The Haks Language Modeling System:
Examples from Buddhist Texts in Sanskrit, Tibetan and Chinese

Christopher Handy, Michael Litchard
Independent scholar, Independent scholar

Conference: North American Research Network in Historical Sociolinguistics (NARNiHS)
21-23 July 2017 – University of Kentucky

We present a custom software package for language modeling, with specific case examples from Sanskrit, classical Tibetan and classical Chinese texts. Our system, which we call Haks, provides information about texts without requiring prior manual tagging of individual words. Haks treats texts as sets of connected syllables, termed n-grams, rather than as discrete words, and performs frequency analysis of n-grams in order to locate common words and phrases.

Haks is a practical method for extracting recurrent strings from digitized texts in cases where grammar, vocabulary and other information about the texts are partly or entirely unknown. Our method involves building concordances of words and phrases from digitized input sets of texts, using a simple but effective pattern recognition algorithm. We demonstrate this process using texts from three major languages of the Buddhist tradition: Sanskrit, classical Tibetan and classical Chinese

The above three languages of the Buddhist literary tradition lack word boundary delimiters in their traditional manuscripts. A person familiar with these languages can identify individual words in such manuscripts by means of context clues, but the texts have no spaces between words. Due to this lack of orthographic spacing, digital concordances for Sanskrit, Tibetan and Chinese often require some amount of manual part-of-speech tagging before performing further linguistic processing automatically. Numerous studies of Sanskrit, Tibetan and Chinese texts using hybrid methods (e.g., Huet 2006, Hackett 2000, Zhan et al. 2006) have yielded useful information, but are unable to deal with unknown texts without human intervention. These tools are therefore limited by the tagging efforts of human researchers. This limit is problematic in the large scale analysis of Buddhist texts, in which we frequently find that we have only partial information about the contents of a set of texts. Consequently, an ongoing issue in automated semantic analysis and translation projects for these languages is the problem of finding word boundaries. Haks overcomes this problem by treating texts as strings of syllables instead of discrete words.

Haks is a modular system for language analysis created in Haskell. It divides each text into individual syllables based on rules for that text's particular language. If the source language is unknown, Haks can also employ a generic division rule. After constructing these initial syllables, Haks then analyzes syllables in sets called n-grams, where 'n' represents the number of syllables in the set. By sorting these n-grams according to the number of times they appear, common words and phrases naturally bubble to the top, with n-grams below a threshold value being discarded. After building a database of frequently observed n-grams, the software can determine further complex relationships between individual n-grams as well as between sets of n-grams, allowing for genre classification and other kinds of semantic analysis.

This system works on any language, human or otherwise, since it does not require any knowledge about the rules or meaning of the source texts in order to find patterns within those texts. The example texts used in our demonstration come from free Internet databases, so that our results can be verified easily. We also provide source code for the project, to make this technique available to others. The modular nature of the system allows for features to be added easily for the analysis of particular aspects of specific languages, and we plan to continue developing this idea further to include other types of string analysis.