A Morphosyntactically Tagged Corpus for Plains Cree

Antti Arppe, Katherine Schmirler, Atticus G. Harrigan, Arok Wolvengrey

This paper presents the underlying resources that comprise a morphologically and syntactically tagged corpus of Plains Cree. Morphosyntactically tagged corpora contribute to linguistic descriptions and language maintenance in a variety of ways. Corpora can be used to supplement online dictionaries with examples of forms in natural language use and allow for systematic quantitative analyses to be performed on much larger scales than previously possible, without extensive experience in computational techniques. Such analyses can then further inform qualitative analysis and benefit descriptions overall.

For Plains Cree, the tagged corpus involves a number of resources; these include the texts, the computational models, the hand-verified "gold standards", and the corpus search interface. Together, these form a morphosyntactically tagged corpus of Plains Cree. The texts included in the present iteration of the corpus are herein described as the Ahenakew-Wolfart corpus and represent a selection of Cree texts collected in the late 20th century. Two levels of computational linguistic analysis are briefly described in this paper; these are the morphological model, which allows for automatic analysis of inflectional morphology and morphophonology, and the syntactic model, which currently disambiguates word forms with multiple analyses (based on contextual information) as well as identifies the arguments of verbs and demonstrative-noun pairs. For each of these models, portions of the texts have been hand-verified for evaluation and further development of the models; the process of annotation and the coverages statistics for each model are presented herein. Finally, the online corpus interface is briefly described.

THE TEXTS

The collection of texts presently available in the morphosyntactically tagged corpus for Plains Cree are drawn from those collected and edited by Freda Ahenakew and H.C. Wolfart in the 1970s to 1990s. Hereafter referred to as the Ahenakew-Wolfart texts, published versions of these texts are available in the following volumes: Ahenakew (2000), Bear et al. (1992), Kâ-Nîpitêhtêw (1998), Masuskapoe (2010), Minde (1997), Vandall and Douquette (1987), and Whitecalf (1993).[1] This collection contains various genres, including dialogues, personal narratives, retold stories, lectures, and speeches.

Altogether, these texts contain currently 142,192 tokens (20,503 types), of which 80,221 are word tokens of Plains Cree (16,532 word form types), with an additional 761 Plains Cree word fragments (436 types, typically initial parts of Cree words which are not finished, or restarted after hesitation, explicitly transcribed in the versions of the texts that we have), 8,791 non-Cree word tokens (1,253 types) including English, French, Michif, Arabic and Roman numerals, proper names, etc., 3,293 yet unanalyzed tokens (2,255 types all from Ahenakew 2000), as well as 49,127 punctuation mark tokens (26 punctuation types).[2] Compared to available corpora for a number of majority languages, a corpus of this size is rather small. However, more comparable corpora are those for Indigenous languages, which, where available, may range from several thousands of words to several million (e.g. Inuktitut). Additionally, the present Plains Cree corpus is not comprised of all available texts of substantial size; the corpus will be bolstered by texts collected by Bloomfield in the early 20th century (Bloomfield 1930, 1934), the Cree Prayer Book (Demers et al. 2010, collected in the late 19th century), as well as multiple

Bible translations from different time periods. Together, these texts offer a corpus of at least two hundred thousand words, and potentially several hundreds of thousands. However, even with the smaller present corpus, meaningful analyses can be undertaken and tools can be made available to communities and researchers alike.

The Two Levels of Linguistic Analysis

The corpus is automatically tagged by means of two computational models, one for the inflectional morphology and related morphophonology and one for disambiguation and the assignment of syntactic relationships. Both models are continually under development and the tagging will thus be improved as the models improve. Descriptions for earlier versions of each of these models are available in Snoek et al. (2014), Harrigan et al. (2017), and Schmirler et al. (2018).

The Morphological Model

The morphological computational model for Plains Cree is based in Finite State Transducer (FST) technology (e.g. Beesley and Karttunen 2003a) and is able to both analyze and generate word forms. The current iteration of the Plains Cree morphological model is able to analyse word class, person, number, animacy, obviation, tense, possession, and direction, as well as identify preverbs, reduplication, and two derivational processes, the formation of diminutive nouns and comitative verbs.

The inflectional component makes use of information drawn from grammars of Plains Cree such as Wolfart (1973) and Okimâsis (2004), as well as the expertise of one of the authors, A. Wolvengrey. The lexicon of nouns, verbs, pronouns, and particles is drawn primarily from the database underlying Wolvengrey (2001), a bilingual dictionary of Plains Cree and English. An example input and its analyzed output are given in (1). The model recognizes *ê*- and *nitawi*- as preverbs, *wî*- as the future intentional prefix,[3] *nîmihitow* as the lemma form of the verb stem which is animate intransitive, and -*yân* as the first person conjunct suffix.

(1) *ê-wî-nitawi-nîmihitoyân*

PV/e+PV/nitawi+nîmihitow+V+AI+Cnj+Fut+Int+1Sg

"I will go and dance"

Morphophonological changes are also covered at specified junctures. These make use of the TWOLC (TWO-Level Compiler) formalism, which applies the set of morphophonological rules in parallel, in comparison to rewrite rules that are applied in an ordered sequence (e.g. Beesley and Karttunen 2003b). Due to their parallel nature, larger numbers of TWOLC rules can become difficult to specify without rule conflicts, though this can be applied rather straightforwardly to the relatively regular phonology of Cree. The TWOLC rules used in a former version of the model are detailed in Harrigan et al. (2017); though we have since simplified and restructured these rules, they still represent the basic sound changes required for Plains Cree inflectional morphology.[4]

The output of the morphological model results in a string of morphological tags, as in (1), which then become tags accessible in the corpus interface. These tags can then be searched individually, or as groups of features, such that a researcher may investigate a feature or group of features of interest using straightforward search capabilities, outlined in greater detail below.

The Syntactic Model

The syntactic computational model for Plains Cree uses the Constraint Grammar formalism (e.g. Karlsson 1995a, 1995b) to 1) disambiguate word form analyses and to 2) identify relationships between words in a phrase. The Constraint Grammar version used is the VISLCG-3 compiler (e.g. Bick and Didrikson 2015). Constraint Grammar operates on principles similar to those in more theoretical dependency-based frameworks, using the morphological feature tags output by the morphological model and the syntactic context in which they occur to select the appropriate analysis and to determine the relationships between words in a sentence.

The syntactic model operates in two stages: first, disambiguation constraints are applied to the results of the morphological analysis, and second, function constraints are then applied to label syntactic roles. A form that is ambiguous for animacy or number, for example, may be disambiguated by looking at adjacent demonstratives or nearby verbs: an animate demonstrative or animate intransitive verb may allow for the ambiguous form to be identified as animate. Similarly, an animate nominal in the same phrase as an animate intransitive verb would be tagged as the actor of that verb (depending on number, obviation, and person agreement features), while an inanimate noun in the same phrase would not be marked as an argument of the same verb.[5] An example sentence is given in (2). Here, the ambiguous demonstrative *ôhi* (either inanimate plural or animate obviative) is determined to be inanimate because it is immediately adjacent to an inanimate plural noun (its context). Then, the transitive inanimate verb allows for an animate actor and an inanimate goal,[6] both of which can be found within the sentence. These are tagged appropriately, with arrows pointing towards the verb that allowed the syntactic feature/tag to be assigned. The goal is not explicitly marked on the verb and so any inanimate nominal, regardless of number, may be selected as the goal of a transitive inanimate verb. However, the features of the verb and actor must agree—in this sentence, the animate singular noun agrees with the third person singular marking on a verb that selects animate actors. Verbs are also tagged for reference in the corpus interface.[7]

(2) *iskwêw ayamihtâw ôhi masinahikana*.

iskwêw	ayamihtâw	ôhi
iskwêw+N+AN+Sg	ayamihtâw+V+TI+Ind+Prs+3Sg	**ôma+Dem+IN+Pl**
		~~awa+Dem+AN+Obv~~
woman	s/he reads s.t.	these, this
@ACTOR>	**@PRED-TI**	**@N>**

masinahikana

masinahikan+N+IN+Pl

books

@<GOAL

‘The woman is reading these books.’

The current version of the model identifies relationships between verbs and their nominal arguments and between nouns and associated demonstrative pronouns. The focus of the current model is on identifying relationships rather than maximizing disambiguation. These goals are achieved using only morphological features at present, though lexical semantic and other features can also be made use of within the Constraint Grammar formalism. Further development of each of these aspects is underway.

Syntactic tags can be used in tandem with morphological tags in corpus searches to investigate patterns of both (groups of) morphological features and the basic word order patterns with which they occur. Examples of corpus search results are discussed below.

THE GOLD STANDARDS

Both models have undergone some degree of testing using hand-verified portions of the corpus. These are referred to as "gold standards" for each model. This section details the annotation process, the coverage of each model after the first stage of development, and the types of errors that were encountered.

The Morphological Gold Standard

For the morphological gold standard, 18,646 unique word types (not individual tokens) were verified. One text (Ahenakew 2000) was not included in the gold standard in order to allow for testing later, though over 86% of the tokens of that text were represented by forms in the gold standard and so had otherwise already been verified. These types were evaluated using the morphological model, resulting in correct analyses, incorrect analyses, and unanalyzed forms. For each of these 18,646 types, two researchers, the second and third authors, then hand-verified the analyses, correcting, removing, and adding analyses where possible. At this stage of verification, forms were analyzed in isolation, and so ambiguous analyses are included and no disambiguation is attempted.

The verification process served two main purposes. One, the verified and corrected analyses are those included in the corpus, resulting in much higher accuracy rates than would be supplied by the morphological model alone, as it is still under development. Two, the corrections and additions allowed us to determine shortcomings and errors in the model. Issues we encountered primarily included differences in vowel length marking, lemmas missing from our lexicon, minor or archaic morphological features not yet implemented in the model (e.g. inanimate actor, *h*-preterit, initial change), vowel sandhi altering preverbs and stems, and errors that arose in the model, particularly the model's overzealous application of morphophonological rules.

A detailed analysis of the coverage of the model at the time the morphological gold standard was first performed is available (for verbs) in Harrigan et al. (2017). However, the morphological gold standard is frequently reviewed by the annotators and so has improved considerably. In the current version, 437 of the 18,646 word form types remain unanalyzed; these include partial words (i.e. due to hesitation), speech errors, or word forms with stems that could not be satisfactorily identified (e.g. for verb class) without context provided by translation, and so require further consideration. Inter-annotator agreement was assessed using two overlapping portions of altogether 200 types from the corpus which had been evaluated by both annotators, and for which, after discussion, a consensus was reached for all analyses; a full account of the inter-annotator agreement can be found in Harrigan et al. (2017). The morphological gold standard is of great benefit to not only an online corpus of Plains Cree and to all analyses drawn from it, but also to the further development and improvement to the morphological model for the analysis of more texts for addition to the corpus and for development of additional tools, such as the syntactic model.

The Syntactic Gold Standard

Unlike the morphological gold standard, which arose after significant development of the morphological model, the syntactic gold standard was developed as a testing tool for both building and testing the syntactic model.[8] A much smaller portion of the overall corpus (one text, Vandall and Douquette 1987, approximately 3,200 words) was tagged for selected syntactic functions having a central role in Plains Cree: these are actors and goals associated with verbs and demonstratives associated with nouns; other syntactic functions were not yet tagged. Where possible, disambiguation based on context and translation was also performed. However, as the model cannot make use of translations and only grammatical features tags are used at this point, much of the disambiguation that can be performed by hand cannot be replicated by the model. The syntactic gold standard will be further expanded as constraints for more syntactic relationships are added to the syntactic model.

Fortunately for modelling, however, the corpus contains a considerable number of word forms that are assigned only one analysis by the morphological model, and so do not require disambiguation in the syntactic model. Before the disambiguation constraints were applied, 65,046 (77.2%) of word form tokens in the entire corpus had a single analysis; after the disambiguation constraints were applied, 75,134 (89.2%) of the tokens had a single analysis.[9] This means that disambiguation reduced 10,088 word tokens (12.0%) that previously had multiple analyses to just a single analysis (which may be correct, or not). While these numbers are promising, it is a feature of Constraint Grammar that at least one analysis will always remain after disambiguation, so these numbers are no guarantee that the correct analysis is selected. Therefore, the model's disambiguation can be compared to the manual disambiguation performed in the syntactic gold standard. This comparison returns a recall rate of 62% and a precision rate of 90%. Firstly, this means that, in the syntactic gold standard, the constraints removed 62% of those multiple alternative analyses that were removed by hand, leaving still 38% of the originally ambiguous word form tokens with either more than one analysis (which may include the contextually correct analysis, or not), or with a single but incorrect analysis. Secondly, this entails that of those analyses the constraints removed, 90% matched those removed manually in the syntactic gold standard; thus, for 10% of the word form tokens with ambiguous analyses, an analysis that should have been removed was not, or the analysis that was removed by the constraints would in fact have been the correct one (one per each word form token, based on context). For further details, see Table 1 in Schmirler et al. (2018). Following automatic disambiguation, we make use of a crude heuristic to reduce all remaining tokens to a single analysis; this heuristic selects the analysis with the fewest morphological feature tags (i.e. the simplest analysis), resulting in one analysis per word form in the corpus.

The Corpus Interface

For the online corpus, we make use of the Korp interface. This interface is based on the open-source tools for corpus search and indexing in the IMS Open Corpus Workbench (Evert and Hardie 2011). The Korp interface is a concordance search tool used by the Språkbanken (Swedish Language Bank) research group at the University of Gothenburg (Borin et al. 2012; https://spraakbanken.gu.se/korp/). It is used to manage the large Swedish language corpus and integrates the IMS Open Corpus Workbench tools with a web interface. This was adapted by our Norwegian collaborators, the Giellatekno and Divvun research teams at UiT Arctic University of Norway, for the morphologically rich indigenous Sámi languages, and thus we have found it suitable for languages such as Cree and Odawa.

For our Plains Cree version, we have incorporated the morphological feature tags, syntactic disambiguation, and syntactic function tags to create a morphosyntactically tagged online corpus. These analyses are supplemented with English glosses for the Plains Cree lemmas, where available, from Wolvengrey (2001). These glosses are also found in an intelligent online dictionary (http://altlab.ualberta.ca/itwewina/) based on the same resource; dictionary entries are linked to the corpus so that the lemmas may be seen in context as well as in the dictionary. Full sentence translations are not yet available, though the inclusion of the translations available in the Ahenakew-Wolfart text corpus to create a parallel Plains Cree-English corpus is among our development plans for the near future.

Here, we briefly detail some of the search capabilities of this corpus interface. However, as this publication format is not conducive to screenshots, examples will instead be available on the Alberta Language Technology Lab website (http://altlab.artsrn.ualberta.ca). The Korp interface (http://altlab.ualberta.ca/korp) has three levels of search capabilities: simple, extended, and advanced. Simple search allows for whole words or sequences of characters at the beginning or end of a word form to be searched. Extended search includes dropdown menus that allow a researcher to search for a form containing a sequence of characters, a particular morphological feature or set of features, a particular syntactic function tag, a sequence of words with particular features, or combinations thereof. Advanced search is similar, though makes use of CQP (corpus query protocol) search format, which includes regular expressions and therefore allows for more precision in the searches. A CQP tutorial is linked directly from the search page and, combined with basic understanding of regular expressions, the protocol allows for reasonably powerful searches.

The interface is able to display search results in a number of formats. The default setting is the KWIC (Key Word In Context) concordance display, which shows the words that fit the search criteria in the centre of each line with the sentential context on either side. The results may also be displayed in paragraph context. On the righthand side of these results, details of the corpus, the text, and the word attributes (morphological feature tags, syntactic functions, lemma gloss) are given when a word is selected. Results may also be displayed in terms of frequency in the corpus. Further refinement of search capabilities is under development.

These search capabilities and display functions allow researchers to explore features or syntactic relationships of interest in the corpus before extracting data for a quantitative study. A morphosyntactically tagged corpus within a user-friendly corpus interface simplifies and streamlines such investigations.

RESULTS

In this section, we present some of the results we have drawn from our morphosyntactically tagged corpus of Plains Cree to exemplify the types of straightforward quantitative investigations made possible by the corpus, though far more complex questions can also be explored. Simple frequency searches for word classes or subclasses are possible; examples of these using earlier versions of the morphological model and gold standard were presented in Harrigan and Arppe (2015) and Schmirler and Harrigan (2016). Frequencies for different noun and verb classes and subclasses are given in Table 1.

<table 1>

Morphological complexity has also been investigated using the Plains Cree corpus. Wolvengrey (2015) and Arppe et al. (in press) investigated the complexity of preverbs, Schmirler et al. (in press) explored the complexity of derivational morphology, and Harrigan et al. (2017) discussed the complexity of inflectional morphology. As these are generally more recent investigations using versions of the corpus quite close to its present state, these statistics are not repeated here. However, the results of these studies lean towards one key conclusion: there is some upper limit to the maximum complexity (e.g. number of preverbs) at any level of Plains Cree morphology, and this is considerably less than the maximum that can be theoretically constructed.

The syntactic model, developed more recently, has been used in fewer investigations. Results can be found, along with more details regarding the development of the Constraint Grammar parser, in Schmirler et al. (2018). General trends, which align well with previous descriptions of Plains Cree sentences (e.g. Dahlstrom 1991, 1995), include the high frequency of verbs that occur with no lexical items as arguments (~50%), and that obviative arguments (e.g. less topical) are more likely to appear as lexical items than proximate arguments, though proximate arguments are more likely to occur earlier as lexical items in a clause than obviative arguments. The clause structures in the Ahenakew-Wolfart corpus can be summarized in broader terms as well: the syntactic function tags indicate that there are 20,726 clauses containing verbs; in these, there are 4,418 actors manifested as lexical items, 3,709 goals manifested as lexical items, and 2,341 demonstratives associated with adjacent nouns.

The corpus interface also allows for searches of morphological features and their syntactic relationships. For example, one can search for animate nouns tagged as actors immediately preceding VTAs—this results in 11 examples. However, a more complex search using regular expressions allows one to search for 1) animate nouns tagged as actors that are 2) followed by a VTA in the same clause allowing for 3) any intervening material that is neither punctuation or another verb. This search results in 24 clauses, and lets us see that where intervening material occurs, this is either a demonstrative or a particle. Further searches into features of nouns or verbs in such clauses may return interesting results for further investigation. For example, nearly all of the 11 clauses of actors immediately followed by their VTA governors were instances of dependent kinship nouns—a coincidence, or is there something else at play?

CONCLUSION

The morphosyntactically tagged Ahenakew-Wolfart Plains Cree corpus is the first step in a larger corpus of Plains Cree that will include over a century of Plains Cree texts collected across Alberta and Saskatchewan. The corpus represents four major components: the texts themselves, the morphological model and gold standard, the syntactic model and gold standard, and the online corpus interface. The morphosyntactic tags and the user-friendly corpus interface make exploration of Plains Cree texts much easier than previously possible, offering a starting point for the much more in-depth quantitative analysis also made possible by this tagged corpus. The gold standards for both the morphological and syntactic models allow for testing and further development of the models themselves, but the morphological gold standard, as it represents most of the Ahenakew-Wolfart texts, ensures that the current morphological feature tags are more accurate than is possible with current model.

The online corpus is also linked with an intelligent online dictionary for Plains Cree, giving speakers, teachers, and students the opportunity to make use of both tools to explore natural language use in educational and everyday contexts. Though currently restricted to researchers during the development phase, the tagged corpus will soon be available for such applications. Additionally, the search capabilities in the online interface, as well as the models and gold standards underlying the tagged corpus, will be further developed and improved. Other existing and new texts will be morphologically and syntactically analysed and implemented in the corpus, and available English translations will be implemented to create a parallel corpus of Plains Cree and English, further increasing the usefulness of the corpus resources.

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*. Edited by H.C. Wolfart. Winnipeg: University of Manitoba Press.

Arppe, Antti, Katherine Schmirler, Miikka Silfverberg, Mans Hulden, and Arok Wolvengrey. (in press). Insights from computational modeling of the derivational structure of Plains Cree. In *Papers of the 48th Algonquian Conference*.

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

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

Beesley, Kenneth R., and Lauri Karttunen. 2003b. Two-Level Rule Compiler. Xerox PARC. https://web.stanford.edu/~laurik/.book2software/twolc.pdf.

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

Bloomfield, Leonard. 1930. *Sacred Stories of the Sweet Grass Cree*. New York: AMS Press.

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

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

Borin, Lars, Markus Forsberg, and Johan Roxendal. 2012. Korp – the corpus infrastructure of Spräkbanken. In *LREC*, pp. 474-478.

Dahlstrom, Amy. 1991. *Plains Cree morphosyntax*. New York: Garland.

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

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

Evert, Stefan, and Andrew Hardie. 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 conference*, University of Birmingham, UK.

Harrigan, Atticus G., and Antti Arppe. 2015. oswêw êkwa ê-tipiskâk, mâka mâcîw âhpô ê-mâcît? Paper presented at the *47th Algonquian Conference* in Winnipeg, Manitoba, October 22-25, 2015.

Harrigan, Atticus G., Katherine Schmirler, Antti Arppe, Lene Antonsen, Trond Trosterud, and Arok Wolvengrey. 2017. Learning from the Computational Modelling of Plains Cree Verbs: Analysis and Generation Using Finite State Transduction. *Morphology*, *27*(4), 565-598.

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

Karlsson, Fred. 1995a. Designing a parser for unrestricted text. In F. Karlsson, A. Voutilainen, J. Heikkilae, & A. Anttila (Eds.), *Constraint Grammar: a language-independent system for parsing unrestricted text* (Vol. 4) (pp. 1-40). Walter de Gruyter.

Karlsson, Fred. 1995b. The formalism and environment of Constraint Grammar Parsing. In Fred Karlsson, Atro Voutilainen, Juha Heikkilä, and Arto Anttila (eds.), *Constraint Grammar: a language-independent system for parsing unrestricted text* (Vol. 4), 41-88. Walter de Gruyter.

Masuskapoe, Cecilia. 2010. *piko kîkway ê-nakacihtât: kêkêk otâcimowina ê-nêhiyawastêki*. Edited by H.C. 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*. Edited by Freda Ahenakew and H.C. Wolfart. Edmonton: University of Alberta Press.

Okimâsis, Jean. 2004. *Cree: Language of the Plains / nēhiyawēwin: paskwāwi-pīkiskwēwin*. Regina: Canadian Plains Research Center.

Schmirler, Katherine, Antti Arppe, Trond Trosterud, and Lene Antonsen. 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 in Montreal, Quebec, October 27-29, 2017.

Schmirler, Katherine, and Atticus G. Harrigan. 2016. Word class frequencies according to corpora. Paper presented at the 3rd Prairie Workshop on Language and Linguistics in Regina, Saskatchewan, March 5, 2016.

Schmirler, Katherine, Antti Arppe, Trond Trosterud, and Lene Antonsen. 2018. Building a Constraint Grammar Parser for Plains Cree Verbs and Arguments. In: Calzolari, Nicoletta et al. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pp. 2981-2988. European Language Resources Association (ELRA), ISBN 979-10-95546-00-9.

Schmirler, Katherine, Atticus G. Harrigan, Antti Arppe, and Arok Wolvengrey. (in press). Plains Cree verbal derivational morphology: A corpus investigation. In *Papers of the 48th Algonquian Conference*.

Snoek, Connor, Dorothy Thunder, Kaidi Lõo, Antti Arppe, Jordan Lachler, Sjur Moshagen, and 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, and Joe Douquette. 1987. *wâskahikaniwiyiniw-âcimowina / Stories of the House People, Told by Peter Vandall and Joe Douquette*. Edited 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*. Edited by H.C. Wolfart and Freda Ahenakew. Winnipeg: University of Manitoba Press.

Wolfart, H. Christoph. 1973. *Plains Cree: A Grammatical Study*. Transactions of the American Philosophical Society New Series, vol. 63 (5). Philadelphia: The American Philosophical Society.

Wolvengrey, Arok. 2001. *nēhiyawēwin: itwēwina / Cree: Words*. Vols. 1 & 2. Regina: Canadian Plains Research Center.

Wolvengrey, Arok. 2005. Inversion and the absence of grammatical relations. In C. de Groot & K. Hengeveld, *Morphosyntactic Expression in Functional Grammar*, pp. 419-46. Berlin: Walter de Gruyter.

Wolvengrey, Arok. 2015. Preverb combinations, co-occurrences, and sequences: Preliminary findings from a preliminary Plains Cree corpus. Paper presented at the 47th Algonquian Conference in Winnipeg, Manitoba, October 22-25, 2015.

[1] We are indebted to Dr. Wolfart for access to digital versions of these texts for our purposes of modeling and corpus development.

[2] The digital versions of texts that constitute our corpus follow the conventions of Standard Roman Orthography (SRO) for Cree. As is the general corpus linguistic practice, word tokens were defined as character strings separated by white-space, and punctuation characters other than word-internal hyphens are considered word tokens as well. Thus, verbal/nominal prefixes, preverbs/prenouns, stems and suffixes are considered to together form single word tokens. Moreover, the counts presented here may vary somewhat from ones we have presented earlier, in part due to us including now the Ahenakew (2000) text (which was reserved for testing previously), for those words in this particular text which occur elsewhere in the corpus and for which we consequently already have a hand-verified morphological analysis, as well as generally tokens that we had not been earlier able to analyze as Plains Cree words, but now can. Therefore, as we

are revising our morphological and syntactic models, the exact counts may yet change slightly.

[3] This prefix could also be considered a preverb as well, incorporating in one morpheme elements of both tense and aspect, and in future development we might treat it as such.

[4] The source code for the Plains Cree morphological model can be found on-line at: https://victorio.uit.no/langtech/trunk/langs/crk/src/

[5] Such an inanimate noun may of course be an adjunct to such a verb, or in the case of a phrase containing a VTA, the second object of a benefactive, for example; these relationships which require the addition of lexical semantic features to verbs and nouns will be implemented in a later version of the model.

[6] We have chosen to use *actor* and *goal* over *subject* and *object* for a number of reasons; primarily, we look to Algonquian tradition (e.g. Bloomfield 1946) and to theoretical argumentation for the lack of grammatical relations in favor of semantic ones (e.g. Wolvengrey 2005).

[7] The tags in this example serve several different purposes. @ACTOR and @GOAL tags are assigned to nominal elements depending on their features and those of nearby verbs; the position of the verb relative to the argument is indicated with < or >. Verbs are tagged as @PRED for predicate, and specified as II, AI, TI, or TA on the basis of the feature tags output by the morphological model. Finally, demonstratives, when associated with adjacent nouns, are marked @N for "dependent on a noun" and the relative position of the noun with respect to the demonstrative is indicated.

[8] The source code for the syntactic model can be found on-line at: https://victorio.uit.no/langtech/trunk/langs/crk/src/syntax/

[9] The yet not analyzed tokens from Ahenakew (2000) were included in the overall token count in the calculation of these disambiguation performance figures.