Expanding the coverage of a computational model for an endangered language with a derivational component – the case of Plains Cree

Antti Arppe, Katherine Schmirler (University of Alberta, Canada), Arok Wolvengrey (First Nations University of Canada, Canada), Miikka Silfverberg and Mans Hulden (University of Colorado, USA)

In the case of many endangered languages, dictionaries and lexical databases, if any exist at all, are typically substantially smaller in the extent of their lexical content – often in the range of 5-20 thousand lexical entries – in comparison to such resources for majority languages (e.g. English, Swedish, Finnish), which will usually contain at least 50-100 thousand lexical entries if not substantially more. Take for example Plains Cree (crk, Algonguian: Western Canada and United States), which among the endangered languages of the world is exceptionally fortunate in having four contemporary dictionaries, namely *The Student's Dictionary of Plains Cree* (Wolfart and Ahenakew, 1998, with 5,000 lexical entries), the *Maskwacîs Cree Dictionary* (Miyo Wahkohtowin Education, with 8,986 lexical entries, *nēhiýawēwin: itwēwina / Cree: Words* (Wolvengrey, 2001, with 16,453 lexical entries, cf. altlab.ualberta.ca/itwewina), and the *Alberta Elders' Cree Dictionary* (Waugh et al., 2002, with 23,117 lexical entries). Now contrast these with the *Oxford English Dictionary* (with more than 600,000 words, cf. www.oed.com), *Svenska Akademiens ordlista över svenska språket* (SAOL, 14th edition, Svenska Akademien 2015, with 126,000 lexical entries), or the *Kielitoimiston sanakirja [Dictionary of the Finnish Language Board]* (Grönroos et al. 2016, with over 100,000 lexical entries).

But is the core vocabulary of Plains Cree truly substantially smaller than that of majority languages? What should be noted is that the three larger Plains Cree dictionaries mentioned above overlap only to a limited though varying degree. Of the 8,986 lexical entries in the Maskwacîs Cree Dictionary, a minority of 2,845 (31.6%) cannot be traced back to lexical entries in Wolvengrey (2001), or their inflected forms, using a morphological parser (Snoek et al. 2014; Harrigan et al. 2017) based on the lexical content of Wolvengrey (2001). In contrast, as many as 15,623 (67.5%) of the 23,117 lexical entries in the *Alberta Elders' Cree Dictionary* cannot be computationally derived as inflected forms of the lexical entries of Wolvengrey (2001). Based on our preliminary scrutiny of the non-overlapping lexical entries among these three dictionary resources, some are due to inconsistencies in orthography, and some represent genuine dialectal and areal differences in vocabulary. However, a large proportion of the non-overlapping entries can be considered part of the shared, core vocabulary of Plains Cree, and thus we estimate the overall vocabulary for Plains Cree, as manifested in these three dictionaries, to exceed 30,000 lexical entries. But might this start approaching the upper bound of contemporary Plains Cree vocabulary? Nowadays, in the case of majority languages, we are able to exploit ever increasing corpora to extract new vocabulary, but in the case of Plains Cree the corpus of the major known works excluding Bible translations and other religious texts, namely the texts compiled and edited by H. C. Wolfart and Freda Ahenakew as well as by Leonard Bloomfield, add up to only some 156,483 word tokens representing 18,605 word types, of which only a small fraction (364 word types) are currently not analyzable as Plains Cree word types. However, Plains Cree lexemes are almost entirely composed of native Cree morphemes, instead of loans from other Indigenous and majority languages, and moreover the derivation of Plains Cree words by concatenating these native morphemes continues to be a fairly regular and productive

morphological process, with the meaning of the resultant derived words being quite transparent, in contemporary Plains Cree (and generally for related Algonquian languages). Thus, we can use a computational model of derivational word/stem formation in Plains Cree to substantially expand our potential vocabulary coverage, though the practical usefulness of such a derivational model depends on how well we can incorporate information on what derivational morpheme combinations are the most likely ones, and consequently being able to rank the derivational decompositions of words as to their plausibility (Arppe et al. 2016).

Plains Cree stem derivation involves three subclasses of morphemes, initial or root morphemes, medial morphemes, and final morphemes. A stem usually involves an initial root morpheme; roots are generally not restricted to use in certain stem classes and so may often occur in nouns, verbs, and particles, though some roots are more restricted. Roots are followed by optional medial suffixes, which often occur with more concrete meanings and, like roots, are often not restricted to particular stem classes; for example, *-âpisk(w)* 'metal' can be used in nouns, as in *pîwapisk* 'piece of metal' and in verbs, as in *kipâpiskaham* 'he closes it with metal, locks it'. Final suffixes are more restricted and are used to determine the class of a stem and, in verbs, the subclass of the verb: transitive, or intransitive, number of animate or inanimate arguments. The concatenation of a root, an optional medial, and a final morpheme constitutes a primary stem. This primary stem may then undergo further derivation, followed by an optional medial and a final morpheme, creating a secondary stem, which may also undergo further derivation (Wolfart, 1973, 1996).

While the computational modelling of inflectional morphology has resulted in a set of morphophonemic rules that can often be applied to Cree derivation, there are some other changes that are relevant to the modelling of derivational morphology. For example, *-i-* is inserted between two consonants and word-final *Cw* becomes only *C*. An understanding of historical environments is also needed; for example, *t* may be palatalized to either *c* or *s* before *i*, or may remain unchanged. The conditioning factors have been obscured by sound change: *t* may be a reflex of either **θ* or **t*, and *i* may be a reflex of either **i* or **e*, which dictate the now-unpredictable alternations (Wolfart, 1996).

The lexical database underlying Wolvengrey (2001) presents an exhaustive derivational decomposition for some 10,363 verbs, consisting of combinations of 1,784 unique initial-like morphemic elements, 308 medial-like elements, and 457 final-like elements. This information together with the above morphophonemic rules can be used to create a general computational model as a finite-state transducer (Lindén et al. 2011) of how the derivational morphemes can be concatenated to form contemporary Plains Cree verb stems. For instance, in (1) we can see the possible derivational decompositional analyses for the stem *acâhkosiwi-* 'it is a star; s/he is a star'. However, the analyses are unweighted, with equal status, not allowing us to know that */atâhkw-/-is/-iwi/* is the most likely analysis. Fortunately, we can use the pairings of stems and their derivational decompositions provided in Wolvengrey (2001) as training data in order to weight all the morpheme sequences in the finite-state transducer as to their likelihood of occurrence (Mohri 1997; Pirinen 2014). As a result, this model can be used to provide all the possible derivational analyses of verb stems, which are ranked as to the overall likelihood of their constituent morpheme sequences, as well as generate the resultant stem form for any given allowable derivational morpheme sequence (2). Therefore, the correct derivational analysis also receives the best ranking with the weighted model in (2).

(1) Unweighted analyses	(Un-weighted)	(2) Weighted analyses	Weights
/at-/-âh-/-kw-/-is/-iwi/	0,000000	**/atâhkw-/-is/-iwi/**	**17,958754**
/at-/-âh-/-kw-/-is/-iwi-/	0,000000	/atâhkw-/-is/-iwi-/	19,567070
/at-/-âhkw-/-is/-iwi/	0,000000	/atâhkw-/-is/-i/-win/-i/	27,347473
/at-/-âhkw-/-is/-iwi-/	0,000000	/at-/âhkw-/-is/-iwi/	29,869141
/at-/âhk-/-w/-is/-iwi/	0,000000	/at-/-âhkw-/-is/-iwi/	30,886749
/at-/âhk-/-w/-is/-iwi-/	0,000000	/at-/âhkw-/-is/-iwi-/	31,477455
/at-/âhk-/w-/-is/-iwi/	0,000000	/at-/-âhkw-/-is/-iwi-/	32,495064
/at-/âhk-/w-/-is/-iwi-/	0,000000	/at-/âhk-/w-/-is/-iwi/	37,466354
/at-/âhkw-/-is/-iwi/	0,000000	/at-/-â-/-hkw-/-is/-iwi/	37,544746
/at-/âhkw-/-is/-iwi-/	0,000000	/at-/-âh-/-kw-/-is/-iwi/	37,544762
/atâhkw-/-is/-iwi/	**0,000000**	/at-/-â/-hkw-/-is/-iwi/	37,640060
/atâhkw-/-is/-iwi-/	0,000000	/at-/âhk-/-w/-is/-iwi/	38,042145

In this paper, we have explored how this weighted computational model can be used as an exploratory tool in the derivational analysis of previously undocumented Plains Cree verb stems. We are currently quantitatively evaluating with the non-overlapping forms from the three major Plains Cree dictionaries how well the derivational computational model fares in generally ranking the most plausible decompositional analyses. Such a derivational computational model could eventually be incorporated as a "guesser" within a general morphological analyzer to provide possible analyses for potentially grammatical words for which the full stem is lacking from the lexicon, which could substantially extend the coverage of such analyzers. Therefore, we hope to provide a case example of how the lack of extensive collections of lexicalized forms that are available for majority languages through extensive lists of lexicalized word forms in dictionaries as well as within large corpora can, in the case of less-resourced endangered languages that primarily use native morphemes for new word formation, such as Plains Cree, be compensated by a smaller lexical database, when that resource is comprehensively enriched with detailed derivational information, which allows for both generally specifying the computational derivational model as well as weighting its analyses based on the co-occurrence likelihood of derivational morphemes.

References

Arppe, A., K. Schmirler, M. Silfverberg, M. Hulden & A. Wolvengrey (2016). Computational modeling of the derivational morphology of Plains Cree words. *48th Algonquian Conference*, Milwaukee, Wisconsin, 14-16 October 2016.

Lindén, K., E. Axelson, S. Hardwick, M. Silfverberg & T. Pirinen (2011). HFST – Framework for Compiling and Applying Morphologies. Proceedings of *Second International Workshop on Systems and Frameworks for Computational Morphology (SFCM)*, 67-85.

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

Pirinen, T. (2014). *Weighted Finite-State Methods for Spell-Checking and Correction. Ph.D dissertation*. Department of Modern Languages, University of Helsinki.

Grönroos, E-R., R. Klemettinen, L. Joki, T. Heinonen, M. Haapanen & L. Nuutinen (Eds.) (2016). *Kielitoimiston sanakirja [Dictionary of the Finnish Language Board]*. Helsinki: Institute for the Languages of Finland.

Harrigan, A., K. Schmirler, A. Arppe, L. Antonsen, S. N. Moshagen, T. Trosterud & A. Wolvengrey (submitted). Learning from the Computational Modeling of Plains Cree Verbs. *Morphology*.

LeClaire, N., G. Cardinal, E. H. Waugh & T. J. Chalifoux (Eds.) (2002). *Alberta Elders' Cree Dictionary / alperta ohci kehtehayak nehiyaw otwestamakewasinahikan.* Edmonton: University of Alberta Press.

Snoek, C., D. Thunder, K. Lõo, A. Arppe, J. Lachler, S. Moshagen & T. Trosterud (2014). Modeling the Noun Morphology of Plains Cree. In Proceedings of *ComputEL: Workshop on the use of computational methods in the study of endangered languages*, 52nd Annual Meeting of the Association for Computational Linguistics, 34-42, Baltimore, Maryland, 26 June 2014.

Svenska Akademien (2015). *Svenska Akademiens ordlista over svenska språket [The Academy of Sweden's word list for the Swedish Language], 14th edition.* URL: http://www.svenskaakademien.se/svenska-spraket/svenska-akademiens-ordlista-saol

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

Wolfart, H. C. & F. Ahenakew (1998). *The Student's Dictionary of Literary Plains Cree, Based on Contemporary Texts*. Algonquian and Iroquoian Linguistics, Memoir 15.

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

Wolvengrey, A. (2001). *nêhiyawêwin : itwêwina / Cree: Words*. Regina: Canadian Plains Research Center.