Chapter 1
Introduction

1.1 Overview

When language communities develop or revise a writing system for their language, they must not only make decisions about how to represent the different sounds of their language. They must also decide how to help readers process the meaning of the text by joining and separating sounds to mark smaller and larger linguistic units. For writing systems in which the phoneme is the lowest level of representation, this means joining these phonemes into syllables, morphemes, words, phrases, clauses, sentences, or paragraphs.

Most writing systems use blank spaces to visually mark off linguistic units. European-derived writing systems, as well as many others, use spaces primarily to mark word boundaries. But spaces can be used to mark other levels as well. For instance, Thai, Burmese, Lao, and other Brahmi-based scripts in Southeast Asia use spaces to mark clauses and other larger units. Some languages, such as Vietnamese, Lahu, or Lisu, use spaces to mark syllables rather than words.

Depending on their sociolinguistic context, some groups may take it for granted that spaces will be used to separate words, rather than larger or smaller units. But in areas where the sociolinguistic context surrounding the use of spaces is more diverse, such as in Mainland Southeast Asia, groups often have a difficult time deciding what linguistic level spaces should demarcate. Moreover, even if a particular group agrees, for instance, that spaces should mark out "words", it can be quite challenging to decide exactly what a "word" is in their language. The same may occasionally be true for defining what a "syllable" is for a given language. Is there anything that an outsider can do to help groups with such decisions?

In order to consider possible answers to that question, we must first give an overview of the criteria that tend to influence groups' decisions as they create a writing system for their language.

1.2 Principles of writing system development

William Smalley, a pioneer in the field of writing system development (or "orthography" development as it is often called), gave the following criteria in order of importance as he viewed it (1964:38):

- Maximum Motivation
- Maximum Representation
- Maximum Ease of Learning
- Maximum Ease of Transfer
- Maximum Reproduction

Malone (2004:38) gives a similar list, saying that the ideal writing system is "a writing system that:

- is acceptable to the majority of the Mother Tongue (MT) speakers of the language;
- is acceptable to the government;
- represents the sounds of the language accurately;
- is as easy as possible to learn;
- enables MT speakers to transfer between the minority and majority languages; and
- can be reproduced and printed easily."

Clearly, the most important criteria are that the writing system be acceptable to most members of the language community, and that it make them motivated and excited to read and write. If a system is acceptable to the speakers of the language and motivates them to read and write, it will be successful, whatever its shortcomings in other ways. An obvious example that readers of this thesis are experiencing at this very moment is the failure of the English spelling system to "represent the sounds of the language accurately" or be "as easy as possible to learn". Yet in terms of usage, the English writing system is wildly successful. Since sociolinguistic criteria tend to be dominant in the minds of native speakers as they make orthography decisions, they naturally and rightfully trump other, more "technical" criteria (Sebba 2009, Unseth 2005, Karan 2006, Cahill 2014).

That said, there are many cases in which sociolinguistic pressures are not strong for a given decision, and the other criteria listed by Smalley and Malone come into play. When considering decisions about spacing, the main technical criterion involved is the ease of learning and using a system. Spaces do not change the way the sounds of the language are represented. Rather, the function of spaces is to segment text into units that allow readers to understand the meaning of the text more quickly, easily, and accurately.

1.3 How do we know what is easiest?

Unfortunately for language groups developing a writing system, languages do not come with manuals describing the "easiest" spacing style for reading and writing. How can we know whether different ways of using spaces will be easier or harder for a given language?

First of all, we must distinguish between ease of reading and ease of writing. Whereas readers are rarely aware of exactly what their eyes and brains are doing as they read, writers tend to be more consciously aware of the writing process. We may find that the easiest style to read may or may not be also easy to write, since the visual representation of language that allows the fastest reading and greatest comprehension may not necessarily follow simple spelling rules or be intuitive for writers.

If we choose to focus first on ease of reading, we could empirically test differences in readability by measuring the time and comprehension of readers who read different spacing styles. Ideally, such a situation would involve spacing styles that readers are equally familiar with, using natural text that readers would typically encounter in their own language. If a difference can be found between different spacing styles, then the style that results in faster reading and greater comprehension would be easier to read, all else being equal.

A major problem with this approach is that readers are not blank slates. They are already used to certain ways of using spaces (or not using spaces), and their familiarity or lack thereof may influence their reading times and comprehension. For instance, Rayner et al. 1998 found that when readers of English read text with spaces removed, they read 40% slower than normal on average. This result alone does not, however, guarantee that interword spaces are inherently beneficial for English (though they probably are, as we will see in Chapter 2). It could be that the unfamiliarity of the unspaced format is disrupting readers, and that if they practiced reading unspaced text, they might become equally or more proficient in reading it than they are in reading the familiar format with interword spaces. Of course, if readers read an *unfamiliar* format even faster than a familiar format, this is clear evidence that the unfamiliar format would increase reading speed.

One possible solution to this conundrum is to test reading speed for a language that currently has no writing system at all, so that no particular spacing format is "familiar" or "unfamiliar" for that language. However, there are two major problems with such an approach. One is that readers of an entirely new writing system may exhibit patterns in reading that would change as they become proficient and experienced readers in that system. In other words, what may hold true about "optimal" spacing for new or beginning readers may not necessarily hold true once they become proficient readers. Secondly, even these readers are not blank slates, since, if they can read at all (which they must if they are to be tested for reading speed and comprehension!), the spacing styles of the languages they already read may influence how they read in their mother tongue.

In the end, as with many questions in social science, finding ideal experimental conditions is impossible. We could never find a language with readers who are all equally familiar and proficient with two or more different spacing styles, having both beginning and advanced readers, so that one could test the difference between spacing styles without any influence of familiarity or other confounding variables. In this thesis, I will do my best to minimize such factors, and to acknowledge their possible influence whenever possible.

In addition to empirically testing reading speed and comprehension, another valuable source of information is the intuition and evaluations of mother tongue readers themselves. They may have a sense of the difficulty of reading different styles that could go undetected by empirical means due to a lack of statistical power.

1.4 Research objectives

As mentioned above, writing systems in Mainland Southeast Asia are diverse in their use of spaces, ranging from spaces between syllables to no spaces at all. Fairly few linguists have published guidelines on spacing decisions for orthography development, but those who have done so mainly work in Africa, where the use of spaces and the typical morphological profile are both quite different. It is not surprising, then, that these guides have tended to take for granted the assumption that groups will be using word spacing, and the only question is how to define a "word" in a given language (Kutsch Lojenga 2014, Eaton & Schroeder 2012, Cahill & Karan 2008, Karan 2006, Van Dyken & Kutsch Lojenga 1993).

Many studies have compared the reading of text with interword spaces to unspaced or phrase-spaced text (e.g. Bai et al. 2008, Bassetti & Masterson 2012, Sainio et al.

2007, Winskel et al. 2009). However, the comparison of word-spaced text to syllable-spaced text is neglected in the literature. The only published studies in English on the subject are of Chinese pinyin, which, although quite valuable, are complicated by the dominance of the character-based hanzi system for Chinese (Bassetti 2009, Bassetti & Masterson 2012). As a result, language groups considering questions of readability and spacing have relatively little empirical information to go on if they are considering syllable spacing as an option.

In order to remedy this situation, the main objective of this thesis is to compare the reading of syllable spacing and word spacing for languages in Mainland Southeast Asia. The thesis is designed not only to compare syllable spacing and word spacing generally, but also to compare the effect of spacing style on different types of words (for example, monomorphemic vs. polymorphemic words, semantically opaque vs. transparent compounds, etc.), as well as different types of readers (young vs. old, more vs. less proficient readers, etc.).

In order to accomplish these objectives, I will focus on Hmong Daw, or White Hmong, a Hmong-Mien language with speakers in China, Vietnam, Laos, Thailand, the United States, and elsewhere. The results from this study will help us to answer the following specific research questions:

1. Is word spacing optimal for Hmong Daw?
2. What word-related factors predict the choice between syllable spacing and word spacing by Hmong writers?
3. What factors predict any differences in reading speed between syllable spacing and word spacing in Hmong Daw?

For the first question, "optimal" is defined both as it relates to reading and to writing. An optimal system for reading would lead to the fastest reading times and highest comprehension. An optimal system for writing would be the most intuitive and simplest for writers to spell, so that spelling variation would be minimized. The experiments in this thesis focus primarily on answering the question of optimality for reading. However, a discussion of relevant issues from the literature, as well as results from this thesis, provide indirect evidence to the question of optimality for writing as well.

1.5 Limits to this thesis

The goal of this thesis is to help language groups who are making decisions about the use of spaces in their language. This help is limited to those for whom it is relevant—that is, for groups who do not have strong sociolinguistic preferences related to spaces, and would like to know how different spacing styles would affect their reading. It is also limited to comparing word-spaced text to syllable-spaced text, since little research has been done in this area (although a review of the literature in Chapter 2 may be helpful to those interested in comparing other spacing styles).

The thesis is limited to languages using the Latin script. For the sake of methodological simplicity, the study is also limited to comparing the presence or absence of spaces, and does not consider the use of other segmentation devices such as hyphens or apostrophes. This is not because such segmenting punctuation marks might not be helpful to groups—indeed, some research suggests that using secondary segmentation devices in addition to blank spaces can facilitate reading in certain contexts (Häikiö et al. 2011, Bertram et al. 2011, Bertram & Hyönä 2013).

Because of its diverse use of spaces, Mainland Southeast Asia provides a fertile ground for studying the differences between syllable spacing and word spacing. Many languages in the region are generally isolating languages, with a high number of monosyllabic morphemes and monomorphemic words, and with frequent use of compounding to form words (Enfield 2005). The writing systems of national languages in the region are syllable-based rather than word-based, and literacy instruction in the region tends to focus on the syllable level (Page 2014). For all these reasons, it seems to be the most likely area in which a benefit for syllable spacing might be found, or where word spacing would show no benefit over syllable spacing. While a benefit for syllable spacing found in this region would not necessarily imply a benefit elsewhere, it would at least demonstrate that word spacing is not universally beneficial, but should instead be empirically tested for different languages. Conversely, a benefit to word spacing in this region would strongly suggest that word spacing is beneficial across most or perhaps all languages.

Most of the testing in this thesis uses Hmong Daw. The results from pilot testing in Akha and in Lahu Si, two Tibeto-Burman languages found in China, Myanmar, and Thailand, provide further context and refinement of methodologies.

1.6 Structure of the thesis

Chapter 2 reviews the literature on spaces and reading, and describes spacing practices in Mainland Southeast Asia. Chapter 3 describes the pilot testing in Lahu Si and Akha, including the methodologies, results, and suggested improvements to the methodology. Chapter 4 describes spacing practices in Hmong Daw, and includes an analysis of the factors that influence how Hmong writers choose to use spaces for different types of Hmong words. This analysis will help us answer the second research question about the factors affecting the choice of spacing by Hmong writers.

Chapter 5 describes the methodology and results of experiments comparing syllable spacing and word spacing using connected text in Hmong Daw. Chapter 6 gives the methodology and results of an experiment using isolated words in Hmong Daw, which shows the effect of spacing style on reading speed and comprehension for different kinds of words. Chapters 5 and 6 together will help us answer the first and third research questions comparing syllable and word spacing for Hmong Daw, and seeing which factors influence the effect of spacing style. Chapter 7 concludes the thesis, discussing the implications of all the results for orthography decisions.

After the Bibliography section, Appendix A gives the stories tested in Lahu Si, and Appendix B contains the Akha stories tested. Appendix C provides the spacing frequency data from Hmong texts analyzed in Chapter 4. Appendix D gives the sources that indicate word status for all polysyllabic words in Hmong Daw used for this thesis. Appendix E has the Hmong stories tested, as presented to Hmong readers in the US. Appendix F gives the same Hmong stories as Appendix E, but with some modifications for readers in Thailand, presented in sentence-by-sentence format. Appendix G provides the Hmong word list presented to readers in the US and Thailand. Finally, Appendix H provides the models, results, and SPSS syntax for the statistical models used throughout the thesis.