Chapter 2
Literature Review

2.1 The cognitive process of reading

Psycholinguists have done a great deal of research on how our eyes and brains see and interpret text while reading. In particular, they have found that our eyes do not move smoothly and regularly across the page. Rather, they make jumps, called "saccades", from one point to another, collecting images as they go that our brains then decode into language and meaning. Each saccade is followed by a "fixation", typically lasting around 200 to 250 milliseconds. The saccade itself lasts around 30 ms, but the overall time it takes for the brain to plan and execute a saccade is around 175 to 200 ms. Most saccades move forward in the text, but we also skip backwards around 10-15% of the time (Rayner & Pollatsek 2013:443).

Our eyes see things differently depending on how close to the center of our gaze objects are. In the fovea, a region within 2 degrees of the center of our gaze, we can see most clearly. The parafovea reaches from 2 to 5 degrees away from the center, while our peripheral vision is least detailed. Because reading requires attention to minute visual details, most reading occurs within the fovea. Eye movements occur in order to bring new words into the foveal region (Rayner & Pollatsek 2013:443–4).

2.1.1 How language and orthography affect reading

While the basic process of reading is the same no matter what language or orthography one is reading, some things do vary. For instance, the distance moved in a saccade depends on many factors, including the difficulty of the text, the reader's ability, and the language and orthography being read. Readers of English typically move their eyes around 7 to 9 letter spaces horizontally. Readers of Chinese, whose characters pack visual and linguistic information much more densely, move their eyes around 2 to 3 characters. Readers of languages that are written left-to-right pick up more information to the right of their main focus, while Hebrew and Arabic readers pick up more information to the left (Rayner & Pollatsek 2013:443–4, Jordan et al. 2014).

For alphabetic languages using spaces between words, saccades seem to target a spot halfway between the beginning and the middle of the fixated word. This spot is called the "preferred viewing location" (Rayner & Pollatsek 2013:448). Long words are more likely to be fixated twice, and short words are often not fixated on, but are read while fixating on the word before or after (Rayner & Pollatsek 2013:443).

It is not entirely clear whether readers who are reading languages without interword spaces also have a preferred viewing location near the middle of the word or not. Clearly, without spaces to mark word boundaries, readers have less information to help them decide where the word boundaries are, and therefore where the middle of the next word is. Some studies (Li et al. 2011, Li & Shen 2013) argue that Chinese readers do not target a preferred viewing location based on word boundaries, although other studies (Reilly et al. 2011, Yan et al. 2012, Reilly 2013) disagree, and Li et al. (2011) say their data is inconclusive on the matter. Japanese readers do, however, use word boundary information to target saccade fixations, but their landing site is typically much closer to the beginning of the word than in English (Kajii et al. 2001; Sainio et al. 2007). Perhaps Japanese readers are able to target saccades based on word boundaries because Japanese writing typically involves multiple scripts together, with the change from one script to another providing some information about word boundaries that is not present in Chinese. When interword spaces are added to pure *hiragana* (syllabary) text in Japanese, the preferred viewing location shifts toward the center of the word, as with English (Sainio et al. 2007). In Thai, an alphabetic script with no interword spaces, readers tend to target the preferred viewing location just to the left of the center of each word, as in English (Reilly et al. 2005, Winskel et al. 2009, Reilly et al. 2011). This implies that Thai readers must be picking up on some visual cues to determine word boundaries as they read, even without interword spaces.

Although the use of word boundaries for saccade targeting is not universal, it seems that words are nonetheless important units of reading for most or all languages. In addition to the studies on Japanese and Thai mentioned above, several studies have found that Chinese readers do not simply process text character by character (that is, syllable by syllable or morpheme by morpheme), but that words influence the reading process (Gu & Li 2015, Li et al. 2014, Li et al. 2013, Li et al. 2009, Bai et al. 2008, Li & Ma 2012, Li et al. 2012). In other words, the claim that some linguists have made that Chinese "has no words," or that the word is a meaningless category for Chinese (see Duanmu 1998:135), does not match how Chinese readers process text. The fact that skilled Chinese and English readers use remarkably similar

processes to read (Li et al. 2014), and can process information at roughly similar rates (Sun & Feng 1999), further adds to the overall case that "reading shares a word-based core and may be fundamentally similar across languages with highly dissimilar scripts" (Li et al. 2014:895).

2.1.2 The paraorthographic linkage hypothesis

Clearly, there are both similarities and differences in how readers of different languages process text. On the one hand, all readers use saccades and fixations to gather visual information, and it seems that all readers also use information both at the grapheme level and at the word level to help them determine optimal eye movements. On the other hand, scripts and languages can differ in precisely *how* saccades, fixations, and saccade landing sites are determined. This is perhaps best summed up in the concept of the "paraorthographic linkage hypothesis," presented by Gary Feng (2008).

Feng defines the "paraorthography" as "a set of graphic symbols and conventions of a writing system that does not directly transcribe linguistic information but exists to ensure faithful transmission of the written message" (Feng 2008:403). Punctuation, the spaces between words and characters, capitalization, paragraph breaks, and the formatting of title or subject headings all help readers to understand the ways that letters, words, sentences, and ideas are connected and separated. As such, they do not provide the same kind of linguistic information as graphemes do, but they nonetheless communicate important and meaningful information that is of great help to readers.

All orthographies have some way of presenting characters in a visual context, but writing systems vary considerably in the linguistic units they represent orthographically. For instance, most alphabetic systems represent phonemes fairly clearly and regularly, but not all alphabets represent syllables clearly, and few give unambiguous morpheme boundary information. Chinese characters represent one-syllable morphemes (except in a very few cases), but the Chinese writing system does not typically indicate phoneme or word boundaries (Mok 2009:1040). A few systems, such as the Korean Hangul writing system and Canadian Aboriginal Syllabics, represent units as small as phonetic features (Lee & Ramsey 2000:43, The Unicode Consortium 2003:149). Hangul is fascinating in that it marks so many linguistic units clearly in one system—features, phonemes, syllables, words (by interword spaces), and phrases or sentences (by punctuation).

Since different languages have different phonological, morphological, and syntactic structures, we would also expect the importance of denoting certain levels to vary from language to language. For instance, it would not be important to separately mark syllables and morphemes in Mandarin, since they correspond so closely. Similarly, if we have a system for clearly marking syllables, we would expect there to be less cost (if any) associated with failing to also mark word boundaries for Vietnamese, where most words are monosyllabic, than for Inuktitut, where words are much longer than one syllable on average.

The eye movement studies above indicate that differences in paraorthography can cause readers to process text differently. Chinese, English, Japanese, Arabic, and Thai all have different ways of presenting, joining, and separating characters, words, and sentences, and these paraorthographic differences affect how people read. Feng's hypothesis is that "readers constantly adapt and optimize their reading behaviors," so that mature readers "are well-adapted to their own writing systems" (Feng 2008:414). In other words, readers learn to utilize as efficiently as possible whatever clues the orthography and the paraorthography give them.

2.2 History of spacing

Of course, not all paraorthographies are created equal when it comes to ease of reading. In his book, *Space Between Words: The Origins of Silent Reading*, Paul Saenger argues that it was the development of spaces between words in medieval Latin that enabled readers to switch from group oral recitation to silent, private reading (1997). Although ancient Latin texts did in fact separate words, just as the Semitic, vowelless orthographies around them in the Mediterranean did, Latin scribes abandoned word separation in the second century AD, imitating the Greek style of *scriptura continua*. Saenger argues that *scriptura continua* went hand-in-hand with the development of vowel symbols (1997:10). Greek and Latin writers could have imitated the word separation used in Semitic scripts, but did not do so, since the inclusion of vowels made word separation redundant for the style of reading most valued by them—namely, the recitation of particular classic texts by an elite class of readers. Ease of reading, silent reading, and mass literacy, he argues, were not their goals, and therefore they considered interword spaces a literal waste of space.

This apparently started changing in Ireland and Britain in the late 7th century, when scribes less familiar with Latin began to use spaces, albeit erratically, to help them read more efficiently (Saenger 1997:32). The use of spaces in Latin spread and

became more consistent over the medieval period, and by the 11th century, manuscripts in some areas were using interword spacing similar to modern European writing (Saenger 1997:44). The influence of Arabic, which uses word spacing, also increased in this time period (Saenger 1999:12–13). The introduction of interword spaces, Saenger argues, enabled readers of Latin and European vernaculars to switch to the style of rapid, silent reading that is common today. Although we cannot directly test the reading speed and style of ancient Romans compared to modern Italians or Americans, the deleterious effect of removing spaces for modern Western readers (Fisher 1975, Rayner et al. 1998) suggests that the introduction of interword spacing in Latin and in European vernaculars did indeed assist in this transition to rapid, silent reading.

Interword spacing continued to spread throughout Europe and beyond. Greek writers adopted interword spaces in the latter 16th century, and Cyrillic became word-separated by the early 17th century (Saenger 1997:13). The practice then spread throughout Indian scripts by the 18th century (Saenger 1997:13, note 65). Finally, Korean writers, influenced by Western orthographic practices, first used interword spaces in 1896, and the practice spread throughout society in the 20th century (King 1998:39–40, Traulsen 2011:104,122).

2.3 But what is a word?

Saying that an orthography uses "interword spaces", of course, assumes that the concept of a word in all these languages is straightforward. As it turns out, defining exactly what a word is can be quite challenging.

If you ask most English readers to define a "word", at some point they will likely refer to the unit marked by spaces when writing. But for a language first deciding how to use interword spaces, this is clearly a circular definition. Is a word an actual linguistic unit, or is it simply an orthographic convention that has gained psychological reality for those who use it?

Orthographic conventions are not fully reliable in defining linguistic units. Is there any relevant linguistic difference between "hot dog" and "greenhouse" that would lead one to be separated and the other joined in English? The fact that even native English speakers often spell the same compound with or without spaces (Kuperman & Bertram 2013) also attests to the unreliability of orthography on its own as a definition of a word.

One approach to defining a word linguistically is through phonology. For some languages, phonological patterns such as tone, intonation, stress, nasalization, or vowel harmony operate on a word level rather than a syllable, morpheme, or phrase level. For instance, Turkish vowel harmony operates at the phonological word level (Julien 2006:617), as does tone in Risiangku (Matisoff 2001:311). Similarly, certain compound words in Hmong Daw show tone sandhi operating within the word, but not across words (Ratliff 2010). Since a word by definition must be equal to or larger than a syllable, phonotactics also helps determine phonological word status. Contractions in English such as *he's* and *I'd* form single phonological words, even though they are syntactically two separate elements. Because the *'s* and *'d* elements are phonologically less than a syllable, they cannot qualify as phonological words.

Another definition of word relies on grammatical criteria. A grammatical word is a unit between a morpheme and a phrase in the syntactic hierarchy, meaning that it can "occur in isolation" and has "independent distribution within the phrase or the clause" (Julien 2006:621). This contrasts with affixes, which are restricted in which parts of speech they can attach to.

Most of the time, this overlaps with the phonological definition above, but not always (Dixon & Aikhenvald 2003). For instance, in the example above, *I'd* consists of a single phonological word, but two separate grammatical words, since a phrase like *My wife and I'd like a vacation* (where the *'d* relates to the whole noun phrase and not just *I*) would be grammatically normal. The common term for a morpheme that acts as a grammatical word but is phonologically bound is a "clitic". This lack of overlap between the two definitions helps explain why clitics are a major gray area for language groups deciding on word boundaries.

If clitics push the lower limits of a word, then compound words push the upper limits. Linguists might define "evening chemistry class" as a single compound word (and, indeed, German orthographic convention would concur), but most English readers would balk at this. Indeed, the words in English that are most commonly seen in multiple formats—spaced, unspaced, and also sometimes hyphenated—are compound words (Kuperman & Bertram 2013). An example in Hmong Daw would be *cua dab cua dub*, "storm" (literally "wind-yellow-wind-black"), an elaborate expression showing clear lexicalization and exhibiting semantic and phonological evidence of word formation (Ratliff 2009). Yet most Hmong Daw speakers would intuitively say this expression is "too long" to be considered a word.

However difficult to define on the edges, the eye movement studies in diverse languages mentioned above support the idea that words are cognitively real. Readers naturally segment text into meaningful units, with or without help from spaces. Words, therefore, are indeed a useful linguistic unit, even if there are borderline cases that defy strict classification. After all, words are not the only linguistic category with fuzzy boundaries. Do the English words "fire" and "toil" have one syllable or two? Do "mulberry" and "twilight" have one morpheme or two, since "mul" and "twi" have no meaning on their own, but "berry" and "light" are clearly morphemes? Despite borderline cases, then, the utility of the word in understanding how readers process text makes it a meaningful unit in linguistics.

For the purpose of determining the optimal spacing for ease of reading, what matters most is not the precise phonological or grammatical definition of a word as it applies to a given language. The question we need to answer is: How can text be segmented in such a way that matches most closely what the reader would do in their minds anyway? That is, how can the paraorthography bear as much of the burden as possible in segmenting text on behalf of the reader, so that the reader can ascertain meaning as quickly as possible?

2.4 How do our minds segment text?

Let us now explore the research on the ways our minds use phonological, morphological, syntactic, and semantic information to segment text while reading.

2.4.1 Reading compound words

One well-developed area of research is the way that the word and morpheme levels interact within compound words. Compound words provide a good way of comparing the word and morpheme levels, since they can be either joined as a single word or separated by their morpheme components. If readers tend to access each component morpheme of a compound word separately, then we would expect that a space between components would facilitate this access. On the other hand, if readers tend to access compound words as a whole, then we would expect separation of components to hinder this access.

Various experiments have provided evidence that, in fact, compound words are read *both* as separate units *and* as a single whole. Pollatsek et al. 2000 describes this process as a "race," where the reader's mind initiates both the morphological decomposition route and the direct lexical access of the whole word, and whichever

route is first successful in decoding the word "wins." This model helps make sense of research that seems to point to differing routes of processing for different readers, word types, and spacing styles.

For instance, Sandra 1990 found that semantic priming did not affect reaction times for a lexical decision experiment with semantically opaque compounds. That is, seeing "bread" beforehand doesn't make recognizing "buttercup" any faster. With semantically transparent compounds, though, there was a priming effect. In other words, in this experiment, readers seemed to read opaque compounds as whole words without accessing the individual morphemes, but they did access the morphemes for transparent compounds. According to the "race" model, morphological decomposition wins for semantically transparent compounds, but whole word lexical access wins for opaque compounds. Ji et al. 2011 found a similar effect for a lexical decision task. Spaces between constituents of normally unspaced opaque compound words slowed readers down in the experiment, but transparent compounds were not affected.

Another study by Frisson and colleagues (2008) showed that semantic transparency or opacity had no effect on any eye movement parameters for normally unspaced English compounds. Using the race model as a framework, one could say that the whole word route is winning for all the words in the study. The authors were not certain as to why semantic transparency influenced the effect of semantic priming in Sandra 1990, but did not influence eye movement parameters in their study. Interestingly, though, they found that if a space is added to a normally unspaced compound, transparent compounds were read a little faster than opaque compounds. They argue that the space is "forcing" a morphological decomposition route for the compounds, which is successful for the semantically transparent compounds, but not for the opaque ones. Readers then have to reanalyze opaque compounds as a single unit, ignoring the space indicating separation. Frisson et al. 2008, Ji et al. 2011, and Sandra 1990 all give us reason to believe that semantically transparent and semantically opaque compounds are sometimes processed differently. When the meaning of a compound is not a straightforward product of its constituents, readers tend to conceive of the word as more of a single unit than when the meaning can be easily determined from its parts.

Age and reading experience also seem to affect the route of processing for compound words. Häikiö et al. 2011 compared the eye movements of Finnish elementary school children reading compounds that were either unspaced or

hyphenated. The study found that all the 4th and 6th grade students, as well as the fastest 2nd grade students, read the unspaced compounds faster than the hyphenated compounds. In contrast, slower 2nd grade students read the hyphenated compounds faster. The authors say that this indicates a difference in the route the students take to processing the compounds. The more experienced readers are processing the compound as a whole, so a hyphen slows them down. The less experienced readers are processing each constituent separately, so a hyphen helps them. This fits with previous research indicating that younger children tend to have smaller perceptual spans (Rayner 1986, Häikiö et al. 2009), shorter and more frequent saccades, and more refixations of words than experienced adult readers (Rayner 1998). With smaller perceptual spans and more time needed to process words, it is not surprising that younger students are more likely than older students or adults to process compound words by their constituents rather than as a whole.

Hyönä 2012 reviews research on Finnish compound words, describing how word length and orthographic cues affect processing. Longer compounds tend to be processed by morphological decomposition, while short compounds are read as a whole (Bertram & Hyönä 2003). Readers also use linguistic cues such as vowel harmony rules to help them parse compound words. In Finnish, vowel harmony operates on the phonological word, and compounds can be formed from multiple phonological words. So readers may note a morpheme boundary within a compound by seeing the vowel harmony rules violated within the compound. Bertram et al. 2004 found that compounds in which vowel disharmony indicates morpheme boundaries are read faster than compounds without vowel disharmony. They also found that this effect was stronger for long compounds than for short compounds. Certain word-internal consonant combinations could also provide clear morpheme boundary clues to readers, if they are never found except at morpheme boundaries. In these cases, Bertram et al. 2004 also found a significant effect on reading speed. Compounds containing the morpheme boundary clues read faster than those where the boundary was more orthographically ambiguous. Finally, a similar study of Dutch compound words also found that consonant clusters that unambiguously mark a morpheme boundary helped readers process compounds faster (Lemhöfer et al. 2011).

Another major factor in compound processing is the whole word frequency and the constituent frequency. Research is clear that whole word frequency is important; more frequent compound words are processed faster (Juhasz & Berkowitz 2011). The effect of constituent frequency on processing is less clear.

All of this research supports the idea that readers are flexible in how they approach compound words. Spacing, word length, morpheme boundary information, semantic transparency, their reading experience, and presumably other factors determine whether readers find the morphological decomposition route or the whole word processing route to be more efficient. Since different routes are taken for even the same words at times, Pollatsek et al. 2000's more flexible "race" model best makes sense of these patterns. Readers do not choose beforehand which route to take; they simply try both simultaneously and see which one reaches the goal of lexical access first.

2.4.1.1 Chinese compound processing

The research above only describes the reading and writing of compound words in European, Latin script orthographies where interword spaces are the norm. Can we generalize our conclusions to include readers of a language like Chinese, where no interword spaces exist, and where compounds and morphemes are both typically much shorter?

Written Chinese has a great number of two-syllable, two-morpheme compound words—72% of all words—as well as several compounds of three or four syllables (Li et al. 2014:896). Since Chinese compounds tend to be shorter both visually and in terms of number of syllables and morphemes than languages like Finnish or Dutch, we might expect a greater likelihood that most or all Chinese compounds would be processed as whole words.

Janssen et al. 2008 gives some indirect support to this idea. It found that whole word frequency was predictive in how long it took Mandarin speakers to name an object in a picture when the name was a compound word. The frequencies of the individual constituents of the compound, however, were not predictive of picture naming latencies. This indicated that, at least when accessing a word mentally rather than orthographically, Mandarin speakers store meaning of compound words as a whole, rather than as a sum of their parts. The study also found the same pattern for native English speakers.

Other studies, however, have shown that in actual reading, Chinese readers access both word-level and morpheme-level information. For instance, Mok 2009 found that all Chinese compounds showed a "word-superiority effect"—that is, characters shown briefly were remembered more easily when part of a word. However, this effect was much stronger for semantically opaque compounds than for transparent

ones. This suggests the possibility that, like the European languages studied, opaque compounds in Chinese are usually processed as a unit, while transparent compounds are usually processed by their constituents. In this study, partially opaque compounds (that is, compounds with one opaque and one transparent element, such as "jailbird" in English) showed the same effect as fully opaque compounds. Mok also brings up the interesting point that semantic transparency is not dichotomous, but should be thought of as a graded variable. Not all semantically opaque compounds are equally opaque, and not all transparent ones are equally transparent. This should be kept in mind when discussing semantic transparency of compounds.

Another study indicated that polysyllabic, monomorphemic words are processed differently by Chinese readers than compound words (Cui, Drieghe et al. 2013). The study measured the "parafoveal-on-foveal" effect of different Chinese words and phrases (that is, a character further away from the center, in the parafovea, affecting the fixation time for the character in the foveal region of viewing). They found that the second character in a two-character string affected fixation time of the first character when the two characters formed a monomorphemic word, such as 玫瑰 "rose", where neither character on its own has meaning. However, they did not see any parafoveal-on-foveal effect either for compound words or phrases. This suggests that polysyllabic monomorphemic words are processed as a unit more than compound words are. This is not surprising, since the first element of a monomorphemic word is not meaningful on its own (except perhaps as a misleading homophone or homograph of a true monosyllabic morpheme), whereas compound words, even opaque ones, have meaningful elements as their first character.

Several studies have indicated that morphology, phonology, and orthography all play a role in processing for Chinese readers, although the morphological elements seem dominant overall. Zhou et al. 1999 found a morpheme priming effect over and above any semantic or orthographic effects. In other words, the study found that being primed beforehand with a morpheme from a word causes faster recognition of the word. This effect is even stronger than when a semantically similar character is shown, or when a homographic character with a different meaning is shown. There was no phonological priming effect—that is, priming with homophonic but not homographic characters did not speed up recognition of a word. This indicates that Chinese readers process not only orthographic and semantic information, but also specifically morphemic information when reading.

Zhou & Marslen-Wilson 2009 found a similar result in a lexical decision test, when replacing real two-character words with "pseudohomophones"—that is, nonwords with one or both characters changed but with the same sound as a real word. An example in English would be "brane" instead of brain. If a pseudohomophone shared one character with a real word, the study found faster lexical decision rates than for random nonwords. But a pseudohomophone with *both* characters changed did no better than random nonwords. This indicates that phonology purely on its own is not particularly helpful for Chinese readers, but it can be helpful in conjunction with the morphemic information in characters. The same result was found using fMRI scans instead of eye movements (Zhan et al. 2013).

Overall, we can conclude that morphemes and graphemes certainly do play some role when Chinese readers process compound words. Given the lack of parafoveal-on-foveal effect for compound words in Cui, Drieghe et al. 2013, we would suspect that when reading, as opposed to accessing lexemes from a picture cue, Chinese readers would process compound words by morpheme rather than as a unit. However, this may be a result of this experiment's manipulation of the text size so as to ensure that the second character lay in the parafovea rather than the fovea (Cui, Drieghe et al. 2013a:406). Since the perceptual span in Chinese is typically four to five characters, and saccade lengths are typically just under three characters (Inhoff & Wu 2005), it seems that having the second character in the parafovea rather than the fovea would not be common when reading a normally sized font. As Bertram and Hyönä (2003) describe, having a whole word in the fovea gives a "visual acuity benefit," making a whole word processing route more likely. In contrast, having the word extend out into the parafovea increases the likelihood that more than one fixation will be required, making the decomposition route more likely.

In the end, there is still much room for debate as to whether the shorter length of Chinese compound words blocks a morphological decomposition route or not. One possible reason why Chinese might employ a morphological decomposition route even for short compounds, while European alphabetic languages would not, is that Chinese spacing is different from the typical European language. In languages like Finnish or Dutch, compound words are joined together visually, with no clear morpheme boundaries. In Chinese, on the other hand, morpheme boundaries are visually distinct, but there is no word boundary information given by the paraorthography. Thus, we should expect that Chinese readers would be more likely than Finnish or Dutch readers to take the decomposition route for compounds. When vowel disharmony or consonant clusters in Finnish and Dutch gave clear

morpheme boundaries, readers were in fact more likely to take the decomposition route (Bertram et al. 2004, Lemhöfer et al. 2011). If indeed Chinese readers are more likely to take the decomposition route for shorter compounds than Dutch or Finnish readers would be, then this would be a fitting example of Feng's "paraorthographic linkage hypothesis"—Chinese, Finnish, and Dutch readers all process compound words as efficiently as possible, given the spatial information available to them in the orthography.

2.4.1.2 How should compounds be segmented to facilitate reading?

We can be confident that mature readers will find ways to maximize their use of the paraorthographic information given them. But is there any indication that some ways of presenting compound words would overall be easier to read than others?

Clearly, the studies above indicate that compound words are not a single category, and their processing depends on many factors. Length, semantic transparency, word frequency, as well as the age and reading experience of the reader, can all influence the processing route readers take. So we may expect that different kinds of compounds would have different optimal spacing style, and that different readers even might have different optimal spacing styles for different compounds. But beyond that, are there any generalizations that can be made?

Inhoff et al. 2000 studied German compounds in different spacing formats. The compounds all had three constituents, and were presented either 1) unspaced as is standard in German orthography (e.g. Datenschutzexperte), 2) unspaced but with a capital letter at the beginning of each constituent (DatenSchutzExperte), or 3) spaced by constituent (Daten schutz experte). The study found that separating constituents by spacing helped readers in a naming recognition task, but did not have any significant effect when reading the words embedded in a sentence. This indicates that separating constituents helped readers to process a single isolated word, but hindered the final process of understanding the meaning of the compound within the sentence, resulting in no net advantage for spacing. It is worth noting, however, that German orthography normally does not have spaces between these compounds, so the unfamiliarity of the spaced format may have slowed readers down somewhat.

Juhasz et al. 2005 studied similar questions in English. Unlike German, English has both compounds that are normally joined (e.g. softball) and those that are normally

separated (e.g. front door). The study looked at how changing the spacing format affected both noun-noun compounds and adjective-noun compounds. As with Inhoff et al. 2000, spaces between the constituents helped with lexical decisions and first fixations while reading sentences, regardless of the standard way of writing these compounds. Unlike Inhoff et al. 2000, adding a space into normally unspaced compounds disrupted processing significantly, whereas taking out an existing space for normally spaced compounds had no effect. The effect for normally unspaced compounds appeared for both noun-noun and adjective-noun compounds, but was larger for the adjective-noun compounds. We see here that several factors are important: spacing style, whether the spacing style is going with or against the standard orthography, and compound type—in this case, noun-noun or adjective-noun. One wonders whether the penalty for adding a space into normally unspaced compounds would decrease or disappear as readers became more familiar with this format, or if it would persist despite increasing familiarity. Also, it is unsurprising but still important to note that the noun-noun compounds were less disrupted by space than the adjective-noun compounds. In English, at least, a noun-noun compound has no obvious syntactic alternative interpretation regardless of spacing (e.g. basketball vs. basket ball), whereas an adjective-noun compound (e.g. greenhouse) could easily be misinterpreted as a phrase (green house) when spaced separately.

One reason why spaces between constituents of compound words may have slowed readers down in Juhasz et al. 2005, even with noun-noun compounds, is brought up by a study on how semantic plausibility within a sentence affects the reading of spaced noun-noun compound words in English (Staub et al. 2007). The authors made a set of sentences with spaced compound words. All the compounds were plausible in context, but the first constituent was either plausible or implausible on its own as the object of the previous verb. The compounds were an average of 12.9 letters long. They found that when the first constituent was implausible, readers were slowed down. This indicates that the spaces between the constituents caused readers to interpret the first constituent as the object, leading to a semantic disconnect. This delayed readers until they read the next word and were able to understand the two constituents as a whole, resolving the implausibility problem.

Cherng (2008) performed a study in English similar to Inhoff et al. 2000 and Juhasz et al. 2005, but comparing an unspaced format to a hyphenated format rather than a spaced format. All the compound words in her study could be found naturally in both unspaced and hyphenated formats, but they varied as to their frequency of

being found unspaced vs. hyphenated. All had two constituents; most were two syllables, and a few were longer. As with Inhoff et al. 2000 and Juhasz et al. 2005, Cherng found that first fixations were longer with the unspaced compounds. Overall, there was no significant difference of total gaze duration between unspaced and hyphenated compounds. However, when looking only at compounds more frequently written as unspaced, Cherng found that hyphenation slowed readers down. Again, this suggests an overall pattern of separation helping word recognition, but hindering comprehension in connected text. The fact that a negative effect was found for hyphenation, when hyphenation gives both morpheme boundary and word boundary information, is somewhat surprising. It may be a combination of their relatively less familiar format for those compounds, plus the fact that most of the words were fairly short and therefore likely to be viewable in one fixation. If so, adding a hyphen would slow down the whole word route and favor an otherwise unnecessary decomposition route.

Bertram et al. 2011 found a different result for hyphens when looking at three-constituent compound words in Finnish and Dutch. Both Finnish and Dutch compounds are normally written unspaced, with no hyphens. The study found that adding a hyphen at the major constituent boundary (for example, airport-taxi, not air-porttaxi) made processing of the first part of the word faster, similar to the studies above. Although the hyphenated format was unfamiliar to readers, both Dutch and Finnish readers showed a learning effect, where their speed with hyphenated compounds increased as they were exposed to more examples. By the end of the experiment, Dutch readers were reading the hyphenated compounds as quickly as the unspaced, and Finnish readers were reading hyphenated compounds even faster than unspaced. When hyphens were added to minor constituent boundaries, however, words were read slower than for the unspaced format.

This intriguing study suggests a few things. One, it seems that readers can fairly quickly get used to a new way of writing if it fits the linguistic structure of the language. Perhaps with even more practice, the benefit of hyphens at major boundaries for Finnish readers would increase even more, and Dutch readers might also see a benefit emerge. Two, in Finnish at least, and perhaps for Dutch also, inserting hyphens at major constituent boundaries of long compound words facilitates reading. This fits with what we saw earlier, that Finnish readers read long compound words via the morphological decomposition route. Hyphens clearly mark morpheme boundaries, while leaving the interword spacing that marks word boundaries intact.

Why are the results of this study different from Cherng 2008? The most obvious cause is that the Finnish and Dutch compound words in Bertram et al. 2011 were much longer than the English words in Cherng's study, in terms of number of letters, number of syllables, and number of morphemes. Since the route of processing seems to depend on word length, this result is not surprising. Indeed, another study on Finnish compounds by Bertram and Hyönä (2013) found that hyphens facilitate reading of long compounds (12.2 letters on average), but hinder the reading of short compounds (7.3 letters on average). They explain this, quite reasonably, by reference to the "visual acuity principle"—that if a word extends beyond the foveal region into the parafovea, it cannot be processed in a single fixation, but the first constituent usually can be. Therefore, the decomposition route gets a "head start" on the whole word route, and it wins the race.

One final study on compound word spacing is Ulfers & David 2000 on the Karang language in Cameroon. Ulfers and David presented different categories of compound words to readers in either spaced or unspaced format, and analyzed the error rates. They found that semantically opaque noun-noun compounds and adjective-noun compounds had fewer errors when written unspaced. In contrast, semantically transparent noun-noun compounds and compounds with a verbal element showed no difference in error rates. We see, then, that both semantic predictability and the internal grammatical structure of compounds may influence their processing.

We have now seen comparisons of spaced vs. unspaced compounds (Inhoff et al. 2000, Juhasz et al. 2005, Ulfers & David 2000), and unspaced vs. hyphenated compounds (Cherng 2008, Bertram et al. 2011). Unfortunately, there do not seem to be any studies yet comparing spaced compounds to hyphenated compounds. One possible clue is given by Bertram and Hyönä (2013), who say that the disruption seen from implausible first constituents of long compound words in Staub et al. 2007 did not occur in their study. In other words, hyphens in Bertram & Hyönä 2013 helped readers avoid misparsing in a way that spaces in Staub et al. 2007 did not. This may indicate that longer compound words would be read more easily with hyphens rather than spaces between constituents. Bertram et al. 2011 found that for short compound words in Finnish, unspaced words were read faster than hyphenated. Since both spaces and hyphens disrupt whole word access, but spaces also remove the indication of the semantic unity of the compound, we might expect spaced compounds to make reading even harder than hyphens for such words.

In summary, it seems that for experienced readers of alphabetic languages that typically use interword spaces, short compound words would be most easily read in unspaced form, whereas long compounds may benefit from hyphens at major constituent boundaries. However, we must keep in mind that younger and more inexperienced readers, with shorter perceptual spans and a higher number of fixations per word, may benefit from having hyphens or spaces between constituents even for shorter compounds (Häikiö et al. 2011). Also, as we saw in Section 2.4.1.1, languages with clear morpheme boundaries but unclear word boundaries may show a benefit to spaces between constituents even for some shorter compounds.

2.4.2 Compounds from the writer's perspective

Although compound processing is usually considered from the reader's point of view, Kuperman & Bertram 2013 provides valuable perspective by looking at how English writers spell compound words, and what that might tell us about preferences for processing. Looking diachronically within a large corpus at compounds that vary in their presentation (unspaced, hyphenated, or spaced), the authors found that more frequent compounds tend to be unspaced. This finding is in line with the idea that more frequent compounds are processed faster (Juhasz & Berkowitz 2011, Janssen et al. 2008), and that direct whole word access tends to be faster than decomposition when easily available. Longer words are also more likely to be spaced, especially when their first constituent is longer, indicating that the benefit of constituent separation to processing long compounds (Bertram & Hyönä 2003, Juhasz et al. 2005) may be intuitively known. 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). The large overlap between the factors that influence the writing and reading of compound words, according to Kuperman and Bertram (2013:940), suggests that "to a large extent...spelling preferences in writing are motivated by the cognitive demands of online word recognition".

One somewhat unexpected result of their analysis is that, contrary to previous assumptions, lexicalization in English does not go through a route of spaced to hyphenated to unspaced formats. Rather, it either goes from spaced straight to unspaced, or goes from hyphenated to spaced, then to unspaced. Words of medium frequency are more likely to be spaced than hyphenated; only the lowest frequency words were most often found hyphenated. This suggests a certain dislike of hyphens among English writers. Given the potential value of hyphenation for longer

compound words seen in Bertram et al. 2011 and Bertram & Hyönä 2013, we might expect hyphenation to be more commonly used in English. One possible explanation that Kuperman and Bertram (2013) give is that for writers (that is, typers, since their corpus was from entirely online sources), it is easier to hit a space bar than to type a hyphen. Since the most frequently used words are those most likely to succumb to forces of economy of effort, hyphenation is only typically found among the least frequent words. Since lexicalization is associated with increasing frequency, hyphenation is only seen at the beginning of this process. This explanation is plausible, but far from proven. It could be as simple as an English cultural bias against hyphen use. Or it could be that English writers are subtly responding to reading preferences, and that hyphens really aren't as helpful as Bertram et al. 2011 and Bertram & Hyönä 2003 suggest.

2.4.3 Segmentation of affixed words

We have looked extensively into how our minds separate and integrate the morphemes of compound words. What can we learn about words with affixes? Do bound morphemes behave the same way in processing text as the free morphemes in compound words?

A great deal of research has been done to understand how our minds process affixed words, and the terms of debate within the research closely mirror that of compound word processing. As with compounds, some research points in the direction of morphological decomposition (Taft & Forster 1975, Duñabeitia et al. 2008, Yan et al. 2014), while other research argues for whole word access being dominant (Manelis & Tharp 1977, Haspelmath & Sims 2010). Within the research favoring decomposition, priming effects have been found not only for morphological stems (e.g. Rastle et al. 2004, McCormick et al. 2008), but also for prefixes (Domínguez et al. 2006) and suffixes (Duñabeitia et al. 2008). Given the conflicting results, a model including a role for both morphological decomposition and whole word access seems likely (see Giraudo & Voga 2014), just as the dual-route "race" model came about within compound word processing research.

Are affixed words any different from compound words when it comes to optimal spacing style? One possible reason they might be is suggested by Ji et al. 2011. The authors argue, following Libben 2005, that the meaning of compounds is more difficult to understand from the meaning of their parts than the meaning of affixed words, because affixes are a "closed-set class," whereas compounds are a

theoretically limitless open-set class, with "no reliable heuristic for parsing" (Ji et al. 2011:407). If compounds are more difficult to interpret from their parts, then we would expect adding a space between compounds to be more detrimental than a space separating an affix from its stem. Conversely, since affixes are a smaller, closed set, we would expect morphological decomposition to occur more easily without the help of spaces, and therefore spaces would give less of an advantage to affixed words than to compounds. Here we might expect semantic opacity to play a role, since semantically transparent compounds presumably cause less difficulty in parsing than opaque compounds.

One other possible piece of evidence comes from Andrews 1986. In this study, suffixed words on their own did not show evidence of being morphologically decomposed by readers. However, a list containing the same suffixed words but also including compound words did show decomposition, both for the compounds and the suffixes. These results indicate that compound words are more commonly decomposed in reading than affixed words, since the compounds provided the "trigger" for decomposition to the affixed words, which were not decomposed on their own. If so, we might expect a space between compound words to be more beneficial, or less detrimental, than a space between a stem and affix.

2.4.4 Monomorphemic, polysyllabic words

Research comparing monomorphemic words to morphologically complex words has confirmed what is intuitive: words with only one morpheme are processed as a single whole more often than words with multiple morphemes. Affixed words and compounds provide additional morphological information that helps readers to access meaning more quickly, whereas monomorphemic words can only be accessed as a whole, or, if they must be accessed in parts, can only use phonological information in integration rather than morphemic information. Worse than that, single-morpheme words with internal spaces could be easily misinterpreted as being polymorphemic. When the syllables of a monomorphemic word happen to also have unrelated meaning on their own (such as the "car" and "pet" in "carpet"), syllable spacing could easily lead readers down false trails of interpretation.

Ji et al. 2011 found that readers completed a lexical decision task faster for compound words than for monomorphemic polysyllabic words. Juhasz 2006 compared compound words and monomorphemic polysyllabic words with the same whole word frequency. The study found that compounds with a high frequency first

constituent were read faster than monomorphemic words, but compounds with a low frequency first constituent were not read any faster. Duñabeitia et al. 2008 compared suffixed words to monomorphemic words, and found a priming effect for suffixes that did not occur for the final letters of the monomorphemic words. Muncer et al. 2014 found that word naming and lexical decision tasks were faster for affixed words than for monomorphemic words. These studies confirm that morphological decomposition may play a beneficial role for processing morphologically complex words that is unavailable to monomorphemic words.

2.4.5 Syllable boundaries and reading

Languages vary in terms of the clarity of their syllable boundaries. For instance, languages like Spanish or Japanese have clear syllable boundaries, whereas English has many words where boundaries are less clear. Not only this, but orthographies vary as to how clearly they mark syllable boundaries. English does nothing to mark syllable boundaries per se, whereas in Japanese, Korean, and Chinese, syllable boundary marking is orthographically obligatory.

Nevertheless, it seems that the syllable plays a role in processing for all these languages, albeit to varying degrees. Several studies have shown clear syllable effects for Spanish (Álvarez et al. 2001, Álvarez et al. 2004, Taft et al. 2007). Syllable effects on processing have also been shown in German (Conrad & Jacobs 2004), French (Mathey & Zagar 2002, Chetail & Mathey 2009), Greek (Aidinis & Nunes 2001), Kannada, which uses an alphasyllabary (Nag & Snowling 2012), and even in English (Ashby & Rayner 2004), although the results for English were more nuanced. The number of syllables in a word affects English readers even when word length is accounted for (Muncer et al. 2014). Japanese uses a syllabary writing system, and native Japanese speakers reading English are much more likely than English readers to show effects of segmenting English text by syllable boundaries (Taft 2002). Taft et al. 2007 found that Spanish readers were more inhibited by adding a space at a non-syllable boundary than English readers. This suggests that the phonological clarity of syllable boundaries may influence the extent to which readers utilize syllable boundary information in processing. The results for Kannada and Japanese suggest that orthography may influence readers' syllabic awareness and use of syllabic information. If so, we would expect readers of languages with phonologically or orthographically clear syllable boundaries to use syllable information more heavily.

2.4.6 The syllable and psycholinguistic grain size theory

The syllable is clearly a more accessible phonological unit for nonreaders and beginning readers than the phoneme. Ziegler and Goswami (2005) describe a range of studies showing that young readers and illiterate adults are often aware of syllables but not phonemes. Page 2014 shows that this phenomenon occurs for beginning readers of all scripts studied so far, not only for syllabaries and the Chinese morpheme-syllabary system, but also for alphabetic scripts and alphasyllabaries.

Ziegler and Goswami (2005) extend this comparison across multiple phonological levels. They argue that awareness of phonological units decreases as one descends to smaller levels, so that young readers are most aware of words, followed by syllables, then onset/rimes, then phonemes. Based on this insight, they develop what they call the "psycholinguistic grain size theory": that different languages with different orthographies have different phonological levels to which readers most easily map symbols. They give three main factors that influence the psycholinguistic grain size for a given language and writing system:

- *availability*, or the accessibility of different phonological levels within the orthography and the reader's awareness of those levels
- *consistency*, or how regularly a given level is represented in the orthography
- *granularity*, reflecting the fact that larger grain sizes involve learning a larger number of possible symbols or symbol combinations.

Since syllable awareness is more common among young and beginning readers than phoneme awareness, a writing system that consistently represents syllables and syllable boundaries would help young readers more than one that does not. This is confirmed by Asfaha et al. 2009, which compares two Ge'ez script syllabary orthographies with two Latin script orthographies in Eritrea. Since all the languages were taught in the same country with the same curriculum and similar teacher training, any differences should be the result of orthographic factors. The readers of Ge'ez script orthographies were reading faster at a younger age than the readers of Latin script orthographies, despite the larger number of symbols in the Ge'ez script.

Since the central question of this thesis is whether there is a difference between syllable-spaced text and word-spaced text in a Latin script orthography, phonemes will always be relatively available to readers orthographically. Depending on the language's phonotactics and orthography, syllables will also be fairly accessible, although spacing or some other explicit syllable marker would make them even

more accessible. However, it remains unclear whether separation of syllables by spaces, even for polysyllabic words, would help or hinder beginning or advanced readers. Ziegler and Goswami's theory suggests that there may be a tradeoff if only words or only syllables are clearly represented (Ziegler & Goswami 2005). Words may be even more accessible to beginning readers than syllables (particularly syllables in monomorphemic, polysyllabic words where the syllables themselves have no meaning), but there are a larger number of words to be represented than syllables. Their theory also suggests that what may be ideal for beginning readers, who typically have phonological awareness of larger units but not smaller, may be different from the ideal for advanced readers who have mastered the alphabetic principle and have high phonemic awareness. Ideally, according to their theory, *all* phonological units would be clearly and consistently represented—words, syllables, onset/rimes, and phonemes—so that readers can access whatever phonological levels are most helpful to them given their stage of development and given the phonological structure of their language.

One major factor missing from Ziegler and Goswami's original theory is morphology, as they acknowledged (Goswami & Ziegler 2006). As will be described more fully below, Bassetti and Masterson (2012) performed a study on Chinese adults and children reading both alphabetic pinyin and morphosyllabic hanzi. They showed that the morphological information in hanzi was responsible for the significantly increased reading rates for hanzi in both children and adults. This, as well as other studies mentioned in Goswami & Ziegler 2006, suggests that morphology can intersect with phonology in important ways, and that clear and consistent representation of morphemes should also be a goal in orthography development.

2.5 Tests of interword spacing

The use of interword spaces, as we have seen, is neither universal today nor in the past. Although we have looked at tests involving the separation or joining of the constituents of compound words, we will now describe research done on the use of interword spaces in general, mainly in comparison to the *scriptura continua* style currently in use by Chinese, Japanese, Thai, and some other Asian writing systems.

2.5.1 Removal of normally occurring interword spaces

A great deal of research has established that removing interword spaces in English disrupts reading. Significant reductions in reading speed have been found for text

without interword spaces compared to the same texts with spaces, as normally written (Fisher 1975, Spragins et al. 1976, Epelboim et al. 1994, Epelboim et al. 1997, Rayner et al. 1998, Rayner et al. 2013, McGowan et al. 2014). This effect varies from reader to reader, but typically causes a 40% to 70% reduction in reading speed (Rayner et al. 2013:354).

Disruption to reading has also been found when spaces were filled with various letters, digits, or other symbols (Fisher 1975, Spragins et al. 1976, Malt & Seamon 1978, Pollatsek & Rayner 1982, Morris et al. 1990, Epelboim et al. 1997, Sheridan et al. 2013, McGowan et al. 2014), with the most disruption occurring with the most letter-like symbols. Removing spaces disrupts readers of all ages, including elementary students (Spragins et al. 1976), younger adults, and older adults (Rayner et al. 2013, McGowan et al. 2014), although it seems to disrupt older adults more than younger adults.

Clearly, part of this disruption is due to the fact that unspaced text is an unfamiliar format for English readers. Interestingly, Malt and Seamon (1978) studied the effect of readers practicing the same style of disruption (in their case, filling spaces with either black or red letter-like symbols) for ten days. They found that those practicing with black symbols improved their reading over the period, but still read more slowly than normal. Those reading with the red filler symbol, however, started out faster than the black symbol group but showed no improvement, so that their final rate was roughly equal, and still below the rate for normally spaced text. This implies that there may be a ceiling for unspaced or disrupted-space text that would be difficult to surpass even with longer practice.

Eye movements are also significantly disrupted when spaces are removed or filled. In particular, saccade landing positions are closer to the beginning of the word than with normally spaced text (Rayner et al. 1998). Since English readers have been shown to utilize word spacing information to plan saccades, they are no longer able to plan saccades to land at or near the center of words as before. Removal of interword spaces also has a much more disruptive effect for low frequency words in English than for high frequency words (Rayner et al. 1998).

Rayner et al. 1998 found that unspaced text was more disruptive for isolated sentences than for passages (with a 50% and 40% reduction in reading speed, respectively). It seems that the increased context of passages compared to isolated sentences helps English readers cope with the removal of spaces.

Is the disruption caused by space removal due to the actual physical gap between words, or is it simply a result of clearly marking word boundaries? To explore this question, Perea & Acha 2009 presented Spanish text to native speakers in three different formats: with normal interword spaces, with spaces removed, or with alternating**bold**format**to**mark**word**boundaries. The alternating bold format was significantly faster than the unspaced format, but slower than the normal spaced format by 31% (compared to a 44% reduction in reading speed for the unspaced format). Since some of the reduction in speed for both the alternating bold and the unspaced format can be attributed to unfamiliarity, it is unclear exactly how much disruption for unspaced text to attribute to either removing the physical gap or unclear word boundaries. But we certainly can say that at least some of the problem with unspaced text for Spanish readers is the lack of word boundary clarity, which the alternating bold format helps to alleviate. This study also helps us generalize the results from English to another language using the Latin script with a standard of interword spacing.

2.5.2 Addition of normally absent interword spaces

Until recently, little work had been done to test the effect of interword spacing on any language other than English. Thankfully, the last two decades has seen a great deal of new research on languages with a tradition of not using interword spaces, with Chinese being the language most studied.

2.5.2.1 Chinese hanzi

Many recent studies have tested the effect of spacing on reading with Chinese characters, or hanzi. When comparing the traditional unspaced format to a word-spaced format in normal reading conditions, most studies have found no significant difference in reading speed in either adults (Inhoff et al. 1997, Hsu & Huang 2000b, Bai et al. 2008, Bassetti 2009, Liu & Li 2014) or children (Shen et al. 2010, Zang et al. 2013). A few studies have found benefits to interword spacing for highly ambiguous sentences, in unusual display formats, or for learning new vocabulary (Hsu & Huang 2000a, Hsu & Huang 2000b, Shieh et al. 2005, Blythe et al. 2012, Bai et al. 2013), while one found an overall disruptive effect for both adults and children reading hanzi with interword spaces on a sentence-picture matching task (Bassetti & Masterson 2012). Both studies by Hsu and Huang (2000a, 2000b) showed a benefit for half-character width spaces over full-character width spaces,

suggesting perhaps that those studies using only full-character width interword spaces might have seen a slight benefit from half-width spaces.

Overall, it seems that Chinese hanzi text with interword spacing is typically read at the same rate as normal unspaced text. However, two studies have found that both unspaced text and text with interword spaces is read faster than text with spaces between every character or spaces within polysyllabic words, both with adults (Bai et al. 2008) and children (Shen et al. 2010). The same pattern in both studies was also observed when spaces were replaced with highlighting, using an alternating graybackgroundbehindeveryotherword. The fact that interword spacing does not slow down readers, whereas spacing in other ways does, suggests that interword spacing *does* have some facilitative effect, but it is counterbalanced by other negative effects, most likely the unfamiliarity of the format. The fact that this effect is also seen in the highlighting condition suggests that it is not simply a matter of more spatially distributed text moving characters further into the parafoveal and peripheral areas, although the highlighting may have other negative effects on processing as well.

As Bai et al. 2008 notes, the fact that adding interword spaces to Chinese text does not slow readers down is quite surprising, given the lifetime of experience readers have had reading without interword spaces. This contrasts strongly with the results from English, where readers are significantly disrupted by the removal of interword spaces. Epelboim et al. 1994 argues that much of the reduction in reading speed for English readers reading unspaced text is a product of the unfamiliarity of the format, and readers could reduce the disruption with more practice. This is undoubtedly true to some extent, as Malt & Seamon 1978 hints at. But the same idea can be applied to Chinese readers, which suggests that if Chinese readers had more practice with interword spacing, they might eventually become faster at reading word-spaced text than unspaced text.

2.5.2.2 Japanese

The Japanese writing system uses a combination of Chinese characters, or *kanji*, and syllabic characters, called *hiragana*, in an unspaced format. In Sainio et al. 2007, an eye movement study on Japanese, interword spaces were added to both mixed kanji-hiragana text as well as to pure hiragana text, which is an uncommon format. The study found that adding interword spaces to the normal mixed text neither sped up or slowed down readers, whereas interword spaces in the pure hiragana text did

show a facilitative effect. Another much smaller eye movement study found the same results (Matsuda 2001). Sainio and colleagues attribute the difference between the spacing effect for mixed kanji-hiragana text and pure hiragana text to the fact that in the normal mixed text, words often start with a kanji character and end with hiragana, providing visually salient word boundary information that makes interword spaces fairly redundant. In the hiragana text, however, no such word boundary information is present. The hiragana text is also less familiar to readers, so interword spaces are presumably less disruptive, and the benefit in word recognition can be seen more clearly.

Although the traditional mixed kanji-hiragana text already provides some word boundary information, the fact that adding interword spaces did not disrupt readers, just as with Chinese hanzi, is somewhat surprising given their unfamiliarity to Japanese readers. It may thus be argued, as with Chinese, that Japanese readers who were given the chance to practice reading with interword spaces might actually see an overall benefit over time, even with mixed kanji-hiragana text.

2.5.2.3 Thai

Like Chinese and Japanese, the Thai writing system does not use interword spacing. Unlike Chinese and Japanese, however, Thai uses an alphasyllabary, where consonant letters contain an inherent vowel, and where individual consonant and vowel letters are arranged by syllable. Although Thai does not use spaces between words, it does use spaces to mark phrase and clause boundaries, similar to how commas and periods are used in English. Since it shares some (but not all) of the features of alphabetic languages like English and Spanish that have shown a benefit for interword spaces, it is an intriguing writing system to study in terms of the effect of the insertion of interword spaces.

As mentioned above, eye movement studies on Thai have found that readers tend to show a preferred viewing location just to the left of center of the word, quite similar to readers of English, German, or other European languages with interword spaces (Reilly et al. 2005, Winskel et al. 2009). These studies, along with Kasisopa et al. 2013, show that Thai readers effectively utilize character frequency information and other aspects of the orthography to target word centers and mark off word boundaries.

One initial study on the effect of spacing in Thai was done by Kohsom and Gobet (1997), who found a marginally significant benefit for adding interword spaces.

However, their results may be skewed by the fact that, unlike in normal Thai writing, the unspaced paragraphs they presented had no spaces whatsoever. Interestingly, they also found that native Thai speakers who were bilingual readers of English did much worse when reading English with no interword spaces (a 55% increase in reading time) than normal English text. The Thai bilinguals were more affected by the removal of spaces in English than the native English readers in their study. In other words, the skills they developed to determine word boundaries in Thai did not help them with English, presumably because they were relying on highly orthography- and language-specific cues in Thai that did not transfer to English reading.

Winskel et al. 2009 used a sentence-based eye movement test to study the effects of interword spacing in Thai. Unlike Kohsom & Gobet 1997, the study found that spaced text was read 5% more slowly, with a marginally significant effect. It also replicated the effect shown in Kohsom & Gobet 1997, that Thai bilinguals who read English are highly disrupted when interword spaces are removed. Another study compared spaced, unspaced, and alternatingbold formats for Thai (Winskel et al. 2012). It found the alternating bold format to be significantly slower than spaced and unspaced text, but the spaced and unspaced text did not statistically differ.

Kasisopa 2011 and Kasisopa et al. 2013 both found no significant difference in reading time for adults with spaced and unspaced text. Kasisopa 2011 did, however, find that younger children benefited from interword spacing. This is not surprising, since Thai elementary students typically read text with interword spacing until second grade (Kasisopa et al. 2013).

Overall, we see a similar effect of interword spacing on Thai as with Chinese and Japanese. Interword spaces seem to neither slow down nor speed up reading in Thai overall, despite it being a relatively unfamiliar format. Thai readers may, however, be slightly more familiar with interword-spaced text than Chinese and Japanese because of the use of interword-spaced text in early elementary school. It is possible, but by no means proven, that Thai readers given time to practice reading text with interword spaces would end up with a net positive effect once the relative unfamiliarity is overcome.

2.5.3 Chinese pinyin: Word and syllable spacing compared

In addition to spacing research on Chinese hanzi, a few studies have looked at pinyin, the official Latinization system for Chinese, to compare the effects of using

spaces between words versus spaces between syllables. Bassetti 2009 had native Chinese readers read texts in pinyin with both spacing styles, and found that there was no difference in reading speed. This was somewhat surprising, since the official standard for pinyin is interword spacing, and pinyin is most typically seen in such a format. In addition, Bassetti 2009 expected that interword spacing may give an advantage to readers of pinyin due to its disambiguation of the many homophones that are possible when a single pinyin syllable is read. For instance, if the sentence *diàn shì zhèng zài bō sòng xīn wén* is written with syllable spacing, the syllables on their own have from 4 to 29 possible homophones the readers must choose from. Interword spacing, however, as in *diànshì zhèngzài bōsòng xīnwén*, reduces this to a single possible alternative for the first word, and no homophones for the other words (Bassetti 2009).

Similarly, Bassetti & Masterson 2012 found that word spacing and syllable spacing for pinyin text showed no difference for adult readers. However, children read syllable-spaced pinyin faster than word-spaced, with an average speed advantage of 16%. This may be due to the same effect we saw with Finnish elementary school children in Häikiö et al. 2011, where younger and less proficient readers benefited from hyphenation of compound words, but older students did not, presumably because of younger readers' shorter perceptual span.

Another study found that when Chinese adult readers were asked to read a text in pinyin and then rewrite the text in hanzi from memory, their hanzi production was more accurate when they read pinyin in syllable-spaced format than in word-spaced (King 1983). This clearly cannot be a result of a shorter perceptual span as with the children in Bassetti & Masterson 2012. It may, however, be a product of the readers' relative unfamiliarity with pinyin, as pinyin instruction was much less universal in 1983 than it is today.

2.5.4 A cultural side note

In the research on interword spacing, there is sometimes a subtle cultural battle occurring between those favoring interword spaces and those seeing them as irrelevant or unnecessary. One the one hand, writers such as Paul Saenger promote the idea of the spread of interword spacing from the British Isles, through Europe and beyond, enabling silent reading and all the cultural benefits that come with it

(1997).[1] It would not be difficult to read into this a form of "cultural imperialism", where Asian writing systems lacking interword spaces are seen as somehow "underdeveloped" or "primitive". I imagine most (hopefully all) researchers considering the benefits of interword spacing in Chinese would be appalled by the use of such terms. Nevertheless, such ideas exist (for a particularly egregious example, see Hannas 2011), and research on interword spacing in Chinese does not take place in a cultural vacuum.

On the other side, those resisting such ideas tend to view interword spaces more negatively, and point out a double standard in research agendas. For instance, Bassetti (2009) downplays Saenger's theory of an evolution toward increased interword spacing. Although her tests found quite similar results overall as Bai et al. 2008, she presents them in a much more negative light in terms of the utility of interword spacing for Chinese. In the mid-1990s, there was a fairly vigorous debate regarding the importance of interword spaces in reading, with undertones of the debate referring to the adequacy or lack thereof of unspaced writing systems for languages such as Japanese and Thai (Epelboim et al. 1994, Rayner & Pollatsek 1996, Epelboim et al. 1996, Epelboim et al. 1997, Rayner et al. 1998). Bassetti and Masterson (2012:19) decry the "English-centric" research on reading, and note that

> [T]here is research on the presumed facilitative effects of adding interword spacing to Chinese or Thai, but little or no research on the potential facilitative effects of adding morpheme boundary markers in English; and there is much more research on phonological than morphological processes in reading, although most writing systems represent both phonological and morphological information to some extent.

Bassetti and Masterson make an important point, one which is underlined by their main findings: Chinese readers, both adults and children, read hanzi texts roughly three times faster than the same texts written in pinyin. Bassetti 2009 also found the same result for adult readers. Since most of Chinese students' initial literacy instruction takes place through pinyin and not hanzi, this cannot be due simply to a lack of exposure to pinyin for students. Instead, Bassetti and Masterson argue that it is the lack of morphological information in pinyin that makes the system slower to read than hanzi. Sun 1993 also found a similar pattern, where children and adults read hanzi 2.5 to 4 times faster than pinyin. Clearly, then, it would not be beneficial

[1] I should in fairness note that Saenger himself gives a high view of the Chinese writing system, even stating that "many skilled adult Chinese readers are able to achieve a proficiency in rapid, silent reading perhaps unequaled in modern occidental languages" (1997:2).

for Chinese writers and readers to abandon hanzi and use the Latin script, as Mao Zedong had proposed (Zhou 2003:155), and as Western critics such as Hannas (2011) still argue.

Thus far, this review has considered the essentially null results from Chinese, Japanese, and Thai interword spacing tests to be positive evidence for the potential benefit to reading speed of interword spacing in these writing systems, if readers became more familiar with this format. The evidence from Bassetti 2009 and Bassetti & Masterson 2012 provide valuable context to this hypothesis. We should not assume that all helpful innovations go from West to East. Perhaps interword spaces might indeed be helpful to some East Asian languages. But Western or Latin-script writing systems might also benefit from Eastern orthographic features such as clear morpheme or syllable boundaries.

We also cannot consider paraorthographic devices in isolation from the orthography as a whole, or from the linguistic structure of the language being written. What may work well for one language, with its orthographic tradition and linguistic structure, may not work as well for others. The intriguing result for syllabic spacing of pinyin is an example of this. Few if any would expect that syl la ble spa cing in En glish would be be ne fi cial to rea ders, especially given English's extreme orthographic irregularity. Yet Bassetti 2009 and Bassetti & Masterson 2012 point to a possible benefit to Chinese readers, both children and adults, of syllable spacing for pinyin. This is quite plausible, given the high correlation of syllables to morphemes in Mandarin, the importance of the morphosyllabic hanzi in Chinese literacy, and the relatively short words of Mandarin.

One example of a possible benefit for English orthography that can be seen in an Asian writing system is pointed to by experiments using extra space between phrases and clauses, as Thai does. Several experiments have shown a benefit in reading time or comprehension when extra space is added at phrase boundaries, particularly for less proficient readers (Bever et al. 1991; Jandreau & Bever 1992, Magloire 2002). It seems that, when done with care not to create too many counterbalancing disruptive factors (Keenan 1984), and with attentiveness to the particular structure of the language and orthography, adding more linguistic information to the text can be helpful to readers, whether on the syllable, morpheme, word, or phrase level.

2.6 Text segmentation and learning to read

We have already seen that children's perceptual span in reading is smaller than adults (Rayner 1986, Häikiö et al. 2009), and that children have shorter and more frequent saccades and more frequent refixations of words (Rayner 1998). Since children are still learning the basic elements and structures of their orthography, they need more time to process text. Given the limitations on human short-term memory, this means that young or less proficient readers need to process text in shorter chunks; hence the shorter saccades.

The results of tests on spacing and segmentation cues that we have seen so far fits well with this understanding. Häikiö et al. 2011 found that less proficient 2nd graders read hyphenated compounds more easily than unspaced compounds, whereas older children and more proficient 2nd graders read unspaced compounds more easily. Bassetti & Masterson 2012 found that adults read pinyin with syllable spacing and word spacing equally well, whereas children (7-10 years old) read the syllable-spaced pinyin more easily. Kasisopa 2011 found that 1st and 2nd grade Thai readers benefited from spaced text, whereas 5th and 6th grade students and adults read both equally well (although this is partly due to Thai instructional materials through 1st grade being written in a word-spaced format).

In addition to needing time to master the basic processes of reading, young readers of orthographies with interword spaces also take time to develop an awareness of words in their language. Despite the "obviousness" of the category for adults reading word-spaced languages, children do not develop awareness of the orthographic word until fairly late in literacy development. Morris 1993 showed that children only develop an understanding of the printed word marked by spaces after significant literacy instruction, typically at around the second grade (Ferreiro 1999, Chaney 1989). When writing, younger children will often leave out spaces between words entirely, or use spaces only at major phrase boundaries (Flanigan 2007, Ferreiro 1999). When asked to count spoken words, however, children often count syllables rather than words (Ferreiro 1999). So, apart from literacy instruction, it seems that the main units children are aware of are sentences and phrases on the one hand, and syllables on the other. This matches the experience of Chinese adult readers when asked to separate Chinese text by words, who often consider a word to be an entire phrase (Hoosain 1992). For English-speaking children as well as for Chinese adults, content words are more easily recognized as units than functor words (Chaney 1989,

Hoosain 1992). English-speaking children often continue to join functors to the main content words for some time (Chaney 1989).[2]

The fact that it takes time to develop word awareness, just as it takes time to develop phoneme awareness and syllable awareness, raises a crucial point for languages deciding on whether and where to use word breaks. Whatever benefit interword spaces may bring to reading speed and comprehension must be balanced by the pedagogical load of learning the "right" way to space text. Limited writers may even choose not to write at all if they are intimidated by spacing decisions and afraid of making mistakes, although this can be mitigated if a community accepts spacing variation as normal and valid (Karan 2014). Since word awareness is not fully developed for preliterate children and adults, and since there are borderline cases that present challenges even to trained linguists (Duanmu 1998, Kutsch Lojenga 2014), every writer must learn the spacing conventions of their own language. This takes time and effort (see for example Morlaeku & Wang 2010:16).

Of course, syllable awareness also takes some time for children to develop, especially for languages where syllables are not as phonologically dominant, such as English. The general consensus of literacy literature, however, seems to be that syllable awareness typically occurs before word awareness (Ferreiro 1999, Chaney 1989). Particularly if a language has clear syllable boundaries, the additional educational burden of learning where word breaks occur should be considered as a factor in orthography decisions.

Finally, there is evidence that languages differ in how easy or hard it is to segment speech into words. For example, English has been found to be much easier to segment into words than Korean (Daland & Zuraw 2013) or Japanese (Fourtassi et al. 2013) because of differing phonotactic restraints. To the extent that orthographies are phonemically regular, these differences may translate into differences in the difficulty of segmenting written text into words as well. In other words, some languages may be naturally easier for writers to divide up into words than others, which may have implications for how long it takes writers to learn word awareness and how beneficial interword spaces would be in these languages.

[2] Both the tendency to consider functor words as not true words, and the tendency to break words into syllables or morphemes, is seen in a word game my preschool-age daughter likes to play. She replaces the first consonant of each word with the same consonant, say, /b/. When thinking of the sentence, "Baby Silas is the silliest in the world," she said: "Baby Bilas is the billiest in the borld," leaving the functor words unchanged. The word "understand" became "bunderband," and the sentence "It's not a museum" became "Bit's bot a bubeum." For English, this might be best understood as relating to phonological stress, since functor words tend to be unstressed, and unstressed syllables like "der" in "understand" or "las" in "Silas" remained unchanged as well.

2.7 Morphology and word formation in Mainland Southeast Asia

There is significant variation within languages and language families in Mainland Southeast Asia in the morphological and phonological structure of words, but some generalizations can nonetheless be made. Enfield 2005 describes the region's languages as "the closest we have" to the "isolating and analytic morphological type" described by Sapir 1921. Enfield continues:

> These are languages in which the number of morphemes per word approaches one, morphemes are modified neither by affixes nor internal changes, and the basic unit for the productive construction of meaningful complexes is the phrase, not the word. (2005:187–188)

Although no real language reaches this extreme, many languages of the region are marked for their isolating and analytic morphology, particularly in relation to the range of languages seen around the world. Nonetheless, it would be an exaggeration to say that MSEA languages lack morphology of any kind. In particular, we see certain types of morphological word formation processes quite frequently.

2.7.1 Compounding

A simple definition of a compound word (or a compound lexeme) is a lexeme consisting of "two or more simple(r) lexemes" (Matthews 1991:37). Compound words are extremely common in many languages of MSEA. Because most languages in MSEA lack any inflectional morphology, compound words are sometimes difficult to distinguish from phrases. We will consider four different types of compound words:

1. phonologically unified compounds
2. semantically opaque compounds
3. coordinate compounds (also referred to as "semantically reduplicated" compounds)
4. elaborate expressions

2.7.1.1 Phonologically unified compounds

In MSEA, compounds can sometimes be distinguished from phrases by some type of phonological joining of the compound's constituents that does not universally occur across syllable boundaries. For example, the Hmong Daw word *nees nkaum* shows

the spreading of nasalization from the morpheme *nees*, "two", to the second morpheme *kaum* "ten" (Ratliff 2009). Hmong Daw also shows a large number of compounds marked by tone sandhi, where the tone of the second element is changed as a result of interaction with the first element, in a way that does not occur between morphologically unrelated syllables. Examples are:

- *dab taws* "ankle," from *dab* "neck" and *taw* "foot"
- *taub dag* "pumpkin," from *taub* "gourd" and *das* "yellow"
- *plaub hau* "hair on the head," from *plaub* "hair" and *hauv* "head" (Ratliff 2009)

An example of phonological unity from Thai is the word น้ำมัน "oil", from the constituents /náam/ "water" and /man/ "fat". When normally pronounced, however, the first constituent is shortened and truncated, resulting in the form /ná.man/. This process is not seen in a phrase such as น้ำขวด /náam.khùat/ "bottled water" (Slayden 2015).

2.7.1.2 Semantically opaque compounds

Another way of distinguishing compound words from phrases is the semantic opacity of the whole meaning relative to the meaning of the constituent parts. For instance, the Thai word for butterfly is ผีเสื้อ /pʰǐi.sûa/, from /pʰǐi/ "ghost" and /sûa/ "shirt". In Hmong Daw, a rainbow is a *zaj sawv*, literally "dragon-rise", while the word *dab tuag*, literally "ghost-dead", can mean "ugly" or "sloppy" (Ratliff 2009). Although each part of a semantically opaque compound is a recognizable word, the meaning of the whole cannot be easily be ascertained as a sum of its parts. As mentioned above, there is a continuum from semantic opacity to semantic transparency (Mok 2009), so sometimes it is difficult to distinguish between a slightly opaque compound and a phrase, especially if there are no phonological, inflectional, or other markings to assist in the determination. We should also be quick to note that semantic opacity is relative to the speakers of the language itself, not to outsiders with a different semantic and cultural frame. What may seem like an opaque compound to an outsider could be a transparent phrase to insiders.

2.7.1.3 Coordinate compounds

Many languages in MSEA exhibit a form of compounding where two synonyms, antonyms, or otherwise strongly related words are paired together to create a

compound, sometimes generalizing the meaning. For instance, in Hmong Njua, a variety of Hmong closely related to Hmong Daw, we have the following compounds:

- *caij nyoog* "time," literally "time-time"
- *muag nug* "siblings," literally "sister-brother"
- *qab npua* "small livestock", literally "chicken-pig"
- *nub mo* "all the time," literally "day-night" (Mortensen 2003)

Since they are often highly semantically transparent, it is sometimes more difficult to separate these types of constructions from phrases. However, Mortensen 2003 notes that these items are strongly lexicalized, so that the particular choice of constituents cannot be substituted for by a different synonym or pair of synonyms. The order of the constituents cannot be reversed, and there can only be two elements. There are also restrictions on phonological symmetry, such that the two elements must have the same number of syllables. These restrictions, and similar restrictions in other languages of the region, help to form a discrete class of compounds. Vietnamese has a similar class of coordinate compounds, the formation of which David Thomas (1962) refers to as "semantic reduplication."

2.7.1.4 Elaborate expressions

Many linguists of MSEA languages have also described a word class known as "elaborate expressions" (Haas 1964, Matisoff 1973, Hanna 2013). These are typically four-syllable, four-constituent constructions that share much in common with the class of coordinate compounds mentioned above. Indeed, Mortensen 2003 argues that elaborate expressions are simply a subset of the class of coordinated compounds, in that they 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. They also often consist of one or even two syllables that have no meaning on their own, and are only found in these set expressions. In such cases, they either share phonological properties with a corresponding rhyming element, forming a "euphonious" element (e.g. for Lao, Enfield 2007:306), or they are ancient synonyms with another element that have since lost their independent usage. Examples from Hmong Daw are:

- *khwv iab khwv daw* "arduous toil", literally "toil-bitter-toil-salty" (Jarkey 2010)

- *pog koob yawg koob* "ancestors," literally "grandmother-great-grandfather-great" (Ratliff 2009)
- *ua qoob ua loo* "agriculture," literally "do-crop-do-crop?", where *loo* has no current meaning in Hmong Daw except as part of this phrase or the compound *qoob loo* "crops", but is still used to refer to crops in related languages (Heimbach 1979, Mortensen 2003)

Mortensen 2003, Matisoff 1973, Hanna 2013, and Ratliff 2009 all describe these expressions as compounds formed by morphological processes, and hence they would typically be included in the linguistic definition of a "word". Certainly, the restrictions placed in many languages in the region on how such expressions can legally be formed indicates a degree of lexicalization, and the frequency of rhyming, alliteration, and tone patterns also argues for this view. It may be best, then, to treat many or even most of these expressions as long compound words, although their length may create problems when dealing with questions of optimal spacing for orthographies.

2.7.2 Reduplication

Reduplication is a very common word formation process in MSEA (Goddard 2005, Nguyen & Ingram 2006, Enfield 2005). Reduplication is often thought of as a phonological process, whereby a suffix is created that duplicates the stem (Cahill 2008). This would then place reduplication within the category of affixation. However, Inkelas & Zoll 2005 shows from examples in various languages that reduplication can be both a phonological process and a syntactic process. Therefore, rather than treating reduplication either as a subset of affixation, or as a subset of compounding, I will consider the process separately from both.

Reduplication can be either full or partial. In full reduplication, the entire word is copied and remains unmodified by any phonological or syntactic processes. Examples of this are the Thai word เด็กเด็ก /dèk.dèk/ "children, from /dèk/ "child"; and Hmong Daw *ntau ntau*, "very much", from *ntau* "much." Both of these exhibit a common pattern in reduplication, where noun reduplication results in pluralization, while adjective reduplication causes intensification (Goddard 2005).

Partial reduplication in Mainland Southeast Asia mainly involves duplication of a word with phonological modification. It can also refer to the reduplication of only part of the stem, as in the Malay *le-luhur* "ancestor", from *luhur* "noble" (Alikamal 2012), but this is less common in most MSEA languages. More often, the word is

changed in its tone, vowel, or other phonological properties. Examples of this type of reduplication are the Hmong Daw *ntxoov ntxoo* "shade," from *ntxoov* "shade" (Ratliff 2009), Thai ดีๆ ดี /dii.dii/ "very good" from /dii/ "good" (Iwasaki & Horie 2005:36); Vietnamese *vội vàng* "in a great hurry", from *vội* "in a hurry" (Thomas 1962); or *bùi ngùi* "very moved (emotionally)" from *ngùi* "moved (emotionally)" (Thompson 1987:158).

2.7.3 Two-syllable ideophones

Martha Ratliff, in several articles and books, describes a category of two-syllable words involving the same kind of phonological similarities between syllables as with partial reduplication in Hmong Daw (Ratliff 1986, Ratliff 2009, Ratliff 2010, Ratliff 2013). These are not clear cases of partial reduplication, however, because the individual syllables have no meaning on their own, but are only found together. A few have more normal "definitions," while others are onomatopoeic "ideophones" used for the sound of different objects or actions (Ratliff 2010). For example, we have *khauj khaum* "shell, crab" (Ratliff 2009), where only the tone changes; or *cij coj*, the sound of chicks chirping (Ratliff 2010), where the vowel changes. These could be considered monomorphemic words, but clearly there is a phonological relationship between the syllables akin to that seen with partial reduplication, and, as Ratliff 2010 demonstrates, there is often an iconicity of tone and vowel that go along with certain classes of sounds. Enfield's examples in Lao, such as *thii1 lii1* "running madly", or *cuun1 phuun1* "heaped up in a pile", demonstrate that this phenomenon is found in other MSEA languages as well (2007:299; see also Williams 2013 for more examples in the region).

2.7.4 Affixed words

As noted above, inflectional affixation is rare in MSEA (Enfield 2005), but there are exceptions. Akha has a rich set of affixes for an MSEA language, including both inflectional and derivational affixation. There are prefixes that negate verbs, form adjectives, or mark a verb phrase as reiterative, prohibitive, or imperative. There are suffixes that form adverbs or nouns, make adjectives comparative or absolutive, or mark verbs as ablative, passive, accusative, or present or past tense (Kya Heh 2002).

Very few languages in the region have this type of inflectional affixation. There are, however, some examples of derivational affixes of various kinds. There are nominalizers, such as Hmong Daw *-tswv* (Ratliff 2009), Thai ความ- /kʰwaam/ or การ-

/kaan/, or the Khmer infix *-n-* (e.g. *cuəl* "to rent", *cnuəl* "rent") (Enfield 2005). Khmer also has other infixes that increase the valence of a verb, derive adjectives from verbs, or derive reciprocals (Enfield 2005).

Hmong Daw also has a set of noun class prefixes, such as *qhov-* "things with holes", or *pob-* "ball-like things" (Ratliff 2009). Thai has noun class prefixes such as ผล- /má/ "fruit" or แมล- /mɛɛŋ/ "many-legged, bug-like" (Iwasaki & Horie 2005:27–28). In the past, these noun class prefixes were probably independent words that were productively used in compounds, but which then later lost their independent distribution. Overall, affixation plays a much less central role in word formation for languages in MSEA than in many languages of the world, but it certainly does exist in the region.

2.7.5 Fossilized morphemes

The noun class prefixes noted above are examples of what probably used to be independent words productive in compounding, but now are bound morphemes forming a class of nouns. Many other morphemes in MSEA languages, however, have lost their independent distribution without forming a clear class of words, perhaps only leaving one instance of its existence within a former compound. In such cases, native speakers usually have no awareness of the morpheme's meaning apart from the meaning of the word as a whole. Examples in English would be words like "raspberry", "mulberry", or "twilight". Terms like "mul-" and "twi-" have no meaning on their own to English speakers, although "-berry" and "-light" are transparently related to the full words.

Hmong Daw has a fairly large number of such fossilized morphemes (underlined below), according to Ratliff 2009:

- *tsiaj txhu* "animal" (animal-?)
- *cua nab* "worm" (?-snake)
- *hnub qub* "star" (sun-?)

Some fossilized morphemes in Hmong Daw are two syllables, possibly themselves the prior product of compounding or some other word formation process:

- *kab laug-sab* "spider" (bug + ?-?)
- *yoov tshaj-cum* "mosquito" (fly + ?-?)

Examples of fossilized morphemes in Thai tend to have what looks like a meaningless suffix attached to a free morpheme, creating a synonym with the free morpheme (Slayden 2015):

- เสบียงกรัง /sa.biaŋ-kraŋ/ "provisions" (provisions-?)
- ผ่อนปรน /pʰɔ̀:n-pron/ "lenient" (lenient-?)

Clearly, fossilization is not so much a word *formation* process as a morpheme decay process. We include it, however, because of its importance in synchronic morphological analysis for languages in the region, and its relevance to the question of optimal spacing.

2.7.6 Monomorphemic words

Finally, although Enfield notes that languages in MSEA "tend toward monosyllabicity" (2005:186), there are certainly examples of morphemes longer than one syllable. Often, these are loanwords from other languages, where any internal morphology of a polysyllabic word would become opaque in the new language. Such examples from Hmong Daw are (Ratliff 2009):

- *nab kuab* "ice" (from Lao /naam-kɔɔn/ "water-block")
- *taj laj* "market" (from Lao /ta.laat/, originally from Khmer)
- *muas lwj* "deer" (from Chinese /mǎ-lù/ "horse-deer")

Other monomorphemic Hmong Daw words have no evidence of being loanwords (Ratliff 2009):

- *liv nyug* "vulture"
- *leeb nkaub* "parrot"
- *viv ncaus* "sister"

This last entry, *viv ncaus*, is interesting in that it used to be a compound word meaning "elder.sister-younger.sister," but neither term is used elsewhere anymore (Ratliff 2009). In other words, it is a doubly fossilized compound, and is therefore now a single morpheme. This shows that internal morphology can be lost not only by borrowing, but also within a language over time.

Many languages in the region have a significant number of polysyllabic, monomorphemic words of Sanskrit or Pali origin. A few examples from Thai are:

- ภาษา /pʰaa.sǎa/ "language"

- ศาสนา /sàat.sa.năa/ "religion"
- ปัญญา /pan.jaa/ "wisdom"

2.7.7 Sesquisyllabicity in MSEA

One other exception to Enfield 2005's description of MSEA languages approaching the isolating, analytical type is the existence of minor syllables, also called "sesquisyllabic" structures. In MSEA languages, these occur before a major syllable, are phonologically bound to the major syllable, and are in some way phonologically reduced relative to the major syllable. They often carry no tone or restricted tone, restricted vowels (often only /ə/), and no or few final consonants (Thomas 1992, Herr 2011). This type of structure is seen especially in Mon-Khmer languages (Enfield 2005), but can also be seen among other language families in the region, such as Thai (Bennett 1995, Thomas 1992), Lemi Chin (Herr 2011), and Cham (Thomas 1992). Sesquisyllabic words can be formed in a variety of ways. They can have a prefix as the minor syllable, they can be formed by the addition of an epenthetic vowel into a consonant cluster in a loanword, or they can be a compound word with the first element reduced and fossilized (Tebow & Lew 2013).

Sesquisyllables present a certain challenge to questions of spacing for the region. When spacing by word, the minor syllable is always connected to the major. However, if a group wants to space their text by syllable, they must then decide if minor syllables "count" as true syllables or not. Since they are phonologically bound to the major syllables, they cannot be properly pronounced on their own. This makes intersyllabic spaces somewhat more problematic for sesquisyllabic words than for disyllabic or other polysyllabic words. None of the languages tested for this thesis have minor syllables, so these issues are left for others to grapple with.

2.8 Spacing practices in Mainland Southeast Asia

As noted in Chapter 1, one reason why the question of spacing comes up often in Mainland Southeast Asia is because the region has such a diverse set of spacing practices across different orthographies.

2.8.1 No spacing, clause or phrase spacing

On the one end, we have the Chinese hanzi system, which does not use blank spaces at all. It does, however, clearly demarcate syllable boundaries with small intercharacter spaces, and punctuation is used for sentence and clause breaks. No

language outside of China uses the hanzi writing system alone, nor do most non-Sinitic languages in China, but the system is nonetheless highly influential in the region.

Close to this end are the Brahmi-based scripts of the region: Burmese, Thai, Lao, Khmer, plus other regional or minority language scripts such as Tai Tham, Cham, Tai Viet, and others. The standard orthographies for these scripts do not use interword spaces, but they do typically use spaces to separate sentences or phrases (Daniels & Bright 1996). Although they are not true syllabaries, many orthographies based on these scripts provide unambiguous syllable boundaries. Thai is somewhat of an exception, although even with Thai, most syllable boundaries can be unambiguously determined from knowledge of the language's phonotactics and orthography. A few minority language orthographies using Brahmi scripts follow the standard Brahmi convention of clause or phrase spacing, such as three related Thai script orthographies for Northern Pwo Karen (Cooke et al. 1976:217, Phillips 2009), and the Burmese script orthographies for Sgaw Karen (Gilmore 1898) and Geba (O.J. Gamache, personal communication, July 20, 2011).

2.8.2 Syllable spacing

On the other end of spacing styles, we have alphabets based on the Latin script that use spaces between every syllable. Examples are Vietnamese, the Akha Baptist orthography (Kya Heh & Tehan 2000), the Lahu Na Protestant and Catholic orthographies (Matisoff 2006), and the Lahu Si orthography (Cooper 2002). The Lisu orthography developed by James Fraser, which uses a derivative of the Latin script, also puts spaces between every syllable (Morse & Tehan 2000). Muak S-aak uses spaces between every syllable, including phonologically reduced presyllables, while another language in the region uses spaces between full syllables, but attaches presyllables to the following main syllable (Page 2013:472; Ellie Hall, personal communication, August 7, 2014). The Latin-based Unified Mien orthography for Iu Mien primarily uses syllable spacing, although it uses hyphens when tone sandhi links two syllables within a word phonologically (Arisawa 2013).

Although these systems are quite different from the Chinese hanzi system, it is clear that hanzi's marking of syllable (usually also morpheme) boundaries, and no marking of word boundaries, has influenced these Latin script orthographies. Indeed, despite the fact that Chinese pinyin is officially written with interword spacing, the standard practice for minority languages in China using the Latin script is to space

by syllable (McLaughlin 2012:88). In addition to hanzi's influence, the alphasyllabaries of the region, with their lack of interword spaces, make the syllable more prominent and the word less prominent than in Western writing systems.

Thai script orthographies typically do not space by syllable, but the Thai orthography for Northern Pwo Karen, mainly used by Christians, does use syllable spacing (Phillips 2009), including separating minor syllables. Interestingly, syllable spacing was chosen over word spacing for this Northern Pwo Karen orthography because word boundaries were more difficult to determine than syllable boundaries (Audra Phillips, personal communication, March 8, 2015). The Pahawh writing system for Hmong also uses spaces between every syllable (Ager 2014).

The Latin script RPA orthography for Hmong Daw does not have a standard when it comes to syllable spacing or word spacing; spacing depends on the writer and the word. Chapter 4 discusses Hmong Daw RPA spacing practices in greater detail.

2.8.3 Interword spacing

There are also orthographies in the region, using both the Latin script and Brahmi-derived scripts, that use interword spacing of one type or another. Clearly the most widely used of these is pinyin, the official Latinization system for Mandarin Chinese. Pinyin is officially spaced by words, although the actual spacing that Chinese writers use can vary due to inconsistent determination of word breaks in Chinese (Bassetti 2009). Although Chinese has four-syllable words that are processed as individual units when reading (Li et al. 2009), the official rules for segmenting pinyin call for breaking up words of four or more syllables into two-syllable parts whenever possible. Spacing pinyin by words allows the disambiguation of most homophones. However, according to Bassetti 2009 and Bassetti & Masterson 2012, disambiguating homophones is either not helpful to Chinese readers, or not helpful enough to overcome other negative factors of interword spacing relative to syllable spacing.

Chinese language policy, both official and unofficial, exerts strong pressure on minority languages to use pinyin as a basis for their writing systems (Zhou 2003). It is therefore surprising in a way that this pressure does not include a pressure toward the official interword spacing of pinyin—indeed, the opposite seems to be the case (McLaughlin 2012:88). This is presumably best explained by the high status of hanzi in China, with its clear syllable boundary marking.

Besides pinyin, there are many Latin script orthographies elsewhere in the region that use interword spacing. Pekon Kayan uses spaces corresponding to phonological words (Manson 2008). The Common Akha Orthography uses interword spaces (Morlaeku & Wang 2010), as does the Hlai orthography (Somsonge 2004) and the Falam Chin orthography (Bibles International 2008). Kachin (or Jingpho), Ngawn, Tedim, and Zotung Chin also use interword spacing (Bible Society of Myanmar 2006), as do Lemi Chin, Mro-Khimi Chin, Makuri Naga, and Geba Karen (O.J. Gamache, personal communication, September 10, 2014).

In addition to Latin script orthographies, there are several languages using the Thai or Lao script that use interword spaces. Many of these are Mon-Khmer languages with complex consonant clusters, so that syllable boundaries are less clear orthographically. These include Northern Khmer (Thomas 1990), Bru Khong Chiem (Green & Van der Haak 2002), Bru Tri (Miller & Miller 2005), So (Migliazza 2002, Markowski 2009), and Kmhmu' (Miller 2013). Some non-Mon-Khmer languages using the Thai script are also reported to use interword spaces, such as Urak Lawoi (Hogan 1976), Kensiw or Maniq (Bishop & Peterson 2002), and Bisu (Person 2008). Lew (2014:37) notes that although the Kmhmu' orthography is officially written with word spacing, in practice, writers often use spaces between the syllables of polysyllabic words as well. The same tendency of writers to break up certain polysyllabic words into syllable components is seen in word-spaced Hmong Daw text (see Section 4.4).

2.9 Sociolinguistics of spaces in MSEA

Kirk Person's description of the Bisu community's deliberation over spacing (2008) illustrates some of the competing pressures facing languages using one of the Brahmi scripts. Bisu is a Tibeto-Burman language found in Thailand and elsewhere, and the Bisu community in Thailand was developing a Thai script orthography for their language. According to Person (2008), since Bisu has more polysyllabic words than Thai, the lack of interword spacing was a greater hindrance to Bisu readers than it would be for readers of Thai. Person and another linguist, Liz Foerster, proposed the use of interword spaces to the Bisu orthography committee. The committee members found the text with interword spaces easier to read, but since interword spacing is only used in Thai for children's books through first grade, they rejected interword spacing, feeling it would "cause the language to appear childish to Thai observers" (Person 2008:7).

Some years after this decision, the Bisu from Thailand made contact with Bisu speakers in Myanmar, who call themselves Pyen. The Pyen were already using a Latin script orthography with interword spaces. Three Bisu linguistics interns worked on a literacy project for the Myanmar Bisu, and found that the interword spaces in the Latin script orthography helped them read more easily. They respectfully suggested that the Thai script orthography also adopt this convention (as well as other changes inspired by the Myanmar orthography), and this time the orthography committee accepted the proposal. As of 2008, Person (2008:12) wrote that "concerns about the social acceptability of word breaks...seem to have dissipated".

In this case study, we see that the sociolinguistic factors for Bisu in Thailand were pushing them toward clause spacing, as Thai uses, but readability factors were pushing toward word spacing. Meanwhile, for the Latin script orthography used in Myanmar, the sociolinguistic factors were different, and word spacing was adopted. Ultimately, the practical benefit associated with word spacing seems to have outweighed sociolinguistic objections. The example of the Latin script orthography with interword spacing clearly had some influence on the Thai-based orthography for Bisu.

Moving in the opposite direction, Morse & Tehan 2000 proposed a shift for Lisu from syllable spacing to word spacing, along with a general shift toward a more English-like orthography using standard Latin script letters and characteristics. According to Bradley 2006, some Lisu writers have implemented some of the letter changes from Morse & Tehan 2000 when writing emails, but few if any have adopted word spacing, and no traditional materials have been reprinted with the reforms implemented. There seems to be have been some felt need for greater convenience when writing emails, but no interest in radically changing the orthography for more established uses.

The choice of script has a significant effect on the spacing style options typically chosen in the region. Brahmi script orthographies sometimes follow the standard of clause or phrase spacing, although they also sometimes add more spaces at the word or syllable level. In contrast, no Latin script orthography in the region currently uses clause or phrase spacing. Syllable spacing is much more common with Latin script orthographies than with Brahmi script. We should not go so far as to say that the spacing practices of each script's dominant languages determine minority language spacing practices—there are far too many counterexamples to that notion. But the

dominant languages for each script tend to provide a starting point from which further variation occurs. We should also note that when variation from the dominant orthography occurs, it is invariably in the direction of adding more spacing. The reasons cited for adding more spaces nearly always include ease of reading, if any reasons are given (Hogan 1976, Person 2008, Bishop & Peterson 2002).

2.10 Syllable-based pedagogy

One other factor affecting the sociolinguistics of spacing in MSEA is the way that pedagogy in the region focuses primarily on the syllable, not the word or phoneme, as the dominant phonological unit. Page 2014 describes how common syllable-based approaches to literacy are in the region, and how replacing phoneme-based primers and literacy instruction with syllable-based instruction significantly improved the learning and the social acceptance of the primers and literacy instruction. The ubiquity of syllable-based chanting in the region, from temples, churches, and schools, further supports this claim. It could be that syllable-based methods are more successful in the region because of the Brahmi-based alphasyllabaries and the way they encode syllable information in the writing systems, as Page 2014 indicates. Even for Latin script writing systems in the region, though, the pedagogical and psychological dominance of the syllable may play a role in promoting syllables to a higher level of awareness and importance than would be seen in the West. If so, this may have relevance to the question of spacing and the optimal segmentation of text for languages in the region, particularly for beginning readers but perhaps for advanced readers as well.