Language Resources and Evaluation (2022) 56:1315–1332
https://doi.org/10.1007/s10579-022-09596-2

PROJECT NOTES

The Hmong Medical Corpus: a biomedical corpus for a minority language

Nathan M. White[1]

Accepted: 16 May 2022 / Published online: 15 July 2022

Abstract

Biomedical communication is an area that increasingly benefits from natural language processing (NLP) work. Biomedical named entity recognition (NER) in particular provides a foundation for advanced NLP applications, such as automated medical question-answering and translation services. However, while a large body of biomedical documents are available in an array of languages, most work in biomedical NER remains in English, with the remainder in official national or regional languages. Minority languages so far remain an underexplored area. The Hmong language, a minority language with sizable populations in several countries and without official status anywhere, represents an exceptional challenge for effective communication in medical contexts. Taking advantage of the large number of government-produced medical information documents in Hmong, we have developed the first named entity-annotated biomedical corpus for a resource-poor minority language. The Hmong Medical Corpus contains 100,535 tokens with 4554 named entities (NEs) of three UMLS semantic types: diseases/syndromes, signs/symptoms, and body parts/organs/organ components. Furthermore, a subset of the corpus is annotated for word position and parts of speech, representing the first such gold-standard dataset publicly available for Hmong. The methodology presented provides a readily reproducible approach for the creation of biomedical NE-annotated corpora for other resource-poor languages.

Keywords Biomedical corpus · Named entity recognition · Language models · Machine learning · Hmong · Minority languages

✉ Nathan M. White
nathan.white1@my.jcu.edu.au

[1] Language and Culture Research Centre, James Cook University, 14-88 McGregor Rd., Smithfield, QLD 4878, Australia

1 Introduction

Effective biomedical communication with minority linguistic groups remains elusive, due to a lack of awareness regarding appropriate means to express various medical concepts. Natural language processing (NLP) can provide a solution through automating the identification of equivalent expressions in documents composed in the language, especially where these documents are translations of English or other majority-language originals. Named entity recognition can be exploited for the development of translation standards for medical terminology, which are necessary for medical practitioners and interpreters to ensure effective communication. Likewise, information extraction involving named entities is an important step in the development of question-answering systems for minority-language speakers and automated translation services for practitioners who lack access to a qualified, licensed interpreter.

Nevertheless, biomedical NLP applications involving named entities generally remain limited to national languages, with a strong bias in favor of English, and several other official languages are at best minimally represented. This is due to a virtually non-existent supply of medically-annotated data that could be used in NLP contexts, a problem that naturally applies to the Hmong language.

The Hmong are an ethnic group of approximately 4.5 million native to Southeast Asia, with major population centers in China, Vietnam, Laos, and Thailand (Lemoine, 2005). In the past 35 years, a Hmong refugee diaspora has emerged in several developed Anglophone countries, including the United States, Canada, and Australia, with an ethnic population of more than 260,000 in the United States alone (Pfeifer et al., 2010).

The Hmong refugee community is markedly affected by communication problems between English-speaking medical practitioners and Hmong patients (Fadiman, 1998; Thornburn et al., 2012). This especially involves difficulties in finding appropriate terminology or other expression of medical concepts in the Hmong language (Johnson, 2002).

The Hmong Medical Corpus addresses this problem by providing an annotated corpus of medical information documents translated by native speakers in an official, government-recognized capacity from English originals. These documents are annotated for parts of speech (POS) as well as named entities (NEs) of several categories most affected by the communication issues mentioned above: body part/organ/organ component terms, names of diseases/syndromes, and names of signs/symptoms.

This paper makes three contributions: (1) the presentation of the construction of the Hmong Medical Corpus, including its part-of-speech and named entity tags; (2) the presentation of a viable, reproducible methodology for producing annotated biomedical corpora for resource-poor minority languages; and (3) the release of the corpus for general use.

The remainder of the paper is comprised of ten sections. Section 2 provides a brief review of related work, while Sect. 3 provides a typological overview of the Hmong language to orient the reader. This is followed by an overview of the

corpus and data collection in Sect. 4, details of the annotation scheme in Sect. 5, the process of annotation development in Sect. 6, and details regarding the word position/POS tagger model in Sect. 7. Section 8 provides a brief description of the limitations of the project, while Sect. 9 gives corpus statistics. Section 10 gives an overview of corpus contributions, and Sect. 11 provides a summary of conclusions.

2 Related work

We briefly review corpora containing named entity annotations in the medical domain, focusing on both English and non-English corpora.

An array of English corpora involving named entity recognition has emerged in the last twenty years. An early corpus is GENIA (Kim et al., 2003), which provides annotations of biological entities. CADEC (Karimi et al., 2015) provides annotations for diseases, symptoms, drugs, adverse effects, and findings found in 1,253 medical forum posts. The NCBI corpus (Doğan et al., 2014) contains a set of 793 abstracts derived from PubMed with annotations for disease concepts and names. The CRAFT corpus (Bada et al., 2012) is a set of 97 biomedical journal articles with annotations for an array of named entity types. The CHQA corpus (Kilicoglu et al., 2018) is a set of 2,614 health-related questions from both web and email sources, with annotations for named entities and several other semantic types. The CHEMDNER corpus (Krallinger et al., 2015) includes 10,000 abstracts sourced from PubMed with annotations that focus on drugs and chemical compounds. The i2b2 Shared Tasks have also involved several corpora (Uzuner et al., 2010, 2011; Stubbs & Uzuner, 2015), which provide medical named entity annotations in a number of areas, including medications and their administration as well as health conditions associated with heart disease.

Several non-English corpora have also emerged in recent years. A recent Spanish corpus, the PharmaCoNER corpus (Gonzalez-Agirre et al., 2019), is a collection of 1,000 clinical case studies with annotations for chemicals, proteins, genes, and other biomedical or clinical substances. Two other major Spanish corpora, the DrugSemantics corpus (Moreno et al., 2017) and the IxaMedGS corpus (Oronoz et al., 2015), provide support for a range of named entity types, especially drugs and diseases. Recent French corpora include the MERLOT corpus (Campillos et al., 2018), which is comprised of 500 documents and provides named entity annotations for symptoms and disorders among several other categories, as well as the Quaero corpus (Névéol et al., 2014), which provides annotations for several medical categories based on the Unified Medical Language System (UMLS; Lindberg et al., 1993; Bodenreider, 2004). Other languages recently represented in medical corpora with named entity annotations include Chinese (Gao et al., 2019) and Romanian (Mitrofan et al., 2019).

To date, there have been no prior annotated medical corpora for a low-resource, minority language such as Hmong.

3 Typological overview

The Hmong language possesses a number of relatively special typological features that merit discussion to orient the reader here.

Typologically, Hmong tends toward a one-to-one correspondence between syllable, morpheme, and word, with a relatively limited number of affixes and some compounding. For example, one relatively technical Hmong text sampled by White (2020) contains 805 grammatical words with 727 of these monosyllabic (at 90.3% of the total). Furthermore, Hmong possesses two phenomena that present a challenge for marking word boundaries as they are technically intermediate between a word and a phrase given their behavior in syntax: coordinate compounds and four-syllable elaborate expressions (White, 2020; cf. Wälchli, 2005). Hmong likewise contains a number of combinations that cohere as grammatical words but where the parts might be otherwise predicted to be independent words on their own, such as *ib-tug* 'one-CLASSIFIER:ANIMATE' (White, 2020; cf. Ratliff, 2009). The issues these phenomena raise and the solution pursued are referenced in Sect. 5 below.

In addition, as with a number of other Southeast Asian languages, Hmong exhibits a tendency toward a lack of obligatoriness, where grammatical categories (aspect, mood, etc.) are typically not obligatorily marked (Bisang, 2015, *inter alia*). There exists instead a high degree of reliance on pragmatic inference and multifunctionality, where the same form or construction may have several functions which are differentiated by context alone (Bisang, 2015). This presents a challenge to straightforward classification of parts of speech, and led to the method of asking critical questions to determine part-of-speech class in potentially ambiguous cases, as mentioned in Sect. 6 below.

Finally, several Hmong part-of-speech classes are relatively unusual. Adjectives as a cohesive part-of-speech class are a relatively limited set of eight words (Bisang, 1993), with the other property concepts (term following Post, 2008) represented by stative verbs; this affects the part-of-speech results for adjectives provided in Sect. 9 below. Hmong also contains classes common for Southeast Asian languages but relatively uncommon elsewhere such as nominal classifiers (Bisang, 1993; White, 2019), verb classifiers (Gerner, 2014), and localizers (Xiong & Cohen, 2005).

4 Overview of corpus and data collection

The U.S. state governments of Wisconsin[1] and Minnesota[2] have produced a significant number of medical informational documents in the Hmong language produced by native speakers translating from English into Hmong. The Hmong Medical Corpus is a collection of 105 of these documents. These documents were taken from a range of government-sourced documents on medical, health insurance, and other health-related topics, obtained through a web crawler where permitted. The

[1] https://www.dhs.wisconsin.gov/.

[2] https://www.health.state.mn.us/communities/translation/hmong.html.

documents were selected based on their coverage of disorders or pathogens, the associated symptoms, and their treatments. Given their focus on specific illnesses, these documents are all genre-specific to the biomedical domain.

The annotation process took the form of two components: combined word position/POS tagging and named entity tagging. First, we obtained combined word position and POS tags. As standard Hmong orthography places spaces between syllables rather than words, word segmentation is less than trivial. The word segmentation task was therefore treated as a sequential tagging task, as previously done for Vietnamese (Nguyen et al., 2006; Dinh et al., 2008), which exhibits the same syllable-based spacing. As the POS-tagging task is likewise a sequential tagging task, the two tasks were combined using two tags separated by a hyphen, as has been done for Chinese (Kruengkrai et al., 2009; Shao et al., 2017) and Vietnamese (Takahashi & Yamamoto, 2016; Nguyen et al., 2017). These combined tags were assigned by syllable rather than by word.

Second, named entity tagging took the form of one tag per syllable, based on three sets of labels derived from semantic types found in the UMLS Semantic Network: body part/organ/organ component terms, names of diseases/syndromes, and names of signs/symptoms.

5 Annotation scheme

The combined word position/POS tag scheme involves two components separated by a hyphen, where the word token annotation is the first element, and the POS tag the second. The word position portion takes one of three values: B (beginning of a word), I (inside of a word), or O (other). The POS tag portion takes one of the values in Table 1, based on the POS categories specifically identified in Hmong as part of ongoing analytical work. The tagset is significantly adapted from that of the Penn Chinese Treebank (Xue et al., 2005).

As Hmong orthography exhibits syllable spacing and community practices show a tendency to experiment with word-based spacing with the category of "word" ill-defined in practice, POS tags in this scheme are specific to the morpheme, rather than the word. For example, the verb *sib txawv* 'differ from one another' is composed of a verb-modifying derivational prefix *sib–* 'RECIPROCAL' and the verbal root *txawv* 'differ'; this is tagged in the scheme as *sib/B-AD txawv/I-VV*.

An example of a sentence with the resulting annotations is as follows: *Cia/B-VV tus/B-CL me/B-NN nyuam/I-NN nyob/B-VV hauv/B-LC tsev/B-NN es/B-CS txhob/B-AD mus/B-VV kawm/B-VV ntawv/B-NN ./O-PU* ("Let the child stay at home and not go and study.").

The named entity tags likewise have two parts separated by a hyphen: an IOB tag portion that indicates position in the named entity of one of four types (B, I, E, or O, where E is 'end'), and a named entity category selected (if the position is not O) from those found in Table 2.

The three categories as shown in Table 2 were selected for the following reasons: (1) the vast majority of documents selected had at least two of these categories robustly represented, and (2) these categories provide the best basis to create

Table 1 POS tag categories

Tag	Category	Examples
CL	Nominal classifier	*lub* 'GENERAL CLASSIFIER'
CV	Verbal action classifier	*vuag* 'TIME(S)'
NN	Common noun	*kab mob* 'illness'
PN	Pronoun	*nws* 'he, she, it'
NR	Proper noun	*Mes kas* 'America'
JJ	Adjective	*me* 'little'
VV	Verb	*kis* 'to contract'
DT	Demonstrative	*no* 'this'
QU	Quantifier	*ib* 'one'
LC	Localizer	*ntawm* '(at) nearby'
RL	Relative-like modifier	*twg* 'which'
AD	Verbal modifier (adverb-like)	*yuav* 'IRREALIS'
PP	Preposition	*rau* 'to, for'
CC	Coordinating conjunction	*thiab* 'and'
CS	Subordinating conjunction	*tias* 'COMPLEMENTIZER'
CM	Clause-final marker	*ne* 'THEMATIC MARKER'
ON	Onomatopoeia	*hawb* 'COUGHING SOUND'
FW	Foreign word	*bleach*
PU	Punctuation	,

Table 2 NE tag categories

Tag	Category	Example
BPOC	Body part, organ, or organ component	*plab* 'stomach' *pob txha* 'bone'
DSYN	Disease or syndrome	*mob ntsws qhuav* 'tuberculosis' *mob ruas* 'leprosy'
SOSY	Sign or symptom	*hnoos* 'cough' *kub ib ce* 'have a fever'

an automated question-answering system to improve Hmong community access to medical information—a longer-term goal of the Hmong Medical Corpus project.

Named entities of the above three categories can be nested in Hmong: for example, a disease name often contains a body part term, as in *ntsws* 'lung' in *mob ntsws qhuav* 'tuberculosis', or a symptom can include a body part term, as in *tob hau* 'head' in *dias tob hau* 'be dizzy'.

This nesting is handled through the creation of separate files, where each contains NE labels of exactly one category. This allows for layering and combining of labels as appropriate for downstream NLP tasks. Other possible situations that could result in tag conflicts, such as overlap of named entities, did not present an issue.

An extended sample from the Hmong Medical Corpus with both word position/POS tags and NE tags is provided in Appendix A.

Table 3 Performance of NE tagging algorithm compared against the final gold-standard annotations

Metric	BPOC Position Tags	DSYN Position Tags	SOSY Position Tags	All Tags
Accuracy	0.9412	0.9789	0.9263	0.9488
Micro Precision	0.9412	0.9789	0.9263	0.9488
Macro Precision	0.2459	0.7084	0.2698	0.4394
Micro Recall	0.9412	0.9789	0.9263	0.9488
Macro Recall	0.2439	0.8562	0.2753	0.4585
Micro F1	0.9412	0.9789	0.9263	0.9488
Macro F1	0.2428	0.7713	0.2721	0.4456

6 Annotation development

The combined word position/POS tag annotations were developed through a two-stage process. In the first stage, a linguistic expert provided the initial POS tags; only one expert was available given the resource-marginalized status of the language. This process was guided by an ever-evolving annotation guidelines document, with revisions made to the full set of documents as necessary changes in the guidelines were identified. Potentially ambiguous cases were then checked with Hmong community collaborators, who were asked critical questions based on part-of-speech criteria specific to Hmong to determine the best possible tag. The linguistic expert then revised the tags based on the answers.

In the second stage, a tagger with a Bidirectional Long-Short Term Memory (BiLSTM; Schuster & Paliwal, 1997; Hochreiter & Schmidhuber, 1997) model architecture (described below) was trained on the documents tagged in the first stage. New documents were then tagged using the automated tagger, and these results were manually verified by the linguistic expert in consultation with Hmong community collaborators, as in the first stage. As new documents were tagged, the BiLSTM tagger was retrained with the new data, and the result applied to additional documents.

The named entity tags were generated based on the creation of curated lists for body part/organ/organ component terms, names of diseases/syndromes, and names of signs/symptoms. The curated lists of diseases/syndromes and signs/symptoms were obtained in raw form through an algorithm. These were initially drawn from semi-structured sections of those medical information documents that encoded this sort of information in a relatively consistent way across the documents. These lists were then modified and expanded manually to represent more general cases. The list for body part/organ entities was developed through the review of a range of Hmong dictionaries, published linguistic research, and consultation with Hmong community collaborators. These lists were then used to algorithmically tag the full corpus. The results were manually verified by an expert in the language.

The algorithm's performance versus the final gold-standard annotations is presented in Table 3, both in terms of the correct choice of position tags for each named entity category (e.g., B-BPOC, I-BPOC, E-BPOC, O for the BPOC category) and overall.

Table 4 Performance metrics of the final version of the BiLSTM model on the final document

Metric	Result
Accuracy	0.9544
Precision:	
Micro	0.9544
Macro, non-predicted = 0	0.8457
Macro, non-predicted = 1	0.9409
Recall:	
Micro	0.9544
Macro, non-true = 0	0.8045
Macro, non-true = 1	0.8522
F1 score:	
Micro	0.9544
Macro	0.8198

The relatively weaker performance for BPOC terms is likely due to the high number of homophones in Hmong involving pairs of named entity terms and non-terms, such as *rau* '(finger/toe)nail' versus *rau* 'six', or *siab* 'liver' versus *siab* 'be high'. SOSY terms are likewise affected since they often contain BPOC terms in Hmong; this is in addition to the homophone issue, such as with *raws* in *raws plab* 'have diarrhea' versus *raws* 'be according to'.

7 Word position/POS tagger

The combined word position/POS tagger used to automate the annotation work described above was trained as a BiLSTM model with a hidden BiLSTM layer of size 256 and trained for 50 epochs. The models were trained using Word2Vec embeddings (Mikolov et al., 2013) of size 150 pretrained on the soc.culture.hmong Usenet Corpus (Mortensen, 2015), consistent with the approach first proposed by Wang et al. (2015). The early coding approach for creating the model was inspired in part by an approach by Ivanov (2018).

The set of hand-annotated documents from the first stage of tagging served as the initial training set, and the model was retrained with additional data from subsequent documents after expert checking of the tags, as described in Sect. 6 above. For reference, performance metrics of the final version of the BiLSTM model on the 11th (final) document appear in Table 4 below. "Non-predicted" and "Non-true" values of 0 and 1 are specified for those cases where the model predicted a label not truly present in the document or failed to predict a label that was present, which would otherwise result in zero division when calculating the Macro Precision and Macro Recall scores.

8 Limitations of the study

There are two limitations of the Hmong Medical Corpus project worthy of note. The first is the presence of "translationese" (see Volansky et al., 2015) in the corpus, due to translation from English originals. This is described in more detail in Sect. 8.1 below.

The second is the limitation of annotators to a single expert annotator, as stated in Sect. 6. This prcluded the use of inter-annotator agreement as a metric, which would otherwise prove useful in evaluating the quality of the annotations (cf. Fort, 2016). This was despite extensive effort and collaboration with two community organizations to find additional community-based annotators, though several members of the community were willing to provide collaborative effort as described in Sect. 6 above.

8.1 Issue of translationese

As a corpus that has been primarily translated from English, there is the expectation that some "translationese" phenomena would be present. Ideally, one would pursue a computational analysis approach along the lines of Volansky et al. (2015), *inter alia*, to perform an analysis comparing native Hmong text with the translation-based text found in the corpus. The issue here, however, is the lack of a corresponding corpus of the same genre and register, meaning that quantitative comparisons between the only other publicly available Hmong corpus, the soc.culture.hmong Usenet corpus (Mortensen, 2015), and the Hmong Medical Corpus would produce numbers that fail to control for these other factors (cf. Baker, 1993).

However, some qualitative observations can be made. First, the extensive use of equivalent English phrases appear in parentheses following their Hmong equivalents, as if to enhance reference or clarify for the reader a Hmong expression by repeating its English original. Examples include *tus khaub thuas (Influenza)* "lit., the influenza (Influenza)", *tshuaj tua kab mob (antibiotics)* "lit., medicine [that] kills pathogens (antibiotics)", and *teb chaws Meskas Sab Hnub Tuaj Qaum Teb (Northeastern United States)* "lit., North-East American country (Northeastern United States)". Second, the use of circumlocutions to translate English words without Hmong equivalents appears, sometimes in combination with the original English word in parentheses, as with *tej chaw uas nyob ib ncig yus (environment)* "lit., the places that are around you (environment)". Overly literal translations that are not standard terms in Hmong also appear, such as *cov pas dej loj (Great Lakes)* "lit., the big lakes (Great Lakes)". Parenthetical explanations in Hmong also appear with some uncommon expressions, as with *av hmo ntuj (night soil) (cov av uas xyaw tib neeg cov quav)* "lit., night soil (night soil) (the soil that mixes [with] people's feces)". The use of a vague Hmong expression with an English term for clarification also occurs, as with *tshuaj pleev ib ce (lotions thiab cream)* "lit., medicine [with which one] smears a body (lotions and cream)" and *tus kab mob swimmer's itch* "lit., the pathogen swimmer's itch".

Table 5 General statistics of the corpus

Content	Total
Documents	105
Tokens	100,535
Sentences	3552
Tokens/document	957.5
Tokens/sentence	28.3
Punctuation	8548
Punctuation/document	81.4
Punctuation/sentence	2.4

Table 6 Statistics of the POS-tagged subset

Content	Total
Documents	11
Tokens	10,152
Full words	8152
Tokens/full word	1.25
Full words/document	741.1

This is in addition to the widespread phenomena in the Hmong diaspora of calquing and code-switching in general (White, 2021).

9 Corpus statistics

The Hmong Medical Corpus is composed of 105 medical informational documents. Basic statistics regarding its content appear in Table 5, and those of the word position/POS-tagged subset in Table 6. The corpus contains 100,535 tokens, of which 10,152 belong to the word position/POS-tagged subset in 8152 full words.

Table 7 provides total counts and percentages for the POS tags associated with syllable-based tokens in the POS-tagged subset of the Hmong Medical Corpus. The largest categories by far are common nouns (18.95%) and verbs (27.79%), while non-content word categories such as nominal classifiers (11.98%) and verbal modifiers (7.51%) comprise a significant portion of the total—a situation which reflects the general syntactic properties of Hmong. The large number of verbs (VV) and common nouns (NN) is not particularly surprising, given their relatively high frequency of use cross-linguistically. Classifiers (CL), on the other hand, are relatively frequent in Hmong as they have a wide range of uses, including to indicate definite reference (Simpson et al., 2011). Localizers (LC) have a relatively high frequency of use (2.46%) as compared to demonstratives (DT; 0.92%) given their greater semantic specificity as regards deictic reference. Of particular note, only two adjectives (JJ) appear among the POS-tagged documents in the corpus, given the unusual nature of Hmong adjectives as a part-of-speech class, as described in Sect. 3 above. Foreign

Table 7 Distribution of POS-tagged tokens

Tag	Category	Total tagged	Percentage (%)
CL	Nominal classifier	1216	11.98
CV	Verbal action classifier	1	0.01
NN	Common noun	1924	18.95
PN	Pronoun	203	2.00
NR	Proper noun	2	0.02
JJ	Adjective	2	0.02
VV	Verb	2821	27.79
DT	Demonstrative	93	0.92
QU	Quantifier	276	2.72
LC	Localizer	250	2.46
RL	Relative-like modifier	38	0.37
AD	Verbal modifier (adverb-like)	762	7.51
PP	Preposition	338	3.33
CC	Coordinating conjunction	404	3.98
CS	Subordinating conjunction	404	3.98
CM	Clause-final marker	1	0.01
ON	Onomatopoeia	2	0.02
FW	Foreign word	527	5.19
PU	Punctuation	888	8.75
	Total	10,152	100.01[a]

[a]This value deviates from 100.00% solely due to rounding errors

Table 8 Distribution of NEs by semantic type

Tag	Category	Total tagged	Total tokens with NE Tag	Total tokens with O tag per NE class	Percentage tokens tagged as NE
BPOC	Body part, organ, or organ component	1788	2507	98,028	2.49
DSYN	Disease or syndrome	1431	4558	95,977	4.53
SOSY	Sign or symptom	1335	4391	96,144	4.37
	Total	4554			

words (FW) feature strongly (5.19%) as a part of the overall corpus, likely as a result of typical code-mixing that features in diasporic Hmong language as well as the number of English terms that are carried over in the translations into Hmong, as described in Sect. 8.1.

The total number of tagged named entities is shown by named entity type in Table 8. The number of tagged body part references is higher than the number of each of the other two semantic groups; this reflects the tendency for body part terms to appear in the names of diseases and symptom expressions in Hmong. The total

of tagged named entities is provided in the table, while the amounts in the other categories cannot be combined into totals as an individual token can have more than one tag and thus a total would count the same token multiple times.

10 Corpus contributions

The Hmong Medical Corpus contributes significantly to natural language processing efforts for low-resource languages in the following ways:

(1) it is the first NE-annotated biomedical corpus for a non-official minority language;
(2) it provides the first publicly-available POS-tagged dataset on the Hmong language;
(3) it is the first biomedical corpus with named entity tags for Hmong, which will enable additional NLP work in the biomedical domain for the language;
(4) it provides a successful means of how to handle the tokenization issues involving syllable-spacing in Hmong;
(5) it is publicly available, both to search and to download; and
(6) it represents an effective, replicable paradigm for the development of NE-annotated corpora for other minority and low-resource languages.

11 Conclusions

We presented the Hmong Medical Corpus, a new biomedical corpus for the Hmong language. The documents comprising the corpus contain annotations of two kinds: POS tags and named entity tags. This dataset represents the first time a publicly-available annotated corpus has been released for a non-official minority language in the biomedical domain. It addresses the prior lack of annotated data for the Hmong language, enabling a range of possible Hmong-specific NLP applications of either a biomedical or general nature. The Hmong Medical Corpus is publicly available to access and download online.[3]

Appendix: Extended sample text with POS and NE tags

An extended sample text with POS and NE tags is provided below. The top line represents the original text, the second line the word position/POS tags, the third line the BPOC NE tag class, the fourth the DSYN NE tag class, and the fifth the SOSY NE tag class.

[3] http://hmong-medical-corpus.org/.

Tus	mob	no	yog	tau	los	ntawm	cov	kab	mob
B-CL	B-NN	B-DT	B-VV	B-VV	B-VV	B-LC	B-CL	B-NN	I-VV
O	O	O	O	O	O	O	O	O	O
O	O	O	O	O	O	O	O	O	O
O	O	O	O	O	O	O	O	O	O

xws	li	cov	enteroviruses	(	tuag	npab	tuag	ceg	thiab
B-VV	B-PP	B-CL	B-FW	O–PU	B-VV	I-NN	I-VV	I-NN	B-CC
O	O	O	O	O	O	B-BPOC	O	B-BPOC	O
O	O	O	O	O	O	O	O	O	O
O	O	O	O	O	B-SOSY	I-SOSY	I-SOSY	E-SOSY	O

tsis	yog	tuag	npab	tuag	ceg	)	thiab	cov	kab
B-AD	B-VV	B-VV	I-NN	I-VV	I-NN	O–PU	B-CC	B-CL	B-NN
O	O	O	B-BPOC	O	B-BPOC	O	O	O	O
O	O	O	O	O	O	O	O	O	O
O	O	B-SOSY	I-SOSY	I-SOSY	E-SOSY	O	O	O	O

mob	aviviruses	xws	li	tus	kab	mob	West	Nile	Virus
I-VV	B-FW	B-VV	B-PP	B-CL	B-NN	I-VV	B-FW	B-FW	B-FW
O	O	O	O	O	O	O	O	O	O
O	O	O	O	O	O	O	O	O	O
O	O	O	O	O	O	O	O	O	O

,	tus	kab	mob	Japanese	Encephalitis	virus	,	los	yog
O–PU	B-CL	B-NN	I-VV	B-FW	B-FW	B-FW	O–PU	B-CC	B-CC
O	O	O	O	O	O	O	O	O	O
O	O	O	O	B-DSYN	E-DSYN	O	O	O	O
O	O	O	O	O	O	O	O	O	O

tus	kab	mob	St.	Louis	encephalitis	virus		Lwm
B-CL	B-NN	I-VV	B-FW	B-FW	B-FW	B-FW	O–PU	B-QU
O	O	O	O	O	O	O	O	O
O	O	O	B-DSYN	I-DSYN	E-DSYN	O	O	O
O	O	O	O	O	O	O	O	O

cov	kab	mob	uas	tej	zaum	yuav	ua	rau	muaj
B-CL	B-NN	I-VV	B-CS	B-AD	I-AD	B-AD	B-VV	B-PP	B-VV
O	O	O	O	O	O	O	O	O	O
O	O	O	O	O	O	O	O	O	O
O	O	O	O	O	O	O	O	O	O

tus	mob	AFM	yog	cov	herpesviruses	(	piv	txwv	li
B-CL	B-NN	B-FW	B-VV	B-CL	B-FW	O–PU	B-VV	I-VV	B-PP
O	O	O	O	O	O	O	O	O	O
B-DSYN	I-DSYN	E-DSYN	O	O	O	O	O	O	O
O	O	O	O	O	O	O	O	O	O

,	tus	kab	mob	cytomegalovirus	,	Epstein	Barr	virus	)
O–PU	B-CL	B-NN	I-VV	B-FW	O–PU	B-FW	B-FW	B-FW	O–PU
O	O	O	O	O	O	O	O	O	O
O	O	O	O	O	O	O	O	O	O
O	O	O	O	O	O	O	O	O	O

thiab	cov	adenoviruses	.	Txawm	tias	tus	mob	AFM	yuav
B-CC	B-CL	B-FW	O–PU	B-AD	B-CS	B-CL	B-NN	B-FW	B-AD
O	O	O	O	O	O	O	O	O	O
O	O	O	O	O	O	B-DSYN	I-DSYN	E-DSYN	O
O	O	O	O	O	O	O	O	O	O

ua	rau	koj	tej	npab	los	yog	tej	ceg	tsis
B-VV	B-PP	B-PN	B-CL	B-CL	B-CC	I-CC	B-CL	B-NN	B-AD
O	O	O	O	B-BPOC	O	O	O	B-BPOC	O
O	O	O	O	O	O	O	O	O	O
O	O	O	O	O	O	O	O	O	B-SOSY

muaj	zog	los	,	cov	mob	raws	keeb	cag	ces
B-VV	B-NN	B-CS	O–PU	B-CL	B-NN	B-VV	B-NN	I-NN	B-CC
O	O	O	O	O	O	O	O	O	O
O	O	O	O	O	O	O	O	O	O
I-SOSY	E-SOSY	O	O	O	O	B-SOSY	O	O	O

,	cov	mob	tau	los	ntawm	tej	khoom	los	yog
O–PU	B-CL	B-NN	B-VV	B-VV	B-LC	B-CL	B-NN	B-CC	I-CC
O	O	O	O	O	O	O	O	O	O
O	O	O	O	O	O	O	O	O	O
O	O	O	O	O	O	O	O	O	O

tej	chaw	muaj	tshuaj	lom	uas	nyob	ib	ncig	yus
B-CL	B-NN	B-VV	B-NN	B-NN	B-CS	B-VV	B-QU	B-CL	B-PN
O	O	O	O	O	O	O	O	O	O
O	O	O	O	O	O	O	O	O	O
O	O	O	O	O	O	O	O	O	O

thiab	tus	mob	Guillain	Barré	syndrome	kuj	ua	rau	muaj
B-CC	B-CL	B-NN	B-FW	B-FW	B-FW	B-AD	B-VV	B-PP	B-VV
O	O	O	O	O	O	O	O	O	O
O	B-DSYN	I-DSYN	I-DSYN	I-DSYN	E-DSYN	O	O	O	O
O	O	O	O	O	O	O	O	O	O

cov	tsos	mob	zoo	ib	yam	thiab	.
B-CL	B-NN	I-VV	B-VV	B-QU	B-CL	B-AD	O–PU
O	O	O	O	O	O	O	O
O	O	O	O	O	O	O	O
O	O	O	O	O	O	O	O

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. Partial financial support was received from a Completion Grant from the College of Arts, Society and Education at James Cook University and a Postgraduate Research Scholarship from James Cook University.

Data availability The material is available online at http://hmong-medical-corpus.org.

Code availability Code can be found at https://github.com/nathanmwhite/hmong-medical-corpus/.

Declarations

Conflict of interest The author served as advisor to the board of directors of a company of which two employees and a board member freely agreed to serve as community collaborators. There were no financial interests on the part of the author, the community collaborators, or the company in this relationship.

Ethical approval Research ethics approval was obtained for the larger project of which the corpus project was one component.

Consent to participate Hmong community collaborators all agreed to serve as collaborators on the corpus project, as this best suited their role on the project as stated in the manuscript. They are named as such on the corpus website.

Consent for publication The sole author of the current manuscript consents to its publication.

References

Bada, M., Eckert, M., Evans, D., Garcia, K., Shipley, K., Sitnikov, D. … Hunter, L. E. (2012). Concept annotation in the CRAFT corpus. *Bmc Bioinformatics*, *13*, 161. https://doi.org/10.1186/1471-2105-13-161

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In G. Francis, M. Baker, & E. Tognini-Bonelli (Eds.), *Text and Technology: In Honour of John Sinclair* (pp. 233–252). John Benjamins. https://doi.org/10.1075/z.64.15bak

Bisang, W. (1993). Classifiers, quantifiers and class nouns in Hmong. *Studies in Language*, *17*(1), 1–51. https://doi.org/10.1075/sl.17.1.02bis

Bisang, W. (2015). Problems with primary vs. secondary grammaticalization: The case of East and mainland Southeast Asian languages. *Language Sciences*, *47*, 132–147. https://doi.org/10.1016/J.LANGSCI.2014.05.007

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Res. 2004 Jan 1; 32*(Database issue), D267–270. https://doi.org/10.1093/nar/gkh061

Campillos, L., Deléger, L., Grouin, C., Hamon, T., Ligozat, A., & Névéol, A. (2018). A French clinical corpus with comprehensive semantic annotations: Development of the Medical Entity and Relation LIMSI annotated Text corpus (MERLOT). *Language Resources & Evaluation*, *52*, 571–601. https://doi.org/10.1007/s10579-017-9382-y

Dinh, Q. T., Le, H. P., Nguyen, T. M. H., Nguyen, C. T., Rossignol, M., & Vu, X. L. (2008). Word segmentation of Vietnamese texts: A comparison of approaches. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pp. 1933–1936

Doğan, R. I., Leaman, R., & Lu, Z. (2014). NCBI Disease Corpus: A Resource for Disease Name Recognition and Concept Normalization. *Journal of Biomedical Informatics*, *47*, 1–10. https://doi.org/10.1016/j.jbi.2013.12.006

Fadiman, A. (1998). *The Spirit Catches You and You Fall Down: A Hmong Child, Her American Doctors, and the Collision of Two Cultures*. Farrar, Straus and Giroux

Fort, K. (2016). *Collaborative annotation for reliable natural language processing: Technical and sociological aspects*. Wiley-ISTE

Gao, Y., Gu, L., Wang, Y., Wang, Y., & Yang, F. (2019). Constructing a Chinese electronic medical record corpus for named entity recognition on resident admit notes. *BMC Medical Informatics and Decision Making*, *19*, 56. https://doi.org/10.1186/s12911-019-0759-2

Gerner, M. (2014). Verb classifiers in East Asia. *Functions of Language*, *21*(3), 267–296. https://doi.org/10.1075/fol.21.3.01ger

Gonzalez-Agirre, A., Marimon, M., Intxaurrondo, A., Rabal, O., Villegas, M., & Krallinger, M. (2019). PharmaCoNER: Pharmacological Substances, Compounds and proteins Named Entity Recognition track. In *Proceedings of the 5th Workshop on BioNLP Open Shared Tasks*, pp. 1–10. https://doi.org/10.18653/v1/D19-5701

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Ivanov, G. B. (2018, September 8). *Build a POS tagger with an LSTM using Keras*. Natural Language Processing for Hackers. Retrieved March 26, 2021, from https://nlpforhackers.io/lstm-pos-tagger-keras/

Johnson, S. K. (2002). Hmong Health Beliefs and Experiences in the Western Health Care System. *Journal of Transcultural Nursing*, *13*(2), 126–132. https://doi.org/10.1177/104365960201300205

Karimi, S., Metke-Jimenez, A., Kemp, M., & Wang, C. (2015). CADEC: A corpus of adverse drug event annotations. *Journal of Biomedical Informatics*, *55*, 73–81. https://doi.org/10.1016/j.jbi.2015.03.010

Kilicoglu, H., Ben Abacha, A., Mrabet, Y., Shooshan, S. E., Rodriguez, L., Masterton, K., & Demner-Fushman, D. (2018). Semantic annotation of consumer health questions. *Bmc Bioinformatics*, *19*(1), 34. https://doi.org/10.1186/s12859-018-2045-1

Kim, J. D., Ohta, T., Tateisi, Y., & Tsujii, J. (2003). GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, *19*, i180–i182. https://doi.org/10.1093/bioinformatics/btg1023

Krallinger, M., Rabal, O., Leitner, F., Vazquez, M., Salgado, D., Lu, Z., Leaman, R., Lu, Y., Ji, D., Lowe, D. M., et al. (2015). The CHEMDNER corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics*, *7*(Suppl 1), S2. https://doi.org/10.1186/1758-2946-7-S1-S2

Kruengkrai, C., Uchimoto, K., Kazama, J., Wang, Y., Torisawa, K., & Isahara, H. (2009). An Error-Driven Word-Character Hybrid Model for Joint Chinese Word Segmentation and POS Tagging. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 513–521

Lemoine, J. (2005). What is the actual number of the (H)mong in the world? *Hmong Studies Journal*, *6*, 1–8

Lindberg, D. A., Humphreys, B. L., & McCray, A. T. (1993). The Unified Medical Language System. *Methods of Information in Medicine*, *32*(4), 281–291. https://doi.org/10.1055/s-0038-1634945

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *ICLR Workshop Papers*

Mitrofan, M., Mititelu, V. B., & Mitrofan, G. (2019). MoNERo: A Biomedical Gold Standard Corpus for the Romanian Language. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pp. 71–79. https://doi.org/10.18653/v1/W19-5008

Moreno, I., Boldrini, E., Moreda, P., & Romá-Ferri, M. T. (2017). DrugSemantics: A corpus for Named Entity Recognition in Spanish Summaries of Product Characteristics. *Journal of Biomedical Informatics*, *72*, 8–22. https://doi.org/10.1016/j.jbi.2017.06.013

Mortensen, D. (2015, May 29). *soc.culture.hmong Usenet (SCH) corpus*. My-hm Listserv. Retrieved January 19, 2022, from http://www.davidmortensen.org/corpora/sch_corpus-2.zip

Nguyen, C. T., Nguyen, T. K., Phan, X. H., Nguyen, L. M., & Ha, Q. T. (2006). Vietnamese Word Segmentation with CRFs and SVMs: An Investigation. In *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation*, pp. 215–222

Nguyen, D. Q., Vu, T., Nguyen, D. Q., Dras, M., & Johnson, M. (2017). From Word Segmentation to POS Tagging for Vietnamese. In *Proceedings of the 15th Annual Workshop of the Australasian Language Technology Association*, pp. 108–113

Névéol, A., Grouin, C., Leixa, J., Rosset, S., & Zweigenbaum, P. (2014). The QUAERO French Medical Corpus: A Resource for Medical Entity Recognition and Normalization. In *Proceedings of the Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing*, pp. 24–30

Oronoz, M., Gojenola, K., Pérez, A., Díaz de Ilarraza, A., & Casillas, A. (2015). On the creation of a clinical gold standard corpus in Spanish: Mining adverse drug reactions. *Journal of Biomedical Informatics*, *56*, 318–332. https://doi.org/10.1016/j.jbi.2015.06.016

Pfeifer, M. E., Sullivan, J., Yang, K., & Yang, W. (2012). Hmong Population and Demographic Trends in the 2010 Census and 2010 American Community Survey. *Hmong Studies Journal*, *13*(2), 1–31

Post, M. (2008). Adjectives in Thai: Implications for a functionalist typology of word classes. *Linguistic Typology*, *12*, 339–381. https://doi.org/10.1515/LITY.2008.041

Ratliff, M. (2009). White Hmong vocabulary. In M. Haspelmath, & U. Tadmor (Eds.), *World Loanword Database*. Max Planck Digital Library

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, *45*(11), 2673–2681. https://doi.org/10.1109/78.650093

Shao, Y., Hardmeier, C., Tiedemann, J., & Nivre, J. (2017). Character-based Joint Segmentation and POS Tagging for Chinese using Bidirectional RNN-CRF. In *Proceedings of the 8th International Joint Conference on Natural Language Processing*, pp. 173–183

Stubbs, A., & Uzuner, Ö. (2015). Annotating longitudinal clinical narratives for deidentification: The 2014 i2b2/UTHealth Corpus. *Journal of Biomedical Informatics*, *58(Suppl.)*, S20–S29. https://doi.org/10.1016/j.jbi.2015.07.020

Takahashi, K., & Yamamoto, K. (2016). Fundamental Tools and Resource are Available for Vietnamese Analysis. In *2016 International Conference on Asian Language Processing*, pp. 246–249. https://doi.org/10.1109/IALP.2016.7875978

Thornburn, S., Kue, J., Keon, K. L., & Lo, P. (2012). Medical mistrust and discrimination in health care: A qualitative study of Hmong women and men. *Journal of Community Health*, *37*(4), 822–829. https://doi.org/10.1007/s10900-011-9516-x

Uzuner, Ö., Solti, I., Xia, F., & Cadag, E. (2010). Community annotation experiment for ground truth generation for the i2b2 medication challenge. *Journal of the American Medical Informatics Association*, *17*(5), 519–523. https://doi.org/10.1136/jamia.2010.004200

Uzuner, Ö., South, B. R., Shen, S., & DuVall, S. L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, *18*, 552–556. https://doi.org/10.1136/amiajnl-2011-000203

Volansky, V., Ordan, N., & Wintner, S. (2015). On the features of translationese. *Digital Scholarship in the Humanities*, *30*, 98–118. https://doi.org/10.1093/llc/fqt031

Wang, P., Qian, Y., Soong, F. K., He, L., & Zhao, H. (2015, November 1). *A Unified Tagging Solution: Bidirectional LSTM Recurrent Neural Network with Word Embedding*. Computing Research Repository, arXiv. Retrieved March 26, 2021, from https://arxiv.org/abs/1511.00215

White, N. M. (2019). Classifiers in Hmong. In A. Aikhenvald, & E. Mihas (Eds.), *Genders and classifiers: A cross-linguistic typology* (pp. 222–248). Oxford University Press. https://doi.org/10.1093/oso/9780198842019.003.0008

White, N. M. (2020). Word in Hmong. In A. Aikhenvald, R. M. W. Dixon, & N. M. White (Eds.), *Phonological word and grammatical word: Across-linguistic typology* (pp. 213–259). Oxford University Press. https://doi.org/10.1093/oso/9780198865681.003.0008

White, N. M. (2021). Language and variety mixing in diasporic Hmong. *Italian Journal of Linguistics/ Rivista di Linguistica*, *33*(1), 157–180. https://doi.org/10.26346/1120-2726-172

Wälchli, B. (2005). *Co-compounds and natural coordination*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199276219.001.0001

Xiong, Y., & Cohen, D. (2005). *Student's practical Miao-Chinese-English handbook*. Yunnan Nationalities Publishing House

Xue, N., Xia, F., Chiou, F. D., & Palmer, M. (2005). The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, *11*(2), 207–238. https://doi.org/10.1017/S135132490400364X