



# David R. Mortensen

Language Technologies  
Institute

**Friday, 4/19/19**

2:30 pm Talk, DH 2315  
4:00 pm Snacks, LTI 5<sup>th</sup> Floor

## HMONG ELABORATE EXPRESSIONS: CONSTRUCTIONAL AND DISTRIBUTIONAL-SEMANTIC PERSPECTIVES

**David R. Mortensen** is a Systems Scientist in the Language Technologies Institute. Prior to coming to Carnegie Mellon University, he was an Assistant Professor in the Linguistics Department at the University of Pittsburgh.

He earned an MA and PhD in Linguistics at the University of California, Berkeley. He has diverse research interests. **His work is multilingual and features a special interest in low-resource languages, especially languages of South and Southeast Asia.** He specializes in computational phonology, morphology, and data resource development. He is currently working on research projects involving morphological disambiguation, historical linguistics, and distributed representations of linguistic units.

**Fluent Hmong speech and writing are full of elaborate expressions**, idioms like *tuav riam tuav phom* ‘wield knife wield gun; wield weapons’ and *poob ntsej poob muag* ‘lose ear lose eye; lose face.’ **This talk argues that these expressions are interesting in two different ways.** First, they are simultaneously based upon a general pattern of coordination and on specific coordinate compounds (words like *ntsej-muag* ‘ear-eye; face’). Second, the words that can occur in these coordinate compounds are predictable and follow a single, general pattern. They seem to sometimes be composed of synonyms (as in *quaj-nyiaiv* ‘cry-cry; cry’, sometimes antonyms (as in *hnuv-hmo* ‘day-night; day and night; all the time’) and sometimes representative members of some class. **However, this talk hypothesizes that they are always distributionally similar.**

The hypothesis is tested against a 13 million word corpus of Hmong newsgroup text. I show that a classifier based on the cosine similarity between the second and fourth word (in Word2vec embeddings) better predicts when a four-gram is an elaborate expression than a strong rule-based baseline. **This finding has implications for other kinds of coordination** (whether in Hmong or in other languages).