A Hmong Corpus with Elaborate Expression Annotations
David R. Mortensen, Xinyu Zhang, Chenzuan Cui, Katherine J. Zhang
Carnegie Mellon University

Introduction
This paper describes an almost 12 million-word corpus of Hmong—a Hmong-Mien language of Southeast Asia—derived from posts to the soc.culture.hmong Usenet group and annotated for ELABORATE EXPRESSIONS. We validate the dataset with ANALOGY COMPLETION and SEQUENCE LABELING tasks.

About the Language
Exonyms White Hmong, Green Hmong
ISO 639-3: hmg, hmn
Family Hmong-Mien
Autonyms Hmoo Dawb (Hmong Daw), Moob Ntsuab (Mong Njua), Moob Leeg (Mong Leng)
Speakers 2.7 million
Primary China, Vietnam, Laos, Thailand (dialect continuum)
Other Myanmar, United States, Canada, France (incl. French Guiana), Germany, Argentina, Australia

Phonology
Rich consonant inventory, moderate vowel inventory incl. nasalized vowels, 7–8 tones. CV(C) syllables.
Morphology
No inflection morphology, one derivation affix, rich compounding
Syntax
Primarily head-initial (SVO clause order, prepositional phrases, most modifiers after nouns) but classifiers before nouns, numerals before classifiers.
Discourse
Extensive use of parallelism and paratactic structures, significant tradition of persuasive oratory and polemic reflected in written work

Orthography
Romanized Popular Alphabet (RPA) Developed by American and French missionaries and their Hmong collaborators (1951–1953).
• Uses the 26 letters present on an American typewriter
• Final letters indicate tones.
• Each syllable is typically written as a “word” (delimited by whitespace).

Experiments

Data Summary

The Hmong language data are posts from the soc.culture.hmong (SCH) Usenet group, posted from 1996–2016.

Tokens 11,822,652
Sentences 858,035
Elaborate Expressions 24,574
Tokens inside EEs (count) 98,296
Tokens inside EEs (%) 0.8

Theoretical Motivation

Claims:
1. all phonology is phonetically grounded like nasal place assimilation in English Hayes and White (2013), Becker et al. (2011);

2. is coordinate compound (CC)
3. is specifier specifier (SS)
4. is coordinate compound (CC)

 широко принят, но все еще силен

References