Carnegie Mellon University
Language Technologies Institute

LREC 2022
Marseille

A Hmong Corpus with Elaborate Expression Annotations

David R. Mortensen, Xinyu Zhang, Chenxuan 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 `mww`, `hmj`

Family Hmong-Mien

Autonyms Hmoob 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

Typology

Phonology Rich consonant inventory, moderate vowel inventory incl. nasalized vowels, 7–8 tones. C(C)V syllables.

Morphology No inflection morphology, one derivation affix, rich compounding

Syntax Primarily head-initial (SVO clause order, prepostional 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).

Elaborate Expressions

Elaborate expressions are four-part parallel constructions with repeating element (A)

(1) a. tag siab tag ntsws
finish liver finish lung
A B_1 A B_2
‘with all one’s soul; satisfied’

b. kawm ntaub kawm ntawv
study cloth study paper
A B_1 A B_2
‘study; pursue education’

Order of AB_1 and AB_2 does not affect morphology, syntax, semantics.

Elaborate Expression Ordering Generalization

The order of B_1 and B_2 can be predicted, in most cases, based on a tonal hierarchy:

Order	Orthography	IPA	Description
1	-j	˥˨	high falling
2	-b	˥	high
3	-m	˨˩	low creaky
4	-s	˨	low
5	-v	˨˦	rising
6	-g	˦˨	falling breathy
7	-∅	˧	mid

Experiments

Word 1	Word 2	Word 3	Word 4 (Ref)	Reasonable Predictions for Word 4
niam ‘mother’	txiv ‘father’	ntxhais ‘daughter’	tub ‘son’	***tub***, *vauv* ‘son-in-law’
siab ‘high’	qis ‘low’	ntev ‘long’	luv ‘short’	(none)
hluas ‘old’	laus ‘young’	me ‘small’	loj/niag ‘large’	***niag*** ‘great, large’
luag ‘laugh’	quaj ‘cry’	zoo ‘happy (good)’	nyuaj ‘sad (difficult)’	*khauvxwm* ‘pity; pitiful’
ze ‘near’	deb ‘far’	no ‘here’	ub ‘there’	(none)
hnub ‘day’	hmo ‘night’	dawb ‘white’	dub ‘black’	***dub***
noj ‘eat’	mov ‘food (rice)’	haus ‘drink’	dej ‘water’	*coffee*, *pepsi* ‘soda’, *cawv* ‘liquor’, *npias* ‘beer’
hlob ‘senior’	yau ‘junior’	laus ‘old’	hluas ‘young’	***hluas***
loj ‘large’	dav ‘wide’	me ‘small’	nqaim ‘narrow’	(none)
pom ‘see’	saib ‘look at’	hnov ‘hear’	mloog ‘listen to’	***mloog***
qab ‘tasty’	tsuag ‘bland’	ntse ‘sharp’	npub ‘dull’	(none)
nkauj ‘youth (female)’	ntxhais ‘girl’	nraug ‘youth (male)’	tub ‘boy’	***tub***, *vauv* ‘son-in-law’
pem ‘up there’	nram ‘down there’	nce ‘ascend’	nqes ‘descend’	(none)
toj ‘hill’	roob ‘mountain’	zos ‘village’	nroog ‘city’	***nroog***

Out of 14 analogies, the embeddings trained on the SCH Corpus correctly predicted the gold standard completion (@10) in 7 cases and produced plausible predictions in two more.

Experiment 1: Analogies

Hypothesis: Word embeddings trained on the corpus can complete word analogies.
Word 1 : Word 2 :: Word 3 : ???

Methodology: Train a 100-dimension word2vec skip-gram model and manually evaluate the top 10 predictions for 14 example analogies.

Experiment 2: EE Labeling

Hypothesis: A neural sequence labeling model can learn to identify elaborate expressions in context and out perform simple baselines.

Methodology: Evaluate the trained model on a held-out test set, which has no overlap with elaborate expressions in the train set.

Model	Precision	Recall	F1
AB_1AB_2 Baseline	26.15	100.00	41.32
+ regex parsable	32.83	100.00	49.24
+ wv. sim. thresh	50.29	77.99	60.99
+ tonal scale	59.37	76.56	66.66
BiLSTM	66.12	84.36	74.10
CNN	87.36	94.52	90.79

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	858,635
Elaborate Expressions	24,574
Tokens inside EEs (count)	98,296
Tokens inside EEs (%)	0.8

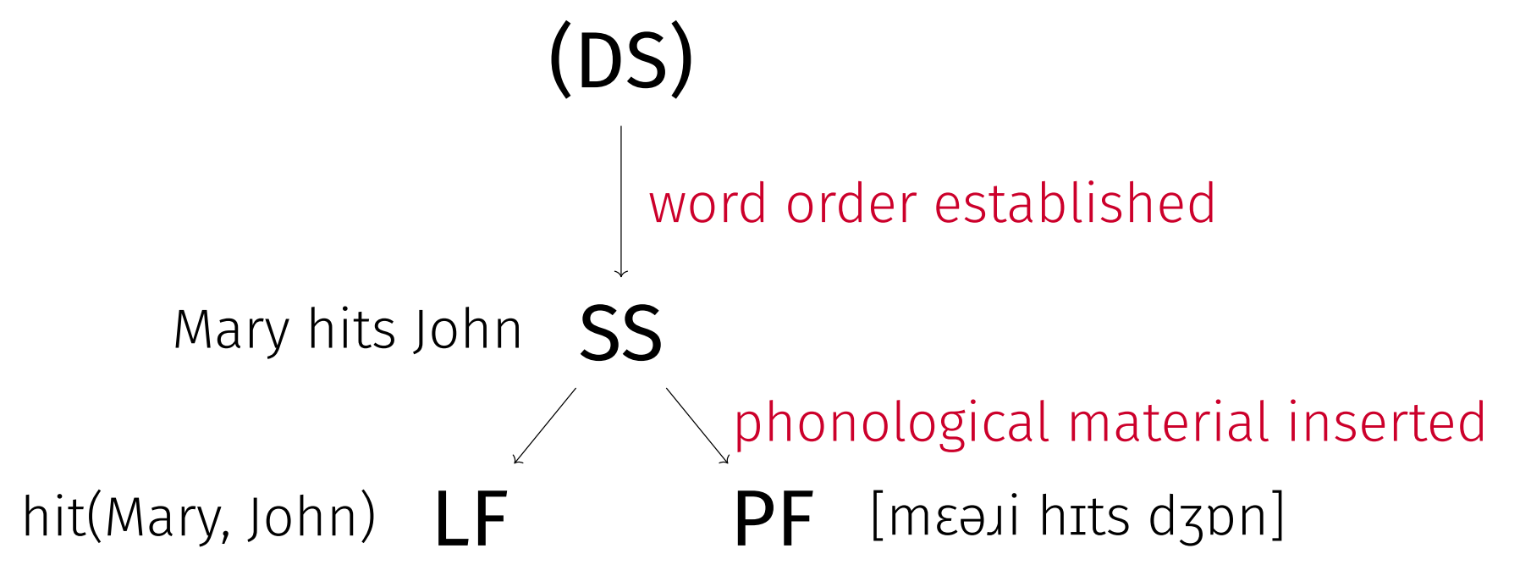
Theoretical Motivation

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

LABIAL	i[m]possible
CORONAL	i[n]tractable
VELAR	i[ŋ]glorious

Widely accepted, but still contested (Hyman, 1970; Anderson, 1981; Hale and Reiss, 2000; Moreton and Pater, 2012a,b).

CLAIM: **phonology cannot determine word order** Chomsky (1981, 1995), because word order is determined before phonological forms are inserted:



Chomsky (1981, 1995)

Contested (Ross, 1967; Kwon and Masuda, 2019; Shih and Zuraw, 2017).

Hmong EE ordering pattern appears to contradict both claims!

Data Processing

1. Quoted text was removed.
2. Plain text was extracted from HTML.
3. Text was segmented into sentences (NLTK Punkt tokenizer).
4. Tokenized (NLTK 3.6.3 `word_tokenize` function).
5. Structured into a CONLL-like format.
6. Documents included if over 60% Hmong RPA according to regex.

```
tias   O
cov    O
laus   O
no     O
tsi    B
txawj  I
tsi    I
ntse   I
thiaj  O
li     O
coj    O
tsis   O
```

Annotation Criteria

1. B_1B_2 is coordinate compound (CC)
2. $R_{syn}(B_1, \text{context}) \cong R_{syn}(B_2, \text{context})$
3. $R_{sem}(B_1, \text{context}) \cong R_{sem}(B_2, \text{context})$

```
IF (1) THEN yes ELSE IF (2) AND (3)
THEN yes ELSE no
```

References

Stephen R Anderson. 1981. Why phonology isn’t “natural”. *Linguistic inquiry*, 12(4):493–539.

Michael Becker, Nihan Ketrez, and Andrew Nevins. 2011. The surfeit of the stimulus: Analytic biases filter lexical statistics in turkish laryngeal alternations. *Language*, pages 84–125.

Noam Chomsky. 1981. *Lectures on Government and Binding: The Pisa Lectures*. Studies in Generative Grammar. de Gruyter.

Noam Chomsky. 1995. *The Minimalist Program*. MIT Press.

Mark Hale and Charles Reiss. 2000. “substance abuse” and “dysfunctionalism”: current trends in phonology. *Linguistic inquiry*, 31(1):157–169.

Bruce Hayes and James White. 2013. Phonological Naturalness and Phonotactic Learning. *Linguistic Inquiry*, 44(1):45–75.

Larry M Hyman. 1970. How concrete is phonology? *Language*, pages 58–76.

Nahyun Kwon and Keiko Masuda. 2019. On the ordering of elements in ideophonic echo-words versus prosaic dvandva compounds, with special reference to Korean and Japanese. *Journal of East Asian Linguistics*, 28(1):29–53.

Elliott Moreton and Joe Pater. 2012a. Structure and substance in artificial-phonology learning, part i: Structure. *Language and Linguistics Compass*, 6(11):686–701.

Elliott Moreton and Joe Pater. 2012b. Structure and substance in artificial-phonology learning, part ii: Substance. *Language and Linguistics Compass*, 6(11):702–718.

John R. Ross. 1967. *Constraints on variables in syntax*. Ph.D. thesis, Massachusetts Institute of Technology.

Stephanie S Shih and Kie Zuraw. 2017. Phonological conditions on variable adjective and noun word order in tagalog. *Language*, 93(4):e317–e352.