Mon-Khmer Studies

VOLUME 44

The journal of Austroasiatic languages and cultures

Author: Seth VITRANO-WILSON
Title: *Variability in the use of spaces by writers of Hmong Daw*
Pages: 1-18

Received: 24 March 2015; Accepted 23 April 2015

Volume 44 Editors:
Paul Sidwell
Brian Migliazza

ISSN: 0147-5207
Website: http://mksjournal.org

Published by:
Mahidol University (Thailand)
SIL International (USA)

Variability in the use of spaces by writers of Hmong Daw

Seth VITRANO-WILSON
Payap University
<seth_vitrano-wilson@sil.org>

Abstract
Hmong Daw is a Hmong-Mien language that primarily uses the Latin script. Both syllable-spaced and word-spaced formats are used, with spacing varying by writer and by word. Using a 15-million word corpus in Hmong Daw, a list of 96 polysyllabic words was analyzed to see how often each word was written in syllable-spaced and word-spaced formats. The results show that most polysyllabic Hmong Daw words are usually written with syllable spacing, but that spacing varies with the orthographic, morphological, syntactic, and lexical properties of words. Just as Kuperman & Bertram 2013 found with English compounds, the patterns in this variation suggest that writers are most likely to use spaces where they benefit readability most. The results here also suggest that a purely linguistic definition of a word is less useful for orthography decisions than a definition that takes into account the variables that affect reading for different types of words.

Keywords: Hmong Daw, syllable spacing, orthographic variation
ISO 639-3 language codes: mww, hnj

1. Introduction[1]

Traditionally, most Latin script orthographies use blank spaces to mark word boundaries. However, in mainland Southeast Asia, some Latin script orthographies, such as Vietnamese, Lahu, and Akha, use spaces to separate every syllable. Meanwhile, many Brahmi-based alphasyllabaries in the region such as Thai, Burmese, Lao, and Khmer use phrase or clause spacing, with letters organized by syllables. Chinese hanzi has no spaces at all, apart from the small spaces that separate each monosyllabic character. In all of these orthographies, syllable boundaries are much clearer than is typical for Latin script orthographies, and word boundaries are less clear. Although some Latin script orthographies in the region do use word spacing, it is likely that the relative dominance of the syllable level in Chinese hanzi and the Brahmi-based alphasyllabaries of the region has led several groups in mainland Southeast Asia to consider using syllable spacing for Latin script orthographies, even though word spacing is the norm internationally for the Latin script.

In Hmong Daw (or White Hmong) writing, both syllable-spaced and word-spaced text can be found, depending on the writer. In addition, since the definition of a “word” in Hmong Daw is not intuitively obvious nor universally agreed upon, different writers will define “word spacing” differently. Some writers will even vary their spacing of the same word at different times. As a result, in addition to having both writers who use syllable spacing and writers who use word spacing, there is a great deal of diversity in spacing style from word to word.

Although most reading research focuses on eye movement studies and other studies of readers, Kuperman & Bertram 2013 provides a valuable perspective by looking at what factors influence the way writers spell compound words in English. Looking diachronically within a large corpus at compounds that vary in their format (unspaced, hyphenated, or spaced), the authors found several factors that seem to influence the spacing style of English writers, whether consciously or unconsciously. For instance, in their study, more frequent compounds tend to be unspaced. This

[1] I am very grateful for the tremendous help of my thesis committee, Dr. Isara Choosri, Dr. Elissa Ikeda, and Dr. Wyn Owen, who helped shape and refine the research upon which this article is based, and helped me communicate my ideas more clearly. Thank you to Dr. David Mortensen, who graciously provided his processed version of the 15-million word online corpus from the soc.culture.hmong Usenet group. Finally, thank you to all the Hmong writers who, through their collective intuition, revealed the insights to be found herein. May you always enjoy reading and writing, and may your words (and syllables) be filled with grace toward each other and others. Ua tsaug ntau ntau!

VITRANO-WILSON, Seth. 2015. Variability in the use of spaces by writers of Hmong Daw. *Mon-Khmer Studies* 44:1-18

finding is in line with the idea that more frequent compounds are processed faster, and that direct whole word access tends to be faster than decomposition when easily available. They also found that semantically transparent compounds are more often found in the spaced format, again in line with research indicating that transparent compounds are more easily separated than opaque compounds (Sandra 1990, Frisson et al. 2008, Mok 2009).

Longer words in Kuperman & Bertram 2013 are also more likely to be spaced, especially when their first constituent is longer. Previous research on reading shows that long compound words benefit more from constituent separation than short compound words (Bertram & Hyönä 2003, Juhasz et al. 2005). Bertram & Hyönä 2003 explains this effect by referencing the "visual acuity principle," namely, that long words are more likely to extend beyond the foveal region of the retina. Because readers are unable to process these words in a single fixation, they process the words one constituent at a time instead. The fact that longer compounds in English in Kuperman & Bertram 2013 are more likely to be separated with spaces indicates that English writers may intuitively know that processing long compound words is made easier by separating constituents.

It should be noted that Kuperman & Bertram 2013 do not directly address the question of whether the general level of spacing used in English is optimal for readability. Rather, they find that when spaces are used by writers to separate compounds, they tend to be used more often in words that research suggests would show a higher relative benefit to spaces, such as transparent compounds, longer compounds, or more frequent compounds, and less often in other types of words. This says nothing, for example, about whether the greater use of spaces in English versus their sparser use by German writers, or their even more frequent use by Hmong writers, would lead to any differences in readability between these three orthographies. Rather, the results of Kuperman & Bertram 2013 suggest that writers are sensitive to the differential effect of spaces on the reading of different types of words.

By analyzing the spacing practices of English writers, Kuperman & Bertram 2013 provides valuable context to the study of English readers and the cognitive process of reading. This study hopes to do the same for Hmong Daw, while also considering the implications of the results for orthography development.

2. Hmong linguistics, sociolinguistics, and orthography

Hmong Daw is a language in the Hmong-Mien family, closely related to Hmong Njua, or Green Hmong. There are roughly 1.7 million speakers of Hmong Daw, mainly in Vietnam, China, Laos, the United States, and Thailand. The roughly 32,000 Hmong Daw speakers in Thailand are mainly found in the north and north central parts of the country (Lewis et al. 2014). In the US, there are roughly 180,000 speakers of Hmong Daw, and another 110,000 speakers of Hmong Njua (Joshua Project). Most Hmong in the US live in California, Minnesota, and Wisconsin (Moua 2010).

The Hmong Daw syllable structure is CV(V)T. Onsets can be highly complex phonetically, such as in the word *nplooj* [mblɔ̃^{52}] 'leaf', but are traditionally analyzed as a single phoneme in Hmong linguistics (Jarkey 1987, Ratliff 2010). Hmong Daw has no final consonant phonemes, according to Heimbach 1979, although there is a final [ŋ], which Heimbach analyzes as being phonemically part of the nasalized vowels. Orthographic final consonants represent tones and not consonants, and final [ŋ] is treated as part of the vowel in the RPA orthography.

The great majority of Hmong Daw morphemes are one syllable. No morphemes are less than a syllable, except for one meaningful tone change from <m> (low creaky tone) to <d> (low rising), which turns a spatial preposition into a demonstrative noun, as in *pem* 'up' vs. *ped* 'up there' (Ratliff 2010:112). No <d> tone words appear in this study. A few morphemes are polysyllabic, though most of these seem to be loanwords, compounds whose original morphemes have been lost, or onomatopoeic expressives such as *cij coj*, the sound of chicks chirping (Ratliff 2010:222). In addition to having mainly monosyllabic morphemes, words in Hmong Daw are also mostly monomorphemic and monosyllabic (Golston & Yang 2001). Polysyllabic words can be formed through compounding, reduplication (either partial or full), affixation, or by starting with polysyllabic morphemes (Ratliff 2009).

The most widely used orthography among Hmong Daw readers is called the Romanized Popular Alphabet, or RPA. This orthography was developed by Catholic and Protestant missionaries and Hmong speakers in the 1950s in Laos (Smalley et al. 1990:151). The RPA uses final consonant letters to mark tones. In China, many Hmong varieties use the Chuanqiandian orthography, in which syllable spacing is standard (McLaughlin 2012). However, for RPA, no standard spacing style exists. Both syllable spacing and word spacing are common, and for word-spaced text, there is no standard list of words to be joined orthographically.

The RPA is used to write both Hmong Daw and Hmong Njua, but there are a few systematic spelling differences between the RPA systems used for the two varieties, given in **Table 1**.

Table 1: Spelling correspondences between Hmong Daw and Hmong Njua

Hmong Daw	**Hmong Njua**
a	aa
ia	a
d	dl / ndl
dh	dlh / ndlh

3. Spacing in different communities

As stated above, spacing in the Hmong Daw RPA orthography varies from writer to writer, though some generalizations can be made. For instance, most materials published by Protestants use some form of word spacing, whereas Catholic publications tend to use syllable spacing. The Catholic Hmong Daw Bible translation (Bertrais 2002), as well as Hmong Catholic websites such as hmongrpa.org (Hmong RPA 2012) and aumoneriehmong.fr (Hmoob Kav Tos Liv Fab Kis Teb 2015) use syllable spacing. The Hmong Daw Bible translation used by most Protestants, in contrast, joins many syllables together into polysyllabic words (United Bible Societies 2000), and many Protestant websites follow this pattern (e.g. Hmong District 2015, Hmong Baptist National Association 2015). For instance, the name "Jesus" is written <Yexus> in the Protestant-based United Bible Societies New Testament (United Bible Societies 2000), but <Yes Xus> in the Catholic Bible (Bertrais 2002) (the tone difference on the first syllable is unrelated to the difference in spacing style). Similarly, the monomorphemic word /ʃɐɨ^{24}dɐɨ^{22}/ 'everyone' is spelled <sawv daws> in the Catholic Bible and <sawvdaws> in the UBS New Testament, while the word /kɨ^{24}ti^{52}/ 'brothers, relatives' (literally 'younger.brother-older.brother') is spelled <kwv tij> in the Catholic Bible, but <kwvtij> in the UBS New Testament. Hmong Njua texts exhibit the same difference in spacing styles between Catholic and Protestant materials.

The difference in spacing between Catholic and Protestant materials seems to go back to early linguists involved in Hmong Daw. Jean Bertrais, a Catholic priest involved in the development of the RPA orthography, used syllable spacing for his dictionary of Hmong Daw (Bertrais 1964). Meanwhile, Ernest Heimbach, a Protestant missionary linguist, joined at least some syllables into words in his dictionary (Heimbach 1979). He was also involved in the translation of several books of the New Testament into Hmong Daw, which were using some form of word spacing since at least 1965 (Heimbach 1965) and possibly as early as 1955 (Heimbach et al. 1955). It is not clear why this difference came about, or whether other writers, especially Hmong writers, also influenced the process of standardization within Catholic and Protestant communities.

Secular texts show both patterns of spacing. For instance, a series of literacy primers for Hmong students in the US (Moua & Vangay 1989a, 1989b, 1989c) uses a space between every syllable, regardless of morphology, apart from a few English names and a few onomatopoeic animal sounds with hyphens. Syllable-spaced Hmong Daw text can also be found in an online dictionary (Xiong 2014), on a website about Hmong religion (Temple of Hmongism 2013), and in various government or health websites and documents (Wisconsin Department of Human Services 2004a, 2004b; U.S. Department of Health and Human Services 2007). On the other hand, both of the printed Hmong Daw-English dictionaries consulted for this study join at least some syllables together into words (Heimbach 1979, Xiong 2005). Joined syllables can also be found in secular news articles (Moua 2012), medical sites (U.S. Department of Health and Human Services 2009; Northpoint Health & Wellness Center 2014), and government services websites (Australian

Government 2014; Wisconsin Court Interpreter Program 2006). Some documents are not internally consistent, even on the same word (e.g. <menyuam> vs. <me nyuam> 'child' in Waisman Center 2006). Even single-author dictionaries vary at times in their spacing style (e.g. <Fab Kis>, <Fabkis> 'France, French' in Xiong 2005:59; <mis niv>, <misniv> 'minute' in Xiong 2005:165).

4. Spacing for different words

Although many Hmong texts and websites join some syllables into larger word units, there is by no means complete uniformity as to exactly which words are joined from text to text, or even within a single text. It also seems that there are no spell-checking programs available for online use, so even if there were a standard, it would spread only by diffusion from writer to writer. This study attempts to quantify this variation at the word level, using a limited list of 96 polysyllabic words in Hmong Daw. The list contains words of a variety of different morphological types, as listed below:

Table 2: List of morphological word types used in word list test

Polysyllabic word type
Monomorphemic words
Semantically opaque compounds
Words with affixes
Coordinate compounds
Tone sandhi compounds
Reduplicated words (either partially or fully reduplicated)
Four-syllable elaborate expressions

The words described as "coordinate compounds" in this study are a particular class of compounds, often found in languages of mainland Southeast Asia, where a pair of synonyms, antonyms, or otherwise semantically connected words are joined to create a more general meaning. Examples in Hmong Daw are *zaub mov* 'food' (literally 'vegetable-rice'), and *nus muag* 'siblings' (literally 'brother-sister'). Thomas 1962 describes such compounds as products of "semantic reduplication."

Tone sandhi in Hmong Daw is both phonologically and lexically conditioned. It only occurs with certain tone combinations, and only at morpheme boundaries within certain compound words that have these tone combinations (Ratliff 2010). Since it only occurs with morphologically related elements, it helps to indicate the word status of the compounds.

All words included in this study were two-syllable words, except for the elaborate expressions. Elaborate expressions in Hmong Daw are four-syllable, four-constituent constructions that, like coordinate compounds, typically involve synonyms or other related word pairs. They show evidence of lexicalization, the order cannot be reversed, they must be four syllables long, they often repeat elements in an ABAC or ABCB structure, and they often share segmental or tonal symmetries (Mortensen 2003). For these reasons, elaborate expressions are often considered words in Hmong varieties (Ratliff 2009, Mortensen 2003) as well as in other languages in the region (Matisoff 1973, Hanna 2013). Some elaborate expressions consist of four separate morphological words, such as the Hmong Daw *khwv iab khwv daw* 'arduous toil', literally 'toil-bitter-toil-salty' (Jarkey 2010), and *pog koob yawg koob* 'ancestors,' literally 'grandmother-great-grandfather-great' (Ratliff 2009). Others show a more complex morphology, where a two-syllable monomorphemic or sesquimorphemic word is divided in half in the formation of the elaborate expression. Examples from Hmong Daw are:

- *ua dog ua dig* 'to do haphazardly', formed from *ua* 'to do' plus the monomorphemic *dog dig* 'haphazard'

- *ua qoob ua loo* 'to do haphazardly, formed from *ua* 'to do' plus the sesquimorphemic *qoob loo* 'crops' (where *qoob* means 'crops', while *loo* is a fossilized morpheme that was once a synonym of *qoob*)

- *nkhaus niv nkhaus nom* 'curvy, crooked', from *nkhaus* 'bent' plus *niv* and *nom*, two "intensifiers" which do not seem to be independent morphemes (Johns & Strecker 1987, Bertrais 1964)

4.1 Methodology

In order to estimate the frequency with which the polysyllabic words on the list were found separated or joined, I used a large online corpus of Hmong text. David Mortensen of Carnegie Mellon University graciously gave me access to his work compiling the entire 15-million word corpus of the Usenet group soc.culture.hmong, or SCH (now a Google group). Mortensen also processed the text to eliminate nonword text, non-Hmong text, and quoted text, so that the word counts should be a fairly accurate count of original uses for each word.

One problem with using this text is that it contains both Hmong Daw and Hmong Njua text, which may be spelled slightly differently. If a word is spelled the same, then it will have a higher count than if the Hmong Daw and Hmong Njua are spelled differently, so the word frequency numbers would be inaccurate, thus making it more difficult to use word frequency as a factor in modeling spacing variation. Therefore, for any word that is spelled differently in Hmong Njua than Hmong Daw, I used the total for both variants to calculate word frequency. I also included any spelling variants due to tone sandhi. For instance, the Hmong Daw word <taub dag> is spelled <taub dlaag> in Hmong Njua. This word shows tone sandhi, and is sometimes spelled ignoring the tone change, as <taub daj> in Hmong Daw, or <taub dlaaj> in Hmong Njua. I included all four variants, with both joined and separated spacing, for my word frequency count.

Within the SCH corpus, I found all spaced and unspaced instances of each word on the word list. The resulting numbers are found in Appendix A. I then modeled the resulting data using IBM’s SPSS statistics software. SPSS output and syntax for the regressions are found in Appendix B.

4.1.1 Determining word status

Syllable breaks are almost always transparent in the Hmong Daw RPA orthography, but word breaks are usually not. To determine whether a given construction in Hmong Daw should be considered a polysyllabic word or a phrase, I relied on a number of sources, especially Jay Xiong's *Lub Hmoob txhais (Hmong-English dictionary)* (2005), Ernest Heimbach's *White Hmong-English dictionary* (1979), and Martha Ratliff's *Meaningful Tone* (2010) and “White Hmong vocabulary” (2009). These last two sources from Martha Ratliff also give descriptions of what she believes are morphological processes in White Hmong, such as tone sandhi compounding, reduplication, or affixation. Any such process that Ratliff described as a morphological process in White Hmong rather than a syntactic process was assumed to be a word formation process, and the output thereof to be a word rather than a phrase. Also, since both the Heimbach (1979) and the Xiong (2005) dictionaries use the unspaced format only sparingly, any lexeme written unspaced in these dictionaries was assumed to be a word, as was any lexeme where the individual syllables were not listed as lexemes.

Appendix A contains a table of the polysyllabic words used in this study, along with the references used in determining their word status.

4.2 Analysis of the soc.culture.hmong corpus

The main analysis of the 15-million word SCH corpus excluded the four-syllable elaborate expressions in order to be able to examine first and second syllable effects on spacing. The log ratio of unspaced to spaced instances of each word was used as the dependent variable within a linear regression model. Two words (*cheb cheb* ‘to keep sweeping’ and *diav* rawg ‘fork’) only had one instance each in the SCH corpus, so their typical spacing style could not be accurately determined and they were therefore removed from the analysis.

Out of 96 words measured, only five words were found more often in unspaced than spaced format. These five were the monomorphemic *kab tsis* ‘sugarcane’ (60%), *phooj ywg* ‘friend’ (55%), and *taj laj* ‘market’ (a loanword from Lao, 52%), as well as the compounds *tab sis* ‘but’ (literally ‘always-even.though’, 53%) and *kaj ntug* ‘dawn’ (literally ‘bright-sky’, 53%). The average word on the list was unspaced in 15.0% of instances, with a standard deviation of 15.5%.

Several factors were found to significantly influence the choice of spacing style:

- **Number-classifier forms**. Number-classifier forms (all tone sandhi compounds, such as *ib qho* ‘one part’, or *ob tug* ‘two people’) are more likely to be written with a space than other types of words (β_{stand}=-.434, p<.001).
- **Fully reduplicated forms**. Words formed by full reduplication are more likely to be written with a space (β_{stand}=-.380, p<.001).
- **Ratio of first syllable frequency in the target word over total frequency of the first syllable**. Words in which the first syllable nearly always occurs in that particular word, and not by itself or in another word, are more likely to be written unspaced (β_{stand}=.294, p<.001).
- **Number of letters in the first syllable**. A greater number of letters in the first syllable of a word corresponds to a greater likelihood of being written with a space (β_{stand}=-.264, p=.001).
- **Number of morphemes**. Monomorphemic words are more likely to be written unspaced than polymorphemic words (β_{stand}=-.215, p=.004).
- **Number of Hmong Daw/Hmong Njua differences**. Words with a large number of spelling differences between the Hmong Daw and Hmong Njua cognates for a given word are more likely to be written unspaced (β_{stand}=.169, p=.016).

Other effects were considered, but did not improve the model or show a statistically significant relationship to spacing style. These include the total word length in letters, the presence of a final tone letter on the first syllable, the presence of a bound morpheme, semantic opacity, and word frequency. Details on the model used and the variables considered are found in Appendix B.

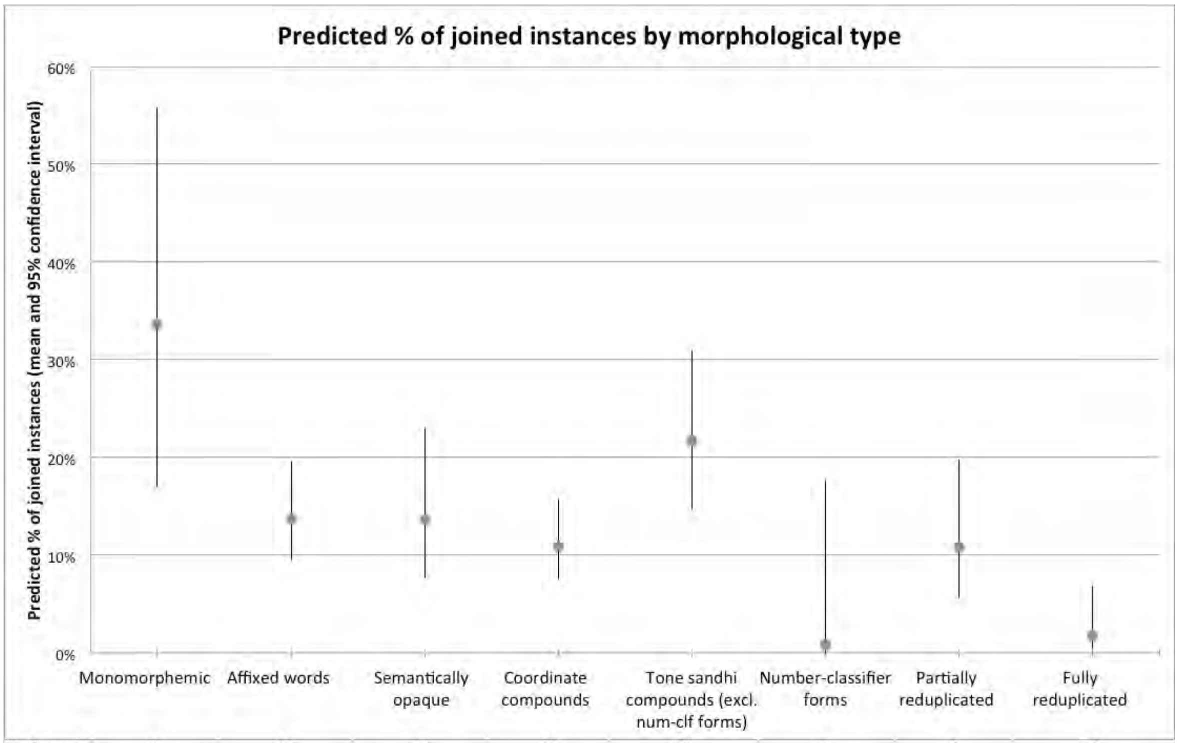
When four-syllable elaborate expressions are included in the analysis, then the number of syllables is highly significant (p<.001), since none of the four-syllable elaborate expressions in the word list are ever written unspaced in the soc.culture.hmong corpus.

4.2.1 SCH results by morphological type

The results for the soc.culture.hmong corpus were also analyzed according to the morphological types of words listed in **Table 2**. The breakdown for each type is found in Figure 1. Elaborate expressions were excluded, as none of them showed any instances of joining, and therefore reliable error estimates could not be made. Results for tone sandhi compounds were split into number-classifier constructions and all other tone sandhi compounds. Reduplicated words were split into partially reduplicated and fully reduplicated words.

The model’s predicted percentage of words in the unspaced format for each word type are given with a circle, and 95% confidence intervals for the means are given with error bars. The results of statistical significance tests (p-values) are also given. Since the dependent variable used is logarithmic, means are not necessarily at the center of the confidence intervals.

We can see that all word types except for monomorphemic words are significantly more likely to be written as spaced than unspaced. The mean percentage of words in the unspaced format is much higher for monomorphemic words than for other types of words, and that the means for number-classifier forms and fully reduplicated forms are much lower. Tone sandhi compounds, excluding number-classifier forms, are moderately more likely to be unspaced than other words.



# of words	11	13	10	16	11	5	4	6
p	0.129	<.001	<.001	<.001	<.001	0.018	0.006	0.001

Figure 1: Predicted % of joined instances out of total instances by morphological type

5. Discussion

5.1 Morphological factors influencing the choice of spacing style

Several morphological factors were found to influence the spacing style that Hmong writers choose.

5.1.1 Number of morphemes

In the SCH corpus, monomorphemic words, such as *phooj ywg* ‘friend’ or *hauj lwm* ‘work’, are more likely to be written unspaced than compounds or affixed words. This effect is not surprising, since the individual syllables of a monomorphemic word have no meaning (or at least no meaning that helpfully relates in any way to the word at hand). Reading research also suggests that monomorphemic words are processed as a single unit in reading more often than polymorphemic words (Ji et al. 2011, Juhasz 2006, Duñabeitia et al. 2008, Muncer et al. 2014).

5.1.2 Fully reduplicated words

Fully reduplicated words, such as *kheej kheej* ‘round’ or *rhiab rhiab* ‘to tickle’, tend to be written with a space in the SCH corpus. This suggests the possibility that a linguistic definition of the word, which includes full reduplication as a word formation process, matches neither the average Hmong writer’s understanding of a word, nor their idea of what makes for optimal reading. The implications of this will be considered below.

5.1.2 Elaborate expressions

Finally, this study did not find a single instance of a four-syllable elaborate expression from the word list written unspaced. Hmong writers’ avoidance of four-syllable orthographic words most likely relates to their perception that Hmong words are usually only one syllable, perhaps two, but certainly not as long as four syllables. Since morpheme breaks within elaborate expressions are often hard to define, it is more natural for writers to simply break them up into their syllable constituents.

5.2 Syntactic factors

One syntactic variable was found to influence spacing choice. Although number-classifier constructions have phonological unity through tone sandhi, they still show a strong tendency to be written separately. These words were included as part of the word list studied, with the belief that their phonological unity was sufficient reason to classify them as words and not phrases. In retrospect, since phonological and grammatical definitions of a word do not always align (Dixon & Aikhenvald 2002), the number-classifier constructions such as *ib tug* 'one person' (derived from *ib* 'one' and *tus* 'CLF.person') are probably best described in Hmong as phrases, not words. Although these particular number-classifier constructions show phonological unity through tone sandhi, all other number-classifier constructions unambiguously contain two phonologically independent syntactic units. The phonological unity of certain of these constructions, then, should not trump the syntactic equivalence of these constructions with other number-classifier constructions that are clearly two separate words. That constructions like *ib tug* 'one person' or *ib qho* 'one part' (from *ib* 'one' and *qhov* 'CLF.part') are ever written as a unit, however rarely, is probably because of their phonological unity. Note, however, that unlike typical clitics, these forms are not phonetically "reduced"; they simply undergo a tone change in the classifier.

5.3 First syllable effects

Two variables relating to the first syllable of disyllabic words influenced the choice of spacing style. First, a given target word was more likely to be written unspaced if the first syllable of the word was usually found as part of that target word, and not by itself or in other words. Contrasting examples in English would be "very" (when readers see the syllable <ver>, they are likely reading the word "very," and not another word like "verdant", "veracity", etc.), versus "bespeckle" (<be> is nearly always found on its own or in other words, and not as part of the word "bespeckle"). The Hmong words akin to "very", such as *ntiaj teb* 'earth' (literally 'surface-earth') or the monomorphemic *phooj ywg* 'friend', tend to be written unspaced. Hmong words that are like "bespeckle", such as *rua lo* 'yawn' (literally 'open-CLF.mouthful') or *qhov rai* 'window' ('Noun.classifier.holes-window'), tend to be spaced. This effect cannot simply be a result of Hmong writers trying to avoid meaningless first syllables of monomorphemic words, since the model included the number of morphemes as a separate variable, but the effect of first syllable frequency in the target word versus other words still existed.

Perhaps the "very" type of words in Hmong tend to be unspaced because the first syllable immediately triggers the whole word in writers' minds, so they are more likely to conceive of the word as a single unit when making spacing decisions. This triggering of the whole word also makes the whole word route faster, which an unspaced format facilitates but a spaced format hinders. In contrast, <be> does not trigger "bespeckled," giving no advantage to the whole word route. So a sensitivity to cognitive processing would favor the unspaced format for "very" type words as well.

Secondly, words with longer first syllables are more likely to be written separately in the SCH data. Examples of words with long first syllables are *tshaib plab* 'hungry' (literally 'hungry-stomach'), or *nplooj ntoos* 'leaf' (literally 'leaf-tree'). Both of these examples have an initial consonant trigraph, a two-letter vowel grapheme, and a final tone letter. The greatest contributor to syllable length in Hmong Daw is the length of the initial consonant grapheme, which can have as many as four letters or be absent entirely (such as in *ib* 'one').

The effect of first syllable length matches Kuperman & Bertram 2013's findings that longer left constituents in English compound words result in a greater likelihood of being written with a space. This may be because short first constituents make it possible for readers to clearly see the syllable boundary, lowering the likelihood that they will need a second or third fixation on the word (Bertram & Hyönä 2003). To the extent that writers are aware (consciously or not) of such processing factors, they would be less likely to separate words with short first constituents, where the syllable boundary is easily available to readers on the first fixation, with or without a space. This may also help explain the effect of first syllable frequency within the word versus on its own or in other words. Clearly, the first constituent is more important than the second in determining both the cognitive processing of morphologically complex words, and the way writers determine spacing.

5.4 Dialectal factors

One dialectal factor influenced spacing style. Words with a large number of spelling differences between Hmong Daw and Hmong Njua are more likely to be unspaced in the SCH corpus. Examples are *me nyuam* ‘child’ (literally ‘small-little’, spelled <miv nyuas> in Hmong Njua), unspaced in 43% of instances, and *kab tsib* ‘sugarcane’ (<quav ntsuas> in Hmong Njua), unspaced in 61% of instances, compared with 15% for the average word on the list. It seems that some writers, at least, are aware when words in Hmong Daw differ significantly from their Hmong Njua counterparts, and they react to these differences by writing these words as unspaced.

5.5 Linguistic versus orthographic “words” in Hmong Daw

As described above, I relied on a variety of sources to determine whether certain Hmong constructions should be considered linguistic “words” or not, particularly Martha Ratliff’s description of Hmong morphology (2009, 2010). Ratliff does not directly address the question of whether certain constructions are words, but only whether the processes unifying them are morphological or syntactic. A more nuanced understanding that included the syntactic relationship in the number-classifier constructions would have treated these constructions as two separate words linguistically, which matches the tendency of Hmong writers to separate these constructions with a space.

While the writing results for number-classifier constructions spurred their reanalysis as phrases and not words linguistically, many Hmong Daw constructions that could truly be considered “words” from a linguistic perspective are also rarely written in an unspaced format, even by writers who are using interword spaces. This includes four-syllable elaborate expressions, which show clear morphological processes, but which Hmong writers seem to think have too many syllables to be joined. This sense reflects what we have seen in the reading research relating to the visual acuity principle (Bertram & Hyönä 2003), and the challenge of parsing long orthographic words that extend beyond the fovea.

Another category of linguistic words that resist being joined orthographically in Hmong Daw is fully reduplicated forms such as *rhiab rhiab*, ‘to keep tickling’. Unlike number-classifier constructions, these function as a single unit both phonologically and syntactically. Also, unlike the four-syllable constructions, some fully reduplicated forms are occasionally found written unspaced in the SCH corpus, such as *dhiadhia* ‘to keep running/jumping’ (unspaced in 3 out of 71 instances, or 4.2%), or *ntauntau* ‘very much’ (unspaced in 25 of 2131 instances, or 1.2%). However, fully reduplicated words as a whole are only unspaced in the SCH corpus an average of 1.4% of instances, compared to 17.8% for other words on average, which the regression shows to be a statistically significant difference. Fully reduplicated words are also never found written unspaced in either the Heimbach (1979) or the Xiong (2005) dictionaries of Hmong Daw, nor in the United Bible Societies New Testament (United Bible Societies 2000).

It seems, then, that full reduplication does not tend to lead to writing as a unit, neither in Hmong dictionaries nor in popular usage. This is despite the fact that reduplication shows phonological unity, which both Ratliff (2009, 2010) and I consider a morphological process (and therefore a word-formation process) rather than a syntactic one. Ease of reading may again be playing a role here in the choice of spacing style by Hmong writers. According to the literature on the cognitive processing of morphologically complex words (Inhoff et al. 2000, Juhasz et al. 2005, Ji et al. 2011, Frisson et al. 2008), the advantage of orthographically joining a polymorphemic word mainly lies in the way it helps readers interpret the semantic unity of the construction, whereas the advantage of separating each constituent with spaces lies in easier access to each constituent (in this case, each syllable). For fully reduplicated forms, both nouns and verbs, there are no competing interpretations that orthographic joining would disambiguate. Meanwhile, adding a space to fully reduplicated forms allows faster access to the individual syllable/morpheme, which contains the main meaning the reader is trying to determine.

The value of isolating the first part of a reduplicated form for quicker access, especially in the case of full reduplication, can be seen in the method used in the Pahawh Hmong script of adding the symbol 𖭃 to indicate full reduplication, much as the symbol ๆ does for Thai. Once readers have read the first element, they do not need to see the entire form represented again, and they certainly do not need to see these two forms joined together in a way that obscures the main morphemic information they need. Rather, they simply need to know that reduplication is occurring, and then interpret the resulting change in meaning accordingly. Perhaps some symbol could function in the same way for Latin script orthographies, either a punctuation mark or a letter which, when standing alone, indicates reduplication. Indeed, some Hmong writers from Laos have at times used a reduplication symbol based on the Lao symbol when writing in Latin script RPA (David Mortensen, personal communication, April 24, 2015), suggesting that they see the value of such a sybol for the RPA orthography. In the absence of such a symbol, though, the data in this study suggests that Hmong writers and lexicographers prefer to use a space to separate the constituents of fully reduplicated forms.

To the extent that writers are making spacing style decisions based on what is easiest to read, the findings of this study are useful for considering what types of words may be read more easily in spaced versus unspaced formats. This is especially true when the patterns in spacing styles by writers match the findings of reading research about the level of cognitive unity of different types of words. Linguists can agree that certain constructions, such as elaborate expressions or fully reduplicated forms, are multimorphemic words and not multiword phrases. However, defining a form as a “word” linguistically says nothing about whether that form would be read more easily if it is written as a single unit in an orthography. While this study did not directly address readability, it found that many factors that influence Hmong writers in their choice of spacing style coincide with the findings of previous research on reading. In the case of elaborate expressions and fully reduplicated words, both previous research on reading as well as this study on writing suggest that considering these constructions “words” for orthographic purposes would not be helpful to readers. Recent research on readers of Hmong Daw confirms that elaborate expressions and fully reduplicated words are read more quickly in syllable-spaced than word-spaced format (Vitrano-Wilson 2015). Similarly, words with long first syllables are both harder to read as a unit (Bertram & Hyönä 2003, Juhasz et al. 2005), and less likely to be written unspaced in this study. Considering only whether a certain form is a “word” by linguistic definitions would not lead to optimal readability in an orthography.

6. Conclusion

We have seen in this study several examples of harmony between the factors that influence the choice of spacing style by Hmong writers and the factors that make reading easier. This mirrors the results of Kuperman & Bertram 2013 that writers are, to some degree, sensitive to factors that make reading easier.

Despite these counterexamples, there is a large degree of overlap between factors that influence reading in previous research and the factors influencing writing in this study (i.e., the number of syllables, the length of the first constituent, the number of morphemes, and the effect of full reduplication). This overlap indicates that writers are not just choosing a spacing style at random; rather, they have multiple factors in mind (consciously or unconsciously), many of which relate directly to the goal of making words as easy to read as possible. As Kuperman and Bertram (2013:940) put it:

> [T]he choice of one orthographic variant over others is not arbitrary, but is co-determined by multiple factors…[T]o a large extent…spelling preferences in writing are motivated by the cognitive demands of online word recognition.

The reality is that both the linguistic definition of a word and an orthographic definition based on optimal ease of reading will be ignored when sociolinguistic forces strongly push toward one style or another. The broad spectrum of spacing styles found in mainland Southeast Asia underlines the power of sociolinguistic forces in orthography development. However, even when sociolinguistic forces are neutral, the criterion of readability conflicts at times with purely linguistic criteria for determining word boundaries.

In short, linguistic definitions of a “word” are useful for linguists, but they should not be relied upon too heavily to determine the use of spaces in an orthography. Native speaker intuition, along with an awareness of the factors that affect how different words are processed in reading, are more valuable sources of information for orthography decisions than purely linguistic criteria.

References

Australian Government, Department of Human Services. 2014. Hmong – information in your language. Online: http://www.humanservices.gov.au/customer/information-in-your-language/hmong.

Bertrais, Yves. 1964. *Dictionnaire Hmong-Français* [Hmong-French dictionary]. Vientiane: Mission Catholique.

Bertrais, Yves. 2002. *Ntawv ntshiab: txhais ua lus Hmoob zaum ob* [Holy Bible: Hmong translation, 2nd edition]. Bangkok: Assumption Printing Press. Online: http://hmongrpa.org.

Bertram, Raymond, and Jukka Hyönä. 2003. The length of a complex word modifies the role of morphological structure: Evidence from eye movements when reading short and long Finnish compounds. *Journal of Memory and Language* 48.615–634.

Dixon, R. M. W., and Alexandra Y. Aikhenvald. 2003. Word: A typological framework. In R. M. W. Dixon and Alexandra Y. Aikhenvald eds., *Word: A Cross-linguistic Typology*, pp. 1–41. Cambridge: Cambridge University Press.

Fortmann-Roe, Scott. 2012. Accurately measuring model prediction error. Online: http://scott.fortmann-roe.com/docs/MeasuringError.html.

Frisson, Steven; Elizabeth Niswander-Klement; and Alexander Pollatsek. 2008. The role of semantic transparency in the processing of English compound words. *British Journal of Psychology* 99(1).87–107.

Golston, Chris, and Phong Yang. 2001. White Hmong loanword phonology. In Caroline Féry, Antony D. Green, and Ruben van de Viver eds., *Proceedings of HILP 5*, pp. 40–57. Potsdam: University of Potsdam.

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

Heimbach, Ernest E. 1955. *Yoha li tsab ntawv lus hmoob dawb* [The first epistle of John in Hmong Daw]. Chiang Mai: Overseas Missionary Fellowship.

Heimbach, Ernest E. 1965. *Yauha ntawv txojlus zoo hais txog Yexu Khito lus Hmoob Dawb* [The gospel of John in Hmong Daw]. Bangkok: Thai Bible House.

Heimbach, Ernest E. 1979. *White Hmong-English Dictionary*. Ithaca, NY: Southeast Asia Program Publications, Cornell University.

Hmong Baptist National Association. 2015. Hmong Baptist National Association: Serving Hmong Baptists. Online: http://hbna.org.

Hmong District. 2015. Hmong district of the Christian and Missionary Alliance. Online: http://hmongdistrict.org.

Hmong RPA. 2012. Hmong RPA. Online: http://hmongrpa.org.

Hmoob Kav Tos Liv Fab Kis Teb. 2015. Hmoob Kav Tos Liv Fab Kis teb: Aumônerie Catholique des Hmong de France [Hmong French Catholics: Catholic Chaplaincy of the Hmong of France]. Online: http://aumoneriehmong.fr.

Inhoff, Albrecht W.; Ralph Radach; and Dieter Heller. 2000. Complex compounds in German: Interword spaces facilitate segmentation but hinder assignment of meaning. *Journal of Memory and Language* 42.23–50.

Ji, Hongbo; Christina L. Gagné; and Thomas L. Spalding. 2011. Benefits and costs of lexical decomposition and semantic integration during the processing of transparent and opaque English compounds. *Journal of Memory and Language* 65(4).406–430.

Joshua Project. n.d. Hmong Daw in the United States. Online: http://joshuaproject.net/people_groups/12112/US

Juhasz, Barbara J.; Albrecht W. Inhoff; and Keith Rayner. 2005. The role of interword spaces in the processing of English compound words. *Language and Cognitive Processes* 20.291–316.

Kuperman, Victor, and Raymond Bertram. 2013. Moving spaces: Spelling alternation in English noun-noun compounds. *Language and Cognitive Processes* 28(7).939–966.

Lewis, M. Paul; Gary F. Simons; and Charles D. Fennig. (eds.) 2014. *Ethnologue: Languages of the world*, 17th edition. Dallas: SIL International. Online version: http://www.ethnologue.com.

Matisoff, James A. 1973. *The Grammar of Lahu*. Berkeley: University of California Press.

McLaughlin, Carey. 2012. A salience scheme for Hmong Soud: Types of foreground and background information in narrative discourse. Dallas: Graduate Institute for Applied Linguistics Master's thesis.

Mok, Leh Woon. 2009. Word-superiority effect as a function of semantic transparency of Chinese bimorphemic compound words. *Language and Cognitive Processes* 24.1039–1081.

Moua, Blia Yao, and Jonas Vangay, trans. 1989a. *Pib nyeem ntawv: A* [Beginning to read: A], *Houghton Mifflin Literary Readers*. Boston: Houghton Mifflin.

Moua, Blia Yao, and Jonas Vangay, trans. 1989b. *Pib nyeem ntawv: B* [Beginning to read: C], *Houghton Mifflin Literary Readers*. Boston: Houghton Mifflin.

Moua, Blia Yao, and Jonas Vangay, trans. 1989c. *Pib nyeem ntawv: C* [Beginning to read: C], *Houghton Mifflin Literary Readers*. Boston: Houghton Mifflin.

Moua, Mai. 2010. 2010 census Hmong populations by state. Hmong American Partnership. Online: http://www.hmong.org/page33422626.aspx.

Moua, Wameng. 2012. Kab mob siab B, kab mob ntxim ntshai [Hepatitis B, an infectious disease threat]. *Hmong Today*. Online: http://www.hmongtoday.com/news/feature/life/4-kab-mob-siab-b,-kab-mob-ntxim-ntshai.html.

Northpoint Health & Wellness Center. 2014. Hmong language. Online: http://www.northpointhealth.org/HmongLanguage/tabid/140/Default.aspx.

Owensby, Laurel. 1986. Verb serialization in Hmong. In Glenn L. Hendricks, Bruce T. Downing, and Amos S. Deinard eds., *The Hmong in Transition*, pp. 237–43. New York: Center for Migration Studies of New York, Inc.

Ratliff, Martha. 2009. White Hmong vocabulary. In Martin Haspelmath and Uri Tadmor eds., *World Loanword Database*. Munich: Max Planck Digital Library. Online: http://wold.livingsources.org/vocabulary/25.

Ratliff, Martha. 2010. *Meaningful Tone: A Study of Tonal Morphology in Compounds, Form Classes, and Expressive Phrases in White Hmong*. DeKalb, IL: Northern Illinois University Press.

Sandra, Dominiek. 1990. On the representation and processing of compound words: Automatic access to constituent morphemes does not occur. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology* 42(3).529–567.

Smalley, William A.; Chia Koua Vang; and Gnia Yee Yang. (eds.) 1990. *Mother of Writing: The Origin and Development of a Hmong Messianic Script*. Chicago: University of Chicago Press.

Temple of Hmongism. 2013. *Temple of Hmongism: The Hmong Religion of the Future*. Online: http://www.hmongism.org.

Thomas, David. 1962. On defining the "word" in Vietnamese. *Văn-hóa Nguyệt-san* 2.519–523.

U.S. Department of Health and Human Services, Centers for Disease Control and Prevention. 2007. *Koob tshuaj tiv thaiv tus kab mob HPV (Human Papillomavirus)*. Translated by the Minnesota Department of Health. Online: http://www.health-exchange.net/pdfdb/hpvHmo.pdf.

U.S. Department of Health and Human Services, Centers for Disease Control and Prevention. 2009. *Koob tshuaj tivthaiv kabmob pneumococcal polysaccharide* [Pneumococcal polysaccharide vaccine]. Translated by Transcend Translations. Online: http://www.immunize.org/vis/hm_pne97.pdf.

United Bible Societies. 2000. Txoj moo zoo uas yog Vajtswv Txojlus: Phau txhais tshiab [The gospel of God: Bible translation]. New York: United Bible Societies.

Vitrano-Wilson, Seth. 2015. Comparing the readability of syllable spacing and word spacing in Hmong Daw. Chiang Mai: Payap University Master's thesis.

Waisman Center. 2006. *2-MBADD: Ib phau ntawv qhia rau cov tsev neeg Hmoob* [2-MBADD: A guide for Hmong families]. Online: http://www.waisman.wisc.edu/2mbadd/general-hg.html.

Wisconsin Court Interpreter Program. 2006. *Hmong Legal Glossary*. Translated by Minnesota Translation Lab. Online: https://www.wicourts.gov/services/interpreter/hmongglossary.htm.

Wisconsin Department of Human Services, Division of Public Health. 2004a. *Qhoob Qhua Pias Soob - Rubella (German measles)* [Rubella disease fact sheet]. Online: http://www.dhs.wisconsin.gov/publications/P4/P42111h.pdf.

Wisconsin Department of Human Services, Division of Public Health. 2004b. *Tus Mob-Listeriosis* [Listeriosis disease fact sheet]. Online: http://www.dhs.wisconsin.gov/publications/P4/ P42069h.pdf.

Xiong, James B. 2014. *Hmong Dictionary Online*. Online: http://hmongdictionary.us.

Xiong, Jay. 2005. *Lus Hmoob txhais (Hmong-English Dictionary)*. 2nd edition. United States: Author.

Appendix A: Spacing frequency data and sources for Hmong words

The table below contains the data used for the analysis of spacing practices by Hmong writers. The first two columns give the number of instances each word was written joined and separated in the soc.culture.hmong corpus. Spelling variants due to tone sandhi or to differences between Hmong Daw and Hmong Njua spellings in the soc.culture.hmong corpus are included in the overall joined and spaced numbers for the first two columns. The complete data set with the number of variants for each word is available upon request from the author.

Abbreviations for word types are as follows. Some words have more than one reason for being considered words.

op	semantically opaque compound
b	bound morpheme
mono	monomorphemic
sesq	sesquimorphemic (first syllable is a fossilized morpheme)
redup	reduplication
ee	4-syllable elaborate expression
ts	tone sandhi compound
orth	normally written unspaced
cc	coordinated compound, “semantic reduplication”
phon	phonological unity (apart from tone sandhi)
V-obj	a verbal compound with a fixed, obligatory object incorporated into the verb

The table below is sorted by the ratio of joined to total instances of each word in the SCH corpus, from highest to lowest.

Word	**Joined**	**Spaced**	**Word gloss**	**Word type**	**Source**
kab tsib	52	34	sugarcane	mono	Ratliff 2009
phooj ywg	7,106	5,920	friend	mono	Ratliff 2009
tab sis	15,704	14,136	but	orth	Xiong 2005:388
kaj ntug	391	353	dawn, daylight	ts	Ratliff 2009
taj laj	212	195	market	mono	Ratliff 2009
poj niam	10,608	11,500	woman	cc	Ratliff 2009
nkag siab	1,217	1,321	understand	op	Ratliff 2009

Word	Joined	Spaced	Word gloss	Word type	Source
me nyuam	5,511	6,691	child	cc	Ratliff 2009
Kaj Siab	439	539	satisfied	op	my own analysis (‘bright-heart’ = satisfied)
pob zeb	777	1,009	stone	b	Ratliff 2009
hauj lwm	4,126	5,586	work	mono	Ratliff 2009
teb chaws	9,054	13,013	country	ts	Ratliff 2009
tswv yim	3,373	5,322	idea, thought	mono	Ratliff 2009
dab tsi	6,321	10,058	what	mono	Ratliff 2009
tam sim	1,844	3,551	now, immediately	mono	Ratliff 2009
yooj yim	847	1,818	easy	mono	Ratliff 2009
tag kis	282	622	tomorrow	mono	Ratliff 2009
hiav txwv	82	215	sea	b	Ratliff 2009
viv ncaus	52	165	sisters	mono	Ratliff 2009
ntiaj teb	1,415	4,854	world, earth	cc	Ratliff 2009
kev cai	1,152	4,057	law, custom	cc	Ratliff 2009
zib mu	3	11	honey	ts	Ratliff 2009
ua si	400	1,585	to play	sesq	Ratliff 2009
huv si	49	230	all	sesq	Ratliff 2009, Heimbach 1979:56
pas dej	89	420	lake	lex?	see below
Ratliff 2009 says *pas dej* (literally ‘lake-water’) is phrasal according to its rules (no tone change, semantically transparent, not a coordinated compound), but then gives it as an example of how the rules are somewhat arbitrary, since *pas dej* is usually thought of as a single unit, and when you ask how to say “lake,” you always get *pas dej* and not just *pas*. Xiong 2005 has it as a single entry, along with just *pas*. So it seems to be a lexicalized compound, albeit a semantically transparent one that does not show any phonological unity or have a coordinate form.					
me ntsis	776	3,875	few, a little bit	cc	Ratliff 2009
xeeb ntxwv	103	523	descendants	phon, b	Ratliff 2009
ko taw	144	742	foot	b	Ratliff 2009
di ncauj	5	26	lips	cc	Ratliff 2009
pob caus	19	100	knot	b	Ratliff 2009
pob txha	110	601	bone	b	Ratliff 2009
sov so	8	47	warm	redup	Ratliff 2010:86
pluag tshais	1	6	breakfast	ts	Heimbach 1979:253
viav vias	2	12	swing	redup	based on Ratliff 2009, 2010
dav hlau	168	1,011	airplane	op	Ratliff 2009
niam txiv	636	3,979	parents; married couple	cc	Ratliff 2009
hauv pliaj	23	145	forehead	cc	Ratliff 2009
caj dab	112	707	neck	b	Ratliff 2009
nom tswv	466	2,959	leaders, officials	cc	Ratliff 2009
nus muag	10	65	siblings	cc	Ratliff 2009
tub nkeeg	45	293	lazy	ts	Ratliff 2009
qaub ncaug	10	70	saliva	ts	Ratliff 2010:191

Word	Joined	Spaced	Word gloss	Word type	Source
zom zaws	64	458	(restricted post-verbal intensifier)	mono	Heimbach 1979:477
taub dag	5	36	pumpkin	ts	Ratliff 2009
tsev neeg	577	4,162	family	op	Ratliff 2009
nees nkaum	4	30	twenty	phon	Ratliff 2009
dab tuag	17	130	ugly; sloppy; ghost	op	Ratliff 2009
plab hlaub	7	55	calf (of the leg)	op	Ratliff 2009
cua nab	2	16	worm	sesq	Ratliff 2009
caj npab	9	75	arm	b	see below
based on Ratliff 2009's description of *caj-* as body part noun class prefix					
qhov ntswg	17	144	nose	b	Ratliff 2009
hauv ncoo	12	106	pillow	phon	Ratliff 2009
nab qa	1	9	lizard	ts, op	Ratliff 2009
tsov rog	75	727	war	op	Ratliff 2009
sawv ntxov	37	360	early morning	op	Ratliff 2009
zaub mov	59	636	food	cc	Ratliff 2009
ntsej muag	260	2,851	face	cc	Ratliff 2009
ris tsho	40	441	clothing	cc	Ratliff 2009
sib ceg	127	1,459	argue	b, ts	Ratliff 2009
qhov rooj	85	1,021	door	b	Ratliff 2009
tiv thaiv	84	1,016	protect	cc	Ratliff 2009
ntxoov ntxoo	2	25	shadow, shade	redup	Ratliff 2009
qhov tsua	21	274	cave	b	see below
based on Ratliff 2009's description of *qhov-* as a noun class prefix					
khov kho	40	530	strong	redup	Ratliff 2010:86
qhov rai	7	94	window	b	Ratliff 2009
aub ncaug	4	60	saliva	ts	Ratliff 2010:164
tsaug zog	58	886	sleep	op	Ratliff 2009
xws li	109	2,042	as, like	cc	my own analysis (see Mortensen 2003)
tshaib plab	18	388	hungry	V-obj	Ratliff 2009
ib leeg	185	4,656	one person	ts	Ratliff 2009
zoo nkauj	96	2,428	beautiful	cc	Ratliff 2009
dhia dhia	3	78	keep jumping/running	redup	Owensby 1986:238
kawm ntawv	168	4,389	to study	V-obj	Ratliff 2009
kawg nkaus	70	2,010	(superlative marker)	lex?	see below
The word *kawg nkaus* is a single entry in Xiong 2005, and is used as a unit in superlative constructions. *Kawg* can mean 'extremity' or 'end', and *nkaus* is a restricted post-verbal intensifier according to Heimbach 1979:153. While there are no clear indicators of its word status, it appears to be lexicalized as a superlative with a fixed form.					
ob tug	135	6,252	two people/animals	ts	Heimbach 1979:326
ntev ntev	9	442	very long	redup	Ratliff 2009
ib los	37	2,160	one mouthful	ts	Ratliff 2010:37
loj loj	13	1,009	very big	redup	based on Ratliff 2009,

Word	Joined	Spaced	Word gloss	Word type	Source
					2010
ntau ntau	25	2,106	very much, a lot	redup	based on Ratliff 2009, 2010
nplooj ntoos	1	96	leaf	ts	Ratliff 2010:179
ib qho	58	7,481	one thing	ts	Ratliff 2009
ib tug	250	34,290	one person/animal	ts	Ratliff 2010:30
nkhaus niv nkhaus nom	0	0	curvy, crooked	ee	Johns & Strecker 1987:106
cheb cheb	0	1	keep sweeping	redup	Owensby 1986:239
diav rawg	0	1	fork	op	my own analysis (‘spoon-chopsticks’ = fork)
khwv iab khwv daw	0	1	hard work	ee	Johns & Strecker 1987:106
tseg tub tseg ki	0	2	bereft of children	ee	Heimbach 1979:82; Johns & Strecker 1987:110
kev mob kev tuag	0	5	sickness	ee	Heimbach 1979:81
zaj sawv	0	10	rainbow	op	Ratliff 2009
pog koob yawg koob	0	17	ancestors	ee	Ratliff 2009
ua qoob ua loo	0	17	agriculture	ee	Heimbach 1979:265; Johns & Strecker 1987:109
rua lo	0	20	yawn	orth	Xiong 2005:370
cua daj cua dub	0	29	storm	ee	Ratliff 2009, Johns & Strecker 1987:109
rhiab rhiab	0	50	keep tickling	redup	based on Ratliff 2009, 2010
kheej kheej	0	59	round	redup	Ratliff 2009
ua dog ua dig	0	370	do haphazardly	ee	Ratliff 2010:162

Appendix B: Models, SPSS Output and Syntax

Independent variables considered

The Hmong spacing practices analysis used a linear regression, and considered the following word-related variables:

- Number of letters
- Number of letters in the first syllable
- Presence or absence of a final tone letter on the first syllable (presence=1, absence=0)
- Having bound morphemes (yes=1, no=0)
- Being a fully reduplicated word (yes=1, no=0)
- Number of morphemes (monomorphemic=1, sesquimorphemic=1.5, dimorphemic=2)

- Number of phonological words (tone sandhi words, reduplicated words, and words with spreading nasalization=1, other=2)
- Semantically opaque (monomorphemic words and semantically opaque compounds=1, all others=0)
- Number-classifier form (yes=1, no=0)
- Noun-noun form (yes=1, no=0)
- Having a verbal constituent (yes=1, no=0)
- Frequency of first syllable in the SCH corpus (number of instances)
- Ratio of instances in the SCH corpus of the first syllable of the word in that target word over total instances (including in other words or on its own)
- Log of instances of word in the SCH corpus
- Number of instances of word in the Catholic Bible
- Number of spelling differences between Hmong Daw and Hmong Njua for target word (counted as # of grapheme substitutions necessary)

Variable selection methods

All the regressions in this study used a backward selection method, with the Adjusted R^2 value as the criterion for selection. If removing a variable resulted in an improvement of Adjusted R^2, or a less than .01 improvement, the variable was removed. Variables were removed more liberally than suggested by Adjusted R^2 because it is known to underpenalize for model complexity (Fortmann-Roe 2012).

Besides removing variables when their retention did not sufficiently improve Adjusted R^2, variables were also removed if significant multicollinearity was found.

Model results

The models that were tested used the log ratio of unspaced instances over spaced instances of the word as the dependent variable. For words that have zero unspaced instances, the negative infinite value of the log ratio was replaced by the following function:

$$f(n) = \ln\left(\mu_{spaced}{}^{\frac{-1}{n+1}} - 1\right)$$

where *n* is the number of total instances (that is, spaced instances) of the word, and μ_{spaced} is the mean ratio of spaced to total instances for all words in the data set, equal to 0.838. This function represents an estimate of the log ratio of unspaced over spaced instances that each word would have in a larger corpus, based on how often words in general are spaced or unspaced in this data set and how many spaced instances occur for a given word.

The independent variables kept in the model were:

- Number of morphemes
- Letters in the first syllable
- Number-classifier form
- Fully reduplicated form
- Ratio of first syllable frequency in target word over total frequency
- Number of Hmong Daw/Hmong Njua grapheme differences

Here is the SPSS syntax and output for the model:

```
REGRESSION
  /STATISTICS COEFF R ANOVA COLLIN TOL
  /DEPENDENT schlnjs
  /METHOD=ENTER zmorphemes zlet1stsyl znumclf zfullredup z1stsylinword zHLspelldiff.
```

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.817	.667	.642	.860466

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	117.377	6	19.563	26.422	.000
	Residual	58.492	79	.740		
	Total	175.868	85			

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-2.193	.093		-23.619	.000
	Zscore(morphemes)	-.306	.104	-.215	-2.954	.004
	Zscore: let 1st syl	-.377	.109	-.264	-3.446	.001
	Zscore(numclf)	-.618	.106	-.434	-5.849	.000
	Zscore(fullredup)	-.581	.104	-.380	-5.562	.000
	Zscore: 1st syl ratio in.word/tot	.421	.105	.294	3.991	.000
	Zscore: HD-HL # grapheme differences	.242	.099	.169	2.451	.016

Coefficients (continued)

Model		95.0% Confidence Interval for B		Collinearity Statistics	
		Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	-2.378	-2.008		
	Zscore(morphemes)	-.512	-.100	.796	1.257
	Zscore: let 1st syl	-.595	-.159	.717	1.395
	Zscore(numclf)	-.828	-.407	.764	1.308
	Zscore(fullredup)	-.789	-.373	.900	1.111
	Zscore: 1st syl ratio in.word/tot	.211	.631	.776	1.289
	Zscore: HD-HL # grapheme differences	.046	.439	.885	1.130