Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4992–5000
Marseille, 20-25 June 2022

A Hmong Corpus with Elaborate Expression Annotations

David R. Mortensen, Xinyu Zhang, Chenxuan Cui, Katherine J. Zhang
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213
United States of America
dmortens@cs.cmu.edu, xinyuzh2@andrew.cmu.edu, cxcui@cs.cmu.edu, kjzhang@cmu.edu

Abstract

This paper describes the first publicly available corpus of Hmong [ISO 639-3: `mww`, `hmj`], a minority language of China, Vietnam, Laos, Thailand, Australia, and various countries in Europe and the Americas. The corpus has been scraped from a long-running Usenet newsgroup called `soc.culture.hmong` and consists of approximately 12 million tokens. This corpus (called SCH) is also the first substantial corpus to be annotated for elaborate expressions, a kind of four-part coordinate construction that is common and important in the languages of mainland Southeast Asia. We show that word embeddings trained on SCH can benefit tasks in Hmong (solving analogies) and that a model trained on it can label previously unseen elaborate expressions, in context, with an F1 of 90.79 (precision: 87.36, recall: 94.52).

Keywords: corpus, Hmong, low-resource, coordination, non-compositional, elaborate expression

1. Introduction

Hmong is a language of Southern China and Southeast Asia with about 2.7 million speakers[1]. A plurality of speakers are located in China, but starting in the eighteenth century, a substantial number of Hmong speakers migrated to Vietnam, then Laos and Thailand (Culas and Micraud, 1997). Following the Indochinese conflict in the middle of the 20th century, many Hmong left Laos as refugees and emigrated to Western nations such as the United States, France, and Australia, carrying their language with them. Hmong remains a substantially under-resourced language. While there are some technologies for Hmong—for example, Google Translate (`https://translate.google.com/`) ostensibly supports it—there are few publicly available resources for the language.

This paper presents the first sizable, publicly available data resource for Hmong: the SCH corpus (`https://github.com/dmort27/sch-corpus/`). It is a collection of 20 years' worth of posts from a Hmong-oriented Usenet group cleaned and filtered by language (using a high-precision regular expression classifier). It consists of almost 12 million tokens and should be of value both to researchers interested in producing language technologies for Hmong and to linguists interested in further documenting the language.

2. The Hmong Language

There is some confusion around the term "Hmong." As used here, it refers to a dialect continuum that includes the Hmong Daw ("White Hmong") and Mong Leng/Mong Njua ("Green/Blue Hmong") varieties that are spoken in Laos and Thailand, as well as the distinct but mutually intelligible varieties that are spoken in Vietnam and China. Sometimes the term has also been used to include all of the members of the Miao nationality in China and this encompasses some lects which, while related to Hmong Daw and Mong Leng, are clearly different languages (by the mutual intelligibility criterion). The current corpus consists entirely of texts from Hmong Daw and Mong Leng speakers. These two lects are written differently (to some degree) but the differences in phonetics, phonology, lexicon, and morphosyntax are actually quite minor (on par with differences between American English lects and British RP).

2.1. Genetic Affiliation and Typology

Hmong is part of the Hmong-Mien language family, which also includes languages like Hmu, Qo Xiong, Ho Ne, Paheng, Biao Min, and Iu Mien (Mortensen, 2017). The broader genetic relationships of this family are uncertain, though a variety of hypotheses have been suggested (Mortensen, 2017). Typologically, these languages have a great deal in common with their neighbors: they are highly tonal, have isolating/analytic morphologies, have primarily head-initial syntax, and make extensive use of serial verb constructions and paratactic constructions (Mortensen, 2019). The most typologically similar major language is likely Vietnamese.

2.2. Orthography

Hmong is written with a variety of different orthographies. There is an official orthography in China based on the speech of Dananshan village in Guizhou (Wang, 1985). There is also an official orthography used in Vietnam (which is less well documented). There are various orthographies used by narrow sectors of the Thai and Lao Hmong language community, including Pahawh Hmong (which was invented by a for-

[1] Here, "Hmong" refers to the set of lects that are mutually intelligible with Hmong Daw [mww] and Mong Leng [hmj] and excludes other Southwestern Hmongic languages. See Section 2

merly illiterate Hmong farmer in 1959) (Smalley et al., 1990). However, most of the published material in Hmong—including this corpus—is written in the Romanized Popular Alphabet (RPA), an orthography developed between 1951 and 1953 by a group of American and French missionaries and their Hmong advisers. By design, it consists only of the 26 letters found on an American typewriter (despite the extraordinarily rich consonant inventory of Hmong). There are no diacritics. Tones are written with a "consonant" letter at the end of a syllable (for example, ⟨-j⟩, for a high-falling tone). The sounds to which Hmong letters correspond differ markedly from those in most other languages with Latin orthographies (for example, ⟨r⟩ indicates a voiceless retroflex stop and ⟨x⟩ corresponds to a voiceless dental fricative). As mentioned above, while the two major dialects from Laos, Hmong Daw and Mong Leng, are very similar, they are written somewhat differently. For example, in words where Hmong Daw has the consonant ⟨d⟩ [d], Mong Leng has ⟨dl⟩ [tl]. These correspondences are largely systematic but mean that a single word will often have two spellings within the corpus ('water' may be spelled as *dej* or *dlej*).

Because the orthography is very well-defined and many of the common character ngrams in Hmong RPA are uncommon in other languages with Latin alphabets (e.g., "ntx"), it is possible to identify Hmong text using a classifier based on a regular expression. We took advantage of this fact in preparing the data (§ 4.3) and also in a baseline for one of the experiments (§ 5.2).

3. Elaborate Expressions

This corpus is unique in that it includes annotations for ELABORATE EXPRESSIONS (EEs). Elaborate expressions (a term introduced by Haas (1964)) are four-part coordinate constructions with a repetitive structure (AB_1AB_2 or B_1AB_2A [2]). Consider the following examples from Hmong:

(1) a. *tag siab tag ntsws*
finish liver finish lung
'with all one's soul; satisfied'

b. *kawm ntaub kawm ntawv*
study cloth study paper
'study; pursue education'

These constructions are exceptionally common in mainland Southeast Asian languages like Hmong, Thai, Lao, Burmese, Khmer, and Lahu (Hanna, 2013; Filbeck, 1996; Johns and Strecker, 1987; Wheatley, 1982; Matisoff, 1973; Pan and Cao, 1972; Watson, 1966; Banker, 1964). In Hmong, they occur in all genres but are especially common in flowery or elevated registers (Mortensen, 2019). Some EEs are idiomatic, but **speakers coin new elaborate expressions freely, and the ability to do so is seen as a mark of a skilled rhetor**. The meaning of such expressions is usually predictable, following a construction-specific principle: the meaning of the whole is a generalization of the meanings of the two parts (in Hmong, typically AB_1 and AB_2). For example, in (1b), *ntaub* 'cloth' and *ntawv* 'paper' are the two main materials on which one writes and from which one reads, so the generalization of learning cloth and learning paper is developing literacy generally. *Ntaub-ntawv* actually exists independently as a COORDINATE COMPOUND (CC) meaning 'writing; literacy' and the non-repeated items in an EE very often (though not always) correspond to CCs.

3.1. Practical Importance

Predictable but non-compositional constructions like EEs present particular challenges to language technologies, both in terms of analysis and generation. Since EEs are so common in Hmong—indeed, since EEs are often the only idiomatic way of expressing a concept—MT and NLG systems for Hmong and other typologically similar languages must handle them fluently. The Hmong model for Google Translate, to cite one example, translates common EEs appropriately (e.g., *tas siab tas ntsws* is translated as 'satisfied' and *kawm ntaub kawm ntawv* is translated as 'study') but novel elaborate expressions are translated inappropriately. Consider (2):

(2) *thov dag thov zog*
beg labor beg strength
'request labor/help'

Google Translate renders this as 'please try hard.' In general, it seems to struggle with infrequent but semantically predictable EEs. And EEs are no less important in NLG tasks where stylistically appropriate text may require producing EEs that were not present in the training data. Being able to evaluate whether a model is generating them in appropriate contexts is equally important. Having a resource with EE annotations is an important step towards addressing this problem. Parallel text in a high-resource language would be a reasonable next step.

Dealing with EEs in Hmong might be seen as a rather niche problem. However, dealing with constructions with characteristics like EEs—semantic predictability without compositionality—is a widespread challenge and evaluating systems on their ability to correctly analyze and generate Hmong EEs can serve as one benchmark for technologies that seek to address these challenges more generally. To facilitate this, additional annotations should be added to the corpus (see § 7 below).

3.2. Theoretical Importance

A significant innovative aspect of this corpus is the annotation of EEs. EEs are not widely known, and have not been widely studied, except among specialists in languages of mainland Southeast Asia. Their theoretical significance may not be immediately obvious.

[2] These are also denoted as $ABAC$ and $ABCB$ in the literature. We use AB_1AB_2 and B_1AB_2A in this paper to highlight the coordinating structure.

The constituent parts of EEs (e.g., AB_1 and AB_2) tend to be similar syntactically and semantically, indicating that the parts *could* be reordered without a change in meaning. Nevertheless, B_1 and B_2 almost always appear in the same order. Earlier work has proposed then that the ordering of these parts (in Hmong and other languages) is determined mainly by phonology (Ting, 1975; Dai, 1986; Mortensen, 2006). Specifically for Hmong, there is a hierarchy of tones in which the tone of B_1 is always higher than the tone of B_2. However, this tonal hierarchy does not appear to be organized on phonetic grounds. Therefore, EEs in Hmong challenge the common assumptions that (1) word order is determined before phonology is applied and (2) phonology must be grounded phonetically.

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

Table 1: Phonetic values of the tones of Hmong Daw, organized according to the EE ordering scale proposed by Mortensen (2006).

Furthermore, previous work (Chomsky, 1981; Chomsky, 1995) has proposed that syntax feeds phonology, but not vice versa, and it is often assumed that phonology does not influence word order. Nevertheless, there is a large body of evidence that it, in fact, does. Examples include heavy NP shift (Ross, 1967), coordinate compounds and echo words in languages such as Japanese and Korean (Kwon and Masuda, 2019) and Jingpho (Dai, 1986), and adjective-noun order in Tagalog (Shih and Zuraw, 2017). This corpus of EEs in Hmong would add to this body of evidence, especially since there are very few exceptions to the phonological patterns found in the EEs.

In addition to assumption (1) discussed above, linguistic theory since the early 20th century has also held that sound patterns in language are based on physical phonetic properties, such as articulatory or acoustic features (Jakobson et al., 1951; Chomsky and Halle, 1968). Additionally, it is often assumed that phonological patterns that are not phonetically natural are more difficult or even impossible for speakers to learn (Hayes and White, 2013). However, more recent artificial language-learning experiments have shown mixed results but have suggested that phonological structure is more important than phonetic substance in learning phonological patterns (Moreton and Pater, 2012a; Moreton and Pater, 2012b). A corpus of EEs in Hmong could also challenge the assumption that phonology must be grounded phonetically, if it could be shown that the order of words within EEs can be predicted from the proposed phonological patterns.

A preliminary study using the SCH Corpus to explore these questions has been completed in Cui et al. (2022). However, a great deal remains to be done in this area.

4. The Corpus

The corpus consists of about 860k sentences (12M tokens) with about 25k EEs. It is freely available at `http://www.github.com/dmort27/sch-corpus`. More detailed numbers are provided in Table 2.

Tokens	11,822,652
Sentences	858,635
Elaborate Expressions	24,574
Tokens inside EEs (count)	98,296
Tokens inside EEs (%)	0.8

Table 2: Statistics for the SCH Corpus.

4.1. Genre and Domain

The Hmong data on which the corpus is based were collected from the Usenet group `soc.culture.hmong` (or SCH). This newgroup, which still exists but which has fallen largely inactive, was used primarily by members of the Hmong-American community (but with participation by members of the Hmong communities in France and Australia). Most posts were one or more paragraphs long and often included extensive quotations from earlier posts by other users. The discourse conventions were highly dialogic and lively threads sometimes continued for months or even years. The participants, at first, were largely younger Hmong who were at educational institutions with access to the Internet. As Internet access broadened, so did the range of interlocutors in SCH. Frequent topics included current events, politics (especially regarding Hmong leaders like General Vang Pao and Dr. Yang Dao—the first Hmong person to receive a PhD and a frequent critic of the general), religion (the competition between traditional animism and various forms of Christianity), family and social issues, and relationships. Discussions were animated and often acrimonious, with personal rivalries extending across the years.

4.2. Data Quality

The data are based on user-generated text in a relatively informal forum. As a result, they are similar in quality to text from a contemporary social media platform—they are noisy and written in a familiar register. Orthographic variation is pervasive. This is increased as a result of competing orthographic standards within the community. In addition to the older, more widespread, orthographic variety in which all prenasalized obstruents are written as a sequence of ⟨n⟩ + obstruent (e.g., ⟨np⟩) some such obstruents are written with a single character (e.g., ⟨b⟩).

Codeswitching with English and Lao is frequent in the data. While whole posts in languages other than Hmong have been largely filtered out, individual words and phrases from words other than Hmong occur very frequently (e.g., the English word *pepsi* for 'soda'). Loanwords and codeswitching occur in the dataset more frequently than in formal written Hmong, but not, impressionistically speaking, more than in casual speech.

4.3. Data Collection

The `soc.culture.hmong` posts were scraped from Narkive (a mailinglist archive) (`https://narkive.com/`). All of the contents of the archive (`https://soc.culture.hmong.narkive.com/`) from 1996 to 2016 were extracted and processed. Processing took four steps:

1. Quoted text (identified using HTML markup) was removed.
2. Plain text was extracted from HTML.
3. The text was segmented into sentences using the NLTK Punkt tokenizer trained on the untokenized corpus (Kiss and Strunk, 2006), tokenized using the NLTK 3.6.3 `word_tokenize` function (Bird et al., 2009), and structured in a CONLL-like format (forms in the first column, annotations in subsequent tab-separated columns, sentence boundaries indicated by empty lines).
4. Documents were filtered by language: documents in which over 60% of tokens were not recognized by a regular expression encoding the orthographic possibilities of Hmong RPA were excluded. In the judgment of the first author, a trained linguist and proficient speaker of Hmong, the vast majority of the remaining documents were written primarily in the subject language (though code-switching into English and other languages still occurs frequently in the corpus).

Individual contributors were not contacted for permission to include their posts. Instead, posts are reproduced and redistributed under the same assumptions as have been made by the original Usenet network and by later distributors of Usenet content including Google and Narchive, namely:

- Users intended their content to be freely available and distributed widely on the Internet.
- Requests from users to remove their content from the corpus/archive should be complied with immediately and thoroughly.

We believe that these data collection policies safeguard the interests of both the authors and potential consumers of the posts.

4.4. Ethical Issues

All metadata and headers were removed from all posts, leaving only the body text. This reduces the usefulness of the corpus for some purposes (e.g., discourse analysis) but it provides considerable anonymity. The individual authors of posts are not marked in the corpus. It is also difficult to determine which posts in a thread were written by the same author. Since the data is already publicly available, and the identity of users that do not use pseudonyms is, in most cases, readily ascertainable from the `narchive.com` and Google Groups archives, no other attempt was made to remove all references to private persons from the text. In order to recover the identities of users from the corpus, when that user is referred to by another user by name, it is necessary to do the following:

1. Identify the name of a user U in a response R. These occur but are not highly frequent.
2. Identify the post P to which a user is responding.
3. Ascertain that the U is believed by the author of R to be the author of P.

We judge it to be unlikely that the publication of the SCH corpus will increase the danger of adverse events to users above that which is already posed by the publicly available `narchive.com` and Google Groups archive. We foresee one possible exception: the publication of a corpus based on the contents of the newsgroup may make some members of the Hmong community who were not previously familiar with the SCH forum aware of it. This may reignite conflicts previously litigated on SCH—especially given the heated nature of much of the discussion there—propagating this resentment into other sectors of the Hmong community and leading to bias against some participants based on their past participation in SCH discussions. We consider this to be unlikely, however, and believe that the benefits of the publication of this corpus to the Hmong community and the scientific and engineering communities outweigh the potential risks. Prior to public release, the corpus has already served as a source for some pedagogical materials for heritage learners of Hmong, including word- and collocation-frequency lists. It can also serve as the basis for fundamental language technologies for Hmong speakers, if it is handled well.

The corpus contains some potentially offensive language (discussion of sexual topics, racist and sexist discourse, and inflammatory accusations against public figures). The decision was made *not* to remove these posts for two reasons:

1. We found it difficult to formulate consistent, culturally neutral standards for determining what language should be included and what language should be excluded.
2. We believe that the study of, and development of NLP tools for, offensive language is valuable (for low-resource as well as high-resource languages and domains). We believe that data like SCH corpus can be an important resource for studying this kind of speech.

Annotation of abusive language would increase the usefulness of the SCH Corpus (see §7). Until then, those who use it should do so with the understanding that models trained on it will reflect the sometimes offensive assumptions and language of the participants who produced the posts.
Fortunately, to our knowledge, the corpus does not contain any language that is, in the strict sense, libelous.

4.5. License

The corpus annotations are distributed under terms of the Creative Commons CC0 1.0 Universal License. The underlying corpus data may be freely distributed and used.

4.6. Annotation

EE annotations were performed on the entire corpus by the first author using custom annotation software. The program scanned the corpus for AB_1AB_2 sequences (the rarer B_1AB_2A sequences were not annotated, since they were considerably more sparse in the corpus) and presented them to the annotator with a four-word context to the left and the right. The annotator made a binary choice regarding whether the AB_1AB_2 sequence was an EE. The tests used to determine whether an AB_1AB_2 sequence was an EE were as follows:

1. B_1B_2 exists independently as a coordinate compound
2. The *syntactic* relationship between B_1 and the context is the same as the relationship between B_2 and the context
3. The *semantic* relationship between B_1 and the context is the same as the relationship between B_2 and the context

If a sequence satisfied (1), both (2) and (3), or all three tests, it was annotated as an EE. Roughly three quarters of AB_1AB_2 sequences were discarded as a result of these tests.
These annotations are represented in a CONLL-like format: Each token is placed on one line in a text file. The fields are separated by tabs. The first field contains the surface form (the Hmong word). The following field contains a `BIO` (begin-inside-outside) tag for EEs. A sentence annotated in this fashion is given in Figure 1.

```
yav     O
tag     O
los     O
nej     O
twb     O
hais    O
tias    O
cov     O
laus    O
no      O
tsi     B
txawj   I
tsi     I
ntse    I
thiaj   O
li      O
coj     O
tsis    O
tau     O
hmoob   O
no      O
nes     O
!       O
```

Figure 1: An annotated sentence from the corpus ('In the past you did say that these elders were unintelligent and therefore incapable of leading the Hmong!'). It includes the EE *tsi txawj tsi ntse* 'not capable not sharp; unintelligent.'

5. Experiments

In order to evaluate the usefulness of the corpus for performing NLP tasks and investigating linguistic hypotheses, we conducted a few experiments. First, we performed a qualitative analysis of skip-gram embeddings trained on the corpus using a set of 14 four-word analogies developed by the first author. This is intended to evaluate the suitability of the corpus as data for training word embeddings and other similar kinds of models. Second, we measure the quality of the elaborate expression annotations by training neural sequence labeling models (with feature extraction performed either by a BiLSTM or a CNN) to add appropriate BIO tags to a held-out test set.

5.1. Evaluating Word Embeddings Trained on the Corpus

We trained a Word2Vec skip-gram model (Mikolov et al., 2013a) on the SCH corpus and manually evaluated the word embeddings with a word analogies task. We trained the model using *Gensim* 4.0's API (Řehůřek and Sojka, 2010; Řehůřek and Sojka, 2011) for 5 epochs to produce 100-dimension word embeddings. Because most Hmong words are compounds, but morphemes are treated orthographically as "words", it is rather difficult to generate word analogies based on individual tokens. As a result, we only analyze the results with 14 examples as in Table 3, instead of reporting accuracy on a large test set. Despite the small size of evaluation data, we believe that this task provides meaningful insight into the quality and limitations of the corpus as training data for word embeddings and neural language models.
An analogy, as we use the term, is the relationship described by "[Word 1] is to [Word 2] as [Word 3] is to [Word 4]". To predict Word 4 given other three words, we can find the word whose embedding is most similar to the vector $e[\text{Word 2}] - e[\text{Word 1}] + e[\text{Word 3}]$, where $e[w]$ returns the embedding of word w (Mikolov et al., 2013b). We consider the top 10 results in our analysis. 7 out of the 14 examples have the exact word included

Word 1	Word 2	Word 3	Word 4	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***

Table 3: 14 example analogies, their English translations, and prediction results for Word 4. Predictions which match the gold standard are given in bold italic (the 7 green rows). Predictions which accurately complete the analogy but which were not in the gold standard are given in italics (2 blue rows). Analogies for which no reasonable prediction was made are labeled with "(none)" (5 white rows).

in the top 10. In fact, 6 of them have the answer in the top 5. 2 more examples elicited words that complete the analogy but do not match the original gold standard we produced. These words, and all of the other results from this experiment, are listed in Table 3. Some interesting observations follow:

1. When *tub* 'son/boy' is expected, *vauv* 'son-in-law' is also among the top words. Especially in the 'mother' : 'father' :: 'daughter' : 'son' analogy, it means that the model also captures the marriage analogy apart from the parent-child relationship and the gender.

2. The word *me* 'small' can occur both before or after nouns. Before nouns it is EVALUATIVE like diminutives in many other languages (often expressing affection). When it occurs after a noun, it refers simply to the physical dimension. It has two antonyms: *niag*, which occurs before nouns and has a (usually derogatory) augmentative meaning and *loj*, which occurs after nouns and refers to physical largeness. The model predicts *niag* but not *loj*. This may reflect the nature of the corpus in which emotional evaluations of things are more common than descriptions of their size, meaning that evaluative relationships are better captured.

3. For things to drink (in the analogy 'eat' : 'food' :: 'drink' : 'water'), the model predicts diverse beverages: *coffee*, *pepsi* 'soda', *cawv* 'liquor', and *npias* 'beer' but not *dej* 'water', perhaps reflecting an overestimation, on the part of the authors, of the health-consciousness of participants in the SCH group.

4. When *nroog* 'city' is the expected word (in 'hill' : 'mountain' :: 'village' : X), many of the other predictions are cities in Laos, which provides further evidence that the embeddings are capturing aspects of the semantics of Hmong.

It is clear that performance on this task with the SCH corpus as training data is not as high as has been reported for other languages with larger, more curated, corpora. On the other hand, the results are clearly much better than chance.

5.2. Tagging Elaborate Expressions in Context

Experiment We investigate whether the `BIO` tags in this dataset can be learned by a neural sequence labeling model, i.e., whether the elaborate expressions can be recognized in context. Since over 99% of all tokens have the `O` label, we report precision, recall, and F1 score instead of accuracy. We split the dataset into training, validation, and test sets such that EEs in the validation and test sets do not overlap with EEs in the training set. This way, the model cannot simply memorize any AB_1AB_2 construction as an EE; rather, it would need to learn the contextual and distributional patterns in order to tag the unseen EEs. We independently produce three such splits to minimize the variance.

Hyperparameter	Value
Word embeding dimensions	100
LSTM hidden dimensions	100
CNN hidden dimensions	200
CNN kernel size	3
Dropout probability	0.5

Table 4: Model configuration hyperparameters.

The sequence labeling model we train consists of a token embedding layer, a feature extractor to process the sentences into features, and a fully connected output layer to predict the `{B, I, O}` tag for each token. We experiment with two feature extractors: 1) a bidirectional LSTM (Hochreiter and Schmidhuber, 1997) and 2) a 4-layer CNN [3] (LeCun et al., 1989; Collobert et al., 2011). Model configuration hyperparameters are listed in Table 4. The model is trained with an SGD optimizer with a momentum of 0.9, batch size of 64, and learning rate of 0.02 for as many epochs as needed until the F1 score stops improving for 10 epochs.

To establish a baseline for comparison, we use a rule-based classifier on each window of four tokens. We report the results by gradually applying the following four filters:

1. The four words are of the form AB_1AB_2
2. The four words are proper Hmong RPA syllables parsable by a regular expression classifier
3. The word vector similarity[4] between B_1 and B_2 is above α. A grid search is performed to find the α with the highest F1 score. We use $\alpha = 0.4$
4. B_1 and B_2 follow the tonal ordering pattern proposed in (Mortensen, 2006)

Results Table 5 shows the sequence tagging results of the baseline and neural models. For the neural models, we report precision, recall, and F1 scores averaged over 9 runs (three independent data splits for three initial seeds each). The simplest baseline model achieves 100% recall, as expected, albeit at the cost of very low precision. As more filters are added, the F1 score begins to improve, and the full baseline model is able to achieve a rather reasonable performance of 66.66 (F1). However, the neural taggers are able to beat the baseline considerably, with higher values for precision and recall. In particular, the CNN feature extractor outperforms the BiLSTM. This is possibly because the CNN kernel is able to capture the AB_1AB_2 structure in the elaborate expressions better than the BiLSTM, which reads in text linearly. Identification of EEs also requires only local context, so it does not benefit from an LSTM's ability to utilize global word context information (Yang et al., 2018). For all models, recall is higher than precision, suggesting that more AB_1AB_2 constructions are mistaken as EEs than actual EEs being mislabeled. We hope the results presented here provides a reasonable starting point that inspires future research in Hmong elaborate expressions.

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**

Table 5: Results on the test set (average over 3 data splits for 3 runs each) for tagging elaborate expressions in context.

6. Discussion

The SCH corpus, while small, is adequate to perform certain useful and interesting NLP tasks and computational linguistics experiments. Minus the annotations, it can be used to train word embeddings that encode meaningful semantic relationships. While these embeddings are not as high quality as those trained on larger corpora (i.e., they cannot solve analogies as accurately), a qualitative analysis shows that the semantics captured by Word2vec embeddings trained on the SCH Corpus are impressionistically reasonable. This suggests that they might contribute meaningfully to realistic NLP tasks. There have been many new methods proposed in recent years to generate better word embeddings (e.g. GLoVE (Pennington et al., 2014), BERT (Devlin et al., 2019)). We chose Word2vec because of its simplicity and its robustness to a smaller training dataset, unlike a transformer-based method.

The SCH Corpus is not large, but compared to corpora for other languages with less than 10 million speakers, it is respectable in both size and quality. It is also important because of the elaborate expression annotations. There is no publicly available corpus with these

[3] Four blocks of *Conv-ReLU-Dropout-Batchnorm* layers

[4] Cosine similarity between 100-dimensional skip-gram embeddings trained on the SCH Corpus.

constructions labeled for any language (including languages with much larger numbers of speakers like Thai, Burmese, and Khmer). We have shown that these annotations are sound enough that they can be used for non-trivial labeling tasks.

7. Future Directions

A small monolingual corpus with limited annotations is still of value when little other data is available. While further work could be done in terms of applying the SCH Corpus to downstream tasks, it is already clear that the corpus can contribute meaningfully to Hmong HLT. However, additional annotations could increase its value significantly. In the future, we plan to annotate part of the corpus using the Universal Dependencies schema (Nivre et al., 2016; Nivre et al., 2020). The annotations will include dependency relations and part-of-speech labels, allowing a variety of other experiments with Hmong. These could include both linguistically oriented investigations and explorations of improved NLP. Furthermore, we hope to annotate the corpus for abusive and offensive language. This will be useful in studying the detection of sensitive language in low-resource settings.

8. Bibliographical References

Banker, E. M. (1964). Bahnar reduplication. *Mon-Khmer Studies*, 1:119–134.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.

Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English*. Harper& Row, New York.

Chomsky, N. (1981). *Lectures on Government and Binding: The Pisa Lectures*. Studies in Generative Grammar. de Gruyter.

Chomsky, N. (1995). *The Minimalist Program*. MIT Press.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12(null):2493–2537, nov.

Cui, C., Zhang, K. J., and Mortensen, D. R. (2022). Learning the ordering of coordinate compounds and elaborate expressions in Hmong, Lahu, and Chinese. In *Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Seattle, USA, July, to appear.

Culas, C. and Micraud, J. (1997). A contribution to the study of Hmong (Miao) migrations and history. *Bijdragen tot de taal-, land-en volkenkunde*, 153(2):211–243.

Dai, Q. (1986). Jingpo yu binglie jiegou fuheci de yuanyin hexie. *Minzu Yuwen*, 1986(5):23–29.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Filbeck, D. (1996). Couplets and duplication in Mal. *Mon-Khmer Studies*, 26:91–106.

Haas, M. R. (1964). *Thai-English Student's Dictionary*. Stanford University Press, Stanford.

Hanna, W. J. (2013). Elaborate expressions in Dai Lue. *Linguistics of the Tibeto-Burman Area*, 36(1):33–56.

Hayes, B. and White, J. (2013). Phonological Naturalness and Phonotactic Learning. *Linguistic Inquiry*, 44(1):45–75, 01.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, 9(8):1735–1780, nov.

Jakobson, R., Fant, G., and Halle, M. (1951). *Preliminaries to Speech Analysis: The Distinctive Features and their Correlates*. MIT Press, Cambridge, Massachusetts.

Johns, B. and Strecker, D. (1987). Lexical and phonological sources of Hmong elaborate expressions. *Linguistics of the Tibeto-Burman Area*, 10(2):106–112.

Kiss, T. and Strunk, J. (2006). Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32:485–525.

Kwon, N. and Masuda, K. (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.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551.

Matisoff, J. A. (1973). *The Grammar of Lahu*. Number 75 in University of California Publications in Linguistics. University of California Press, Berkeley.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In Yoshua Bengio et al., editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Mikolov, T., Yih, W., and Zweig, G. (2013b). Linguistic regularities in continuous space word representations. In Lucy Vanderwende, et al., editors, *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 746–751. The Association for Computational Linguistics.

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

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

Mortensen, D. R. (2006). *Logical and Substantive*

Scales in Phonology. Ph.D. thesis, University of California, Berkeley.

Mortensen, D. R. (2017). Hmong-Mien languages. In *Oxford research encyclopedia of linguistics*. Oxford University Press.

Mortensen, D. (2019). Hmong (Mong Leng). In *The Mainland Southeast Asia Linguistic Area*, pages 609–652. De Gruyter Mouton.

Nivre, J., de Marneffe, M., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R. T., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal dependencies v1: A multilingual treebank collection. In Nicoletta Calzolari, et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

Nivre, J., de Marneffe, M., Ginter, F., Hajic, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F. M., and Zeman, D. (2020). Universal dependencies v2: An evergrowing multilingual treebank collection. In Nicoletta Calzolari, et al., editors, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 4034–4043. European Language Resources Association.

Pan, Y. and Cao, C. (1972). Four-syllable coordinative constructions in the Miao languages of eastern Kweichow. In Herbert C. Purnell, editor, *Miao and Yao Linguistic Studies*, number 88 in Data Papers, pages 211–234. Cornell University Southeast Asia Program, Ithaca, New York.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. `http://is.muni.cz/publication/884893/en`.

Řehůřek, R. and Sojka, P. (2011). Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masary University, Brno, Czech Republic*, 3(2).

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

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

Smalley, W. A., Vang, C. K., and Yang, G. Y. (1990). *Mother of writing: the origin and development of a Hmong messianic script*. University of Chicago Press.

Ting, P.-h. (1975). Lunyu, Mengzi, ji Shijing zhong binglie yu chengfen zhijian de shengdiao guanxi [tonal relationship between the two consistuent of the coordinate construction in the Analects, the Mengtze, and the Book of Odes]. *Bulletin of the Institute of History and Philology, Academia Sinica*, 47(1):17–52.

Wang, F. (1985). *Miao yu jianzhi [Sketch grammar of the Miao language]*. Minzu Chubanshe, Beijing.

Watson, R. L. (1966). Reduplication in Pacoh. Master's thesis, The Hartford Seminary Foundation.

Wheatley, J. (1982). *Burmese: A Grammatical Sketch*. Ph.D. thesis, University of California, Berkeley.

Yang, J., Liang, S., and Zhang, Y. (2018). Design challenges and misconceptions in neural sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3879–3889, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.