Corpus Approaches to Northern Haida

Jordan Lachler and Antti Arppe (University of Alberta, Canada)

Northern Haida (ISO 639: hdn) is a severely endangered language isolate spoken by fewer than 10 speakers in British Columbia and Southeast Alaska. The largest collection of original narrative texts in the language was recorded by John R. Swanton and published in 1908 using a pre-phonemic orthography. As part of a program of language documentation and revitalization in the community, we have been adapting Swanton's original texts into the modern orthography (Hubert et al. 2016) and creating a linguistically-annotated digital corpus of approximately 100,000 words.

While much of Northern Haida grammar and lexicon (Enrico 2003, 2005) is well-documented in comparison to other Indigenous languages, no corpus studies have ever been undertaken on the language. In our project, we are exploring a range of basic questions about Northern Haida grammar to develop a fuller understanding of all aspects of the language.

While word order in Northern Haida is relatively rigid, speakers do have ordering choices in certain constructions, particularly those involving nouns and possessive pronouns (*náay díinaa* vs. *gyáagan náay* 'my house'), as well as the general ordering of postpositional phrases with respect to nominal arguments in the clause. We will report on initial results of analyses of these variable-order constructions based on their frequency and distribution within the corpus.

In addition to improving our grammatical description of the language, we are also using the corpus to conduct forensic dialectology, examining the sub-phonemic alternations captured by Swanton's original transcription. Of particular interest are certain dialectal shibboleths, such as the form of the definite suffix (*-aay* in the Alaskan dialect vs. *-ee* in the British Columbian dialect), which show clear community differentiation at later points in the 20th century, but which showed significant interspeaker and intraspeaker variation within Swanton's corpus, allowing us to capture change in progress. We will discuss several such variations, and report on what insights the corpus allows us to gleen about this earlier stage of the language.

We will conclude our presentation by looking at the real-world application of the corpus in supporting language revitalization in the community, showing how information from our analysis is being incorporated into improved language reference and teaching materials for younger learners.

References

Enrico, J. (2003) *Haida Syntax*. University of Nebraska Press.

Enrico, J. (2005) *Haida Dictionary: Skidegate, Masset, and Alaskan Dialects*. Alaska Native Language Center and Sealaska Heritage Institute.

Hubert, I., A. Arppe, J. Lachler & E. A. Santos (2016). Training & Quality Assessment of an OCR Model for Northern Haida. In N. Calzolari, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis (Eds.) *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016; ISBN 978-2-9517408-9-1, pp. 3227-3234). Portorož, Slovenia, 23-28 May 2016.

Swanton, J. (1908 [1975]). *Haida Texts – Masset Dialect*. Memoirs of the American Museum of Natural History, Volume 14, Part 1. New York: AMS Press.