Chapter 4
Spacing Practices in Hmong Daw

4.1 Choice of main language for testing

My experience with the pilot testing in Akha and Lahu Si gave me greater insight into the ideal characteristics of a language for testing word versus syllable spacing. I needed a language with:

1. Many readers
2. One dominant, established orthography
3. A fairly neutral sociolinguistic situation surrounding spacing and orthography in general
4. A well-studied and described grammar and morphology
5. Fairly monosyllabic word patterns
6. Text commonly found in both syllable-spaced and word-spaced format

The reasons for 1 and 2 are related. Only languages with established orthographies are likely to have enough readers to provide a sufficient sample size for testing a spacing hypothesis. Criterion 2 is also necessary to avoid the complexity of analysis and interpretation, however interesting in itself, that we saw in the Akha situation. Criterion 3 is clear enough—it would be difficult to find enough readers with as unbiased a demographic sampling as possible, and to provide a valuable contribution to the language community itself, if significant internal political issues cloud the orthography situation. Criterion 4 is a prerequisite for meaningful linguistic analysis of spacing issues, as well as a great help to the preparation of texts with different spacing styles.

For criterion 5, a highly monosyllabic language is obviously less affected by the entire question of word vs. syllable spacing. However, since my main interest in the thesis is learn more about whether word spacing is easier to read than syllable spacing, testing a highly monosyllabic language offers the best chance of finding an answer. If even a highly monosyllabic language does better with word spacing, then word spacing is more likely to be beneficial for all languages. On the other hand, detecting an advantage for syllable spacing in a highly monosyllabic language would

be presumably more likely to occur. Thus, the chances of a meaningful result from testing are maximized.[3]

Criterion 6 makes it much easier to interpret the data. If a language that only uses syllable-spaced text is read faster with syllable spacing, it might be due to an inherent benefit of syllable spacing, or it could be merely a product of what readers are more used to. Ideally, all readers would be equally familiar with both formats, although any real situation will only offer at best an approximation of this ideal.

Hmong Daw fits all of these criteria well, and also has a fairly large and easily accessible population in and near Chiang Mai. There are also readers in the US, whom I was able to recruit for test development and for testing itself. Texts are also found easily in both word-spaced and syllable-spaced format, and most readers are exposed to both formats, although, as we shall see, the different ways to create "word-spaced" Hmong text will still cause some difficulties for us. Still, Hmong Daw writing is far closer to the ideal of having readers exposed to both word-spaced and syllable-spaced text than most writing systems, which typically have one style with less variation. For all these reasons, I chose to use Hmong Daw as the main language for my research.

4.2 Hmong linguistics, sociolinguistics, and orthography

Hmong Daw, or White Hmong, 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 (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, but are analyzed as a single phoneme. Hmong Daw has no final consonants. 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; Ratliff 2010:112), which does not

[3] Of course, it may be possible to detect a potential benefit for syllable spacing in a language that is not predominantly monosyllabic, as was the case in Akha testing, in which case the significance would be even greater. However interesting and surprising these results were, I did not want to assume a similarly fortuitous result for another language, especially given the complicating factors which may have caused this result for Akha.

occur in my sample texts. A few morphemes are polysyllabic, though most of these seem to be loanwords, or compounds whose original morphemes have been lost. In addition to having mainly monosyllabic morphemes, words in Hmong Daw are also mostly monosyllabic (Golston & Yang 2001). Polysyllabic words can be formed by 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 (CQD), in which syllable spacing is standard (McLaughlin 2012).

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 2. These differences reflect differences in pronunciation between the two varieties.

Table 2 Spelling correspondences between Hmong Daw and Hmong Njua

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

4.3 Spacing in different communities

Spacing in the Hmong Daw RPA orthography varies from writer to writer. Many books can be found that separate every syllable with a space, while other publications join certain words together. There are some generalizations that can be made, however. 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 (Hmong District 2015, Hmong Baptist National Association 2015). Hmong Njua texts exhibit the same difference in spacing styles between Catholic and Protestant materials.

Secular texts show both patterns of spacing. For instance, the stories on which I based my story and sentence tests in Hmong Daw (see Chapter 5) come from a series of literacy primers for Hmong students in the US. In all these books, there is a space between every syllable, regardless of morphology, apart from a few English names, and a few onomatopoeic animal sounds that use hyphens (Moua & Vangay 1989a, 1989b, 1989c). Syllable-spaced Hmong Daw text can also be found in one 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 I have used for this thesis 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. <me nyuam> vs. <menyuam> in Waisman Center 2006). Even single-author dictionaries vary at times in their spacing style (e.g. <Fab Kis>, <Fabkis> in Xiong 2005:59; <mis niv>, <misniv> in Xiong 2005:165). The variation of spacing styles even in dictionaries reflects a relative lack of concern about the standardization of spacing in word-spaced Hmong Daw text.

4.4 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. Since I failed to find any research on the question of how often particular words in Hmong Daw are written spaced or unspaced, I performed my own study on the matter.

4.4.1 Spacing frequency analysis methodology

The goal of the analysis of Hmong spacing practices was to estimate the frequency with which the polysyllabic words in my tests of Hmong Daw reading (see Chapters 5 and 6) were found separated or joined, and what word-related factors influence the spacing style writers choose. I used three main data sets for this analysis:

1. The online group soc.culture.hmong, or SCH
2. The text of the United Bible Societies translation of the Hmong Daw New Testament (United Bible Societies 2000)
3. The text of the Catholic translation of the Hmong Daw Bible (both Old and New Testaments)

The first two sources have words in both spaced and unspaced formats, whereas in the Catholic Bible translation, all syllables are separated with spaces. The Catholic Bible was not analyzed for spacing practices, but instead used to calculate word frequencies in biblical text.

Dr. David Mortensen 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 edited the text to eliminate nonword 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. 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. The resulting word frequency numbers are therefore frequencies in both varieties of Hmong, not Hmong Daw alone.

For both the SCH corpus and the UBS New Testament, I found all spaced and unspaced instances of each word on the word list (see Chapter 6) and each polysyllabic word in the story and sentence tests (see Chapter 5). The resulting

numbers are found in Appendix C. I then modeled the resulting data using IBM SPSS v.22. Different model types for the two data sets were necessary, because the spacing frequency numbers for the SCH corpus were on a continuous spectrum, whereas in the UBS NT, the words were either separated throughout the text, or joined throughout. The SCH model used a linear regression model, while the second used a logistic regression, which can be used with dichotomous dependent variables. SPSS output and syntax for the regressions are found in Appendix H.

4.4.2 Analysis using the soc.culture.hmong corpus

The main analysis of the 15-million word SCH corpus excluded 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. 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 spacing:

- **Number-classifier forms.** Number-classifier forms (which are all tone sandhi compounds, such as *ib qho* "one part", or *ob tug* "two people") are more likely to be spaced than other types of words ($\beta_{stand} = -.434$, $p < .001$).
- **Fully reduplicated forms.** Words formed by full reduplication are more likely to be spaced ($\beta_{stand} = -.380$, $p < .001$).
- **Ratio of first syllable frequency in target word over total frequency.** 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 unspaced ($\beta_{stand} = .294$, $p < .001$).
- **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 spaced ($\beta_{stand} = -.264$, $p = .001$).

- **Number of morphemes.** Monomorphemic words are more likely to be unspaced than polymorphemic words ($\beta_{stand} = -.215$, $p = .004$).
- **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 ($\beta_{stand} = .169$, $p = .016$).

No other effects were included in the model. The overall regression model had a significance of $p < .001$. The model and data satisfied the criteria of normality and lack of multicollinearity. Details on the model used are found in Appendix H.

When four-syllable elaborate expressions are included in the model, then the number of syllables is also highly significant ($p < .001$), since none of the four-syllable elaborate expressions in my word list were found unspaced in the soc.culture.hmong corpus.

4.4.3 Analysis using the UBS New Testament text

Since spacing in the UBS New Testament tended to be deliberately chosen by the translators and editors, spacing style was nearly always consistent for a given word within the text, though two words did show variation. The ratio of joined instances to total instances was therefore either zero or one for all but these two words. For simplicity, these two words were removed from the model. Any word used in my tests that did not appear in the New Testament text was also removed. This left 61 words from the original 96.

Out of the 61 words analyzed, 29 were written unspaced throughout the UBS New Testament text, 30 were written as spaced, and two words were found in both styles. Modeling the spacing in this text with a logistic regression gives us the following significant factors:

- **Ratio of first syllable frequency in target word over total frequency.** Words in which the first syllable nearly always occurs in that particular word in the soc.culture.hmong corpus, and not by itself or in another word, are more likely to be written unspaced in the UBS New Testament ($\beta_{stand} = 1.26$, $p = .001$).
- **Noun-noun compounds.** Noun-noun compounds are more likely to be written unspaced than other words on average ($\beta_{stand} = 1.22$, $p = .004$).

- **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 ($\beta_{stand} = 1.26$, $p = .014$).

No other effects reached statistical significance. The overall regression had a significance of $p < .001$. In this text, as with the soc.culture.hmong corpus, there were no instances of four-syllable elaborate expressions written as an unspaced unit.

4.4.4 Discussion

Let us compare the results from the SCH corpus and the UBS New Testament, and the degree to which these results overlap.

4.4.4.1 Factors influencing spacing style in both data sets

Comparing the models for the two data sets, we see that three variables influenced spacing style in both models. In both the SCH corpus and the UBS New Testament, 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 probably reading the word "very", and not another word like "verdant", "veracity", etc.), versus "bespeckle" (an instance of the syllable "be" is very unlikely to be part of the word "bespeckle" instead of "be", "become", "begin", etc.). The Hmong words akin to "very" tend to be written unspaced, and the words that are like "bespeckle" 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 SCH 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 writers strongly associate the first syllable with the whole word for these words, so they are more likely to conceive of them as a single unit. Another explanation is that writers are somehow sensitive to the way these types of words are processed cognitively. The first syllable of "very" words is likely to trigger the meaning of the whole word in readers' minds and make 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.

Secondly, in neither the SCH corpus nor the UBS New Testament was there a single instance of a four-syllable elaborate expression from my tests 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.

Finally, in both data sets, a larger number of spelling differences between Hmong Daw and Hmong Njua resulted in a greater likelihood of a word being written unspaced. It seems that some writers, including the editors and translators of the UBS New Testament, 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. This suggests that these writers see word spacing as the easier choice for reading when faced with dialect differences that could cause problems for readers.

One reason why there are not more shared factors that influence spacing style in both data sets relates to the statistical structure of the data. The UBS New Testament represents the consensus view of a group of translators or a translation committee, who, for the most part, use the same spacing style consistently for a given word. The SCH corpus, in contrast, represents probably thousands of Hmong writers who all have their own spacing preferences, and who may be inconsistent within their own writing. There is therefore much more nuanced information in the SCH data that can be used to estimate the factors that influence spacing style than for the UBS New Testament. As a result, more factors show statistical significance for the SCH model.

4.4.4.2 Factor influencing spacing style in the UBS NT only

There is, however, one factor that is statistically significant for the UBS New Testament, but not for the SCH data. The writers of the UBS New Testament seem to be more influenced by the syntactic criterion of whether a word is a noun-noun compound or not. They typically chose to write noun-noun compounds as unspaced. This variable did not affect the SCH data.

4.4.4.3 Factors influencing spacing style in the SCH data only

Several factors influence spacing style in the SCH corpus, but not in the UBS New Testament.

Number-classifier forms

Although number-classifier constructions have phonological unity through tone sandhi, they still show a strong tendency to be written separately. These constructions were counted as polysyllabic words in this thesis, with the belief that their phonological unity indicated they were words and not phrases. In retrospect, since phonological and grammatical definitions of a word do not always align (Dixon & Aikhenvald 2003), the number-classifier constructions such as *ib tug* "one 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 and syntactically independent units. The phonological unity of certain number-classifier 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" 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 due to the neighboring numeral.

Length of first syllable

Words with longer first syllables are more likely to be spaced in the SCH model. 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 spaced. The result for Hmong writers may be because short first syllables make it easier for readers to find the syllable boundary even in unspaced format, 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 syllables, where the syllable boundary is easily available to readers on the first fixation, with or without a space. Clearly, the first constituent is more important than the second in determining both the cognitive processing of polysyllabic words, and the way writers determine spacing.

Number of morphemes

In the SCH corpus, monomorphemic words are more likely to be written as a unit than compounds or affixed words. This 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).

Fully reduplicated words

Finally, fully reduplicated words tend to be spaced. This raises the possibility that the definition of a word used for the Hmong Daw texts, which relies fairly heavily on phonological criteria, 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.

4.4.5 Implications for Hmong testing

As Section 5.2.1.1 describes, 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 is not directly addressing the question of whether certain constructions are words, but only whether they show signs of phonological or morphological fusion. A more nuanced definition of a "word", which includes the syntactic relationship in the number-classifier constructions, would treat these constructions as two separate words linguistically, matching the tendency of Hmong writers to separate these constructions with a space. This minor issue affects a small number of sentences and words in the Hmong tests developed for this thesis. However, Section 5.3.6 reveals that excluding sentences with number-classifier constructions does not change the results for the sentence test, and Section 6.8.3.1 shows that the same is true for the test of isolated words.

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 are "too long" to be joined in writing. 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. Thankfully, four-syllable expressions are

marginal in my test of isolated words, and non-existent in the tests of connected text, so this issue will have little or no effect on the results.

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. None of the fully reduplicated test words were found unspaced in the UBS NT. Fully reduplicated words are also never found unspaced in either the Heimbach (1979) or the Xiong (2005) dictionaries of Hmong Daw.

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 (2010) and I consider a morphological process (and therefore a word-formation process) rather than a syntactic one. We will revisit the issue of fully reduplicated words and how spaces affect their reading in Section 6.9.3.2.

4.4.6 Summary

We have seen in this chapter 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.

There are, of course, other factors that research would suggest influence the reading process that we do not see here, as well as factors seen in this writing analysis that have not appeared in the literature on reading. For instance, Sandra 1990, Frisson et al. 2008, and Mok 2009 suggest that semantically opaque compounds would benefit from an unspaced format more than semantically transparent compounds. However, this study does not find a statistically significant difference relating to the semantic opacity of compound words. Similarly, there has been no research done that suggests that dialectal differences would make an unspaced format easier to read,

yet we see spelling differences between Hmong Daw and Hmong Njua to be a significant factor influencing the spacing style used in both the SCH corpus and the UBS New Testament.

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 (e.g., 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.

To the extent that writers are making spacing style decisions based on what is easiest to read, the findings of this chapter are useful for considering what types of words may be read more easily in spaced versus unspaced formats. Section 7.2 compares the results from this chapter with the results from the test of isolated words in Chapter 6. This will tell us the degree to which the results from this chapter are applicable to questions of readability in Hmong Daw, with implications for people making orthography decisions.