix in question will trigger palatalization in the appropriate context, e.g. after the appropriate *t*. Similarly, special characters are needed to indicate which type of *t*, depending on its origin, occurs in these environments to indicate what the palatalized consonant will be: either *c* or *s*. A TWOLC rule can then target the appropriate sequence (in the above case, a *t* of a particular origin followed by a morpheme boundary which is in turn followed by the trigger) and resolves the morphophonemic change. This process can be seen in (15).[29]

(15) `t3:s <=> _ %>:0 %^TS:0 [ .#. | i ]`

Rules beyond core inflection can be simple, such as allowing the Conjunct prefix *ê-* to occur as *êh* before vowel-initial stems or preverbs (orthographic variation) or selecting the consonant from the following morpheme to complete the reduplicative prefixes. Two of the modelled rules also account for the comitative derivational circumfix *wîci- -m*; one rule inserts *i* before *m* if the preceding segment is a consonant and the second rule inserts *i* after *m* if the following segment is a consonant. Finally, two irregular VTA stems are dealt with using their own specific rules. The small amount of irregularity in Plains Cree allows for specific rules such as these without resulting in an unfeasibly large and complex model.

Reduplication is dealt with quite simply by LEXC specifying two special multi-character symbols `^DUPL1` and `^DUPL2`, which TWOLC rules align with the stem-initial phoneme ((16), (17), (18), and (19)). In the case of consonant-initial stems, these symbols are realized as that initial consonant followed by *a* or *â* (depending on which form of reduplication is happening, whereas in the case of vowel-initial stems these consonants become deleted by default, with light reduplication having an epenthetic *y*.

(16)
```
%^DUPL1:Cx <=> _ [ a y2: | â h ] ( %-: ) %>:0 Cy: ;
where

                Cx in ( c k m n p s t w y )
                Cy in ( c k m n p s t w y )
                matched ;
```

(17)
```
%^DUPL2:Cx <=> _ a y2: ( %-: ) %^DUPL1>:Cy ;
where

                Cx in ( c k m n p s t w y )
                Cy in ( c k m n p s t w y )
                matched ;
```

[29] The notation of (15) is the same as in (14), with the following exceptions:

(a) t3:s represents the palatalization of the underlying [t] as [s] in the surface form (where t3 is used only stem-final in the stems of the verb sub-class where the palatalization occurs).
(b) %^TS:0 represents the context of a suffix beginning with a trigger (to restrict the change to only specific instances within the paradigm), which surfaces as nothing/is deleted.
(c) [.#. | i] restricts the change to instances where, after the trigger, there is either a full word boundary (.#.) or <i>.

```
(18)  y2:y <=> %DUPL1:0 a _ ( %-: ) %>:0 Vow2: ;

(19)  y3:y <=> %DUPL2:0 a _ ( %-: ) %DUPL1:0 ;
```

4.4 Further issues in computational modelling

Although finite state morphology has its origins in the modelling of morphologically complex languages and has proven useful for Plains Cree, there are a number of considerations in making use of a computational model. Orthographic variation, for example, has presented one of the thorniest challenges for this computational model. While common spelling errors are easily dealt with, a Plains Cree model must be able to deal with considerable orthographic variation found throughout the various communities. Although a standard Plains Cree orthography exists (the Standard Roman Orthography (SRO)), the system is used loosely at best. This creates a situation wherein speakers of Plains Cree from different communities, despite speaking mutually intelligible dialects, are unable to communicate with each other easily through writing. These orthographic differences cause problems for our model. Some common issues are the indication of vowel length (some speakers prefer macrons or circumflexes, but many do not mark vowel length at all, e.g. *êwâpamât* vs. *ewapamat* 's/he sees him/her/them'),[30] the use of hyphens for separating preverbs (hyphens are standard, e.g. *ê-wâpamât* vs. *êwâpamât*), and even the writing of reduced forms of words (*tânisi* vs. *tân'si* vs. *tânsi* 'hello, how'). Much of this variation is easily dealt with a relaxation of spelling rules, wherein possible variants based on known orthographical variation are generated with a finite state transcriptor that is combined with the computational morphological model, and any match would be returned (we call this a *descriptive* model in contrast to a *normative* one, as the objective of the former model is to maximize the possibility of finding an analysis for a word form). As an example, we could enter into our descriptive analyzer the string *sakahikan*. This string would result in analyses for two lemmata: a minimal pair of *sakahikan* ('nail') and *sâkahikan* ('lake'). If there were any possibility for hyphenation, such as the non-word **sa-kahikan*, or for any combination of long/short vowels, each permutation would be tested for and, if possible, an analysis would be offered for each form. This permissibility is important for languages such as Plains Cree, where orthographic standards are not used consistently (due to community differences as well as difficulty in typing diacritic characters on a North American keyboard).

5 Evaluation of the computational model

5.1 Motivations for an evaluation (Gold Standard) corpus

A computational model is generally evaluated in terms of how many words forms can be analyzed in a corpus, which ideally would not have been used in the original development of the model. This evaluation is comprised of two measurements: how

[30] Cree vowel length is not simply a matter of duration, but includes contrast through quality (Muehlbauer 2012; Harrigan and Tucker 2015).

many forms are recognized and analyzed (recall), and what proportion of the analyses are actually correct (precision). For the forms which are not recognized/analyzed, we must determine whether the model is missing a morphological or morphophonological process, whether such a process is too restricted, or whether we are missing a root/lemma. For the forms that are recognized/analyzed, we want to assess that the analyses are correct, (or at least plausible if ignoring context-based disambiguation) among a set of possible ambiguous analyses for homographs. This (non-statistical) scrutiny of the accuracy of the analyses also allows us to identify accidental errors in the underlying morphological description (whether in morphotactics or morphophonology). The additional challenge concerning a computational model for a morphologically complex language such as Plains Cree is the extent of morphological productivity and the characteristics of morphemes and their combinatorics (in particular with new word formation through mostly compounding, but also derivation), which can result in semantically impossible, or very unlikely, analyses for truly misspelled or even foreign words. For this reason, we do not merely want a list of words that should be recognized (recall), but also a set of correct linguistic analyses (precision).

For majority languages with huge corpora, we can implement a *Gold Standard* corpus, which is hand verified for a representative sample (however we want to define 'representative'). This Gold Standard can be used as a benchmark which a model must at least meet, and then the remaining texts can be used as test data for the model. For a language such as Plains Cree, available corpora are so small that we can, and need to, use them in their entirety as a reference corpus, as we want to capture as much of the possible morphotactics and morphophonology as possible, and cannot afford to keep much aside for testing.

The creation of a Gold Standard also allows us to assess whether existing morphological or morphophonological descriptions are complete, or if they are potentially incorrect or lacking concerning some process. Further, a manually verified Gold Standard allows for follow-up development of our basic computational morphological model by providing reliable training data that will allow use to weight the computational model. This weighting can be indispensable in ranking analysis results if we increase productivity to increase recall, and can compensate for the increasing amount of less likely analyses by ranking analyses in terms of their likelihood (cf. Arppe et al. to appear).

5.2 Composition of the Gold Standard corpus

The corpus used for developing the Gold Standard Corpus was made up of roughly one hundred thousand tokens[31] of (relatively) contemporary Plains Cree (collected during the 1970s–1990s), collectively known as the Ahenakew-Wolfart texts (the editors of the works).[32] The texts consist of interview transcriptions, wherein the elicitor (Dr. Freda Ahenakew) spoke with participants in Plains Cree and elicited

[31]This figure excludes punctuation tokens, but includes Arabic and Roman numerals, foreign words, and proper names; including punctuation raises the count to approximately 142,000.

[32]Upon request, this corpus is available for research purposes on the corpus server maintained by the Alberta Language Technology Lab (ALTLab) at: http://altlab.ualberta.ca/korp/.

mainly narrative speech. These recordings resulted in a number of publications, including Ahenakew (2000), Bear et al. (1992), Kâ-Nîpitêhtêw (1998), Masuskapoe (2010), Minde (1997), Vandall and Douquette (1987), and Whitecalf (1993). These texts were turned into word form type list (ordered in terms of decreasing token frequency), which had been analyzed with an earlier version of the computational model (in January 2016). This sorted word-list was divided into four roughly equal parts so that each part had an overlap of 100 forms in the beginning and 100 forms in the end, in order to assess inter-annotator agreement.

Three quarters of the corpus were annotated by the second author of this paper, while the remaining quarter was done by the first. Neither annotator is a native speaker of Plains Cree, though both have had several years of university instruction in the language and considerable experience working with the language in multiple communities. The annotators frequently conferred with each other regarding the process and concerning ambiguous cases. Modifications to the analyses were made as the Gold Standard was being annotated, as we learned more about the language.[33]

In order to assess the degree of agreement between the two annotators and the reliability of the analyses, we focused on the first two overlapping sections, for which we had two analyses for each of the 200 word form types from both two annotators; of these, 124 could be recognized as Cree word form types.[34] Since the frequencies of the word form types in the overlapping sections are low (only 1 or 2 tokens), the calculation of inter-annotator agreement statistics would not have been practically meaningful; however, we could assess the proportions of *observed (raw) agreement* in analyses (Artstein 2017) as well as types of divergence. After a first-pass analysis, the two annotators were in full agreement for 105 (84.7%) of the 124 word form analyses. Out of the remaining 19 word form types (15.3%), in 10 cases (6.4%), only one of the annotators had been unable to reliably analyze a word form type, due to uncertainty about some aspect of Cree morphology; in 5 cases (4.0%), both annotators had one analysis in common, but one of the annotators also presented an alternative analysis (due to possible ambiguity) which the other did not provide; and in 4 cases (3.2%) the analyses of the two annotators differed with respect to some morphological feature, due to either error or a different initial interpretation of the analysis scheme. Following this first-pass analysis, the annotators scrutinized and discussed together these 19 cases of non-agreement, and were able to quickly reach a full consensus for the analysis of all word form types in question.

5.3 Evaluation results for the performance of the computational model

The Gold Standard corpus represents an ideal version of how the computational model should function for Plains Cree and is improved as our knowledge of Plains Cree morphosyntax increases. In this section, we evaluate the coverage of the computational model compared to the hand-annotated version.

[33] We were also given a wealth of knowledge directly from last author, Professor of Algonquian Languages and Linguistics at the First Nations University of Canada and author of *nêhiyawêwin: itwêwina / Cree: Words* (2001).

[34] The 76 other word form types were English words or proper names (English or French), Arabic numerals, or non-linguistic codes/abbreviations.

The Gold Standard version of the Ahenakew-Wolfart texts consists of 108,413 tokens representing 18,649 types, of which 75,232 tokens (17,394 types) were actual Plains Cree words.[35] Of these, yet another 14,476 tokens (2,700 types) had some recognizable orthographical error, 1,340 tokens (738 types) were word fragments (typically sequences of verb-initial preverbs lacking a stem and suffixes), and 2,467 tokens (1,398 types) of Plains Cree words which we were unable to accurately analyze. This left 56,964 tokens (12,572 types) of correctly spelled Plains Cree words (according to SRO). Of these, 46,693 tokens (10,863 types) had a single, unambiguous analysis, while 10,271 tokens (1,709 types) remained ambiguous with two or more possible analyses. While our longer term intention is to reduce such ambiguity through context-based disambiguation using the Constraint Grammar formalism (Bick and Didriksen 2015; Bick 2011; Karlsson 1990; Karlsson et al. 1995), for the moment we have applied some simpler heuristics to this end. Thus, we selected any Plains Cree word form for which at least one analysis designated it as a verb (16,605 tokens; 9,999 types), and if there remained multiple analyses (as was the case for 1,343 types), we then chose the one which had the smallest number of preverbs (which for a not insignificant number of cases are similar to light or heavy reduplication morphemes). This left us with 9,983 Plains Cree verb form types having altogether 16,538 tokens, which are the focus of the subsequent analyses and model evaluation presented in this article. In our observations on morphological complexity, we use only Independent or Conjunct forms, which reduce the number of forms slightly to 15,842 tokens representing 9,673 types.

For the 8,731 Plains Cree word types which received at least one verbal analysis from the computational model (representing 14,937 word form tokens), 7,813 types had one or more of the verb analyses which matched one of the hand-validated analyses provided in the Gold Standard, indicating a verb form type recall of 78.3% (7,813/9,983) and a precision of 89.5% (7,813/8,731). Taking into account the frequencies of these word form types (representing altogether 13,535 tokens with one or more matching analyses in the Gold Standard), the performance values were somewhat higher, with a verb form token recognition recall of 81.8% (13,535/16,538) and a precision of 90.6% (13,535/14,937). Among the unrecognized verb form types that we were able to diagnose, 20.6% were due to a verb stems not yet included in our lexicon (as they were not yet known to us), 35.4% because of a morphological form not yet implemented in the computational model, and 4.3% due to both a missing stem and a missing morphological form.[36] The remainder (39.7%) were word forms that can be recognized as Plains Cree verbs, based on their prefixes and suffixes, but which we were unable to reliably fully analyze because no translation was available to confirm verb class information.

[35]The rest include 25,240 punctuation tokens (32 types), 980 tokens (12 types) of Arabic or Roman numerals, 5,915 tokens (906 types) of English, French, or Latin words, 619 tokens (202 types) non-Cree proper names.

[36]Missing morphology generally refers to minor inflectional patterns, frequently obsolete, that are not thoroughly described in the literature or well-represented in the corpus. These include initial change (a conjunct-marking phenomenon), rare inverse forms, and inanimate subject paradigms (e.g. Wolfart 1973). Other morphological issues include verbal derivation such as those processes discussed above.

5.4 Comparisons with other relevant computational modeling word

The best points of comparison concerning the performance of our computational model for Plains Cree verbs would firstly be those for other morphologically complex, less resourced, languages with nevertheless relatively good dictionaries and morphological descriptions, and with at least a certain amount of text corpora, and secondly morphologically complex majority languages, with presumably the best possible resources for model development and testing. With these criteria in mind, we found descriptions of models for Odawa, Arapaho, Nahuatl, Quechua and North Saami in the first category, and for Finnish in the second category.

Bowers et al. (2017) used the finite state formalism to model Odawa, an Algonquian language related to Plains Cree. This Odawa model was tested on a corpus containing 7,578 tokens of 2,685 types. From this corpus, 85% of types were recognized; of the non-recognized forms, 30% of them were due to orthographic errors within the corpus, 28% were due to morphological entries missing in the model, 18% were due to phonological processes not being modelled, 5% were due to dialectical difference, and a final 2% were due to other unspecified errors (Bowers et al. 2017). Kazeminejad and Huldén (2017, 15) similarly present an FST based parser for Arapaho (another Algonquian language), reporting a 98.2% recall (based on paradigm tables, rather than corpora). Arppe et al. (2017) detail the development of an XFST and TWOLC based analyzer for East Cree, though no coverage statistics are presented. An FST-based parser for North Saami (a morphologically rich, Indigenous Uralic language spoken in Northern Scandinavia) was reported to have a 91.4% recognition rate (Johnson et al. 2013). Lindén and Pirinen (2009) present a Finnish FST-based parser which boasts a 99.9% recall and precision rate. Although other attempts have been made in creating linguistic analyzers for under-resourced, morphologically complex languages, e.g. Nahuatl (Martínez-Gil et al. (2012) and Gutierrez-Vasques et al. (2016)), these are often incomplete or used as demonstration cases, or, as in the case of Quechua (Rios 2016), lack a systematic evaluation of performance. In this company, our Plains Cree model can be judged to perform relatively well.

6 Other corpus-based observations of interest

Our Gold Standard can be thought of as an ideal analysis that our computational model, once completed, would provide for each legitimate Plains Cree word. Thus, in the following section we present an investigation of a Plains Cree corpus with hand verification in the style of our computational model. We have focused on morphological complexity because Plains Cree is a polysynthetic language (which are often under-represented in the literature), and so it is empirically interesting to scrutinize *how* polysynthetic Plains Cree word forms actually are. Furthermore, Plains Cree morphology has a number long-distance morphological interrelationships and it is interesting to observe how many parallel dependencies native speakers actually have to deal with. Finally, having an empirically observed understanding of such morphological complexity can be used to inform our computational modeling work.

6.1 Morphological complexity in principle

Both the theoretical models of Plains Cree verb structure as well as our computational model presented above would in principle allow for quite complex forms, if a verb were to include a morpheme in every possible slot and contain multiple parallel long-distance morpheme co-occurrence dependencies/constraints. Examples of theoretically possible forms where every permissible slot in the computational model is used are presented in (20), consisting of four or five morphemes (counting circumfixes as single morphemes and excluding morphemes denoting absolute tense). One should note that the maximal set of slots varies based on the conjugation class of the verb: the comitative/associative forms are only possible for VAI and VTI verbs, whereas direct or inverse directionality is only relevant for VTA verbs with two animate participants. If the verb form had more than one lexical preverb, the maximal number of morphemes could be even greater.[37]

(20) a. ki-wî-nôhtê-wa-wâh-wîci-mîciso-m-âw
2-FUT.INT-want-RDPLL-RDPLH-with-eat.VAI-COM-2SG.SBJ.3SG.OBJ
cî?
Q
'Will you want to be repeatedly eating with him/her?'

b. âhaw,
Okay
ni-wî-nôhtê-wa-wâh-wîci-mîcisô-m-âw
1-FUT.INT-want-RDPLL-RDPLH-with-eat.VAI-COM-2SG.SBJ.3SG.OBJ
'Okay, I will want to be repeatedly eating with him/her'

c. ê-kî-nôhtê-wa-wâh-wâpam-iyan cî?
CNJ-PST-want-RDPLL-RDPLH-see.VTA-2SG.SBJ.1SG.OBJ Q
'Did you want to repeatedly be seeing me?'

d. nâmoya
NEG
ê-kî-nôhtê-wa-wâh-wâpam-it-ân
CNJ-PST-want-RDPLL-RDPLH-see.VTA-INVERSE-1SG.SBJ.2SG.OBJ
'I did not want to repeatedly be seeing you'

These forms also exhibit the maximum number of long-distance morpheme relationships/constraints that a speaker of the language would have to keep track of to produce a correct form. This maximum number in our models above is four, though again their composition varies based on the conjugation class and order of the verb. To recap, these dependencies are the following:

(21) a. i. Independent forms: prefix and suffix components of the circumfix indicating subject/object person/number; OR

[37] Wolvengrey (2015) presents an artificial complex verb form with incorporating a maximal amount of elements, i.e. *nikî-kakwâhyaki-nôhtê-pê-mâci-nipahi-kakwê-mâh-misi-miyo-kitohcikân* 'I had desperately wanted to come start really trying to play music extremely well.'

ii. Conjunct forms: Conjunct preverb and the suffixes indicating person and number for subject and object;

b. Initial phonemes of the light reduplication morpheme and stem;

c. Initial phonemes of the heavy reduplication morpheme and stem;

d. The prefix and suffix components of the comitative/associative circumfix (possible for VAI verbs but not VII, VTI, or VTA verbs)

In addition to the above, one could also consider the *inverse* morpheme as an additional, fifth category of a long-distance relationship, as it forces the reanalysis of the subject vs. object reference of the initial prefix denoting person/number for independent order VTA forms. Beyond these dependency types noted above, our model does not currently presume nor specify interrelationships (i.e. co-occurrence preferences or dispreferences) among lexical preverbs, or between lexical preverbs and the verb stems (or derivational morphemes within the stem). However, there are reasonable grounds to assume that there exist pragmatically-determined, more and less realistic/likely sequences and scopes of modality reflected in preverb order, as well as constraints on the semantic compatibility between lexical preverbs and verb stems.

6.2 Morphological complexity in practice

We can now use our Gold Standard corpus to empirically observe how much of the reported types of morphological complexity, both in terms of long-distance dependencies and simple morpheme sequence lengths, that native speakers normally make use of and have to cognitively deal with in producing correct verb forms.[38] Table 3 presents an overview of the various non-stem morphological elements among the 15,528 verb form tokens representing 9,397 verb form types. As can be seen, a clear majority of verb forms, 75.6% of the tokens and 66.8% of the types, consist of only the person/number marking and grammatical preverbs expressing Conjunct order or absolute tense.

Next, we can scrutinize the extent of co-occurrence of these morpheme categories, both in terms of morpheme length as well as multiple parallel dependencies in verb forms. Interestingly, there are no verb forms in the Gold Standard with the maximal morpheme complexity (with a morpheme from every type described above) that the prior theoretical models and our computational model would allow for. Table 4 lists the morpheme counts according to the design principles of our analyzer, which takes a chunking approach to morphological breakdown. We can see when we count these morphemes decompositionally that we get up to seven morphemes in total, including preverbs, direction, and person-marking suffixes, namely: *ê-kî-pê-isi-ka-kîmôcihitân* 'I had been coming thus to sneak around on/conceal it from you' with Conjunct + 3 x Preverb + Reduplication + Inverse + Person, *kâ-mosci-ka-kitâpamicik* 'that s/he merely looks at me thus' with Conjunct + 2 x Preverb + Reduplication + Inverse + Person, and *ê-kî-wî-kakwê-miy-ôsîhtâcik* 'they had intended to try to make something well' with the Conjunct + 4 x Preverb + Person. Crucially, the clear majority of verb

[38] Underlying details on the individual types of morphological structure for the verbs studied can be found in the Supplementary materials for this issue, or on the authors' website at http://altlab.artsrn.ualberta.ca/.

Table 3 Overall type and token frequencies of various inflectional morpheme categories in independent and conjunct verb forms

Tokens	%	Types	%	Morpheme
10968	69.2	7275	75.2	CNJ (suffix)
10841	68.4	7208	74.5	PV-C:1
5207	32.9	3548	36.7	PV-G:1
2241	14.1	1950	20.2	PV-L:1
2741	17.3	1217	12.6	IND3 (suffix)
2133	13.5	1181	12.2	IND12 (circumfix)
948	6.0	643	6.6	INV
352	2.2	331	3.4	RDPLH:1
348	2.2	302	3.1	RDPLL:1
161	1.0	154	1.6	PV-L:2
29	0.2	28	0.3	PV-H:1
17	0.1	17	0.2	PV-G:2
9	0.1	8.0	0.1	COM
5	0.0	5.0	0.1	PV-L:3

Table 4 Type and token frequencies of inflectional morpheme sequence length

# of Morphemes	Tokens	%	Types	%
1	1555	9.8	396	4.1
2	7660	48.4	4423	45.7
3	5178	32.7	3594	37.2
4	1220	7.7	1042	10.8
5	207	1.3	197	2.0
6	21	0.1	20	0.2
7	1	0.0	1	0.0

forms (6,279 types representing 11,744 tokens) have just one inflectional morpheme of the categories considered here.[39]

With respect to the number of long-distance dependencies that may occur in verb forms, their type and token frequencies are presented in Table 5. Similar to morpheme complexity, we did not observe any verb form in the Gold Standard that exhibited the maximal number of four (or five) long-distance dependencies (as discussed above), though there were 32 types representing 35 tokens with three parallel long-distance dependencies. For example, in *ê-na-nâh-nâkatohkêyân* 'I habitually/really notice people', the dependencies here are between the Conjunct *ê-* and the Conjunct

[39] As there are a number of common grammatical preverbs that can be formally identical to reduplication in particular contexts, e.g. *ka-* 'future/optative: will, shall, ought, should' and *kâh-* 'would, ought to; likely to', and this results in some structural ambiguity in the Gold Standard analyses leading to a certain degree of uncertainty about the numbers and analyses presented here, as deciding which of the theoretically equally possible analyses would require context-based disambiguation or even more detailed scrutiny of the semantic and pragmatic context, which we have not yet had the time to pursue. Since this is key to forming the most accurate picture, we are planning to improve the Gold Standard in this respect in the near future.

Table 5 Type and token frequencies of long-distance dependencies in verb forms in the gold standard

# of Dependencies	Tokens	%	Types	%
0	2594	16.4	1093	11.3
1	11772	74.3	7451	77.0
2	1442	9.1	1098	11.4
3	34	0.2	31	0.3

suffix *-yân*, and between the two reduplicative prefixes and the stem, as they take their initial consonants from the stem. Again, the large majority of verb forms (13,961 tokens representing 8,182 types) incorporated only one long-distance dependency.

7 Conclusion

The complex morphology of Plains Cree has presented both straightforward and challenging phenomena for computational modelling. Though our Plains Cree modelling is ongoing, steps such as understanding historical phonology and hand-verifying a section of the corpus have brought us closer to a model with which we can more thoroughly and accurately examine our corpus and other Plains Cree texts. These steps have also given us a basis for derivational and syntactic modelling. As our model improves, we also have the opportunity to create and improve tools for linguists, teachers, and learners of Cree such as spelling and grammar-checkers, online dictionaries that offer paradigms and recordings of words spoken in isolation or within sentential contexts, and other applications to facilitate language maintenance and learning in communities and classrooms.

Our hand-verified Gold Standard corpus, wherein we have manually implemented the morphological and morphophonemic elements dealt with in the computational model, has allowed us to evaluate not only our model but the underlying theoretical morphological and morphophonemic treatments of Plains Cree as well. Theoretical descriptive models of polysynthetic languages in general, and of Plains Cree in particular, give the impression that verbs in these languages can be extremely complex in terms of number of morphemes and parallel long-distance dependencies. In contrast, our observations from the Gold Standard corpus have shown that there are practical limits to how much of such possible complexity is actually realized on an individual word basis. We observed no case of maximal theoretical complexity, and for the most part (in terms of tokens in running text) Plains Cree verbs exhibit only one long-distance morpheme dependency. From the perspective of the practical utility of prior linguistic descriptions, verb classes and their paradigms prepared by Wolvengrey have proven to work well within our model, allowing us to implement morphological patterns and morphophonemic changes with respect to the classes they affect. However, the Gold Standard corpus has shown that subclasses do not always behave as theoretically expected: for instance, some verbs are expected to always occur as singular or plural on the basis of semantics, yet language use in texts has shown that this need not be the case. Descriptions of elements such as reduplication (Ahenakew and Wolfart 1983) and the comitative (Wolvengrey 2011) allow us to understand the

application of such features as they have been analyzed in descriptions, while the corpus gives us the opportunity to see how they are used in speech and how that may differ from theoretical descriptions.

Works such as Wolvengrey (2011) and Wolfart (1973, 1996) are excellent resources not only for our morphological modelling, but also offer descriptions of the phonological alternations seen in those morphological patterns. We have drawn our TWOLC rules from morphophonemic rules described in all of these sources, taking into account differences across subclasses, and have found these rules to adequately account for much of Plains Cree morphophonology. With the ongoing addition of historical elements to our rules, we hope to simplify our TWOLC rules even more. Many of the orthographical errors encountered in the Gold Standard were the result of a single error, not in rule implementation, but in data extraction, and many more were the result of vowel sandhi (Russell 2008),[40] which is beyond the word-internal morphophonological processes implemented in the model described herein.

With further research into the available descriptive and theoretical works on Plains Cree and the errors encountered in the Gold Standard, we can continue to improve the accuracy of our inflectional model in the analysis of Plains Cree words. Alongside the current model, we can expand to a computational model of the derivation of Plains Cree, to allow for a better analysis of lemmata found in texts that do not appear the lexicon we make use of (cf. Arppe et al. to appear), as well as develop a syntactic analyzer, to allow for investigations of word order and other syntactic functions (e.g Schmirler et al. 2017). A well-structured Plains Cree model can also be adapted to closely-related Cree dialects, with straightforward phonological correspondences and relatively few morphological and lexical differences. Still being developed, these tools represent a significant step forward in the description, documentation, and sustainability of Plains Cree.

Acknowledgements We would like to thank the two anonymous reviewers and the guest editors of this issue for their helpful and insightful comments, as well as Dustin Bowers for his comments and suggestions on an earlier version of this paper. Moreover, we appreciate the comments, feedback and suggestions we received during the Workshop on Computational Methods for Descriptive and Theoretical Morphology organized by Olivier Bonami and Benoît Sagot at the 17th International Morphology Meeting.

This research was made possible by a SSHRC Partnership Development Grant (# 890-2013-0047), a SSHRC Joseph Armand Bombardier Canada Graduate Scholarship (Master's), a Kule Institute for Advanced Study (KIAS) Research Cluster Grant, and a Killam Cornerstones Grant (University of Alberta).

References

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. In H. C. Wolfart & F. Ahenakew (Eds.), *Publications of the Algonquian Text Society Winnipeg*. Winnipeg: University of Manitoba Press.

Ahenakew, F., & Wolfart, H. C. (1983). Productive reduplication in Plains Cree. In *Actes du quatorzième congrès des algonquinistes [Québec, 1982]* (pp. 369–377).

[40] Vowel sandhi refers to a series of phonological changes that occur when a vowel-final word or preverb is adjacent to a vowel-initial word or preverb. Typically, the first vowel is deleted and the second, if short, is lengthened. However, some vowel pairs result in different vowels arising. For example, *kâ-itwêt* 's/he says' is often realized as *k-êtwêt* in the corpus.

Arppe, A., Antonsen, L., Trosterud, T., Moshagen, S., Thunder, D., Snoek, C., Mills, T., Järvikivi, J., & Lachler, J. (2015). *Turning language documentation into reader's and writer's software tools*. Paper presented at 4th International Conference on Language Documentation and Conservation (ICDLC), Honolulu, HI, 26 February–1 March 2015.

Arppe, A., Harrigan, A., & Schmirler, K. (2016a). *So similar in principle, but so different in practice–mixing texts, elicitation and experimentation in the study of the Plains Cree independent and conjunct verb constructions*. Paper presented at the 2nd new ways of analyzing syntactic variation, Ghent, Belgium: Universiteit Ghent.

Arppe, A., Lachler, J., Trosterud, T., Antonsen, L., & Moshagen, S. N. (2016b). Basic language resource kits for endangered languages: A case study of Plains Cree. In C. Soria, L. Pretorius, T. Declerck, J. Mariani, K. Scannell, & E. Wandl-Vogt (Eds.), *CCURL 2016—Collaboration and computing for under-resourced languages – towards an alliance for digital language diversity (LREC 2016 workshop)*, Portorož, Slovenia, 23 May 2016. European Language Resource Association.

Arppe, A., Junker, M. O., & Torkornoo, D. (2017). Converting a comprehensive lexical database into a computational model: The case of east cree verb inflection. In *Proceedings of the 2nd workshop on the use of computational methods in the study of endangered languages, association for computational linguistics* (pp. 52–56). http://aclweb.org/anthology/W/W17/W17-0101.pdf.

Arppe, A., Schmirler, K., Silfverberg, M., Huldén, M., & Wolvengrey, A. (to appear). *Insights from computational modeling of the derivational structure of Plains Cree*. Papers of the 48th Algonquian Conference.

Artstein, R. (2017). Inter-annotator agreement. In N. Ide & J. Pustejovsky (Eds.), *Handbook of linguistic annotation* (pp. 297–313). Netherlands, Dordrecht: Springer. doi:10.1007/978-94-024-0881-2_11, https://doi.org/10.1007/978-94-024-0881-2_11.

Bakker, P. (2006). Algonquian verb structure: Plains Cree. In G. J. Rowicka & E. B. Carlin (Eds.), *What's in a verb? Studies in the verbal morphology of the languages of the Americas* (Vol. 5, pp. 3–27). Uthrecht: LOT, Netherlands Graduate School of Linguistics.

Bear, G., Fraser, M., Calliou, I., Wells, M., Lafond, A., & Longneck, R. (1992). *Kôhkominawak otâcimowiniwâwa/Our grandmothers' lives: As told in their own words*, edited by F. Ahenakew and H. C. Wolfart, Vol. 3. University of Regina Press

Beesley, K. R., & Karttunen, L. (2003). *Finite state morphology*. Center for the Study of Language and Information.

Bick, E. (2011). A barebones constraint grammar. In *Proceedings of the 25th Pacific Asia conference on language, information and computation*, Singapore, December 16–18, 2011 (pp. 226–235).

Bick, E., & Didriksen, T. (2015). Cg-3—beyond classical constraint grammar. In *Proceedings of the 20th Nordic conference of computational linguistics*, NODALIDA 2015, Vilnius, Lithuania, May 11–13, 2015 (Vol. 109, pp. 31–39). Linköping: Linköping University Electronic Press, Linköpings universitet.

Bloomfield, L. (1946). Algonquian. In *Viking fund publications in anthropology: Vol. 6. Linguistic structures of native America*, New York (pp. 85–129).

Bowers, D., Arppe, A., Lachler, J., Moshagen, S. N., & Trosterud, T. (2017). A morphological parser for Odawa. In *Proceedings of the 2nd workshop on the use of computational methods in the study of endangered languages, Association for Computational Linguistics*, Honolulu, Hawaii, USA (pp. 1–9). http://aclweb.org/anthology/W/W17/W17-0101.pdf.

Cenerini, C. A. M. (2014). *Relational verbs: Paradigm and practice in a manitoba dialect of Swampy Cree*. PhD thesis, Faculty of Graduate Studies and Research, University of Regina.

Cook, C. (2014). *The clause-typing system of plains cree: Indexicality, anaphoricity, and contrast*. Oxford; New York: Oxford University Press.

Cook, C., & Muehlbauer, J. (2010). A morpheme index of Plains Cree. https://www.academia.edu/304874/A_morpheme_index_of_Plains_Cree.

Dahlstrom, A. (2014). *Plains cree morphosyntax*. Routledge Library Editions: Linguistics, Routledge.

Ethnologue (2016). Cree, plains. http://www.ethnologue.com/language/crk.

Goddard, I. (1990). Primary and secondary stem derivation in Algonquian. *International Journal of American Linguistics*, *56*(4), 449–483.

Gutierrez-Vasques, X., Sierra, G., & Pompa, I. H. (2016). Axolotl: a web accessible parallel corpus for Spanish-Nahuatl. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the tenth international Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France: European Language Resources Association (ELRA).

Harrigan, A. G., & Arppe, A. (2015). *oswêw êkwa ê-tipiskâk, mâka mâcîw âhpô ê-mâcît?* Paper presented at the 47th algonquian conference, Winnipeg, Canada. University of Manitoba.

Harrigan, A., & Tucker, B. (2015). Vowel spaces and reduction in Plains Cree. *Journal of the Canadian Acoustics Association*, *43*(3), 124–125.

Hulden, M. (2009). Foma: a finite-state compiler and library. In *Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics* (pp. 29–32). Association for Computational Linguistics.

Johnson, R., Antonsen, L., & Trosterud, T. (2013). Using finite state transducers for making efficient reading comprehension dictionaries. In S. Oepen, K. Hagen, & J. B. Johannessen (Eds.), *NEALT proceedings series: Vol. 16. Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)* (pp. 59–71). Linköping, Sweden: Linköping University Electronic Press.

Kâ-Nîpitêhtêw, J. (1998). *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 unrestricted text. In H. Karlgren (Ed.), *Proceedings of the 13th international conference of computational linguistics* (Vol. 3, pp. 168–173). Stroudsburg, PA, USA: Association for Computational Linguistics.

Karlsson, F., Voutilainen, A., Heikkilä, J., & Anttila, A. (Eds.) (1995). *Natural language processing: Vol. 4. Constraint grammar: a language-independent system for parsing unrestricted text*. Berlin: de Gruyter.

Karttunen, L. (2003). Computing with realizational morphology. In A. Gelbukh (Ed.), *Computational linguistics and intelligent text processing: 4th international conference, CICLing 2003 Mexico City, Mexico, February 16–22, 2003 proceedings* (pp. 203–214). Berlin: Springer.

Kazeminejad, G., & Huldén, M. (2017). Creating lexical resources for polysynthetic languages—the case of Arapaho. In *Proceedings of the 2nd workshop on the use of computational methods in the study of endangered languages, Association for Computational linguistics*, Honolulu, Hawaii, USA (pp. 10–18). http://aclweb.org/anthology/W/W17/W17-0102.pdf.

Lacombe, A. (1872). *Dictionnaire et grammaire de la langue crise*. Montreal: Beauchemin and Valois.

LeClaire Nancy, Cardinal, G., Hunter, E., & Waugh, H.E. (1998). *Alberta Elders' Cree Dictionary =: Alperta ohci kehtehayak nehiyaw otwestamâkewasinahikan*. Edmonton: University of Alberta Press.

Lewis, P. M., & Simons, G. F. (2012). Assessing Endangerment: Expanding Fishman's GIDS. *Revue roumaine de linguistique Editura Academiei: Bucuresti*, *55*(2), 103–120.

Lindén, K., & Pirinen, T. (2009). Weighted finite-state morphological analysis of Finnish compounding with hfst-lexc. In K. Jokinen & E. Beck (Eds.), *Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009)* (Vol. 4, pp. 89–95). Northern European Association for Language Technology.

Lindén, K., Axelson, E., Hardwick, S., Silfverberg, M., & Pirinen, T. (2011). HFST-framework for compiling and applying morphologies. In *Proceedings of second international workshop on Systems and Frameworks for Computational Morphology (SFCM)* (pp. 37–85).

Martínez-Gil, C., Zempoalteca-Pérez, A., Soancatl-Aguilar, V., de Jesús Estudillo-Ayala, M., Lara-Ramírez, J. E., & Alcántara-Santiago, S. (2012). Computer systems for analysis of Nahuatl. *Research in Computing Science*, *47*, 11–16. http://aclweb.org/anthology/W/W17/W17-0102.pdf.

Maskwachees Cultural College (1997). *Nehiyaw Pîkiskwewinisa*. Maskwachees Cultural College.

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

Minde, E. (1997). *kwayask ê-kî-pê-kiskinowâpahtihicik/Their example showed me the way*. F. Ahenakew & H. W. Edmonton (Eds.). University of Alberta Press.

Muehlbauer, J. (2012). Vowel spaces in Plains Cree. *Journal of the International Phonetic Association*, *42*(1), 91–105.

Okimâsis, J. L. (2004). *Cree, language of the Plains = nêhiyawêwin, paskwâwi-pîkiskwêwin. University of Regina publications: Vol. 13*. Regina: Canadian Plains Research Center.

Ratt, S. (2016). *Mâci-nêhiyawêwin = Beginning Cree*. Regina: University of Regina Press.

Rios, A. (2016). A basic language technology toolkit for Quechua. *Procesamiento de Lenguaje Natural*, *56*, 91–94.

Russell, K. (2008). Sandhi in Plains Cree. *Journal of Phonetics*, *36*(3), 450–464.

Schmirler, K., Arppe, A., Trosterud, T., & Antonsen, L. (2017). *Computational modelling of Plains Cree syntax: A Constraint Grammar approach to verbs and arguments in a plains cree corpus*. Paper presented at the 49th algonquian conference, Montreal, QC.

Schmirler, K., Harrigan, A. G., Arppe, A., & Wolvengrey, A. (to appear). *Plains Cree verbal derivational morphology: A corpus investigation*. Papers of the 48th Algonquian Conference.

Snoek, C., Thunder, D., Loo, K., Arppe, A., Lachler, J., Moshagen, S., & Trosterud, T. (2014). Modeling the noun morphology of Plains Cree. In *Proceedings of the 2014 workshop on the use of computational methods in the study of endangered languages, Association for Computational Linguistics* (pp. 34–42).

Statistics Canada (2015). Population with an aboriginal mother tongue by language family, main languages within these families and their main provincial and territorial concentrations, Canada, 2011. https://www12.statcan.gc.ca/census-recensement/2011/as-sa/98-314-x/2011003/tbl/tbl3_3-1-eng.cfm.

Stump, G. T. (2001). *Inflectional morphology: A theory of paradigm structure. no*. *Cambridge studies in linguistics: Vol. 93*. Cambridge: Cambridge University Press.

Trosterud, T. (2006). Grammatically based language technology for minority languages. In A. Saxena & L. Borin (Eds.), *Lesser known languages of South Asia* (pp. 293–316). Hague: de Gruyter.

Valentine, R. (2001). *Nishnaabemwin reference grammar*. Toronto: University of Toronto Press.

Vandall, P., & Douquette, J. (1987). *wâskahikaniwiyiniw-âcimowina/Stories of the house people*. F. Ahenakew (Ed.). University of Manitoba Press.

Whitecalf, S. (1993). *kinêhiyâwiwininaw nêhiyawêwin/The Cree language is our identity: The La Ronge lectures of Sarah Whitecalf*. Publications of the algonquian text society/collection de la société d'édition des textes algonquiennes Winnipeg. University of Manitoba Press. Edited and translated by H. C. Wolfart and F. Ahenakew.

Wolfart, H. C. (1973). *Plains Cree: A grammatical study*. Transactions of the American Philosophical Society: new ser., v. 63, pt. 5, Philadelphia, American Philosophical Society, 1973.

Wolfart, H. C. (1996). Sketch of Cree, an Algonquian Language. In *Languages: Vol. 17. Handbook of American Indians* (pp. 390–439). Washington: Smithsonian Institute.

Wolvengrey, A. (2001). *nêhiyawêwin itwêwina = Cree: Words*, bilingual edition edn. Regina: University of Regina Press.

Wolvengrey, A. (2011). *Semantic and pragmatic functions in Plains Cree syntax*. PhD thesis, University of Amsterdam.

Wolvengrey, A. (2012). *The verbal morphosyntax of Aspect-Tense-Modality in dialects of Cree*. Paper presented at the 2nd international conference on functional discourse grammar, Ghent, Belgium: Universiteit Ghent.

Wolvengrey, A. (2015). *Preverb combinations, co-occurrences and sequences: Preliminary findings from a preliminary Plains Cree corpus*. Paper presented at the 2nd prairie workshop on language and linguistics.