Original Article

QJEP

Quarterly Journal of Experimental Psychology
1–17
© Experimental Psychology Society 2017
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1747021817738965
qjep@sagepub.com
SAGE

The role of lexical variables in the visual recognition of two-character Chinese compound words: A megastudy analysis

Chi-Shing Tse[1,2] and Melvin J Yap[3]

Abstract

To examine the effect of lexical variables on two-character Chinese compound word processing, we performed item-level hierarchical regression analyses on lexical decision megastudy data of 18,983 two-character Chinese compound words. The first analysis determined the unique item-level variance explained by orthographic (frequency and stroke count), phonological (consistency, homophonic density), and semantic (transparency) variables. Both character and word variables were considered. Results showed that orthographic and semantic variables, respectively, accounted for more collective variance than phonological variables, suggesting that Chinese skilled readers rely more on orthographic and semantic information than phonological information when processing visually presented words. The second analysis tested interactive effects of lexical variables and showed significant semantic transparency × cumulative character frequency and word frequency × cumulative character frequency interactions. The effect of cumulative character frequency was stronger for transparent words than for opaque words and was stronger for low-frequency words than for high-frequency words. However, there was no semantic transparency × word frequency interaction in reaction time. Implications of the current findings on models of Chinese compound word processing are discussed.

Keywords
Chinese; compound word; megastudy; visual word recognition

Received: 19 January 2017; revised: 31 August 2017; accepted: 6 September 2017

In English, a compound word is often formed by a combination of two constituents that are words on their own (e.g., *snow* and *man*: *snowman*). Similarly, in Chinese, the basic unit of a written form (character) is very often a monosyllabic morpheme that can be a word on its own, and a word is often formed by combining two characters (e.g., 花 "flower" and 園 "park"; 花園 "garden"). Such two-character compound words represent 73.6% of Chinese words (Institute of Language Teaching and Research, 1986; see also Zhou & Marslen-Wilson, 1995). Studies have investigated the effects of various lexical variables on Chinese compound word processing (e.g., Peng, Liu, & Wang, 1999; Taft, Liu, & Zhu, 1999; Zhou & Marslen-Wilson, 2000a). However, these studies have relied on factorial design, where categorically defined experimental variables (e.g., word frequency) are crossed, and extraneous variables are controlled for. The joint effects of these variables on participants' mean reaction time (RT) and accuracy are then assessed. As reviewed below, these factorial-design studies have not always yielded consistent results. Using a megastudy approach (i.e., examining normative lexical decision data using item-level regression analyses), this study addresses two central issues in Chinese compound word processing: (a) how the access of compound words is influenced by character- and/or word-variables and (b) whether character phonology affects their processing. Before reporting our analyses and elaborating on their theoretical significance, we first review prior studies most pertinent to these two issues and explain why the megastudy approach is likely to shed light on them.

---

[1] Department of Educational Psychology, The Chinese University of Hong Kong, Shatin, Hong Kong
[2] Centre for Learning Sciences and Technologies, The Chinese University of Hong Kong, Shatin, Hong Kong
[3] Department of Psychology, National University of Singapore, Singapore

Corresponding author:
Chi-Shing Tse, Department of Educational Psychology, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR.
Email: cstse@cuhk.edu.hk

How is the access of Chinese compound word influenced by character-variables and/or word-variables?

The access of Chinese compound words can be influenced by character-variables (e.g., character frequency), word-variables (e.g., word frequency), or some combination thereof. These possibilities can be investigated by examining the main effects of the character- and word-variables and their interactions in lexical decision performance. In this study, we targeted three character-variable × word-variable interactions: character frequency × word frequency, character frequency × semantic transparency, and word frequency × semantic transparency.

Character frequency × word frequency interaction

Word frequency effects are robust in Chinese compound word processing (e.g., Zhang & Peng, 1992). Participants respond more quickly to high-frequency words, for example, 朋友 "friend," than to low-frequency words, for example, 私仇 "personal enmity," showing that the speed of accessing a compound word depends on how frequent a word appears as a whole. In addition to whole-word frequency, each character within a word varies in frequency.[1] For example, the 山 "mountain" in 山丘 "hill" is a high-frequency character, whereas 丘 "mud" is a low-frequency character. The 仇 "hatred" in 仇人 "enemy" is a low-frequency character, but 人 "person" is a high-frequency character. Some studies have reported that character frequency also plays a role in compound word processing (e.g., Taft, Huang, & Zhu, 1994). Peng et al. (1999) manipulated cumulative character frequency (sum of the first and second character frequencies) and word frequency of compound words. They found a significant interaction between these variables in lexical decision performance. For high-frequency words, participants responded more quickly when cumulative character frequency was higher than when it was lower. In contrast, this simple effect of cumulative character frequency did not occur for low-frequency words. The result that the recognition of high-frequency words, but not low-frequency words, can be facilitated by their characters suggests that high-frequency words are accessed via their characters and low-frequency words are accessed as a whole word.

However, it is noteworthy that the interaction effect reported in Peng et al. (1999) was only marginally significant in participant analyses and not significant in item analyses. Indeed, this interaction was not significant in a virtual study based on Tse et al.'s (2017) normative data when the analyses were restricted to Peng et al.'s stimuli. Other patterns of character frequency and word frequency interaction have also been reported with other restricted stimulus sets (e.g., Taft et al., 1994, found lexical decision performance in the order of high–high = low–low > high–low = low–high in the first–second character frequency of compound words). Hence, it is important to re-examine the character frequency × word frequency interaction with a larger set of stimuli beyond those used in previous studies.

Semantic transparency effect

As defined by Zwitserlood (1994), semantic transparency refers to the extent that a compound word is semantically related to its characters (transparent; e.g., 花園 flower-park "garden") or unrelated to its characters (opaque; e.g., 花生 flower-grow "peanut"). It can be quantified by participants' ratings for semantic relatedness between a compound word and its characters (Mok, 2009). The semantic transparency effect is reflected by a quicker RT to transparent words than to opaque words (e.g., Lu et al., 2001; Myers, Libben, & Derwing, 2004, but see, for example, Frisson, Niswander-Klement, & Pollatsek, 2008; Myers, Derwing, & Libben, 2004).

Based on general dual-route models of morphological processing, the meaning of a compound word is simultaneously and independently accessed via direct retrieval of a whole-word representation and a decomposition-then-composition process of constituent (character) representations (e.g., Baayen, Dijkstra, & Schreuder, 1997; Libben & Jarema, 2007). The meaning of a compound word can be accessed via direct retrieval from semantic memory. When a compound word is accessed via decomposition, its characters' meanings are first activated and then combined to obtain the word's meaning. Lexical access is more time-consuming via the decomposition route than via the direct retrieval route. The contributions of the direct retrieval and decomposition routes in the access of a compound word can be regarded as race-like, that is, determined by the quicker route. For transparent words (e.g., 花園 "garden"), the outputs from direct retrieval and decomposition routes are similar, so their meanings can be accessed via either route. Such words may benefit from decomposition as the compound word and characters are closely connected in semantic memory. In contrast, for opaque words, because at least one of the characters is unrelated to the whole word, there is a conflict between the two routes (e.g., combined meaning of the two unrelated characters, 花 *flower* and 生 *grow* versus meaning of the word, 花生 *peanut*). This conflict may lead participants to engage in post-lexical checking to confirm the lexicality of that word (see, for example, Balota & Chumbley, 1984), thereby delaying the word recognition process for opaque, relative to transparent, words; this yields the semantic transparency effect. How semantic transparency interacts with character or word frequency in lexical decision performance (e.g., Peng et al., 1999) could also identify the boundary conditions under which the access of compound words is influenced by character- and/or word-variables.

Semantic transparency × character frequency interaction

Holding word frequency constant, Peng et al. (1999, see also Wang & Peng, 1999) examined the interaction between semantic transparency and cumulative character frequency of compound words in lexical decisions. They found a significant semantic transparency × cumulative character frequency interaction. For transparent words, participants responded more quickly when cumulative character frequency was higher than when it was lower, but the simple effect of cumulative character frequency was reversed for opaque words (though only marginally significant in the item analyses). In a virtual study using normative data, Tse and his colleagues (2017) also found that transparent words were responded to more quickly when their cumulative character frequency was higher than when it was lower, although they failed to find a reversed effect of cumulative frequency for opaque words. This suggests that opaque words are accessed as a whole word, independent of character frequency. In contrast, transparent words are accessed via their characters, as reflected by the influence of character frequency. However, this interaction was not fully replicated in other studies (e.g., Gao & Gao, 2005), which casts doubt on whether the effect could be generalisable beyond their stimulus set.

Semantic transparency × word frequency interaction

Mok (2009) had participants identify an underlined character that was embedded in a two-character Chinese compound word or nonword, which was briefly presented (40-80 ms) and preceded by a forward mask in each trial. She found that for both high- and low-frequency words, the word superiority effect (i.e., better recognition when the underlined character appeared in a word than in a nonword) was stronger for opaque words than for transparent words. Given that the word superiority effect reflects the extent to which characters are accessed in compound word processing, Mok's results suggest that compound words are accessed via their characters, regardless of word frequency. That being said, the lack of a semantic transparency × word frequency interaction has been questioned by subsequent research. Using a visual lexical decision task, Tse (2010) had participants respond to two-character Chinese compound words distributed equally across the 2 (word frequency) × 2 (semantic transparency) conditions, with words matched on other variables (e.g., number of strokes). Tse found a significant semantic transparency × word frequency interaction in RT, with the semantic transparency effect occurring for low-frequency words, but not for high-frequency words. The absence of a semantic transparency effect for high-frequency words suggests that compound words might not always be accessed through their characters.

Using the perspective that two routes underlie morphological processing, word frequency likely affects the direct retrieval route (i.e., speed of accessing a whole-word's representation) more strongly than the decomposition route. Relative to low-frequency opaque words, high-frequency opaque words are accessed more quickly via the direct retrieval route, such that any influence from the decomposition route (or character activation) is minimised. Therefore, semantic transparency might exert less influence on the lexical decisions of high-frequency words. For low-frequency opaque words, the direct retrieval of their meanings might not be fast enough to bypass the influence of the decomposition route. Thus, the conflict in the outputs from the direct retrieval and decomposition routes might slow down participants' lexical decisions for low-frequency opaque words, relative to those for low-frequency transparent words.

The discrepancy between Tse's (2010) and Mok's (2009) findings could be attributed to the use of different paradigms (word superiority paradigm and lexical decision task). In Mok's paradigm, the compound word was presented very briefly, and to-be-identified characters were underlined. This might emphasise character processing and bias participants to decompose the two-character compound words during the task, thereby weakening the modulation of word frequency on the semantic transparency effect. Nevertheless, both Tse's and Mok's findings were based on a relatively small set of stimuli ($N=200$), so it is important to determine whether the word frequency × semantic transparency interaction would replicate with a much larger stimulus pool.

Does character phonology affect lexical decision for two-character compound words?

According to Perfetti and Tan's (1999) Interactive Constituency Model, phonology is a privileged constituent in Chinese word recognition, and not just a by-product. Chinese word recognition results from a convergence of orthography, phonology, and semantics, and it can be argued that orthography-to-phonology mapping is more reliable than orthography-to-semantics mapping. This is attributed to the fact that there is a nearly one-to-one orthography–phonology relationship (e.g., 表 pronounced as *biu2*) but a one-to-many orthography–semantics relationship (e.g., 表 could mean watch, express, surface, or metre).[2]

If the mapping from orthography to phonology is nearly deterministic and the mapping from form (orthography and phonology) to meaning is under-deterministic, the one-to-one relation can be more quickly established in Chinese word recognition (Perfetti & Tan, 1998). Perfetti and Tan manipulated the stimulus-onset asynchrony of prime–target presentation and the lexical (orthographic, phonological, or semantic) relationship between prime and target to examine the time course of phonological and semantic activation in a primed naming paradigm. They found that the influence of phonological activation precedes that of semantic activation. Although this supports the involvement of phonology in Chinese character recognition, it is noteworthy that inconsistent evidence has also been documented in the literature (e.g., Chen & Shu, 2001; Myers, Taft, & Chou, 2007; Zhou & Marslen-Wilson, 2000b).

In compound word processing, if character phonology were obligatorily processed, one would expect that lexical decisions could be modulated by ambiguity at the phonological level. There could be a potential role of character phonology (e.g., phonological consistency) in modulating lexical decision performance of two-character compound word processing. Although most studies that have tested phonological involvement were based on single characters, a few exceptions have examined phonological consistency effects in lexical decision, using two-character Chinese compound words (Leong & Cheng, 2003).

Phonological consistency refers to whether a character has one (phonologically consistent) or more than one (phonologically inconsistent) pronunciation (e.g., Tan & Perfetti, 1999).[3] For example, 體重 “weight” is phonologically inconsistent because 重 is pronounced as *cung5*, but it can also be pronounced as *zung6* when 重 refers to important thing (e.g., 側重 “emphasise”). Using a visual lexical decision task, Leong and Cheng (2003) demonstrated that words with phonologically inconsistent character in the second position were responded more quickly than phonologically consistent words such as 骨頭 “bone,” where both characters are always pronounced as *gwat1* and *tau4*, respectively. To explain this result, Leong and Cheng argued that inconsistent characters might receive diffused phonological activation across characters that share the same pronunciation (i.e., other homophones) but consistent characters do not. Although it is not very clear why visual lexical decisions could be facilitated by diffused phonological activation and why this occurred only when phonologically inconsistent character was in the second, but not the first, position, this phonological consistency effect provides evidence for the role of character phonology in the recognition of compound words. Indeed, based on a virtual study using normative lexical decision data (Tse et al., 2017) with Leong and Cheng’s stimuli, this pattern of results was replicated. Nevertheless, it is noteworthy that the number of items in Leong and Cheng’s stimuli was rather limited: nine in each of the 2 (phonological consistency) × 2 (character position) conditions. The need to match extraneous variables compelled Leong and Cheng to use very small sets of stimuli, which might limit the generalisability of their results.

Using normative data in a megastudy (Sze, Rickard Liow, & Yap, 2014), Sze, Yap, and Rickard Liow (2015) conducted regression analyses to compare the proportion of unique item-level variance that orthographic (e.g., character frequency), phonological (e.g., consistency), and semantic (e.g., imageability) properties of a Chinese character accounted for the lexical decision performance for Chinese single characters. They found that orthographic and semantic variables, respectively, accounted for more variance than phonological variables, suggesting that skilled readers rely more on orthographic and semantic information than phonological information when processing visually presented characters. This suggests that phonological involvement might not be as strong as claimed in previous studies (e.g., Leong & Cheng, 2003). Nonetheless, it is important to further examine this issue by testing whether phonological variables would account for less variance than orthographic and semantic variables, even in lexical decision performance of two-character Chinese compound words.

The present study

The mixed evidence in Chinese compound word processing literature (see Myers, 2006, for a more detailed review) could be attributed to their use of factorial-design experiments that involved a relatively small set of stimuli constrained to be matched on various lexical variables (see, for example, Tse et al., 2017, for a review of other potential problems of factorial design). These concerns can be mitigated by an alternative strategy, the megastudy approach. Tse et al. recently used Balota et al.’s (2007) megastudy approach to develop a repository of behavioural measures and lexical variables for two-character Chinese compound word recognition where RT and accuracy rate were compiled from 594 university students in a lexical decision task on >25,000 Chinese words (see Balota, Yap, Hutchison, & Cortese, 2013, for a review of the megastudy approach). This study makes use of this large dataset to address the theoretically important questions raised in our literature review. Performing item-level analyses on these data could reduce idiosyncratic effects in word selection, reveal the proportion of unique variance of performance that a lexical variable, whether continuous or categorical, explain after statistically controlling for the effect of other variables, and indicate the relative contribution of these variables in lexical processing (e.g., Sze et al., 2015, for an example).

We first examine the relative influence of seven lexical variables on lexical decisions for two-character compound words. Orthographic variables include the stroke number and frequency of the first and second characters and word frequency.[4] Phonological variables include phonological consistency and homophonic density of the first and second characters. Semantic variables include semantic transparency of the first and second characters. All but phonological consistency are continuous variables. Apart from the main effect of these lexical variables, we investigated three theoretically driven interactions: (a) word frequency × semantic transparency, (b) word frequency × cumulative character frequency, and (c) semantic transparency × cumulative character frequency. Overall, the analyses were motivated by two main questions:

1. If phonology indeed plays a substantial role in two-character compound word processing, phonological variables should account for more variance than orthographic and semantic variables in lexical decision. This should hold even though the lexical decision task does not require participants to explicitly generate phonology.
2. If the way in which compound words are accessed depends on word frequency and semantic transparency, based on previous studies (e.g., Peng et al., 1999) we would expect a significant cumulative character frequency × word frequency interaction (larger effect of cumulative character frequency for high-frequency words than for low-frequency words), cumulative character frequency × semantic transparency interaction (larger effect of cumulative character frequency for transparent words than for opaque words), and semantic transparency × word frequency interaction (larger effect of semantic transparency for low-frequency words than for high-frequency words).

Method

Behavioural data from the Chinese Lexicon Project for two-character compound words (see Tse et al., 2017, for more details) were used in the present analyses. This is a repository of lexical variable and behavioural data (RT and accuracy) for 25,000+ commonly used traditional two-character Chinese compound words obtained from 594 native Cantonese speakers in Hong Kong. Over three sessions on separate days, participants completed 2,810 to 2,812 randomised lexical decision trials (1,405-1,406 words and 1,405-1,406 nonwords created by recombining characters). All compound nonwords consisted of two real Chinese characters. This ensures that participants have to process the compound words at the word level as they cannot make "word" responses based solely on the lexicality of individual characters.

Results

Only responses for experimental trials were analysed. RT from incorrect word responses were first excluded. The remaining responses being <200 ms or >3,000 ms were then excluded. The mean and standard deviation (*SD*) were computed for each participant's word responses. Any correct word response above or below 2.5 *SD* from his or her mean was labelled as outlier scores and excluded. Following Faust, Balota, Spieler, and Ferraro (1999) procedure, these RTs were then transformed into *z* scores for each participant, before averaging across the participants for each word to yield the individual word's zRT. This transformation controls for individual differences in processing speed and variability. The level of significance was set at .05. All analyses reported were conducted at the item level, that is, how fast responses were for each compound word averaged across participants.

Main analyses of the lexical variables using multiple regression analyses

For the orthographic variables, the number of strokes of the first and second characters was based on a pocket dictionary (Que, 2008). Number of strokes is often used as an index of visual complexity (e.g., Xing, Shu, & Li, 2004); characters with more strokes are regarded as being more complex (e.g., 人 "human" versus 鬱 "depressed"). The character and word frequency counts refer to the orthographic print exposure of character and word. Cai and Brysbaert's (2012) log-transformed subtitle frequency norms were used as they predict lexical decision performance better than other competing metrics (Tse et al., 2017). This measure was computed based upon the number of TV shows and films that a particular character/word appears within the subtitles. Note that this character frequency measure refers to character *token* frequency, which to our knowledge was used in all previous two-character Chinese word lexical decision studies and was therefore included in the current analyses. However, it is possible to quantify the character frequency as character *type* frequency (i.e., the number of different compound words in which a character can occur). This latter measure is discussed with details in the "Discussion" section.

Turning to the phonological variables, phonological consistency is defined by whether a character has one (phonologically consistent) or more than one (phonologically inconsistent) pronunciation (e.g., Leong & Cheng, 2003). It is a dichotomous variable, where 1 is consistent and 0 is inconsistent. Homophone density refers to the number of characters that share the same pronunciation. The number of homophones for each character was verified against the Multi-function Chinese Character Database (http://humanum.arts.cuhk.edu.hk/Lexis/lexi-mf/). For the semantic variable, semantic transparency is defined by the extent to which compound words are semantically related (transparent; e.g., 黑板 black-board "blackboard") or unrelated to their characters (opaque; e.g., 東西 east-west "thing"). It was quantified by 20 participants' ratings for the semantic

Table 1. Mean and standard deviation (*SD*) of lexical variables and dependent measures.

Variable (*N* = 18,983)	*M*	*SD*
RT (z-score)	−0.06	0.39
Accuracy rate	0.93	0.07
Number of strokes (C1)	10.64	4.44
Number of strokes (C2)	10.63	4.49
Log token frequency (C1)	3.35	0.51
Log token frequency (C2)	3.41	0.48
Log type frequency (C1)	1.65	0.41
Log type frequency (C2)	1.72	0.42
Log word frequency	1.63	0.80
Phonological consistency (C1)	0.59	0.49
Phonological consistency (C2)	0.58	0.49
Homophone density (C1)	18.57	15.75
Homophone density (C2)	18.61	15.66
Semantic transparency (C1)	0.00	0.49
Semantic transparency (C2)	0.00	0.52

RT: reaction time; C1: first character; C2: second character.

relatedness between a compound word and each of its characters, which showed moderate-to-high inter-rater reliability (Cronbach $\alpha = .82$).[5] Apart from the possibility that both characters are semantically transparent or opaque to the compound words (i.e., the examples given above), some compound words consist of one transparent character and one opaque character (e.g., 麻煩 flax-bother “trouble” as an opaque–transparent case, whereas 信差 letter-difference “messenger” as a transparent–opaque case).

Analyses of the main effects

Out of 22,808 words that yield at least 70% accuracy across participants, values of all lexical variables were available for 18,983 words, which were used in the following analyses. Table 1 presents the mean and *SD* of the lexical variables and the dependent measure for these 18,983 words. Table 2 shows their correlation matrices. Hierarchical regression analyses were run to examine the predictive power of orthographic, phonological, and semantic variables (see Table 3). Orthographic variables (number of strokes, character frequency, word frequency) were entered in the first step, phonological variables (phonological consistency, homophone density) in the second step, and semantic transparency in the third step.[6]

The analyses yielded a number of noteworthy observations. First, orthographic variables accounted for most of the variance in lexical decision performance. To maintain parity with previous work, we treated word frequency as an orthographic variable (but see Baayen, Feldman, & Schreuder, 2006). Word frequency (quicker and more accurate responses for high-frequency words) was by far the most powerful predictor, followed by character frequency (quicker and more accurate responses for compound words with high-frequency characters), then by number of strokes (slower responses for words with more complex characters).[7] Next, though the phonological variables collectively accounted for relatively little variance, words with phonologically consistent characters were recognised more quickly than those with phonologically inconsistent characters, and words with phonologically dense characters (i.e., with more homophones) were recognised more quickly than those with phonologically sparse characters. Finally, compound words were recognised more quickly and more accurately when they contained characters that were more semantically related to them.

Analyses of the interactive effects

We now turn to the three theoretically important interactions described in the Introduction: cumulative character token frequency × word frequency, cumulative character token frequency × semantic transparency, and semantic transparency × word frequency. The variables of interest, along with other control variables, were first z-transformed and entered in the regression, followed by the relevant interaction term in the second step. Across the analyses, the control variables entered, when they were not the variables of interest, were average number of strokes, log word frequency, log cumulative character token frequency, phonological consistency of the first and second characters, average homophone density, and average semantic transparency.

The cumulative character token frequency × word frequency interaction was significant in lexical decision RT, $F_{change}(1, 18{,}974) = 14.42$, $p < .001$. To explore this interaction, simple slopes were then plotted using the effects package (Fox, 2003). As shown in Figure 1 (top panel), the effect of cumulative character frequency was stronger for low-frequency, compared to high-frequency, words. The cumulative character token frequency × semantic transparency interaction was also statistically significant, $F_{change}(1, 18{,}974) = 26.15$, $p < .001$ (see Figure 1 bottom panel for simple slopes). In this instance, the effect of cumulative character token frequency was stronger for transparent, compared to opaque, compound words. Finally, the semantic transparency × word frequency interaction was not significant, $F_{change} < 1$, suggesting that the effect of semantic transparency was similar in magnitude for high- and low-frequency compound words.[8]

Discussion

Summary of the current findings

By using the megastudy approach (e.g., Balota et al., 2007) to analyse the normative data of lexical decision performance on two-character Chinese compound words

Table 2. Correlation matrices of lexical variables and dependent measures.

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	RT (z-score)	–														
2	Accuracy rate	–.66***	–													
3	No. of strokes (C1)	.08***	.01	–												
4	No. of strokes (C2)	.09***	.00	.02*	–											
5	Log token frequency (C1)	–.24***	.12***	–.27***	–.02**	–										
6	Log token frequency (C2)	–.23***	.11***	–.01†	–.27***	.15***	–									
7	Log type frequency (C1)	–.15***	.05***	–.33***	.00	.75***	.06***	–								
8	Log type frequency (C2)	–.14***	.04***	.00	–.36***	.07***	.75***	.05***	–							
9	Log word frequency	–.58***	.42***	–.06***	–.07***	.26***	.24***	.08***	.09***	–						
10	Phonological consistency (C1)	–.01	.01*	.06***	.02**	–.09***	–.03***	–.09***	–.01†	–.02**	–					
11	Phonological consistency (C2)	.01	.01	.00	.09***	–.01	–.11***	.01	–.12***	–.04***	.02*	–				
12	Homophone density (C1)	.00	.00	.02***	.00	–.09***	–.02*	–.09***	–.02**	–.01*	.06***	.00	–			
13	Homophone density (C2)	–.01	–.01	–.01	–.02**	–.03***	–.03***	–.02*	–.01	–.01	.00	.07***	.02**	–		
14	Semantic transparency (C1)	–.03***	.06***	.21***	–.03***	–.24***	–.01	–.23***	.04***	–.03***	.09***	–.01	.04***	.01†	–	
15	Semantic transparency (C2)	.04***	.02**	.03***	.22***	–.07***	–.27***	–.04***	–.29***	–.14***	.02**	.10***	.02*	–.06***	.07***	–

RT: reaction time; C1: first character; C2: second character.
****p* < .001; ***p* < .01; **p* < .05; †*p* < .10.

Table 3. Standardised RT and accuracy regression coefficients from Steps 1 to 3 of the item-level regression analyses.

Lexical variable	RT (z-score)	Accuracy rate
Step 1: Orthographic variables		
Number of strokes (C1)	.028***	.043***
Number of strokes (C2)	.026***	.037***
Log token frequency (C1)	−.081**	.025***
Log token frequency (C2)	−.078***	.013†
Log frequency (word)	−.538***	.417***
Adjusted R^2	*.354****	*.181****
Step 2: Phonological variables		
Phonological consistency (C1)	−.028***	.021**
Phonological consistency (C2)	−.027***	.024***
Homophone density (C1)	−.018**	.007
Homophone density (C2)	−.014*	−.006
Adjusted R^2	*.356****	*.182****
Change in R^2	*.002****	*.001****
Step 3: Semantic variables		
Semantic transparency (C1)	−.067***	.077***
Semantic transparency (C2)	−.064***	.076***
Adjusted R^2	*.364****	*.193****
Change in R^2	*.008****	*.011****

RT: reaction time.
The p value for each R^2 change is represented with asterisks.
***$p < .001$, **$p < .01$, *$p < .05$, †$p < .10$. The regression coefficients reported reflect the coefficients entered in that particular step.

(Tse et al., 2017), this study investigated two issues that have yielded mixed evidence in previous studies that used restricted stimulus sets. First, we tested whether character phonology plays a substantial role in Chinese compound word processing by comparing the proportions of variance in lexical decision performance being accounted for by phonological variables versus orthographic or semantic variables. Similar to Sze et al. (2015), we found that orthographic and semantic variables separately accounted for more variance in lexical decision RT (35.4% and 0.8%, respectively) than phonological variables (0.2%). This was also the case for accuracy data (18.1% and 1.1% versus 0.1%), showing that character phonology plays a relatively modest role in lexical decision performance, compared to the orthographic and semantic characteristics of the compound words.

That being said, it is worth noting that phonological variables *did* significantly predict lexical decision performance. Participants responded more quickly to compound words with characters that share pronunciation with many other characters (i.e., higher homophone density). On one hand, this finding might seem at odds with classic studies addressing the negative impact of homophony in English word recognition (e.g., Rubenstein, Lewis, & Rubenstein, 1971). That is, when the characters within the compound words activate their phonological representations, those with more homophones may be recognised more slowly, because there is more competition between the alternative orthographic structures sharing the same phonological representation (e.g., 朋 "friend" and 憑 "support" are both pronounced as *pang4*).

On the other hand, a positive effect of homophonic density (i.e., quicker RTs for words with more homophones) has been reported in other Chinese lexical decision studies (e.g., Chen, Vaid, & Wu, 2009; Ziegler, Tan, Perry, & Montant, 2000). It is noteworthy that the number of characters/words that share the same pronunciation is much higher in Chinese (11 on average) than in English (one in most cases), as suggested by Tan and Perfetti (1998). According to Chen et al. (2009), the higher homophonic density in Chinese may trigger strong feedback from phonology to orthography, and in turn a heightened level of global orthographic (or lexical) activation (similar to the premise of Grainger & Jacobs, 1996, multiple read-out model), such that a quick response can be made in a lexical decision task. In contrast, in English, most homophones share their pronunciation with only one other word, which may not cause a large increase in global orthographic (or lexical) activity. The single, activated homophone might even compete with the target word during word recognition, and in turn slow down lexical decision.

Using Japanese Kanji words in a lexical decision task, Hino, Kusunose, Lupker, and Jared (2013) reported slower RTs for words with higher homophonic density when they share their pronunciation with only one other word, but quicker RTs for word with higher homophonic density when they share their pronunciation with multiple words. Hence, the apparent contradictory effect of homophonic density in this study (with Chinese stimuli) and classic studies in English could be attributed to systematic differences in homophonic density between the two languages.

Apart from words with higher homophone density, words with phonologically consistent characters (i.e., the same pronunciation regardless of which other character it pairs with) yielded quicker RTs than those with phonologically inconsistent characters (i.e., pronunciation modulated by the other character that it pairs with). This latter facilitatory effect of phonological consistency stands in contrast to the inhibitory effect of phonological consistency reported by Leong and Cheng (2003). Moreover, the effect we observed was quite similar whether the phonological consistency occurred in the first or second character of the compound words, contrary to the position-specific effect reported by Leong and Cheng. However, phonological consistency may be confounded with morphemic/character consistency. When a character is pronounced in two different ways, it is likely that two pronunciations correspond to two different meanings. Hence, it is not clear whether the effect of phonological consistency was solely phonological in nature or could be partially attributed to the two competing meanings.

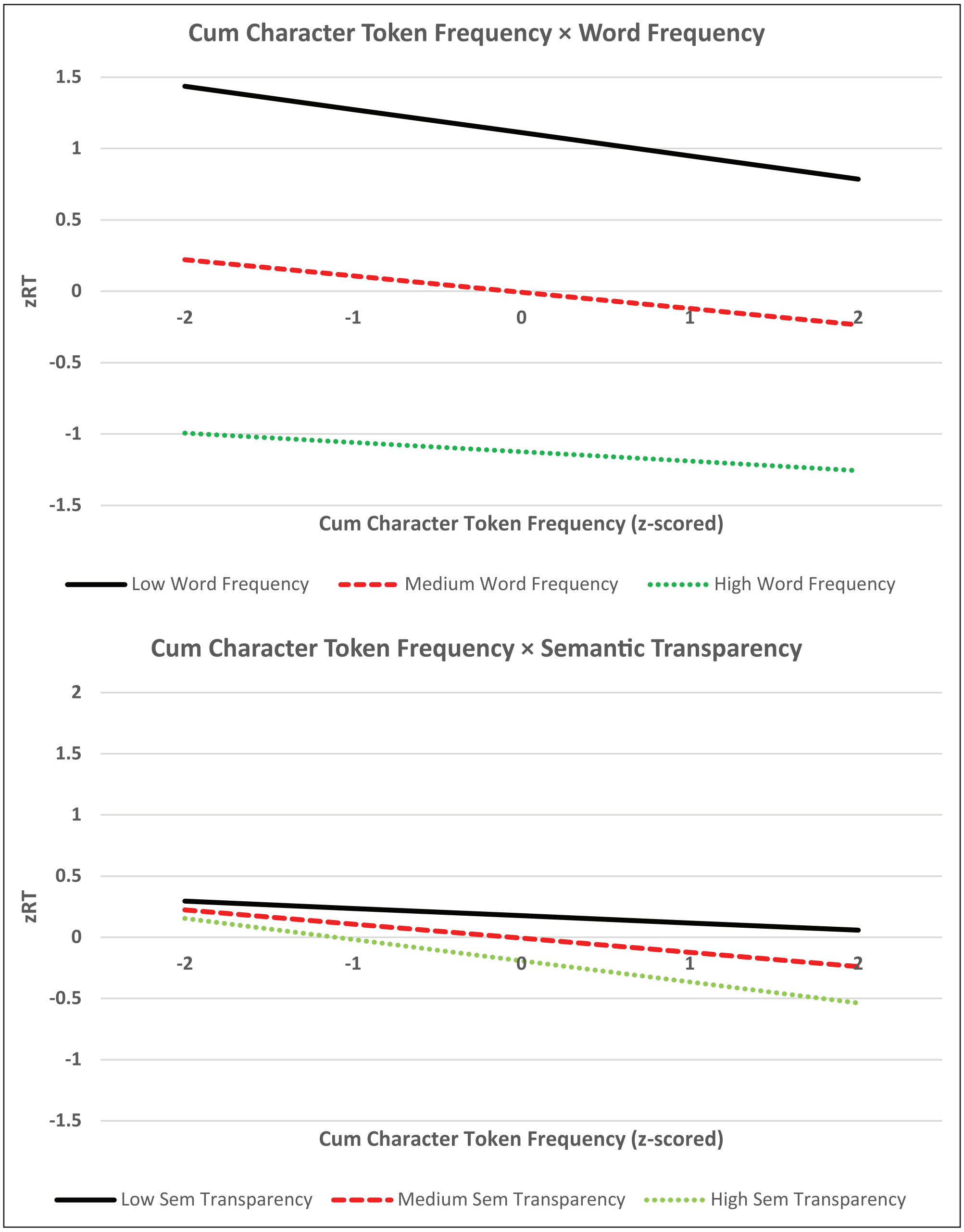

Figure 1. Significant interactions yielded in regression analyses.

Second, we tested the three word-variable × character-variable interactions related to the access of compound words: semantic transparency × cumulative character token frequency, semantic transparency × word frequency, and word frequency × cumulative character token frequency. We obtained a semantic transparency × cumulative character token frequency interaction, with the effect of cumulative character token frequency being stronger for transparent words than for opaque words. This was consistent with Peng et al.'s (1999) results, which were also based on cumulative character token frequency. Although we also obtained a word frequency × cumulative character token frequency interaction, the effect of cumulative character token frequency was stronger for low-frequency words than for high-frequency words, contrary to the pattern reported by Peng et al. Finally, we did not find a semantic transparency × word frequency interaction in RT, inconsistent with Tse's (2010) factorial study, which reported a larger semantic transparency effect for low-frequency, compared to high-frequency, words. The difference in the findings based on the current megastudy and those reported by previous studies (Peng et al., 1999; Tse, 2010) could be attributed to various causes, such as the restricted number of items used. Because these studies

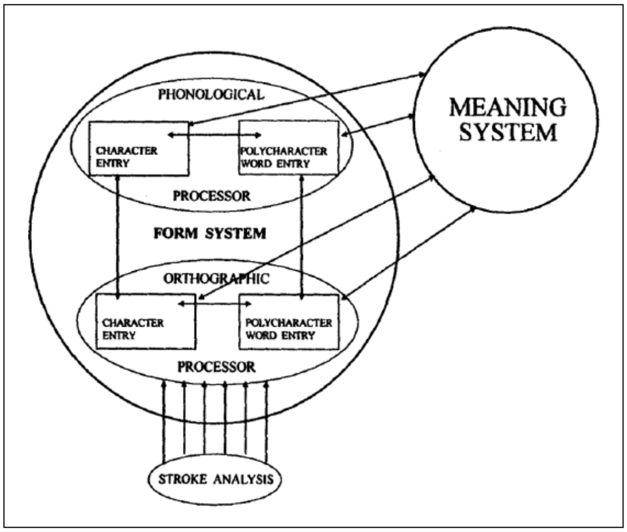

Figure 2. Tan and Perfetti's (1999) model of visual recognition of two-character Chinese words.

used factorial designs to examine the interaction of lexical variables, ensuring that extraneous variables were matched across conditions limited the number of stimuli that could be used, such that the findings may or may not necessarily generalise when a larger set of stimuli is studied.

Before discussing the theoretical implications of the current findings, it is important to highlight the distinction between character token frequency and character type frequency. The character frequency used in all previous studies in Chinese compound word processing (and in the current analyses) was computed by counting the number of times that a character occurs in a specific corpus (*character token frequency*). However, it is also possible to quantify character frequency by counting the number of different compound words in which the character can occur (*character type frequency*). To test whether this latter measure converges with the results from the typical token frequency measure, we computed the character type frequency for each character based on 49,360 two-character Chinese compound words that were initially identified in Tse et al. (2017) prior to the any pre-screening (see Tables 1 and 2 for their statistics). We then re-ran the same analyses as those that we reported for character token frequency.

When type frequency was used in place of token frequency, we essentially obtained exactly the same results for the two interactions that involved character frequency: the character frequency × word frequency interaction and character frequency × semantic transparency interaction. One could argue that if the character of a compound word combines with many different characters (i.e., high in character type frequency), the level of competition could be increased during word recognition, which may in turn counteract the access benefit afforded by high character token frequency. However, as the character type frequency and token frequency are highly correlated (.75 for both first and second characters), it is not appropriate to enter *both* character type frequency and character token frequency in the same regression model. Thus, the effects of these two character frequency measures should be further teased apart in future studies that manipulate them orthogonally (see Taft, 2003, for an example of a study that examined the joint effects of character type and token frequency in character decision [i.e., character or non-character?] and word decision [i.e., does a character exist as a free-standing word?]). In the following sections, we discuss whether the theoretical models of Chinese compound word processing could accommodate the present results, which are based on a much larger stimulus pool (18,893 words), compared with previous studies (mostly fewer than 300 words).

Implications of the current findings on models of Chinese compound word processing

Tan and Perfetti's (1999) model of visual recognition of two-character Chinese words. According to this model (see Figure 2), compound word recognition is determined by a form system with orthographic and phonological processors and a meaning system, which represents character and word meaning. The system first detects and analyses the strokes and positional relations of a two-character word. The detected features then send activation to the two characters' and word's orthographic units simultaneously, but the two characters' orthographic units may reach activation threshold before the word's orthographic unit, which implies the properties of individual characters may influence compound word processing. The characters' and word's orthographic units then send activation to their corresponding phonological units in the phonological processor.

At the same time, characters are combined through an assembly process, which can be influenced by word frequency. The characters of high-frequency words, which co-occur more frequently, are assembled more easily than those of low-frequency words, which co-occur less frequently. This explains the typical word frequency effect. The unitisation of the two characters and the activation of the word proceed independently and simultaneously. Similar assembly processes occur in the phonological processor and the meaning system. Whether the assembly processes are completed before, after, or during the time when the word is fully recognised depends on various factors, such as phonological consistency and semantic transparency of compound words. The inter-connection between the form and meaning systems represents the convergence of phonological, orthographic, and semantic information sources leading to word recognition, which could potentially explain the interactions between character-variables and word-variables on lexical decision

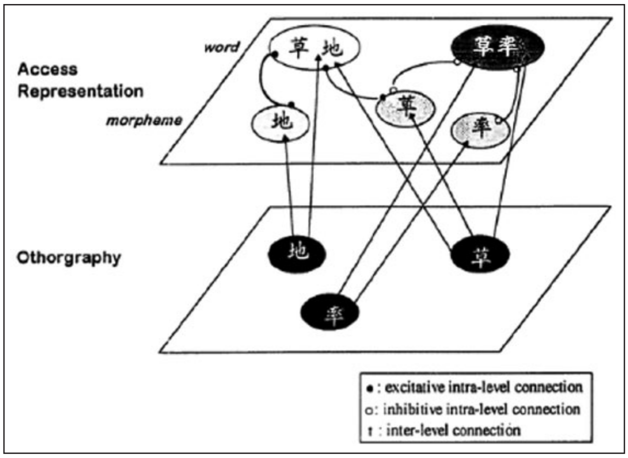

Figure 3. Peng et al.'s (1999) inter-intra connection model.

performance. However, this model has not been fully developed to account for the current findings, that is, the character frequency × word frequency and character frequency × semantic transparency interactions, alongside the additive effects of word frequency and semantic transparency.

Tan and Perfetti's (1999) model postulates a clear role of phonology in the recognition of two-character words. When one character of a compound word is phonologically inconsistent, its two phonological representations are activated and connected with one orthographic entry. The inappropriate phonological representation leads the assembly process down a garden path and in turn slows down readers' lexical decisions. This explains the facilitatory effect of phonological consistency in our analyses. When a character shares its pronunciation with many other characters (i.e., higher homophone density), feedback from the orthographic units counteracts potential interference from the homophones; the stronger phonological activation afforded by homophones facilitates compound word processing, thereby yielding the facilitatory effect of homophonic density. This is consistent with Chen et al.'s (2009) explanation that the quicker lexical decision RT for words with more homophones is due to the strong feedback from phonology to orthography that heightens the level of global orthographic (or lexical) activation.

Given that the orthography-to-phonology mapping is more reliable than orthography-to-semantics mapping (Perfetti & Tan, 1998, see also Perfetti, Liu, & Tan, 2005), this model suggests that phonology plays a fundamental role in word recognition (Perfetti & Dunlap, 2008, see also Perfetti & Liu, 2005, Lexical Constituency Model, for a computational account of the time course of orthography, phonology, and meaning constituents activation during explicit character recognition). In this study, even though our regression analyses did reveal a statistically significant effect of character phonology in compound word processing, the contribution of character phonology (as measured by phonological consistency and homophonic density) was much less than that of character orthography (as measured by number of strokes and character/word frequency). Interestingly, even the semantic transparency of compound words per se accounted for a larger proportion of variance in lexical decision RT and accuracy than the combined contribution of two phonological variables. Hence, the role of character phonology in compound word processing may not be as strong as suggested by Tan and colleagues. Other models of compound word processing, such as Peng et al. (1999) and Zhou and Marslen-Wilson (2000a), which are discussed below, have not explicitly specified the role of phonological representations in the processing of two-character words.

Peng, Liu, and Wang's (1999) inter-intra connection model. According to this model (see Figure 3 and also Liu & Peng, 1997), there are separate storage systems for characters and words. In the access representation, there are intralevel connections between character and whole word units and inputs from orthography can map directly onto access representations via interlevel connections where both words and characters can be activated. The strength of intralevel connections can be affected by word frequency and semantic relatedness between character and words (i.e., semantic transparency), which are consistent with the overall word frequency and semantic transparency effects in the current findings. Based on the pattern of Peng et al.'s character frequency × word frequency interaction, high-frequency words were more likely represented as decomposed components (i.e., facilitated by high, relative to low, character frequency), whereas low-frequency words were more likely represented in a unitary fashion (as reflected by a null effect of character frequency). However, this pattern was not replicated in the analyses of our megastudy, in which we found the opposite: low-frequency words showed a stronger character frequency effect than high-frequency words. Our findings are more consistent with the assumption of some lexical processing models that when presented frequently (i.e., high-frequency words), the two characters of a compound word are encountered together on many occasions, such that the word tends to be processed as a whole word unit (e.g., Caramazza, Laudanna, & Romani, 1988, Augmented Addressed Morphology model).

What accounts for the semantic transparency effect? The strong semantic relatedness between word and characters in a transparent word provides a strong positive connection between the word and its characters that may facilitate the recognition of the whole word. For opaque words, due to the negative relation between the word and its characters, the characters of the word may interfere with or inhibit recognition. Peng et al. (1999) showed a quicker RT for transparent words with higher cumulative character frequency. Relative to the low-frequency characters, the stronger activation (due

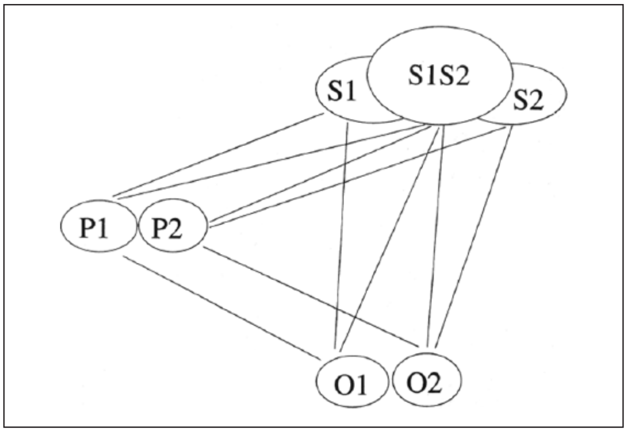

Figure 4. Zhou and Marslen-Wilson's (2000a) model for representation and processing of compound words.

to more frequent exposure or higher familiarity) of high-frequency characters facilitated recognition of transparent words to a greater extent. Peng et al. also found a slower RT for opaque words with higher cumulative character frequency, suggesting that the stronger activation of high-frequency characters, as compared with low-frequency characters, inhibited the recognition of opaque words to a greater extent. However, this latter effect was not replicated in a virtual study based on Tse et al.'s (2017) normative data when the analyses were restricted to the stimuli used in Peng et al. In this study, we also found a quicker RT for transparent words with higher cumulative character frequency, but no difference due to the cumulative character frequency for opaque words (see the bottom panel of Figure 1).

Given that word frequency and semantic transparency could both influence the strength of intralevel connections in the access representation, they are likely affecting a common processing stage during the recognition of compound words; this in turn produces interactive effects of word frequency and semantic transparency in lexical decisions to compound words. (It is noteworthy that the original proposal of Peng et al.'s, 1999, inter-intra connection model did not explicitly derive this prediction.) Contrary to this prediction, we did not find any word frequency × semantic transparency interaction in RT, even when our stimulus pool was much larger than those used in previous studies (e.g., Tse, 2010). Hence, our findings may pose difficulties for aspects of the inter-intra connection model.

Zhou and Marslen-Wilson's (2000a) model for representation and processing of compound words. According to this model (see Figure 4), both character and word representations are semantic in nature and are found at the same level. The semantic representations of compound words and characters possess overlapping semantic features in varying extent, depending on the degree of semantic relatedness between word and characters. As the spoken and written forms of compound words are essentially concatenations of the forms of their characters, the orthographic and phonological representations of compound words can be represented as the combinations of the form representations of their characters. The orthographic representations of characters connect directly to their corresponding phonological and semantic representations, as well as to the semantic representations of words that contain these characters. The semantic activation of words and characters can then feedback to the phonological and orthographic representations in lexical processing.

According to this model, transparent and opaque words are represented in the same way at the orthographic and phonological levels, but their semantic representations differ in the degree of semantic overlap between words and characters, with transparent words sharing more features than opaque words. In other words, semantic transparency is graded in nature. In lexical processing, semantic representations of both words and characters are activated in parallel from their orthographic/phonological representations, resulting in possible competition when the two sets of semantic features are not shared between the word and characters. For transparent words, the facilitation of large semantic overlap between word and characters is larger than the interference from the competition between the unshared semantic features. For opaque words, there is only interference from competition between unshared semantic features, without any facilitation due to the null semantic overlap between word and characters. Hence, transparent words are better recognised than opaque words, producing a semantic transparency effect.

In contrast to the detailed explanations provided for the role of semantic transparency, the effect of word frequency has not been explicitly conceptualised in Zhou and Marslen-Wilson's (2000a) model. However, based on the configuration of the model, it is likely that word frequency is regarded as a semantic variable, rather than as an orthographic variable (because there are no orthographic representations for compound words). This would predict a semantic transparency × word frequency interaction, which was not supported by the current RT findings. On the contrary, this model could predict character frequency × word frequency and character frequency × semantic transparency interactions *only* when character frequency is regarded as a semantic variable. In contrast, if character frequency is regarded as orthographic in nature, neither of these interactions would be predicted, which is inconsistent with the present results from the regression analyses. Hence, without a more detailed specification, it is not clear whether Zhou and Marslen-Wilson's model is able to accommodate our current findings.

Taft et al.'s (1999) interactive activation model. Following the classic connectionist model (Seidenberg & McClelland, 1989), some multi-level interactive models assume that the word and characters of a compound word are represented at

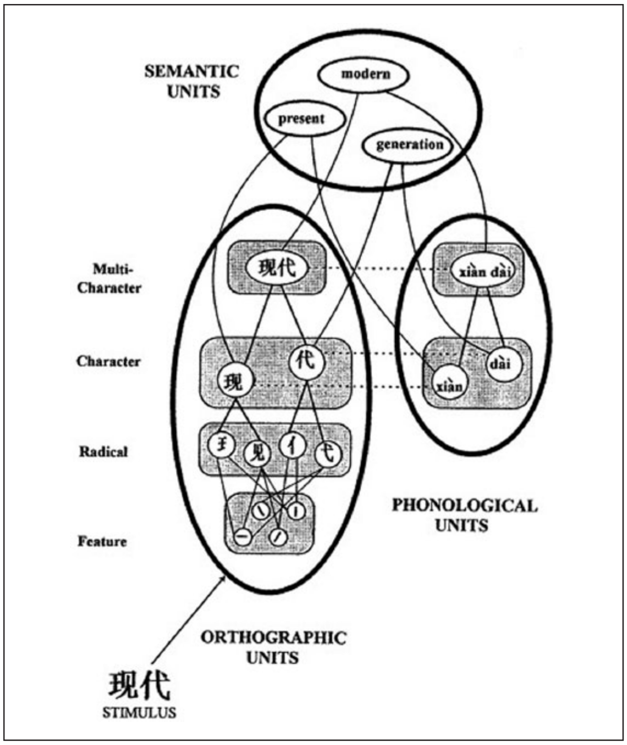

Figure 5. Taft and Zhu's (1995) interactive activation model (figure adapted from Taft, Zhu, & Peng, 1999).

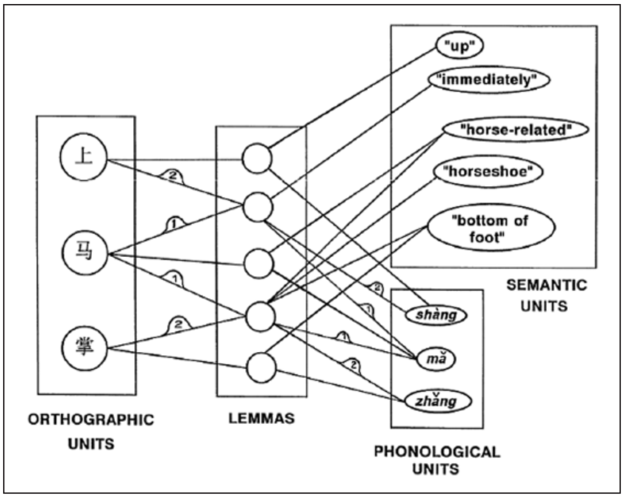

Figure 6. Taft et al.'s (1999) interactive activation model.

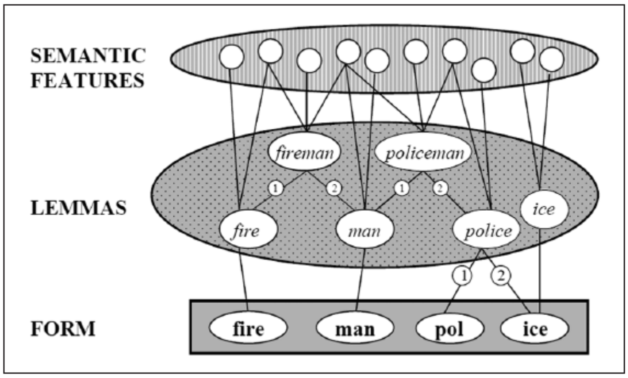

Figure 7. The representation of polymorphemic and polysyllabic words when lemmas are hierarchically represented in Taft (2003).

different levels, in contrast to Peng et al.'s (1999) model and Zhou and Marslen-Wilson's (2000a) model. In an earlier form of these models (e.g., Taft & Zhu, 1995, see Figure 5), the units at the meaning level consist of a large number of semantic features. These can be linked to units at all but the stroke level because radicals, characters, and compound words can all be associated with meaning. At the character and word levels, there are both orthographic and phonological representations. When a word is visually presented, activation moves up from the stroke level to the meaning level via various orthographic units and potentially via the associated phonological units. There are separate character and word levels of orthographic representation, and each of these can be associated with both semantic and phonological representations. Lexical decision responses to compound words are based on whether or not there is an orthographic representation and/or a semantic representation at the word level. Although the phonological representation might play a role, it is not as reliable as other representations because a nonword could be pronounced similarly with a real word, such as heterographic homophones.

In an updated version of this interactive activation model (Taft et al., 1999), instead of having word-level orthographic and phonological representations, Taft et al. proposed a set of modality-free, abstract lemma units that link together orthographic units, phonological units, and semantic units (see Figure 6). Similar to the hidden units that intervene between input and output layers in typical distributed connectionist models of lexical processing (e.g., Seidenberg & McClelland, 1989), the lemma units are developed when orthographic and/or phonological information and semantic information of a word or a character repeatedly co-occur across various contexts. The strength of the lemma units depends on the correlation between the meaning and form of word/character. Recently, Taft and his colleagues proposed an updated lemma model of word recognition (Taft, 2003; see also Taft & Nguyen-Hoan, 2010, for a related account on derived words). This updated model (Figure 7) is quite similar in its structure to Peng et al.'s (1999) inter-intra connection model, except that the access representations are now lemmas (see Figures 3 and 7). The activation of the morpheme (or character) lemmas are activated in parallel with compound-word lemma, with the former being a critical step in activating the latter. Hence, the compound-word lemmas are hierarchically activated via their morpheme/character lemmas when they are semantically related (i.e., in transparent words), but compete with each other when they are unrelated (i.e., in opaque words). Although this model was not specifically designed for Chinese compound words, we can still consider if it is able to accommodate the current findings.

In Taft's lemma models (Taft, 2003; Taft et al., 1999), the strength of connection between units depends on how frequent that connection is used, so high-frequency words are activated more quickly via stronger links than low-frequency words, which accounts for the word frequency effect. Compound words are recognised via the processing of their characters, which is sensitive to the extent to which the activation of the character lemmas contributes to the activation of the corresponding compound-word lemmas. Given that the threshold of activating the high-frequency word lemma is lower than that of activating the low-frequency word lemma, the latter case might be more sensitive to the activation of high- versus low-frequency character lemmas, thus yielding the character frequency × word frequency interaction we observed.

Semantic transparency depends on the degree to which the semantic units associated with the lemma for the word overlap with those associated with the lemmas for its characters. A two-character word is represented at the function-based lemma level, drawing its activation from the representations of its characters. The conflict between the activation of semantic units of characters and word takes place in the lemma units. Specifically, when those lemmas are not linked to overlapping semantic information (i.e., opaque words), the lemma of the compound word receives greater competition from the lemmas of its individual characters. In addition, the activation of semantic units via the lemma for transparent words, but not for opaque words, could feedback to the lemmas of characters, thereby speeding up the recognition of these words. This explains the semantic transparency effect. When the characters are of higher frequency, they would activate their lemma more strongly and thus trigger a stronger competition to the lemma of the compound word. Hence, the semantic transparency effect would be stronger when character frequency is high, consistent with the pattern observed in our findings.

Turning to the word frequency × semantic transparency interaction, while Taft's models (Taft, 2003; Taft et al., 1999) do not provide an explicit prediction for this interaction, it is possible to speculate. The lemmas of high-frequency words should be more resistant to competition from lemmas of their characters. As suggested by the significant cumulative character frequency × semantic transparency interaction, the semantic transparency effect is mediated by the competition from character lemmas, such that it should be smaller for high-frequency words than for low-frequency words. However, we did not observe this interaction in the present RT analyses.

To explain the additive effects of word frequency and semantic transparency, one might need to assume that the activation level of word lemmas provides minimal protection against competition from character lemmas. Given the similarity between Taft's (2003) updated lemma model and Peng et al.'s (1999) model, the latter could also account for the lack of word frequency × semantic transparency interaction in RT with this additional assumption. However, this modification would make it hard for these models to explain the character frequency × semantic transparency and character frequency × word frequency interactions. In other words, it is very challenging to provide a principled explanation for the additive effects of word frequency and semantic transparency, which is at the same time compatible with the character frequency × semantic transparency and character frequency × word frequency interactions.

One possibility is that the amount of competition received by an opaque word during word recognition is more likely to be reflected by the frequency of the word relative to the frequency of its characters. In other words, the higher the ratio of word frequency to cumulative character frequency, the less likely it is that character lemmas, relative to the compound-word lemma, are going to trigger strong competition when the characters are unrelated to the words (i.e., opaque words); this predicts a smaller semantic transparency effect. Thus, one might expect the ratio of word frequency to cumulative character frequency, rather than word frequency per se, to interact with semantic transparency. To address this possibility, we conducted a supplementary analysis where word frequency was replaced with the ratio frequency measure. Although word frequency is highly correlated with ratio frequency ($r = .98$), we found a significant ratio frequency × semantic transparency interaction, wherein semantic transparency effects become smaller as ratio frequency becomes larger. This was true whether cumulative character frequency was controlled ($p < .001$) or not controlled ($p = .003$) in an earlier step in the regression analyses. This interaction was also significant in accuracy ($p < .001$), indicating that semantic transparency effects are larger for words with low ratio frequencies. This intriguing finding is compatible with the predictions made by both Taft's (2003) model and Peng et al.'s (1999) model, which are structurally quite similar. However, to our knowledge, this ratio frequency measure is novel and has not been studied in the literature. It is possible that the degree to which the activation (and thus the competition due to semantic opacity) reflected by frequency might not be in the same scale for compound words and for their corresponding character. With that in mind, it is important to implement the computational form of all these models before the activation dynamics can be more clearly observed and more precise and fine-grained predictions can be explicitly tested.

Regarding the role of phonology in word recognition, Taft et al.'s (1999) model does not explicitly state whether phonological activation is a crucial component for character or word recognition. However, phonology may not be crucial for the semantic activation of a character or word because both orthography→meaning and orthography→phonology pathways are mediated by the same lemma unit. In other words, character/word meaning can be entirely activated via the orthography→meaning pathway, without any involvement of phonological units. Moreover, in Taft's (2003) updated lemma model, the ambiguity due to extensive Chinese homophony could be resolved by the introduction of character orthography, given that different homophonic morphemes are represented by different characters. This might explain why character phonology does not necessarily contribute more than other lexical characteristics of characters in compound word processing.

Overall, although the above models were proposed to explain two-character Chinese compound word processing, they might be optimised to explain some lexical effects, but not others. Indeed, no single model can fully accommodate the current constellation of findings without some degree of modification. Taft's (2003) model does a relatively good job of accounting for effects, especially when a novel word-frequency variable (i.e., ratio of word to cumulative character frequency) was explored in the context of the joint effects of word frequency and semantic transparency. We look forward to the development of implemented versions of the foregoing models in the future that can be then evaluated against the present behavioural benchmarks.

Conclusion

Using the lexical decision megastudy data of 18,983 two-character Chinese compound words (Tse et al., 2017), we performed item-level hierarchical regression analyses and found that orthographic and semantic variables, respectively, accounted for more collective variance than phonological variables. This provides evidence against the view about the stronger contribution of character phonology (compared with other lexical characteristics) on the processing of compound words. We obtained significant semantic transparency × cumulative character frequency and word frequency × cumulative character frequency interactions, but not the semantic transparency × word frequency interaction (at least in RT when the typical word frequency measure was used), which might not be straightforwardly accommodated by models of Chinese compound word processing.

Overall, together with previous studies (e.g., Sze et al., 2015), this study provides another demonstration on how normative data in megastudies can be utilised to test the lexical and semantic effects in word recognition. Moving forward, more sophisticated computational modelling and simulation based on these data can be done to inform and constrain existing and emerging models of Chinese lexical processing.

Acknowledgements

The authors thank Marcus Taft for his suggestions (e.g., an examination of character type frequency) on an earlier version of this manuscript and Nga-Ting Lee for cross-checking the values of the lexical variables.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by General Research Fund (#451113) awarded by Hong Kong Research Grants Council.

Notes

1. It is important to note that character frequency is generally much higher than word frequency, so a relatively high-frequency value for a word can potentially be a relatively low-frequency value for a character. Thus, even though the frequency of a compound word does contribute to the frequency of its characters, a low-frequency character (relative to other characters) could be found in a high-frequency compound word (relative to other compound words).
2. All syllables listed in this article are based on Cantonese pronunciation.
3. This definition is different from the one proposed by Feng, Horng, and Tzeng (1986), which defined consistency at the sub-character level in accord with whether a character is pronounced in the same way as its phonetic radical. As we are interested in replicating Leong and Cheng (2003), we adapted their definition of phonological consistency.
4. Word frequency has often been treated as a lexical variable in psycholinguistic literature (e.g., Reingold, Yang, & Rayner, 2010). However, some researchers (e.g., Baayen et al., 2006) have argued that word frequency is semantic because it reflects conceptual familiarity, which is tightly related to other semantic variables. However, to directly compare our findings of this study with the previous ones based on megastudy data, we followed their lead in terms of classifying character and word frequency as orthographic variables (e.g., Sze et al., 2015).
5. Specifically, the 25,286 words were divided into 18 sets of 1,404 to 1,405 words. Each of these sets was presented to 20 raters to yield semantic relatedness judgments for each word (one for first character-word and one for second character-word) (see Tse et al., 2017, for more details).
6. In the initial regression model, we also entered the interaction term of the semantic transparency of the first and second characters to examine whether the effect of semantic transparency of one character would be moderated by the extent to which the other character is semantically transparent. However, the interaction was not significant ($p = .95$), suggesting that the effects of semantic transparency of the two characters were additive. Hence, we did not consider this interaction term further in our analyses.

7. It is noteworthy that the interpretation of the RT effect of number of strokes might be complicated by the opposite effect in accuracy, that is, more accurate responses for words with more complex characters.
8. We performed the same analyses for accuracy data. The cumulative character token frequency × word frequency interaction was significant, $F_{change}(1, 18{,}974) = 57.29$, $p < .001$. The facilitatory effect of cumulative character token frequency was predominantly observed in low-frequency words. The cumulative character token frequency × semantic transparency interaction was also significant, $F_{change}(1, 18{,}974) = 9.01$, $p = .003$; the facilitatory effect of cumulative character token frequency is most clearly seen for transparent words. These two findings mirrored those we obtained in RT. However, contrary to the RT findings, we obtained a weak but significant semantic transparency × word frequency interaction in accuracy, $F_{change}(1, 18{,}974) = 4.30$, $p = .038$, with the semantic transparency effect being larger for low-frequency words.

References

Baayen, R. H., Dijkstra, T., & Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual route model. *Journal of Memory and Language*, *37*, 94–117.

Baayen, R. H., Feldman, L. B., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, *53*, 496–512.

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance*, *10*, 340–357.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., . . . Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*, 445–459.

Balota, D. A., Yap, M. J., Hutchison, K. A., & Cortese, M. J. (2013). Megastudies: What do millions (or so) of trials tell us about lexical processing? In J. S. Adelman (Ed.), *Visual word recognition, Vol. 1: Models and methods, orthography and phonology* (pp. 90–115). New York, NY: Psychology Press.

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS ONE*, *5*, e10729.

Caramazza, A., Laudanna, A., & Romani, C. (1988). Lexical access and inflectional morphology. *Cognition*, *28*, 297–332.

Chen, H.-C., & Shu, H. (2001). Lexical activation during the recognition of Chinese characters: Evidence against early phonological activation. *Psychonomic Bulletin and Review*, *8*, 511–518.

Chen, H.-C., Vaid, J., & Wu, J.-T. (2009). Homophone density and phonological frequency in Chinese word recognition. *Language and Cognitive Processes*, *24*, 967–982.

Faust, M. E., Balota, D. A., Spieler, D. H., & Ferraro, F. R. (1999). Individual differences in information processing rate and amount: Implications for group differences in response latency. *Psychological Bulletin*, *125*, 777–799.

Feng, S. P., Horng, R. Y., & Tzeng, O. J. L. (1986). Consistency effects in the Chinese character and pseudo-character naming tasks. In H. S. R. Kao & R. Hoosain (Eds.), *Linguistics, psychology, and the Chinese language* (pp. 11–21). Hong Kong: Center of Asian Studies.

Fox, J. (2003). Effect displays in R for generalized linear models. *Journal of Statistical Software*, *8*, 1–27.

Frisson, S., Niswander-Klement, E., & Pollatsek, A. (2008). The role of semantic transparency in the processing of English compound words. *British Journal of Psychology*, *99*, 87–107.

Gao, B., & Gao, F. (2005). The interaction between word frequency and semantic transparency in the recognition of Chinese words. *Psychological Science*, *28*, 1358–1360. (in Chinese)

Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review*, *103*, 518–565.

Hino, Y., Kusunose, Y., Lupker, S. J., & Jared, D. (2013). The processing advantage and disadvantage for homophones in lexical decision tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 529–551.

Institute of Language Teaching and Research. (1986). *A frequency dictionary of modern Chinese*. Beijing, China: Beijing Language Institute Press.

Leong, C.-K., & Cheng, P.-W. (2003). Consistency effects on lexical decision and naming of two-character Chinese words. *Reading and Writing*, *16*, 455–474.

Libben, G., & Jarema, G. (2007). *The representation and processing of compound words*. Oxford, UK: Oxford University Press.

Liu, Y., & Peng, D. (1997). Meaning access of Chinese compounds and its time course. In H.-C. Chen (Ed.), *Cognitive processing of Chinese and related Asian languages* (pp. 219–232). Hong Kong: The Chinese University Press.

Lu, C.-C., Bates, E., Hung, D., Tzeng, O., Hsu, J., Tsai, C.-H., & Roe, K. (2001). Syntactic priming of nouns and verbs in Chinese. *Language and Speech*, *44*, 437–471.

Mok, L. (2009). Word-superiority effect as a function of semantic transparency of Chinese bimorphemic compound words. *Language and Cognitive Processes*, *24*, 1039–1081.

Myers, J. (2006). Processing Chinese compounds: A survey of the literature. In G. Libben & G. Jarema (Eds.), *The representation and processing of compound words* (pp. 169–196). Oxford, UK: Oxford University Press.

Myers, J., Derwing, B., & Libben, G. (2004). The effect of priming direction on reading Chinese compounds. *Mental Lexicon Working Papers*, *1*, 69–86.

Myers, J., Libben, G., & Derwing, B. (2004, July 3). *The nature of transparency effects in Chinese compound processing*. Poster presented at the Fourth International Conference on the Mental Lexicon, Windsor, Ontario, Canada.

Myers, J., Taft, M., & Chou, P. (2007). Character recognition without sound or meaning. *Journal of Chinese Linguistics*, *35*, 1–57.

Peng, D. L., Liu, Y., & Wang, C. (1999). How is access representation organized? The relation of polymorphemic words and their morphemes in Chinese. In J. Wang, A.W. Inhoff & H.-C. Chen (Eds.), *Reading Chinese script: A cognitive analysis* (pp. 65–89). Mahwah, NJ: Lawrence Erlbaum.

Perfetti, C. A., & Dunlap, S. (2008). Learning to read: General principles and writing system variations. In K. Koda & A. Zehler (Eds.), *Learning to read across languages* (pp. 13–38). Mahwah, NJ: Lawrence Erlbaum.

Perfetti, C. A., & Liu, Y. (2005). Orthography to phonology and meaning: Comparisons across and within writing systems. *Reading and Writing*, *18*, 193–210.

Perfetti, C. A., Liu, Y., & Tan, L. H. (2005). The lexical constituency model: Some implications of research on Chinese for general theories of reading. *Psychological Review*, *12*, 43–59.

Perfetti, C. A., & Tan, L. H. (1998). The time course of graphic, phonological, and semantic activation in Chinese character identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 101–118.

Perfetti, C. A., & Tan, L. H. (1999). The constituency model of Chinese word identification. In A. Inhoff, H.-C. Chen & J. Wang (Eds.), *Reading Chinese script: A cognitive analysis* (pp. 115–134). Mahwah, NJ: Lawrence Erlbaum.

Que, D. L. (2008). *Longman Chinese dictionary* (3rd ed.). Hong Kong: Longman.

Reingold, E. M., Yang, J., & Rayner, K. (2010). The time course of word frequency and case alternation effects on fixation times in reading: Evidence for lexical control of eye movements. *Journal of Experimental Psychology: Human Perception and Performance*, *36*, 1677–1683.

Rubenstein, H., Lewis, S., & Rubenstein, M. A. (1971). Evidence for phonemic recording in visual word recognition. *Journal of Verbal Learning & Verbal Behavior*, *10*, 645–657.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed developmental model of word recognition and naming. *Psychological Review*, *96*, 523–568.

Sze, W. P., Rickard Liow, S. J., & Yap, M. J. (2014). The Chinese lexicon project: A repository of lexical decision behavioral responses for 2,500 Chinese characters. *Behavior Research Methods*, *46*, 263–273.

Sze, W. P., Yap, M. J., & Rickard Liow, S. J. (2015). The role of lexical variables in the visual recognition of Chinese characters: A megastudy analysis. *The Quarterly Journal of Experimental Psychology*, *68*, 1541–1570.

Taft, M. (2003). Morphological representation as a correlation between form and meaning. In E. Assink & D. Sandra (Eds.), *Reading complex words* (pp. 113–138). Amsterdam, The Netherlands: Kluwer.

Taft, M., Huang, J., & Zhu, X. (1994). The influence of character frequency on word recognition responses in Chinese. In H.-W. Chang, J.T. Huang, C.-W. Hue & O. Tzeng (Eds.), *Advances in the study of Chinese language processing*. Taipei: National Taiwan University.

Taft, M., Liu, Y., & Zhu, X. (1999). Morphemic processing in reading Chinese. In J. Wang, A. Inhoff & H. C. Chen (Eds.), *Reading Chinese script: A cognitive analysis* (pp. 91–113). Mahwah, NJ: Lawrence Erlbaum.

Taft, M., & Nguyen-Hoan, M. (2010). A sticky stick? The locus of morphological representation in the lexicon. *Language and Cognitive Processes*, *25*, 277–296.

Taft, M., & Zhu, X. (1995). The representation of bound morphemes in the lexicon: A Chinese study. In L. Feldman (Ed.), *Morphological aspects of language processing*. Hillsdale, NJ: Lawrence Erlbaum.

Taft, M., Zhu, X., & Peng, D. (1999). Positional specificity of radicals in Chinese character recognition. *Journal of Memory and Language*, *40*, 498–519.

Tan, L. H., & Perfetti, C. A. (1998). Phonological codes as early sources of constraint in reading Chinese: A review of current discoveries and theoretical accounts. *Reading and Writing*, *10*, 165–200.

Tan, L. H., & Perfetti, C. A. (1999). Phonological activation in visual identification of Chinese two-character words. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *25*, 382–393.

Tse, C.-S. (2010). *Interactive effect of semantic transparency and word frequency on lexical decision for Chinese compound words*. Paper presented at the 51st Annual Meeting of the Psychonomic Society, St. Louis, MO. 21 November 2010.

Tse, C.-S., Yap, M. J., Chan, Y.-L., Sze, W. P., Shaoul, C., & Lin, D. (2017). The Chinese lexicon project: A megastudy of lexical decision performance for 25,000+ traditional Chinese two-character compound words. *Behavior Research Methods*, *49*, 1503–1519.

Wang, C. M., & Peng, D. L. (1999). Effects of semantic transparency and surface frequency on Chinese word processing. *Acta Psychologica Sinica*, *31*, 266–273.

Xing, H., Shu, H., & Li, P. (2004). The acquisition of Chinese characters: Corpus analyses and connectionist simulations. *Journal of Cognitive Science*, *5*, 1–49.

Zhang, B., & Peng, D. L. (1992). Decomposed storage in the Chinese lexicon. In H.-C. Chen & O. Tzeng (Eds.), *Language processing in Chinese* (pp. 131–149). Amsterdam, The Netherlands: North-Holland.

Zhou, X., & Marslen-Wilson, W. (1995). Morphological structure in the Chinese mental lexicon. *Language and Cognitive Processes*, *10*, 545–600.

Zhou, X., & Marslen-Wilson, W. (2000a). Lexical representation of compound words: Cross-linguistic evidence. *Psychologia*, *43*, 47–66.

Zhou, X., & Marslen-Wilson, W. (2000b). The relative time course of semantic and phonological activation in reading Chinese. *Journal of Experimental Psychology, Learning, Memory, and Cognition*, *26*, 1245–1265.

Ziegler, J. C., Tan, L. H., Perry, C., & Montant, M. (2000). Phonology matters: The phonological frequency effect in written Chinese. *Psychological Science*, *11*, 234–238.

Zwitserlood, P. (1994). The role of semantic transparency in the processing and representation of Dutch compounds. *Language and Cognitive Processes*, *9*, 341–368.