Insights from Computational Modeling of the Derivational Structure of Plains Cree stems

Antti Arppe, Katherine Schmirler, Miikka Silfverberg, Mans Hulden, Arok Wolvengrey

The complex inflectional and derivational morphology of Plains Cree and other Algonquian languages has long been considered from both a synchronic and diachronic perspective (e.g. Bloomfield 1946; Goddard 1974; Oxford 2014). While the composition of some modern Plains Cree stems has been obscured by sound change, they can often still be identified by linguists, and for speakers, many morphemes are available to freely derive new stems. Unlike derivational morphology, the inflectional morphology of Cree is quite regular and lends itself to straightforward description and this has translated to a computational model that can analyze inflected forms of Plains Cree lemmata (e.g. Harrigan et al. forthcoming; Snoek et al. 2014). Though the derivational morphology poses more challenges to model, lists of existing derivational morphemes can be extracted from existing sources and various morphophonological rules have been described (Cook and Muehlbauer 2010; Wolfart 1996; Wolvengrey 2001). However, we can make use of the derivational model to assess how well the rules and morphemes given for Plains Cree apply when tested against lemmas included in available dictionaries. This approach, following Karttunen (2006), allows us to test theoretical descriptions against larger data sets than those used to produce the rules: where the human mind can only make sense of so much data at once, a quantitative approach can take thousands of words into account.

In this article, we present the first version of a computational model for Plains Cree derivational morphology, using a weighted finite-state transducer, and discuss its

development, testing, strengths, and shortcomings. Our model is constructed using the morphological breakdowns of the stems given by Wolvengrey (2001) and documented morphophonological rules for Plains Cree (e.g. Wolfart 1996). While the concatenation of existing morphemes is straightforward, and most of the relevant morphophonological changes have been documented, the efficacy of the model is hindered by several factors, such as idiosyncrasies due to obscured sound changes, borrowings, and under-documentation of morphemes and morphophonological rules. We consider the extent to which documented rules and morphemes are sufficient for the computational modeling of Plains Cree derivation.

## BACKGROUND

### *Plains Cree*

Plains Cree is an Algonquian language spoken in Western Canada with several thousands of speakers. The language is still learned by children in many communities, and it is used in many everyday contexts, such as in homes, in television and radio broadcasts, and in books and other written materials. Plains Cree is a member of the Cree-Montagnais-Naskapi dialect continuum; various language revitalization efforts have been undertaken for members of this continuum, such as textbooks or grammars (e.g. Ahenakew 1987; Okimâsis 2004; Wolfart 1973, 1996 for Plains Cree; Ellis 2000 for Swampy and Moose Cree), children's books (e.g. Ahenakew 1988; Lavallee and Silverthorne 2014), and morphological analyzers (e.g. Harrigan et al. forthcoming; Snoek et al. 2014). While existing computational analyzers focus primarily on inflectional morphology, Algonquian languages also display considerable derivational morphology.

Cognate derivational morphemes are apparent across a number of Algonquian languages and many of these are still productive in modern Algonquian languages. In the following subsection, we look at nominal and verbal derivational elements and derivational processes in Plains Cree.

*Derivation*

Derivation in Cree makes use of three types of morphemes: roots or initials, medials, and finals. All stems in Cree will contain at least one root and one final, though phonetically null finals have been posited to maintain this structure; medials are generally optional in the derivation process. Roots are the initial elements of stems and carry considerable semantic content, but are generally free to occur across stem classes (nouns, verbs, and particles), as in (1) (Wolfart 1973:65-6). However, where nouns are formed through a process called primary derivation (see below), they generally occur with roots specific to nouns, rather than those that can occur across word classes (Bloomfield 1946); see examples in (2).[1]

(1) *Root morphemes across stem classes* (Wolvengrey 2001)

a. /wâp-/ 'light, bright'

| | | | |
|---|---|---|---|
| i. | wâpi- | 'see, have vision' | VAI |
| ii. | wâpahta- | 'see s.t.' | VTI |
| iii. | wâpam- | 'see s.o.' | VTA |
| iv. | wâpastim | 'white dog/horse' | NA |

b. /âw-/ 'carry'

| | | | |
|---|---|---|---|
| i. | âwacikan | 'wheelbarrow' | NI |
| ii. | âwacikê- | 'haul things' | VAI |
| iii. | âwah- | 'haul s.o.' | VTA |

c. /âyît-/ 'firm, tight'

| | | | |
|---|---|---|---|
| i. | âyîci- | 'firmly, tightly' | preverb |
| ii. | âyîtina- | 'hold firmly to s.t.' | VTI |

(2) *Nominal stems with zero finals* (Wolvengrey 2001)[2]

a. atim
atimw-
'dog'

b. maskwa
maskw-
'bear'

Medials occur between root and final morphemes, though they are not required in derivation. Like general roots, their occurrence is relatively unrestricted across stem classes. They also tend to have fairly concrete meanings. They may also be derived from other stem classes, such as in forms in (3). Dependent nouns (inalienably possessed body parts, kin terms) are considered medials that occur with zero roots, though body part medials may also occur in verbs, as in (4). Many medials fall into the subclass of classificatory medials, such as those in (5), which serve to denote not a stem class, but a semantic class. Finally, many medials have shorter and longer variants; the latter are known as extended medials (Wolfart 1973:66-8).[3]

(3) *Roots ~ derived medials*

a. /atimw-/ ~ /-astimw-/ 'dog'
b. /masin-/-ah/- ~ /-asinah-/ 'mark, write'
c. /pâhpih-/ ~ /-âhpih-/ 'laugh at s.o.'

(4) *Body part medials*

a. ni**hcikwân** 'my knee' ~ kaski**hcikwân**êhw- 'break s.o.'s knee by shot'
b. ni**cihciy** 'my hand' ~ saki**cihc**ên- 'seize s.o. by the hand'

(5) *Classificatory medials*

a. /-âpisk(w)-/ 'stone, metal' > pîw**âpiskw**- 'piece of metal' (NA)
b. /-âpêk-/ 'rope' > it**âpêk**in- 'hold s.o. thus on a rope' (VTA)

Final morphemes are required for derivation, though they are often phonetically null. Some finals determine the stem class: noun, verb, or particle, and, within verbs, determine the subclass (transitivity and animacy) as well. They are often either more

concrete or more abstract; concrete finals carry easily identifiable semantic information along with stem class information, while abstract finals contain only stem class information (Wolfart 1973:68-75). Examples of nominal and verbal finals, both with more concrete and more abstract semantics, can be seen in (6) and (7) respectively.

(6) *Nominal finals*
  a. *Concrete*
    i. /-âhtikw/ ‘tree’: ayôhkwanâhtikw- ‘raspberry bush’ (NA)
    ii. /-âpoy/ ‘liquid’: maskihkîwâpoy ‘tea’ (NI)

  b. *Abstract*
    i. /-win/ abstract noun: âcimowin ‘story’ (NI)
    ii. /-n/ instrument, product: kistikân ‘grain, wheat’ (NA)

(7) *Verbal finals*
  a. *Concrete*
    i. /-isw/ ‘by heat’: kîsisw- ‘cook s.o.’ (VTA)
    ii. /-(i)n/ ‘by hand’: itin- ‘move s.o. thus by hand’ (VTA)

  b. *Abstract*
    i. /-(i)kê/ general object: nôcihcikê- ‘hunt things’ (VAI)
    ii. /-h/ causative: cîsih- ‘mislead s.o.’ (VTA)

Medials and finals are suffixed to roots in primary derivation, to form primary stems. Primary stems may then undergo further derivation with an optional medial and required final, forming a secondary stem. Secondary derivation may occur several times, forming quite morphologically complex stems, as in (8).

(8) *Complex derived forms*
  a. *Verb*
    ayamihcikêstamâso-

| /ayam-/ | /-i/ | /-htâ/ | /-ikê/ | /-stamaw/ | /-iso/ |
|---|---|---|---|---|---|
| ROOT | FINAL | FINAL | FINAL | FINAL | FINAL |

    ‘read to oneself’

  b. *Noun*
    matwêkahikêwin

| /matwê-/ | /-ikah/ | /-ikê/ | /-win/ |
|---|---|---|---|
| ROOT | FINAL | FINAL | FINAL |

    ‘sound of chopping heard in the distance’

Alongside stem derivation, forms can also be derived through compounding. Noun compounds contain a noun with a prefixed particle or other noun, while verb compounds contain a preverb and a verb stem. Within noun compounds, the prefixed noun generally takes the suffix *-i*, formally identical to the particle final; the same form may occur as a particle, a prenoun, and a preverb (Bloomfield 1946; Wolfart 1973:75-8). Examples of compounds are given in (9).

(9) *Compounds*

a. *Nominal*

| paskwâwi-mostos | |
|---|---|
| paskwâw | mostos |
| prairie | cow |
| 'buffalo' | |

b. *Verbal*

| kâmwâci-pimâtisi- | |
|---|---|
| kâmwâci | pimatisi- |
| quiet | live |
| 'live quietly' | |

In creating a derivational model for Plains Cree, we must recognize both stem derivation and compounding processes, as well as any derivation within the initial members of compounds or unfamiliar preverbs. Below we offer a brief description a finite state transducer and how this has been applied to the inflectional morphology of Plains Cree and the recognition rate of the inflectional analyzer.

## THE DERIVATIONAL MODEL

As the formalism in our computational modeling of Plains Cree derivational word formation, we make use of finite state transducer (FST) tools as described in e.g. Beesley and Karttunen (2003), and in particular the Helsinki Finite-State Transducer (HFST)

software technology suite (Lindén et al. 2011), since it allows for the weighting of the model, the details and benefits of which we will discuss further below. The HFST compiler provides two sub-formalisms which we will use: (1) LEXC, which allows us to specify how morphemes are concatenated, and (2) XFSCRIPT regular expressions, which enable us to define a cascade of ordered SPE-style rewrite rules for implementing the various morphophonological processes, typically occurring at morpheme junctures, but possibly across an entire stem (such as palatalization in conjunction with the diminutive morphemes *-is*, *-isis*, and *-si*). Previously, we have also made use of FST technology to build a computational inflectional model for Plains Cree (e.g. Harrigan et al. forthcoming; Snoek et al. 2014), which analyzes verbal and nominal inflection, such as person, number, tense, possession, etc.[4]

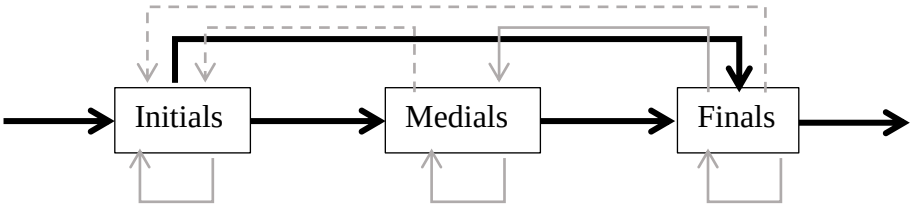
*Constructing the Model*

To analyze the derivational elements of Plains Cree stems, we first determined which morphemes are analyzed in Plains Cree stems. These were drawn from the database underlying Wolvengrey's (2001) Plains Cree dictionary, which contains a morphological breakdown of each recorded stem, noting the overt[5] roots or initials, medials, and finals for each. While there is some homophony in medials and finals, the total number of individual morphemes in the database is 2550. Of these, there are 1784 roots, 308 medials, and 547 finals, which are coded appropriately in a morpheme lexicon to which the FST can refer.

The morphological model to concatenate these morphemes is extremely simple. Using the LEXC formalism, we describe the concatenation of initial, medial, and final

morphemes in various combinations. In **Error! Reference source not found.**, the black arrows indicate the simplest paths of concatenation, initial+(medial)+final. The grey arrows indicate paths of possible recursion.[6]

FIGURE 1
The general derivational model




As presented, this model clearly does not accurately represent Plains Cree stem derivation: we have allowed for infinite initials, medials, and finals in any order, while in practice Cree demonstrates [[[ initial + (medial) + final ] + (medial) + final ] … ]. If we wanted to generate Plains Cree stems, this model would generate far more impossible stems than possible ones. However, the practical goal of such a model is to analyze existing stem forms, both known and unknown, and so the resulting analyses become problematic only in the case of homophonous medials and finals. Furthermore, this design has two main advantages. One, by allowing recursion in the medials and finals, we avoid the need for zero morphemes in the model. Two, if we allow for recursion from medials or finals back to initials, we can allow for compounding of preverbal or prenominal elements with the particle final *-i* followed by the initial morpheme of another stem: e.g. [[ root/stem + *-i* ] + [ root + (medial) + final]]. Moreover, we expect that the weighting of this simple morpheme concatenation model will be able to order the potentially large numbers of resultant analyses so that most likely ones will be ranked first.

Alongside the morpheme concatenation model, we must also implement the relevant morphophonological rules. Since many of the morphophonological elements are no longer productive, we must both make reference to obsolete sounds and allow for the rules to be optional. As noted above for the morpheme concatenation, this does not accurately represent the morphophonology of Cree stems and is not restrictive for the analysis of unknown stems. However, this method is more likely to produce at least one analysis for any given stem, which can then be confirmed as a likely analysis by researchers or speakers when combined with a translation or other contextual information. For the ordered list of rules in (10), we have specified a morphophonological component in the computational derivational model, representing the resulting possible changes, using the XFSCRIPT regular expression formalism for implementing SPE-style rewrite rules.

(10) Morphophonological rules[7]

a. t → c / _ $ [ is | isis | si ] $ (nominal and verbal diminutives)
b. ê → â / $ _ $ n $
c. Palatalization rules
    i. T → t | s / _ $ i | î
    ii. t → t | c / _ $ i | î
    iii. T | t → t / _ $ *e
d. *e → i
e. Ø → i / C $ _ C[8]
f. Vowel-glide-*i* rules
    i. [ a | â ] [ w | y | ý[9] ] $ i → â
    ii. [ ê [ w | y | ý ] $ i → ê
    iii. [ i | î ] [ w | y | ý ] $ [ i | î ] → î
    iv. [ o | ô ] [ w | y | ý ] $ i → ô
g. V-V rules
    i. [ a | â ] $ [ a | â ] → â
    ii. [ ê ] $ [ ê ] → ê
    iii. [ i | î ] $ [ i | î ] → î
    iv. [ o | ô ] $ [ o | ô ] → ô
h. w → o / _ $ i
i. w $ o → [o | ô ] / C _

j. w $ w → [ w | ow] / C _ V
k. [ m | n ] $ → h / _ [ p | t | k | c ]

These rules are ordered to allow for the best possible recognition rates at this time, though further development may result in changes to their order. Many of these rules are straightforward, though some require further comment. First, note the use of <T> and <e>, which refer to the Proto-Algonquian reconstructions *$\theta$ and **e* (which have fallen together with *t* and *i* respectively). While these sounds no longer occur in Plains Cree, they have left their marks in palatalization rules and are coded where possible in the lexicon.[10] However, they do not consistently palatalize as expected, and so the rules must still be optional. Second, we have sets of rules that can easily be summarized into single rules, but for the sake of the model must be written more specifically. These include the vowel-glide-*i* rules, which we must specify for each vowel, but can be stated as simply as to "any vowel followed by a glide and short *i* will collapse to a long vowel of the same quality." Similarly, when vowels of the same quality meet at a morpheme boundary, regardless of length, they become a single long vowel; again, we must write four separate rules for each vowel quality. Further development may allow for many of these rules to be streamlined, as well as for the addition of rules which are known but not yet implemented, such as the hiatus of vowels of different qualities and morpheme-specific rules.

*Training the Model*

The combination of the morpheme concatenation and morphophonological rule components described above provides us with our basic computational derivational

model. Since there are few restrictions on how the multiple morphemes may combine, and there are several single-character morphemes which can combine quite freely with the rest, for almost any stem this results in a large number of structurally possible but for the most part pragmatically improbable, if not entirely impossible, analyses. For instance, the VTA verb *nîhciwêpin-* 's/he throws s.o. down, off' is given altogether six derivational analyses, presented in (11), of which the fourth (11d) /nîht-/wêp-/-n/[11] is the correct one ('down', 'throw', 'by hand (VTA)').[12]

(11) Unweighted analyses

| | | |
|---|---|---|
| a. | /niy-/iht-/wêp-/-n/ | 0.000000[13] |
| b. | /niy-/iht-/wêp-/-i/-n/ | 0.000000 |
| c. | /niy-/iht-/wêp-/-in/ | 0.000000 |
| **d.** | **/nîht-/wêp-/-n/** | **0.000000** |
| e. | /nîht-/wêp-/-i/-n/ | 0.000000 |
| f. | /nîht-/wêp-/-in/ | 0.000000 |

For the non-compound stems in Wolvengrey (2001), the basic derivational analyzer can provide for a single form anywhere from one up to as many as 51,092 alternative analyses, with a median number of analyses being 217. To impose some order into this thicket, weighting the morpheme transitions in the computational derivational model will allow us to rank the multitude of structurally possible analyses, so that the most likely ones will be provided first. We make use of the substantial number of already attested forms and their derivational decompositions in Wolvengrey (2001), starting with the 11,614 non-compounding stems in that resource, to determine which morpheme sequences are most likely. In order to create such a weighted version of the basic derivational FST model (e.g. Mohri 1997; Pirinen 2014), we learn the transducer weights from the aforementioned list of string pairs, consisting of (a) realized stems and (b) corresponding derivational breakdowns, using a simple procedure: we (1) traverse the

states and transitions of the non-weighted FST using each string pair in turn, while keeping track of the number of times each transition was used, (2) normalize these counts (after *add-1 smoothing*) of outgoing transitions in each state to a proper probability distribution, and (3) convert the probabilities to penalty weights, which are the negated logarithms of the overall probabilities of the derivational analyses for a stem.[14] The resulting weighted derivational model can then rank the analyses for e.g. *nîhciwêpin-* (11) as shown in (12). The smaller the weight is, the more likely the analysis is considered. Now, the correct, expected analysis, /nîht-/wêp-/-n/, receives the smallest weight of 14.08, and is thus also the best-ranked one.[15]

(12) Weighted analyses

| | | |
|---|---|---|
| **a.** | **/nîht-/wêp-/-n/** | **14.082328** |
| b. | nîht-/wêp-/-in/ | 16.578869 |
| c. | /nîht-/wêp-/-i/-n/ | 20.568705 |
| d. | /niy-/iht-/wêp-/-n/ | 34.791088 |
| e. | /niy-/iht-/wêp-/-in/ | 38.455750 |
| f. | /niy-/iht-/wêp-/-i/-n/ | 38.871815 |

*Testing the Model*

We next tested the weighted derivational model using the same set of 11,614 non-compounding stems and their derivational analyses from Wolvengrey (2001) which had been used to train the model. In terms of assessing the performance quantitatively, we first observed how many stems received a correct derivational analysis, or none at all, and second, and even more importantly, how the correct derivational analyses were ranked in terms of their weights.

Taking into account documented morphemes and the majority of documented phonological rules, this current derivational model was able to provide the correct

morphological decomposition, among often a plethora of more or less likely analyses, for 80.9% (n = 9,392) of the 11,614 stems analyzed in the database underlying Wolvengrey (2001). Focusing on these 9,392 stems receiving the correct derivational decomposition, for 76.7% this was the top-most ranked analysis, and for 96.2% among the top four ranked analyses (Table 1). Moreover, the poorest ranking for any correct analysis was 43rd, and the median ranking is one.

TABLE 1
Proportions of rankings for the correct derivational decompositional analysis among all analyses for stems in the evaluation (=training) data.

| RANK(S) | PROPORTION (%) | CUMULATIVE (%) | COUNT |
|---|---|---|---|
| 1 | 76.7 | 76.7 | 7205 |
| 2 | 13.8 | 90.5 | 1296 |
| 3 | 4.0 | 94.6 | 378 |
| 4 | 2.0 | 96.5 | 185 |
| 5 | 0.9 | 97.4 | 83 |
| 6 | 0.5 | 97.9 | 43 |
| 7 | 0.4 | 98.3 | 35 |
| 8 | 0.4 | 98.7 | 38 |
| 9 | 0.2 | 98.8 | 18 |
| 10 | 0.2 | 99.1 | 23 |
| 11-43 | 0.6 | 100.0 | 88 |

For the 2,222 stems (19.1%) that did not receive a single correct analysis corresponding with the one provided by a linguist, the breakdowns that the weighted derivational model nevertheless produces allow us to determine where further modifications and extensions to the morpheme set and the morphophonological rules may be needed to improve the model's performance with respect to the training data, and by extension, unknown stems in general. Of course, when applying this model to unknown stems, the recognition rate would drop considerably.

## DISCUSSION

Where this computational derivational model works, it works very well and would be an excellent tool in determining potential compositions of an unknown stem. However, this model still only holds for four-fifths of the data on which it was trained, and so further development is necessary.

Various issues that have resulted in non-recognition can be identified in both the model and the test data. In the model, two main issues are apparent. First, the required morphemes may not be represented in our morpheme lexicon, and so will not be found in test analyses. Second, and perhaps foremost, our morphophonological rules are not yet exhaustive and display some issues with respect to rule ordering that affects their application. Unlike inflectional analysis, the morphophonological rules for derivational morphemes also do not apply as regularly, and several possibilities may present themselves where any two sounds meet. For example, …*Cw*-*w*… (where the hyphen represents a morpheme break) may become either *Cow* or *Cw*, and there is no wider context phonological that influences which surface form occurs. Similarly, where two vowels meet, various changes may take place based on the qualities and quantities of the vowels in question. While many of these rules are documented (e.g. Wolfart 1973, 1996), they still do not necessarily apply categorically and so further scrutiny of the data and rules is required.

Such morphophonological inconsistencies are often the cause of identifiable issues in our results, namely, that some forms are simply synchronically idiosyncratic. While historically we may be able to identify why a stem has a certain shape, we are not able to identify these contexts from synchronic stems. For example, the stem *apîst-* 'to sit near something' is composed of *api-* and *-st*, with no synchronic motivation for the

lengthening of the vowel; this is simply a fact of the morpheme *-st* (Wolfart 1973:74-5). However, as such morpheme-specific rules are not yet implemented, this stem would not be recognized. Idiosyncracies involving particular sounds also occur; for instance, palatalization rules in the literature generally refer to *i*, but we see instances of *t>s* before *o*, or of palatalization not occurring where the context indicates we would expect it. Historical motivations are sometimes identifiable on a case-by-case basis, which could in principle be modeled with morpheme-specific or sequence-specific morphophonological rules, but this would come with the cost of losing generality of application beyond the current morpheme set and training data, and would in practice run the risk of quickly inflating the rule set so that it could become unwieldy to maintain. For example, we can specify that our *Cw-w>Cow* rule always applies for the sequence *-amw-win*, but can vary with the *Cw-w>Cw* rule in other contexts. Occasionally, stems are idiosyncratic not because of obscured phonological context, but due to lexical semantics. For example, the form *apiscawâsis* 'small child', from the elements *apist-* and *awâsis* would not be recognized (even if compounding were implemented satisfactorily) because there is no phonological context (i.e. *i*) or morphological context (i.e. a diminutive suffix) to trigger the palatalization of *t* to *c*. However, this is a case of diminutive palatalization, because *awâsis* 'child' refers to a small person, and therefore frequently occurs with diminutive palatalization, even where an overt suffix is not present. Though idiosyncracies are unavoidable as languages change, these will always be problematic for derivational recognition of unknown stems.

## CONCLUSION

As a first attempt at modeling the derivational morphology of Plains Cree, the model we have presented in this article has been shown to be a good start for further development and understanding of Plains Cree derivation. Though we are able to return correct analyses for more than four-fifths of the training data, we are still left with many avenues for further development, in both the morpheme concatenation and morphophonological rules we have devised. We have noted a set of morphemes that need to be added to our lexicon to improve recognition, and we have different options for modeling the morphophonemic rules, which may yield improved results. Furthermore, our model sheds light on the documentation of morphemes and morphophonological rules in Plains Cree: while they have served as an excellent starting point for a derivational model, the variation that has arisen over time due to sound change and other idiosyncracies that may occur cannot be handled by such a model. The model allows us to test quantitatively how much of the derivational morphology can indeed be modeled by a set of relatively general rules, and how much remains idiosyncratic. In the current instance, it appears that this ratio of regularity vs. irregularity/idiosyncracy is approximately 4/5 vs. 1/5. However, with the results of the model available, we can begin to make improvements which will allow us to examine and describe, and perhaps even model, more of the derivational morphology of Plains Cree. In the realm of Algonquian linguistics, where the corpora available us are relatively limited and cannot thus be an extensive source for extracting a comprehensive lexicon of a language, a well-developed derivational model for Plains Cree morphology would complement the existing inflectional model, allowing us to identify the possible morphemes underlying unknown stems in our analyses, which would improve overall recognition rates not just in current

corpora of Plains Cree but newly created texts as well, and can offer previously undocumented stems to add to dictionaries, teaching materials, and online language-learning tools.

## REFERENCES


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

Ahenakew, Freda. 1987. *Cree Language Structures: A Cree approach*. Winnipeg: Pemmican Publications, Inc.

Ahenakew, Freda. 1988. *wîsahkêcâhk êkwa waskwayak: âtayohkêwin.* [Cree only edition]. Saskatoon: Saskatchewan Indian Cultural Centre.

Beesley, Kenneth R., & Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Publications.

Bear, Glecia, Minnie Fraser, Irene Calliou, Mary Wells, Alpha Lafond, & Rita Longneck. 1992. *kôhkominawak otâcimowiniwâwa / Our Grandmothers' Lives as Told in Their Own Words*. Ed. by Freda Ahenakew & H. Christoph Wolfart. Regina: Canadian Plains Research Center.

Bloomfield, Leonard. 1946. Algonquian. In *Linguistic structures of Native America* (Vol. 6, pp. 85-129). New York: Viking Fund Publications in Anthropology.

Cook, Clare & Jeffrey Muehlbauer. 2010. *A Morpheme Index of Plains Cree*. URL: http://www.academia.edu/304874/A_morpheme_index_of_Plains_Cree

Ellis, C. Douglas. 2000. *Spoken Cree, Level I* (2nd edition). Edmonton: The University of Alberta Press.

Goddard, Ives. 1974. Remarks on the Algonquian Independent Indicative. *International Journal of American Linguistics*, *40*(4), 317-327.

Harrigan, Atticus, Katherine Schmirler, Antti Arppe, Lene Antonsen, Sjur Moshagen, & Trond Trosterud. (forthcoming, June 2017: accepted pending revisions). Learning from the computational modeling of the Plains Cree verb. *Morphology*.

Hulden, Mans. 2009. Foma: a finite-state compiler and library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, 29-32.

Kâ-Nîpitêhtêw, Jim. 1998. *ana kâ-pimwêwêhahk okakêskihkêmowina / The Counselling Speeches of Jim Kâ-Nîpitêhtêw*. Ed. by Freda Ahenakew & H. Christoph Wolfart. Winnipeg: University of Manitoba Press.

Karttunen, Lauri. 2006. The insufficiency of paper-and-pencil linguistics: the case of Finnish prosody. *Intelligent linguistic architectures: Variations on themes by Ronald M. Kaplan*, 287-300.

Lavallee, Ray, & Judith Silverthorne. 2014. *Honouring the Buffalo: A Plains Cree Legend*. Regina: Your Nickel's Worth Publishing.

Lindén, Krister, Erik Axelson, Sam Hardwick, Tommi A. Pirinen, & Miikka Silfverberg. 2011. HFST - framework for compiling and applying morphologies. In *Proceedings of Second International Workshop on Systems and Frameworks for Computational Morphology* (SFCM), 37-85.

Masuskapoe, Cecilia. 2010. *piko kîkway ê-nakacihtât: kêkêk otâcimowina ê-nêhiawastêki*. Ed. by H. Christoph Wolfart and Freda Ahenakew. Winnipeg: Algonquian and Iroquoian Linguistics.

Minde, Emma. 1997. *kwayask ê-kî-pê-kiskinowâpatihicik / Their Example Showed Me the Way: A Cree Woman's Life Shaped by Two Cultures*. Ed. by Freda Ahenakew & H. Christoph Wolfart. Edmonton: University of Alberta Press.

Mohri, Mehryar. 1997. Finite-state transducers in language and speech processing. *Computational Linguistics*, 23:269–311.

Okimâsis, Jean. 2004. *Cree, Language of the Plains*. Regina: University of Regina Press.

Oxford, William. R. 2014. *Microparameters of agreement: A diachronic perspective on Algonquian verb inflection* (Doctoral dissertation, University of Toronto).

Pirinen, Tommi. 2014. Weighted Finite-State Methods for Spell-Checking and Correction. (Doctoral dissertation, Department of Modern Languages, University of Helsinki).

Ratt, Solomon. 2016. *Beginning Cree*. Regina: University of Regina Press.

Snoek, Connor, Dorothy Thunder, Kaidi Lõo, Antti Arppe, Jordan Lachler, Sjur Moshagen, & Trond Trosterud. 2014. Modeling the Noun Morphology of Plains Cree. Paper read at ComputEL: Workshop on the use of computational methods in the study of endangered languages, 52nd Annual Meeting of the ACL, Baltimore, Maryland, 26 June 2014.

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

Whitecalf, Sarah. 1993. *kinêhiyawiwiniwaw nêhiyawêwin / The Cree Language is Our Identity: The La Ronge Lectures of Sarah Whitecalf*. Ed. by H. Christoph Wolfart & Freda Ahenakew. Winnipeg: University of Manitoba Press.

Wolfart, H. Christoph. 1973. *Plains Cree: A Grammatical Study*. Transactions of the American Philosophical Society, n.s., vol. 63, part 5. Philadelphia.

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

Wolvengrey, Arok. 2001. *nêhiyawêwin: itwêwina / Cree: Words*. Regina: University of Regina Press.

---

[1] Abbreviations: VAI = animate intransitive verb, VTI = transitive inanimate verb, VTA = transitive animate verb, NA = animate noun, NI = inanimate noun.

[2] These stems both end in *Cw* clusters, which are resolved in two ways. For *atim*, the *w* is deleted; however, for *maskwa*, a final *-a* is suffixed. This is an obsolete animate singular marker, which is retained here to avoid a monosyllabic noun. As *atimw-* is disyllabic, this *-a* is not retained the cluster simplification applies instead.

[3] As all types of stem morphemes, initials, medials, and finals, can carry considerable semantic content, we may even deviate from traditional terminology and recognize that any of these categories can, in some way, contain root-like morphemes. For instance, when we see derived medials, body part medials, and nominal finals, these can also be considered bound roots that require further morphological material to be used as free forms.

[4] This inflectional model has been used to analyze a corpus of Plains Cree (Ahenakew, 2000; Bear et al., 1992; Kâ-Nîpitêhtêw, 1998; Masuskapoe, 2010; Minde, 1997; Vandall and Douquette, 1987; Whitecalf, 1993). Further development of the derivational analyzer can be used to improve analyses of unknown forms in the corpus.

[5] Recall that zero finals are structurally posited (e.g. Bloomfield, 1946).

[6] The dotted lines in this figure represent pathways that are necessary to the modeling of Plains Cree, but are not yet fully implemented in the model tested in this article.

[7] In this formalism, $ refers to a morpheme boundary and | separates different possible contexts.

[8] For the sake of simplicity, we use C here to represent any consonant, though this is represented in the actual formalism as [ c | h | k | m | n | p | s | t | w | y | ý ].

[9] The symbol *ý* is used by Wolvengrey (2001) to represent reflex of Proto-Algonquian **l* or **r*, which occurs as *y* in Plains Cree, but varies in closely related dialects and allows the dictionary to be used by speakers of these dialects.

[10] As this coding for **e* vs. *i* is not yet exhaustive, and perhaps cannot be due to homophony or lack of evidence, these rules contribute to the model but do not apply perfectly.

[11] The notation used in our derivational FST uses the slash '/' to indicate a morpheme boundary, so in the analysis /nîht-/wêp-/-n/ the model posits three morphemes.

[12] Though not the primary focus of this paper, one can also use the computational derivational model in the inverse direction to provide expected realizations of a sequence of morphemes, taking into account the

known morphophonological rules. In the case of */nîht-/wêp-/-n/*, this generates two alternative stems, *nîhciwêpin* and *nîhtiwêpin*.

[13] A weight of 0.0 provided by a HFST analyzer indicates that the FST is not weighted. When weighting does not occur, there is no systematicity to the order in which analyses are presented.

[14] These negated logarithms of probabilities, i.e. *-log(p)*, are known in computational linguistics as *tropical weights*. These can be interpreted as penalties because a large *-log(p)* indicates a small probability *p*. Additionally, if the computational model were non-deterministic, we would need a tropical semiring to approximate the actual probability, but that does not apply for our model as it is deterministic.

[15] Similar to what was noted earlier in conjunction with the non-weighted basic derivational model, the weighted computational model can just as well also be used in the inverse direction to produce realizations of stems resulting from morpheme sequneces, taking into account the morphophonological rules (and their occurrences in the training data). For /nîht-/wêp-/-n/, the two possible stems are weighted as *nîhciwêpin-* (14,08) and *nîhtiwêpin-* (16,92), suggesting that the palatalization *t>c* is slightly more likely.