Developing Data-Driven Hawaiian Language Vocabulary Lists using Preserved Documents

Kelsea Hosoda, University of Hawai`i at Manoa
Lipyeow Lim, University of Hawai`i at Manoa

The Hawaiian language has a rich history that includes a thriving language boasting the most literate nation in the late 1800s to less than one thousand Native speakers in the 1950s and is now a leading language in revitalization efforts [Donaghy, 1997]. The Hawaiian language has an advantage in its revitalization because of the documentation, preservation, and digitization efforts starting from the 1840s including Hawaiian language newspapers, oral histories and video recordings of elders. These documents are crucial to the revitalization efforts of researchers and students alike to developing the next generation of Hawaiian language speakers and understanding the Hawaiian culture.

The motivation for this project is to use knowledge from Hawaiian language text documents to develop statistically relevant vocabulary lists. Currently there are two major textbooks used to teach Hawaiian--Ka Lei Ha'aheo and Nā Kai ʻEwalu. These two curricula provide an introduction to the language, covering a range of vocabulary. However, the sequence of teaching and relevance of the vocabulary lists are trivial. This study aims to develop data driven vocabulary lists and compare them to current vocabulary lists.

Studies examining vocabulary size and frequency in second language learning suggests that there is a set of useful, high frequent, words and the set of useful words generally covers 70% or more of the words in a vocabulary size of 1000 or more [Francis, 1982; Schmitt, 1997]. Research by Laufer [1988a] suggested that 95% coverage of a vocabulary is needed to allow for reasonably successful guessing of the meaning of unknown words. Research on vocabulary size shows that a 2000 to 3000 word vocabulary provides a good basis for language use [Hirsh, 1992]. The vocabulary lists that are developed in this study are targeted for second language students at an introductory level proficiency.

In this project high frequency vocabulary sets are defined based on corpora analyses of a pre-1950s document, Hiʻiakaikapoliopele [Hoʻoulumahiehie, 2008], a post-1950s corpora of modern Hawaiian language and compare the high frequency vocabulary lists using rank order and content analysis to the vocabulary lists of Nā Kai ʻEwalu and Ka Lei Haʻaheo. Hiʻiakaikapoliopele was chosen because of its significance in the Hawaiian community [Berger, 2008]. The broader impact of developing vocabulary lists from the Hiʻiakaikapoliopele book is to provide statistics for the use of this book as a standard for third year proficient students to read.

Works Cited

Berger, John. “Hiʻiaka finds a voice in English.” Honolulu Star Bulletin. Vol 13. Issue 50. (2008).

Donaghy, Keola. "'Olelo Hawai'i - A Rich Oral History, A Bright Digital Future," Cultural Survival Quarterly, (1997).

Francis, W.N. and H. Kucera. Frequency Analysis of English Usage. Boston: Houghton Mifflin Company. (1982).

Hirsh, David, and Paul Nation. "What vocabulary size is needed to read unsimplified texts for pleasure?." *Reading in a foreign language* 8 (1992): 689-689.

Hooulumahiehie. Nogelmeier, P. (edit). “Hiiakaikapoliopele”. Booklines Hawaii, Ltd. (2008).

Laufer, Batia. "The concept of ‘synforms’(similar lexical forms) in vocabulary acquisition." *Language and Education* 2.2 (1988): 113-132.

Schmitt, Norbert, and Michael McCarthy, eds. *Vocabulary: Description, acquisition and pedagogy*. Vol. 2035. Cambridge: Cambridge university press, 1997.

50 word Summary

Hawaiian language text documents are used to develop statistically relevant vocabulary lists. High frequency vocabulary based on corpora analyses of pre-1950s documents, post-1950s documents are evaluated using rank order and content analysis to the vocabulary lists of two major textbooks used to teach Hawaiian as a second language.