misi-mîkiwâhp pêsêkinosa ohci - A Corpus of Miscellaneous Plains Cree Texts

Like most North American Indigenous languages, corpus resources available for Plains Cree/*nêhiyawêwin* (ISO:crk) are relatively limited in size, scope, and number. These include a 152 405 token morphosyntactically-tagged corpus of personal reminiscences and legends (Schmirler 2022), a roughly 50 000 token corpus compiled for NLP development (Teodorescu et al. 2022), and a database of 20 300 morphosyntactically-tagged Cree terms and phrases (Poulin et al. 2023, forthcoming). However, these corpora have largely made use either of lengthy existing narratives or novelly elicited speech as sources; to date, none have made use of the large volume of Plains Cree translations of governmental documents, such as PSAs, census documents, and court proceedings, published by federal and provincial authorities in Canada, and only Teodorescu et al. (2022) has made use of the highly diverse and numerous small pieces of Plains Cree text spread disparately across blogs, social media platforms, online textbooks, and other decentralized sources across the internet. This paper therefore discusses the creation of a 'miscellaneous' Plains Cree corpus consisting entirely of these existing online Plains Cree texts, which have, for the first time, been centralized, parallelized at the sentence level with English translations, and made searchable in a single repository alongside other Plains Cree corpora. Composed of 428 distinct texts, ranging in length from single sentences to 17 000 token instructional texts, this corpus contains a total of 285 864 tokens of Cree and English text, of which approximately 150 000 tokens are in Plains Cree. We outline in this paper the (ongoing) construction and metadata-tagging process of this corpus, its potential use cases and limitations, and the benefits of taking advantage of small pieces of natural language data, such as those used here, which are often overlooked in language documentation in favor of lengthier, more conventional documentary corpus texts.

References

Poulin, Jolene, Dacanay, Daniel, & Arppe, Antti. (2023, forthcoming). Speech Database (Speech-DB) – An on-line platform for recording, storing, validating, and searching spoken language data. In *the Proceedings of 1st Workshop on NLP applications to Field Linguistics (Field Matters)*.

Teodorescu, Daniela, Matalski, Josie, Lothian Delaney, Barbosa, Denilson, & Demmans Epp, Carrie. 2022. Cree Corpus: A Collection of nêhiyawêwin Resources. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6354–6364, Dublin, Ireland. Association for Computational Linguistics.

Schmirler, Katherine. 2022. *Syntactic Features and Text Types in 20th Century Plains Cree: A Constraint Grammar Approach*. PhD dissertation, University of Alberta. https://doi.org/10.7939/r3-pz87-ye25