What can a small corpus tell? (with the help of computational modeling) - case studies in morphological complexity, syntactic argument structure, and word order in Plains Cree

Antti Arppe
Alberta Language Technology Lab, Department of Linguistics, University of Alberta

For many endangered Indigenous languages, corpora, if existing at all, often at their largest amount only up to several tens or hundreds of thousands of words, and thus appear tiny in comparison to those currently available for the majority languages of the world, with corpora adding up to billions of words, and ever increasing. In corpus linguistics, it has been taken for granted that more and more data is almost always be better, but does that mean that there is not much to be learnt from much smaller corpora?

Plains Cree (crk: nÄ“hiyawÄ“win) is an Algonquian language still spoken by several thousands of people in the Western Plains, primarily in the Canadian provinces of Alberta and Saskatchewan, as well as in Manitoba, and in Montana on the American side of the border. Often mentioned characteristics of Plains Cree (and Algonquian languages in general) are its rich, polysynthetic morphological system, in particular for verbs, with a number of circumfixes and other parallel long-term word-internal morphological dependencies, as well as its four-way classification of verbs according to their transitivity and the animacy of the principal arguments (subject and object). A morpho-syntactic feature of note is the marking of both the subject and the object on the verb, when these arguments are animate, which allows in principle for not explicitly expressing one or both these arguments as nouns or pronouns in the immediate context. As for general syntactic traits, Plains Cree is traditionally described as having a flexible word order.

The largest available corpus for Plains Cree is a collection of contemporary texts adding up to some 100 thousand word tokens, which were recorded, transcribed, translated, and edited by Freda Ahenakew and H. Chris Wolfart in the 1980s and 1990s, thus known as the Ahenakew-Wolfart corpus (Arppe at al., in press; Ahenakew, 2000; Bear et al., 1992; KaÌ‚-NiÌ‚piteÌ‚hteÌ‚w, 1998; Masuskapoe, 2010; Minde, 1997; Vandall & Douquette, 1987; Whitecalf, 1993).

Moreover, thanks to having access to largest lexical database for Plains Cree (Wolvengrey 2001) as well as the accompanying complete basic paradigms (multiple p.c. from Wolvengrey), we have been able to create a finite-state based (e.g. Beesley & Karttunen, 2003) computational morphological model of Plains Cree (Snoek et al. 2014; Harrigan et al. 2017). Furthermore, we have also developed a constraint-grammar-based (Karlsson 1990) surface-syntactic model for morpho-syntactic disambiguation and identifying verbal arguments (Schmirler et al. 2018).

The development of the aforementioned morphological and syntactic computational models, as well as their application on, and testing with, the Ahenakew-Wolfart corpus, has allowed us to undertake a comprehensive and systematic study of how complex are the word-internal morphological dependencies in practice in actual language use, the results of which can be understood from the perspective of cognitive load. In addition, we have also been able to study in detail how often (or rarely) and in which order the syntactic arguments of verbs are realized, and how this relates with certain morphological features of the predicate verbs, which can be interpreted on pragmatic grounds, thus providing us an empirically based, in-depth picture of Plains Cree core argument structure and word order. In my presentation, I will discuss these case studies of Plains Cree, demonstrating that even small corpora can yield valuable results.

REFERENCES

Arppe, Antti, Katherine Schmirler, Atticus G. Harrigan, & Arok Wolvengrey (in press). A morpho-syntactically tagged corpus for Plains Cree. Papers of the 49th Algonquian Conference, Montreal, Quebec, 27-29 October, 2017.

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

Harrigan, Atticus, Katherine Schmirler, Antti Arppe, Lene Antonsen, Sjur N. Moshagen, Trond Trosterud & Arok Wolvengrey (2017). Learning from the Computational Modeling of Plains Cree Verbs. Morphology, 27(4), 565â€“598.

Conor Snoek, Dorothy Thunder, Kaidi LoÌƒo, Antti Arppe, Jordan Lachler, Sjur Moshagen, and Trond Trosterud (2014). Modeling the Noun Morphology of Plains Cree. ComputEL: Workshop on the use of computational methods in the study of endangered languages, 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, Maryland, 26 June 2014.

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

Wolvengrey, A. (2001). neÌ‚hiyaweÌ‚win itweÌ‚wina = Cree: Words, bilingual edition. Regina: University of Regina Press.

ORIGINAL TEXT REFERENCES for the AHENAKEW-WOLFART CORPUS

Ahenakew, Alice. 2000. aÌ„h-aÌ„yiÌ„taw isi eÌ„-kiÌ„-kiskeÌ„yihtahk maskihkiy / They Knew Both Sides of Medicine. Told by Alice Ahenakew ; edited, translated, and with a glossary by H. C. Wolfart and Freda Ahenakew. Publications of the Algonquian Text Society / Collection de la SocieÌteÌ dâ€™eÌdition des textes algonquiennes. Winnipeg: University of Manitoba Press.

Glecia Bear, Minnie Fraser, Irene Calliou, Mary Wells,

Alpha Lafond, and Rosa Longneck. 1992. koÌ„hkominawak otaÌ„cimowiniwaÌ„wa / Our Grandmothersâ€™ Lives As Told in Their Own Words. Told by Glecia Bear, Minnie Fraser, Irene Calliou,
Mary Wells, Alpha Lafond, and Rosa Longneck; edited by Freda Ahenakew and H. C. Wolfart. Saskatoon : Fifth House Publishers.

KaÌ„-NiÌ„piteÌ„hteÌ„w, Jim. 1998. ana kaÌ„-pimweÌ„weÌ„hahk okakeÌ„skihkeÌ„mowina / The Counselling Speeches of Jim KaÌ„-NiÌ„piteÌ„hteÌ„w. Told by Jim KaÌ„-NiÌ„piteÌ„hteÌ„w ; edited, translated, and with a glossary by Freda Ahenakew and H. C. Wolfart. Publications of the Algonquian Text Society / Collection de la SocieÌteÌ dâ€™eÌdition des textes algonquiennes. Winnipeg : University of Manitoba Press.

Masuskapoe, Cecilia. 2010. piko kiÌ„kway eÌ„-nakacihtaÌ„t: keÌ„keÌ„k otaÌ„cimowina eÌ„-neÌ„hiyawasteÌ„ki mitoni eÌ„-aÌ„h-itweÌ„t maÌ„na Cecila Masuskapoe / Thereâ€™s Nothing She Canâ€™t Do : KeÌ„keÌ„kâ€™s Autobiography published in Cree. Exactly as told by Cecilia Masuskapoe ; in a critical edition by H. C. Wolfart and Freda Ahenakew. Algonquian and Iroquoian Linguistics, Memoir 10.

Minde, Emma. 1997. kwayask eÌ„-kiÌ„-peÌ„-kiskinowaÌ„pahtihicik / Their Example Showed Me They Way. Told by Emma Minde; edited, translated and with a glossary by Freda Ahenakew and H. C. Wolfart.

Vandall, Peter and Joe Douquette. 1987. waÌ„skahikaniwiyiniw-aÌ„cimowina / Stories of the House People. Told by Peter Vandall and Joe Douquette ; edited, translated, and with a glossary by Freda Ahenakew. Publications of the Algonquian Text Society / Collection de la SocieÌteÌ dâ€™eÌdition des textes algonquiennes. Winnipeg: University of Manitoba Press.

Whitecalf, Sarah. 1993. kineÌ„hiyaÌ„wiwininaw neÌ„hiyaweÌ„win / The Cree Language is Our Identity : The Laronge Lectures by Sarah Whitecalf. Told by Sarah Whitecalf ; edited, translated, and with a glossary by H. C. Wolfart and Freda Ahenakew. Publications of the Algonquian Text Society / Collection de la SocieÌteÌ dâ€™eÌdition des textes algonquiennes. Winnipeg : University of Manitoba Press.