Plains Cree textual analysis with PCA: Across the Bloomfield and Ahenakew-Wolfart subcorpora

Katherine Schmirler, Antti Arppe

University of Alberta

Schmirler & Arppe (2020) presented a first look at textual analysis in Plains Cree, applying Principal Components Analysis (PCA, e.g. Bryant & Yarnold, 1995) to a subset of the Plains Cree corpus. In this talk, we look at the expanded Plains Cree corpus of 150,000 words, including both the Ahenakew-Wolfart subcorpus and the Bloomfield subcorpus, exploring morphosyntactic differences and similarities and how these relate to the text types involved.

Principal Components Analysis aims to reduce the dimensions of large datasets (here, 138 texts containing 90 distinct morphosyntactic features) to determine how different texts pattern with respect to the frequency of different features. This approach can be used as an element of register analysis (Biber, 1985; Biber et al., 2002; Biber & Conrad, 2019) where the features that characterize different groups of texts (spoken or written) are considered with respect to the situational context: who is speaking, to whom, and for what purpose. Thus, we find that some texts contain far more of some features, and others far more of other features, and these features may be related to the text types identified by Cree speakers: *âcimisowina* (personal narratives) contain more exclusive person features, while *kakêskihkêmowina* (counselling speeches) contain more inclusive person features (also previously reported in Schmirler & Arppe, 2020). The analysis also reveals some variation within narratives: some contain more dialogue, and thus more features associated with first and second persons, while others contain very little, and thus more third person features. Several patterns emerge, often demonstrating variation between texts with respect to less-common typological features found in Cree. Clusivity, as mentioned above, is one of these, as is relative topicality coded in morphology: some texts contain more obviative while some more proximate features, some more direct verbs and others more inverse.

References

Biber, D. (1985). Investigating macroscopic textual variation through multifeature/multidimensional analyses. *Linguistics 23*(2), 337-360.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Biber, D., & Conrad, S. (2019). Register, Genre, and Style (2nd ed.). Cambridge University Press. https://doi.org/10.1017/9781108686136

Bryant, F. B., & Yarnold, P. R. (1995). Principal-components analysis and exploratory and confirmatory factor analysis. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp. 99–136). American Psychological Association.

Schmirler, K., & Arppe, A. (2020). A quantitative look at Plains Cree text types: *âtayôhkêwina* vs. *âcimowina* in Bloomfield's texts and *âcimisowina* vs. *kakêskihkêmowina* in the Ahenakew-Wolfart corpus. Paper presented at the 52st Algonquian Conference at UW Madision [hosted online], October 23-25, 2020.