The Chinese Lexicon Project: A megastudy of lexical decision performance for 25,000+ traditional Chinese two-character compound words

Chi-Shing Tse[1,2] · Melvin J. Yap[3] · Yuen-Lai Chan[1] · Wei Ping Sze[3] · Cyrus Shaoul[4] · Dan Lin[5,6]

Published online: 12 October 2016

Abstract Using a megastudy approach, we developed a database of lexical variables and lexical decision reaction times and accuracy rates for more than 25,000 traditional Chinese two-character compound words. Each word was responded to by about 33 native Cantonese speakers in Hong Kong. This resource provides a valuable adjunct to influential mega-databases, such as the Chinese single-character, English, French, and Dutch Lexicon Projects. Three analyses were conducted to illustrate the potential uses of the database. First, we compared the proportion of variance in lexical decision performance accounted for by six word frequency measures and established that the best predictor was Cai and Brysbaert's (*PLoS One, 5*, e10729, 2010) contextual diversity subtitle frequency. Second, we ran virtual replications of three previously published lexical decision experiments and found convergence between the original experiments and the present megastudy. Finally, we conducted item-level regression analyses to examine the effects of theoretically important lexical variables in our normative data. This is the first publicly available large-scale repository of behavioral responses pertaining to Chinese two-character compound word processing, which should be of substantial interest to psychologists, linguists, and other researchers.

Keywords Chinese · Compound word · Megastudy · Reaction time · Visual word recognition

In English, a compound word is typically formed by a combination of two or more constituent words (e.g., *snow* and *man* in *snowman*) (Dressler, 2006). Similarly, in Chinese, compound words are often formed via a combination of two or more characters that are almost always monosyllabic morphemes (e.g., 雪人 snow-man [snowman]). However, about 73.6 % of modern Chinese words are two-character compound words (Institute of Language Teaching and Research, 1986), showing that compounding is normative, rather than exceptional, in Chinese word formation (Packard, 2000). Adopting the megastudy approach in Balota et al.'s (2007) English Lexicon Project, we collected normative data for participants' lexical decision (i.e., deciding whether a two-character string forms a Chinese word, e.g., 朋友 [friend], or a nonword, e.g., 形忌) performance for more than 25,000 two-character Chinese words varying on various lexical characteristics, such as word frequency. Before elaborating on the details of the present megastudy, we first provide a brief selective review of the literature on Chinese two-character compound word processing, including a discussion of its limitations, and then show how the megastudy approach can complement this body of factorial work and help shed additional light on Chinese compound word processing.

Electronic supplementary material The online version of this article (doi:10.3758/s13428-016-0810-5) contains supplementary material, which is available to authorized users.

✉ Chi-Shing Tse
cstse@cuhk.edu.hk

[1] Department of Educational Psychology, The Chinese University of Hong Kong, New Territories, Hong Kong, China

[2] Centre for Learning Sciences and Technologies, The Chinese University of Hong Kong, New Territories, Hong Kong, China

[3] Department of Psychology, National University of Singapore, Singapore, Singapore

[4] Department of Linguistics, University of Tübingen, Tübingen, Germany

[5] Department of Psychological Studies, The Education University of Hong Kong, Hong Kong, China

[6] Centre for Brain and Education, The Education University of Hong Kong, Hong Kong, China

Compound word processing in Chinese

Character and word frequency

Several 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, 2000). The most well-known psycholinguistic effect, the word frequency effect (i.e., better lexical decision performance for high-frequency words than for low-frequency words), is reported in Chinese compound word processing (e.g., Zhang & Peng, 1992). Subsequent studies investigated how the first and second characters' frequencies interact with word frequency in lexical decision performance. For example, Peng et al. manipulated both the cumulative character frequency (i.e., the sum of the first and second character frequencies) and word frequency of compound words in a lexical decision experiment. Replicating other studies (e.g., Zhang & Peng, 1992), they obtained a significant word frequency effect. However, this effect was moderated by cumulative character frequency. For high frequency words, participants were faster when the cumulative character frequency was higher than when it was lower. In contrast, this simple effect of cumulative character frequency was not observed for low frequency words. The cumulative character frequency × word frequency interaction could suggest that compound word and its constituents (i.e., characters) might be represented at the same level, which is inconsistent with the premise of some multilevel interactive models (e.g., Taft, 1994; Zhou & Marslen-Wilson, 1995). However, in Peng et al., this interaction was only marginally significant in by-participant analyses and not significant in by-item analyses, so using this as evidence for the view that both compound word and characters are represented at the same level is at best equivocal (see Taft, 2004, for a related discussion of polymorphemic words in English). Further complicating the issue, other patterns of character frequency and word frequency interaction have also been reported in the literature (e.g., Taft, Huang, & Zhu, 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).

Semantic transparency

Semantic transparency is defined by whether compound words are semantically related (transparent, e.g., 黑板 black-board [blackboard]) or unrelated to their constituents (opaque, e.g., 東西 east-west [thing]). It can be quantified by participants' ratings for the semantic relatedness between a compound word and its constituents (e.g., Mok, 2009). Transparent words are recognized more quickly than opaque words (e.g., Lu, Bates, Hung et al., 2001; Myers, Libben, & Derwing, 2004a; but see e.g., Frisson, Niswander-Klement, & Pollatsek, 2008; Myers, Derwing, & Libben, 2004b). The evidence for this semantic transparency effect in Chinese has been mixed. Whereas some studies showed that transparent words are recognized more quickly than opaque words (e.g., Myers et al., 2004a; Su, 1998), others did not find any difference between transparent and opaque words in lexical decision reaction times (RTs) (e.g., Myers et al., 2004b). Holding word frequency constant, Peng et al. (1999; see also Wang & Peng, 1999) manipulated semantic transparency and cumulative character frequencies of compound words and investigated whether compound words and their characters were represented at the same level in Chinese mental lexicon. They obtained a significant cumulative character frequency × semantic transparency interaction. For transparent words, participants were significantly faster when cumulative character frequency was higher than when it was lower. In contrast, this simple effect of cumulative character frequency was reversed for opaque words (despite being marginally significant in item analyses). However, this result could not be fully replicated in other studies (e.g., Gao & Gao, 2005), thereby questioning the generalizability of these findings.

Phonological consistency

Phonological consistency is defined by whether a character has one (i.e., phonologically consistent) or more than one (i.e., phonologically inconsistent) pronunciation (e.g., Tan & Perfetti, 1999).[1] In lexical decisions, words with phonologically inconsistent characters (e.g., 曾 in 曾經 [already] is pronounced as *cang4* but it can also be pronounced as *zang1* when 曾 is used as a surname) are responded faster than those with phonologically consistent characters (e.g., both characters in 頭 [bone] are always pronounced as *gwat1* and *tau4*). This could be because inconsistent characters receive diffused phonological activation across characters that share the same pronunciation (i.e., other homophones) but consistent characters do not. The presence of this phonological consistency effect could provide evidence for the role of phonology at the character level in the recognition of compound words. Leong and Cheng (2003) manipulated the phonological consistency of the first or second character of compound words and found that words with phonologically consistent second characters yielded slower lexical decision RTs than those with phonologically inconsistent second characters. However, this effect did not occur when words with a phonologically inconsistent first character were compared with those with a phonologically consistent first character. The position-specificity of the phonological consistency effect cannot be straightforwardly explained without imposing additional assumptions, such as the serial processing of compound words. Nevertheless, Leong and Cheng argued that the significant phonological consistency effect *per se* could provide evidence for

[1] This definition is different from the one proposed by Fang, Horng, and Tzeng (1986), which defined consistency at the sub-character level in accord to whether a character is pronounced in the same way as its phonetic radical.

the role of character-level phonological activation in the recognition of compound words.

Potential problems associated with the use of factorial designs

The ambiguities of the findings in previous Chinese compound word processing studies (see Myers, 2006, for a more comprehensive review) might be attributed to the fact that all of them were based on factorial-design experiments, wherein a relatively small set of stimuli constrained to be matched on various lexical characteristics were used. There are at least two concerns associated with the use of factorial design, and these concerns can be mitigated by an alternative strategy, the megastudy approach (see Balota, Yap, Hutchison, & Cortese, 2013, for a review of this approach).

First, it is difficult to select two sets of words that vary on only one binary dimension as many lexical variables are correlated, e.g., to compare the RTs of high- vs. low-frequency words, one needs to match the two sets of words on other variables like the first and second characters' frequency, as these could also affect RT. The need to match variables has compelled researchers to use undesirably small sets of stimuli, thus potentially limiting the generalizability of their results. This problem can become more severe when the researchers intend to test higher-order interactive effects.

Second, most lexical variables are continuous, not categorical. Setting up a categorical boundary for a lexical variable may reduce the statistical power of the data analyses (Maxwell & Delaney, 1993) and fail to capture the effect of a lexical variable across its full range of values (Baayen, 2004). Researchers also tend to use items with extreme values on the targeted variable (e.g., high- vs. low-frequency words), so participants could be subtly aware of the manipulation, leading to a potential problem of demand characteristics. Moreover, Forster (2000) has argued that researchers might (implicitly) choose words that may likely yield their expected outcome (see also Kuperman, 2015, for a recent discussion). To reduce this experimenter bias, he recommended that researchers randomly draw stimuli from a larger pool of words with targeted lexical characteristics. Echoing this recommendation, by using a bootstrapping approach (i.e., randomly sampling stimuli with properties that vary in critical linguistic variables from megastudy databases), Kuperman investigated the relationship between psychological valence (positivity) of a word and recognition performance of that word, as normed in English Lexicon Project and British Lexicon Project, and tested whether the conclusion drawn based upon limited stimulus sets in previous studies would hold when a much larger sample of stimuli is used. Clearly, doing so requires a comprehensive database that provides information on various lexical variables for widely known words in a language.

Megastudy approach as an alternative to factorial-design experiments

To address the concerns surrounding the use of factorial design, Balota et al. (2007) examined English word recognition in a large-scaled norming study (megastudy) where RT and accuracy rate were compiled from > 1,200 university students in lexical decision and naming tasks on > 40,000 English words (see also Chang, Hsu, Tsai, Chen, & Lee, 2016, for traditional Chinese character naming; Ferrand et al., 2010, for French; Keuleers, Diependaele, & Brysbaert, 2010, for Dutch; Keuleers, Lacey, Rastle, & Brysbaert, 2012, for British English; Liu, Shu, & Li, 2007 for simplified Chinese character naming; Sze, Rickard Liow, & Yap, 2014 for simplified Chinese character lexical decision; Yap, Rickard Liow, Jalil, & Faizal, 2010, for Malay). This megastudy approach has also been used in other experimental paradigms, such as recognition memory (Cortese, Khanna, & Hacker, 2010), semantic priming (Hutchison et al., 2013), and masked priming (Adelman et al., 2014). Performing item-level analyses on these normative data could reduce idiosyncratic effects in word selection, reveal the proportion of unique variance of performance that a lexical variable, whether continuous or categorical, explains after statistically controlling for the effect of other variables, and indicate the relative contribution of these variables in lexical processing (see Sze, Yap, & Rickard Liow, 2015, for a recent example). However, to our knowledge, no megastudy has been reported on two-character compound words, the most common type of word encountered in Chinese reading (Institute of Language Teaching and Research, 1986).

To address this important gap, in the current megastudy we followed Balota et al.'s (2007) procedure and developed a database for descriptive statistics of lexical variables (see Table 1) and lexical decision behavioral measures for more than 25,000 two-character Chinese compound words (averaged across participants). By providing norms for this large pool of words, we hope to cover most if not all two-character words typically known and used by educated young adults.

We focus on Cantonese-speaking young adults and words printed in traditional Chinese characters, which are most commonly used in Hong Kong, the site of the present megastudy. Whereas traditional characters are much more often used in Hong Kong and Taiwan, simplified characters are used more frequently in mainland China (see Lam, 2003, for an overview of the character simplification scheme). In most cases, the meanings of a traditional and a simplified Chinese character share a one-to-one correspondence; that is, two characters share the same meaning but are written in different scripts.[2]

[2] The pronunciations of traditional and simplified characters are independent of the scripts they are written in: speakers read aloud both traditional and simplified characters with the same pronunciation in Cantonese or in Mandarin.

Table 1 Variables listed in the Excel .xlsx file

For Sheet “Word”		
Column	Variable name	Definition
1	Word_Trad	Chinese word in traditional characters
2	Word_Sim	Chinese word in simplified characters
3	Ntrials	Number of participants whose trials were sufficiently reliable to provide the reaction time for that item (maximum being 33)
4	Acc	Mean accuracy rate for each word computed across participants
5	RT	Mean reaction time for each word computed across participants
6	RT-SE	Standard error of the reaction time for each word
7	RT-SD	standard deviation of the reaction time for each word
8	zRT	Mean of the standardized reaction time for each word computed across participants
9 and 10	Stroke-1 and Stroke-2	Number of strokes of the first and second characters based on a pocket dictionary (Que, 2008)
11 and 12	C&B-Subtitle-raw-C1 and C&B-Subtitle-raw-C2	Character frequency of the first and second characters based on Cai and Brysbaert’s (2010) raw subtitle character frequency count
13 and 14	C&B-Subtitle-CD-C1 and C&B-Subtitle-CD-C2	Character frequency of the first and second characters based on Cai and Brysbaert’s (2010) contextual diversity subtitle character frequency count
15 and 16	SS&M-C1 and SS&M-C2	Character frequency of the first and second characters based on Shaoul, Sun, and Ma (2016)
17 and 18	Da-Modern-C1 and Da-Modern-C2	Character frequency of the first and second characters based on Da’s (2004) list of modern Chinese characters
19 and 20	Google-freq-C1 and Google-freq-C2	Character frequency of the first and second characters based on the number of searchable Chinese character entries indexed in Hong Kong traditional Chinese database in Google
21	C&B-Subtitle-raw-W	Word frequency based on Cai and Brysbaert’s (2010) raw subtitle word frequency
22	C&B-Subtitle-CD-W	Word frequency based on Cai and Brysbaert’s (2010) contextual diversity subtitle word frequency
23	SS&M-W	Word frequency based on Shaoul et al. (2016)
24	Da-News-W	Word frequency based on Da’s (2004) list of modern Chinese words based on news sub-corpora
25	Da-Fiction-W	Word frequency based on Da’s (2004) list of modern Chinese words based on fiction sub-corpora
26	Google-freq-W	Word frequency based on the number of searchable Chinese word entries indexed in Hong Kong traditional Chinese database in Google
27 and 28	ST-C1 and ST-C2	Semantic transparency between first/second characters and compound word based on standardized means of participants’ semantic relatedness ratings for the first/second characters in respect to the compound word
For Sheet “Nonword”		
Column	Variable name	Definition
1	Nonword_Trad	Chinese nonword in traditional characters
2	Nonword_Sim	Chinese nonword in simplified characters
3	Ntrials	Number of participants whose trials were sufficiently reliable to provide the reaction time for that item
4	Acc	Mean accuracy rate for each nonword computed across participants
5	RT	Mean reaction time for each nonword computed across participants
6	RT-SE	Standard error of the reaction time for each nonword
7	RT-SD	standard deviation of the reaction time for each nonword
8	zRT	Mean of the standardized reaction time for each nonword computed across participants

For some characters, both traditional and simplified scripts share more visual characteristics (e.g., 記 vs. 记 [remember]), but this is not the case for the others (e.g., 書 vs. 书 [book]). Hence, people who are familiar with one script may not necessarily be able to read the other script. In the literature, compound word processing studies have been conducted using both simplified characters (e.g., Peng et al., 1999) and traditional characters (e.g., Leong & Cheng, 2003). To ensure that the current data reflect the young adults’ lexical processing of the script that is most familiar to them, we presented all stimuli in traditional characters in the current megastudy.

To demonstrate the different applications of this database, we conduct the following analyses. First, we evaluate various word frequency counts (e.g., Cai & Brysbaert’s, 2010, subtitle frequency, see below for their definitions) and determine which would account for the largest proportion of variance in lexical decision performance. Second, we run three virtual studies to test whether our normative data could replicate previous lexical decision findings in the compound word processing literature. The effects to be replicated, which were briefly introduced in the “Compound Word Processing in Chinese” section, are the interaction between cumulative character frequency and word frequency (Peng et al., 1999), the interaction between cumulative character frequency and semantic transparency (Peng et al., 1999), and the effect of phonological consistency (Leong & Cheng, 2003). Third, we run item-level regression analyses to investigate the basic lexical effects (e.g., word frequency and semantic transparency effects) in compound word processing.

Method

Participants

Following Balota et al.'s (2007) procedure, we divided 25,286 two-character Chinese compound words into 18 lists of 1,404–1,405 words and collected data from 594 (i.e., 18 × 33) right-handed participants in order to obtain about 33 lexical decision responses for each word. This number of observations per word, which is consistent with other megastudies for other languages, allowed us to obtain stable point estimate for each word's mean RTs. The participants were all native Cantonese-speaking students from the Chinese University of Hong Kong (CUHK), with self-reported normal or corrected-to-normal vision. All reported Cantonese as their first language. Data from 26 additional participants were replaced due to their failure to participate in all experimental sessions (*N* = 8) or showing overall accuracy rate that was lower than 70 % (*N* = 18). Stimuli were all traditional Chinese characters typically used in Hong Kong and should thus be familiar to participants. Each participant made lexical decision responses for 467–468 words and 467–468 nonwords in each of the three experimental sessions. The presentation order of word/nonword sets in the three sessions was counterbalanced across participants. Each participant was paid HK$150 (~US$19) for his/her voluntary participation. All participants were asked to report their gender (428 female), age (*M* = 19.83, *SD* = 1.50), amount of time they read Chinese materials per week (*M* = 178.96 min, *SD* = 320.31), cumulative grade-point average (*M* = 4.72, *SD* = .72),[3] self-rated knowledge in 7-point scales for traditional characters (*M* = 5.65, *SD* = 1.05) and spoken Cantonese (*M* = 6.11, *SD* = .94), and grades in Chinese language university entrance exam (*M* = 4.85, *SD* = 1.07).[4] Informed consent was obtained at the beginning of the study.

Materials and design

We first identified about 49,360 two-character Chinese compound words from a pocket dictionary (Que, 2008) and from the suggested word pool in Microsoft Word database. Then, we eliminated the words that (a) are not familiar to the first and third authors (both of whom are native Cantonese speakers), (b) are very rarely used alone as they are part of the words with more characters (e.g., 衞道 in 衞道之士 [people who defend traditional moral principles]), or (c) are proper names for a person or a place (e.g., 合肥 [Hefei, a city in China]). We also recruited four raters, CUHK undergraduate students, to rate the familiarity of the remaining words on a 7-point scale, with 7 being most familiar. After several rounds of checking and based on the mean familiarity ratings (Cronbach alpha = .69), we selected 25,286 words to serve as experimental stimuli.

The frequency counts for Chinese words and characters were based on the six frequency measures (Table 1, see the details in the "Comparison of word frequency measures" section). To quantify semantic transparency, we used Mok's (2009) procedure to conduct a norming study to estimate the semantic relatedness between the first/second character and the compound word for all 25,286 words. The words were divided into 18 sets of 1,404–1,405 words. To obtain 20 pairs of semantic relatedness judgments for each word (one for first character-word and one for second character-word) in these 18 sets, we recruited 360 (i.e., 20 × 18) raters, who were from the same population as those in the lexical decision task. None of these raters participated in the lexical decision task. Each rater made self-paced semantic relatedness judgments on a 7-point scale, with 7 = most related (transparent) for 1,404–1,405 words in two sessions. Each set was randomly divided into two lists: A and B. In the first session, the raters gave the judgments for all first character-word pairs in list A and for all second character-word pairs in list B. In the second session, the raters gave the judgments for all second character-word pairs in list A and for all first character-word pairs in list B. This assignment was counterbalanced between the raters. Within each session, the order of character-word pairs was randomized anew for each rater, who was told to distribute the ratings on the scale as widely as possible. Cronbach alphas ranged from .76 to .86 across sets (*M* = .82, *SD* = .03), indicating the moderate-to-high inter-rater reliability of semantic relatedness ratings. The semantic transparency rating of each rater was standardized before averaging to take into account the individual variability in the raters' responses. All values of the six frequency measures and semantic transparency for the words, among other variables (see Table 1), are listed in the .xlsx file appended to this article.

For the lexical decision task, words were randomly divided into 18 sets (about 1,404–1,405 words per set), with a constraint that the proportion of the words with a specific first character in each subset roughly reflects the proportion of

[3] We asked participants to choose one of the six options (3.51–4.00, 3.01–3.50…, 1.00–1.50) to indicate their cumulative grade point average and we converted their option into 6-point scales with 6 = 3.51–4.00 (i.e., the highest grade point average). The mean and *SD* reported here are based on 522 participants who were not first-year students.

[4] Prior to 2012, Hong Kong students needed to take both Hong Kong Certificate of Education Examination (HKCEE) after the completion of Form 5 (about Grade 11) and Hong Kong Advanced Level Examination (HKALE) after the completion of Form 7 (about Grade 13) for university entrance. In these examinations, students' performance was classified by letter grades (A–E, and two failing grades, F and U). Beginning from 2012, Hong Kong students need to take only one examination (Hong Kong Diploma of Secondary Education Examination, HKDSE) for university entrance, in which their performance is classified by 7-point scales (5**, 5*, 5, 4, 3, 2, and 1, with the last two being the failing grades). Since our participants consist of both groups of Hong Kong students, for those who enrolled prior to 2012 we converted their Chinese Language grades into 7-point scales (i.e., A = 7, B = 6…, and U = 1) in HKCEE and in HKALE and averaged them to be one score. For those who enrolled on or after 2012, we converted their top two grades 5** and 5* in Chinese Language into 7 and 6, respectively, such that their university entrance exam score could be compared with those who enrolled prior to 2012.

the words with that specific first character in the overall word pool. For each set, we created a nonword set by recombining characters randomly. This ensured that words' and nonwords' character-level lexical variables (e.g., number of stroke and character frequency) were matched. All these nonwords were checked against a large Internet-based Chinese word corpus (The Tübingen Corpus of Simplified Chinese, TüCoSiC, Shaoul et al., 2016) to make sure that they did not inadvertently form existing words. While a few nonwords may still be orthographically similar to real words, their influence on our findings should be negligible. Indeed, the nonwords in English Lexicon Project (Balota et al., 2007) were pronounceable and created by changing 1–2 letters in a target word, so some of them, especially the long ones, were also orthographically similar to real words. To minimize the repetition of the characters, we presented a different set of nonwords with the word set for each participant. For instance, nonwords that were created from the words of Set 8 were grouped with words of Set 12 within the experiment. All nonwords[5] and participants' performance on them are listed in the .xlsx file appended to this article.

Procedure

Each participant was tested in a quiet cubicle in three sessions separated by no more than a week. The presentation order of four stimuli blocks in the three sessions was freshly randomized for each participant. PC-compatible computers with E-Prime 2.0 (Schneider, Eschman, & Zuccolotto, 2001) were used to display the stimuli and collect RT and accuracy data. Stimuli were in white and were visually presented on a black background, one at a time, at the center of the screen. In each session, 468 (or 469) experimental words and 468 (or 469) experimental nonwords were randomly presented to the participants. Participants typically took between 35 to 50 min to complete each session.

Each trial started with a 500-ms fixation point ***, followed by a 120-ms blank-screen interstimulus interval. Two Chinese characters appeared next to each other without any space in font size 36 and font type 標楷體 [DFKai-SB] were then presented on the screen until the participants responded. Participants decided whether a two-character string formed a Chinese word by pressing L key (word) or A key (nonword) with right and left index fingers as quickly and accurately as possible. They were presented a 250-ms blank-screen for their correct response or a 250-ms auditory feedback (Utopia Default.wav) for their incorrect response. There were three self-paced rest breaks distributed evenly in each session. At the beginning of a session, participants were familiarized with the task demands by responding to five practice words and five practice nonwords, which were randomly intermixed. None of the practice words or nonwords overlapped with experimental words or nonwords.

Results and discussion

Only responses for experimental trials were analyzed. Based on 594 participants, the mean accuracy rates for nonwords and words were 89.75 % (*SD* = 6.01 %) and 88.33 % (*SD* = 4.83 %), respectively. RTs from incorrect word responses were excluded in the following RT analyses. Remaining responses that were faster than 200 ms or slower than 3,000 ms were first excluded. The mean and *SD* were then computed for each participant's word responses. Any correct word response above or below 2.5 *SD*[6] from his/her mean was labeled as outlier scores and excluded, following the lead of the Chinese single-character lexical decision megastudy (Sze et al., 2014). The mean RT for the trimmed correct word trials was 656.76 ms (*SD* = 181.48 ms). Following Faust, Balota, Spieler, and Ferraro's (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. The zRT is more reliable than raw RT (e.g., Ferrand et al., 2010; also see below for reliability analyses done on the current data) as standardization controls for differences in overall RT and variability between participants and reduces noise from the data without artificially reducing item variability. The level of significance was set at .05. The effect sizes of statistics were quoted from SPSS output or computed using the method recommended by Lakens (2013).

In line with previous megastudies (e.g., Ferrand et al., 2010; Keuleers et al., 2010), we determined the reliability of our dependent measures (RT, zRT, and accuracy) by computing the split-half correlation and correcting it for length (i.e., about 33 observations per word) using the Spearman-Brown formula $(2 \times r)/(1 + r)$; this indicates the proportion of variance in a variable that can be explained (i.e., the remainder is noise). The corrected correlation (r_{corr}) between the dependent measures computed on the first half ($N = 16$) of participants

[5] Given that we randomly combined the characters of the words to create the nonwords, there were 193 and four nonwords that repeated twice and three times, respectively, across sets. Nevertheless, it should be noted that these repetitions occurred across sets, so that each participant was presented each nonword once in the whole experiment.

[6] Whereas some megastudies used the same criteria for RT trimming as in the current study (e.g., 2.5 *SD* in Malay Lexicon Project, Yap et al., 2010), other used a more lenient criteria (e.g., 3 *SD* in English Lexicon Project, Balota et al., 2007 and French Lexicon Project, Ferrand et al., 2010). To test whether the elimination of long RT outliers might have distorted our current findings, we re-ran the analyses by using a 3 *SD* cutoff. We obtained the same pattern of results as reported in the main text, except that when based on the subset of words that appeared in both Cai and Brysbaert's raw and contextual diversity subtitle and Google measures, Google measure accounted for slightly more variance than Cai and Brysbaert's raw measure (33.26 % vs. 33.08 %, respectively), though more importantly, Cai and Brysbaert's contextual diversity measure still accounted for the largest proportion of variance in zRT (33.45 %).

who saw the word and the dependent measures computed on the second half (N = 17) of participants who saw the word were .78, .84, and .88 for RT, zRT, and accuracy, respectively. The fact that the reliability of zRTs is higher than the reliability of the raw RTs is in line with Faust et al.'s (1999) view that taking away differences in overall RT and variability between participants may remove noise from the data and does not artificially reduce the variability of the words. Following Sze et al. (2014), we excluded words that yielded lower than 70 % accuracy rate (about 9.78 %), so the following analyses are based on 22,808 words. Using http://translate.google.com (Google Translate), we converted our traditional words to simplified words in order to quote the frequency counts from the word frequency measures that were normed based on simplified characters (see below). Table 1 presents all variables listed in the Excel .xlsx file, which are appended to this article for noncommercial use by researchers.[7]

Comparison of word frequency measures

Word frequency has consistently been shown to be one of the strongest predictors of visual word recognition performance in various languages. Given that this is a central lexical variable to manipulate and control, it is important to determine which word frequency count accounts for the largest proportion of variance in lexical decision performance. A number of Chinese word frequency norms are available. We consider six measures that are publicly accessible in the current megastudy (see Table 2).

First, we adopted two measures from Da's (2004) list of news and general fiction modern Chinese bigram sub-corpora based on harvested electronic texts and digitized hard copies from 16 websites published between 1997 and 2003. These are considered representative of informative and imaginative texts in Modern Chinese, with the former containing 14 million simplified Chinese characters and 0.73 million unique bigrams, and the latter, 18 million simplified Chinese characters and 0.97 million unique bigrams. A bigram is defined as a string with two consecutive characters in a text and can be treated as a close approximation to a two-character word in Chinese.

Second, we adopted two measures from Cai and Brysbaert (2010) based on simplified Chinese subtitles from films and TV shows. Their corpus size was based on DVDs and the two largest Chinese websites supplying subtitles in simplified character (46.8 million characters, 33.5 million words). Two word frequency measures are adopted, subtitle (raw) frequency (i.e., how many times a word actually occurs in token counts) and subtitle (contextual diversity) frequency (i.e., the number of films and TV shows a word occurs in).

Third, we obtained the word frequency measure from the TüCoSiC corpus (Shaoul et al., 2016), which is built from a very large set of simplified Chinese web documents that were automatically cleaned to remove formatting and other HTML tags, redundant boilerplate text and non-Chinese text. The final corpus contains 358 billion word tokens after segmentation extracted from over 400 million documents.

We should acknowledge that we did not include the frequency counts of the Dictionary of Modern Chinese Frequency (Institute of Language Teaching and Research, 1986) because it is relatively outdated as compared with the above five word frequency measures. Also, Sze et al. (2014) demonstrated that the character frequency measures based on this Dictionary accounted for the smallest proportion of variance (12.61 %) of lexical decision performance for 1,273 Chinese simplified characters.

All of the aforementioned word frequency counts were based on simplified Chinese texts. As mentioned earlier, traditional, rather than simplified, characters are more often used in Hong Kong. Despite this, previous studies involving participants in Hong Kong often selected their stimuli using these simplified-character-based frequency measures (e.g., Liu & McBride-Chang, 2010). It is not clear whether these frequency measures are representative enough to reflect the actual frequency count of two-character Chinese compound words in traditional characters. To properly evaluate these popular word frequency counts, we also included a traditional-character- based frequency measure. Specifically, we obtained the frequency counts based on the number of searchable entries indexed in Hong Kong traditional Chinese database in Google, arguably the most popular Internet search engine (see the Desktop Search Engine Market Share information in http://www.netmarketshare.com/search-engine-market-share.aspx?qprid=4&qpcustomd=0), during July and August 2015, that is, the time period when we began to collect normative data.[8] The advantage of doing this is that we could restrict the frequency counts to the websites that are published in Hong Kong and mainly consisted of traditional Chinese characters. Nevertheless, it is noteworthy that this direct search from Google online search engine might yield only very approximate counts for Chinese character and word

[7] Five nonwords (泡疹, 蠔鼓, 針炙, 撤謊, and 褫奪) were mistakenly included as words in the experimental programs. Because their lexical decision data may not accurately reflect compound word processing, the lexical variables and behavioral data of these five items are not listed in the .xlsx file. Hence, the total number of words reported in the files is 25,281, including the words that yielded lower than 70 % accuracy.

[8] It is noteworthy that "Google word frequency" was also quantified in other studies using Google Ngram Viewer that includes word frequency estimates based on the gigantic digitized Google Books corpus including millions of books published since 1500 (e.g., Brysbaert et al., 2011a; Brysbaert, Keuleers, & New, 2011b). It is interesting to test whether "Google character and word frequency" based on Google Books corpus for Chinese (see https://books.google.com/ngrams/info for more details) would yield similar results to those reported in the current study.

Table 2 Comparison of the percent variance of lexical decision zRT and accuracy rate (in R^2) explained by six word frequency measures

	zRT (%)	Accuracy (%)
Da-News	17.04	7.06
Da-Fiction	22.87	8.55
SS&M	28.45	12.37
Google-freq	33.71	18.75
C&B-Subtitle-raw	34.37	16.68
C&B-Subtitle-CD	34.72	17.36

These analyses were based on the restricted set of words that had a frequency count specified in all six frequency measures (*N* = 15,759). Da-News = frequency counts based on Da's (2004) list of modern Chinese bigrams based on news sub-corpora. Da-Fiction = frequency counts based on Da's (2004) list of modern Chinese bigrams based on fiction sub-corpora. SS&M = raw frequency based on a large corpus of Simplified Chinese web pages (Shaoul et al., 2016), Google-freq = frequency based on the number of searchable Chinese word entries indexed in Hong Kong traditional Chinese database in Google. C&B-Subtitle-raw = raw frequency based on TV and film subtitles (Cai & Brysbaert, 2010). C&B-Subtitle-CD = contextual diversity based on TV and film subtitles (Cai & Brysbaert, 2010). All frequency measures are log-transformed

frequencies, which likely vary over times. Blair, Urland, and Ma (2002) in their study that involved English stimuli showed that Internet search engines yielded the most representative word frequency estimates. However, they did not validate their counts against behavioral measures. Subsequent studies (e.g., Brysbaert et al., 2011b) showed that Google American English frequencies, as quantified by Google Books corpus, explain less variance in the lexical decision RTs from the English Lexicon Project (Balota et al., 2007) than English subtitle word frequencies, based on a corpus of 51 million words from film and television subtitles (see also Brysbaert et al., 2011a, for similar results in German). However, for two-character Chinese compound words it remains unclear if a frequency measure based on Internet search engines is a better predictor of lexical decision performance, compared to other measures. Hence, in the current megastudy, we explored whether this Google word frequency measure is able to account for more variance of lexical decision performance than the other five measures (see Table 2).

In the following analyses, we compared the frequency counts across six measures (Cai & Brysbaert's, 2010, raw frequencies and contextual diversity from a subtitle corpus; Shaoul et al.'s, 2016, word frequency from a web corpus; Da's, 2004, word frequency based on news and fiction sub-corpora; and Google word frequency measure). Of the 22,808 eligible words, about 15,759 (69.09 %) had a frequency count specified in all six measures. To ensure a fair comparison among frequency counts, we reported analyses based on a restricted set of words that appeared in all six measures. Each frequency count was first log-transformed and centered based on the means of all available items in the measure. We then computed the proportion of variance in lexical decision zRT and accuracy rate accounted for by each frequency measure (see Table 2).

In all six measures, frequency counts based on film and TV subtitles (Cai & Brysbaert, 2010) and those based on a Google corpus accounted for a larger proportion of variance than Da's (2004) and Shaoul et al.'s (2016) frequency measures. In addition, Cai and Brysbaert's measures accounted for slightly more variance in zRT and slightly less variance in accuracy rate than Google measure. However, these analyses were based on the restricted set that consisted of less than 70 % of our full set of compound words. To expand the restricted set, we performed another regression analyses based on the subset of words that appeared in *both* Cai and Brysbaert's raw and contextual diversity subtitle and Google measures (*N* = 18,983, i.e., 83.23 % of 22,808 words). In these analyses, Cai and Brysbaert's raw and contextual diversity measures still accounted for more variance than the Google measure in zRT (33.36 % and 33.76 % vs. 33.28 %, respectively). However, Google measure accounted for more variance than Cai and Brysbaert's raw and contextual diversity measures in accuracy rate (19.78 % vs. 17.07 % and 17.74 %, respectively). To test whether there were any reliable differences in variance explained by the top three frequency measures, we ran a series of hierarchical regression analyses. As shown in Table 3, Cai and Brysbaert's contextual diversity and raw subtitle frequency measures accounted for more variance in lexical decision zRT, above and beyond Google word frequency measure than the other way around. In contrast, Google word frequency measure accounted for more variance in lexical decision accuracy rate, above and beyond Cai and Brysbaert's contextual diversity and raw subtitle frequency measures than the other way around.

Table 3 Comparison of the percent variance of lexical decision zRT and accuracy rate (in R^2) explained by Cai and Brysbaert (2010) and Google word frequency measures in hierarchical regression analyses

Variables entered in:		Change in R^2 in the second step	
The first step	The second step	zRT (%)	Accuracy (%)
C&B-Subtitle-raw	Google-freq	+7.65	+5.71
Google-freq	C&B-Subtitle-raw	+7.73	+3.00
C&B-Subtitle-CD	Google-freq	+7.41	+5.36
Google-freq	C&B-Subtitle-CD	+7.90	+3.32
C&B-Subtitle-raw	C&B-Subtitle-CD	+0.42	+0.76
C&B-Subtitle-CD	C&B-Subtitle-raw	+0.02	+0.09

The analyses were based on the words that had a frequency count specified in both Cai and Brysbaert (2010) and Google word frequency. *N* = 18,983. Google-freq = frequency based on the number of searchable Chinese word entries indexed in Hong Kong traditional Chinese database in Google. C&B-Subtitle-raw = raw frequency based on TV and film subtitles (Cai & Brysbaert, 2010). C&B-Subtitle-CD = contextual diversity based on TV and film subtitles (Cai & Brysbaert, 2010). All frequency measures are log-transformed

Contrary to our prediction, Cai and Brysbaert's contextual diversity measure, despite being based on film and TV program subtitles in simplified characters, may account for the largest proportion of variance in lexical decision performance when words are presented in traditional characters. Similar findings for the robustness of subtitle frequency measure on predicting lexical decision performance were reported in English (Brysbaert et al., 2011b) and German (Brysbaert et al., 2011a). It is noteworthy that the size of word frequency corpus does not necessarily predict the extent to which the corpus predicts lexical decision performance. For example, Cai and Brysbaert's (2010) subtitle measures accounted for a larger proportion of variance in lexical decision zRTs than Shaoul et al.'s (2016) measure (34.37 % vs. 28.45 %, see Table 2) even though the former measure was based on a much smaller corpus than the latter measure (33.5 million vs. 358 billion). Other factors, such as the extent to which the sources of word frequency count may be differentially accessible to the population, should also be considered. In our example, Shaoul et al.'s measure was based on simplified Chinese web documents, which might not be as accessible to our current sample (i.e., university students in Hong Kong) as Cai and Brysbaert's subtitle frequency measures, which were based on the simplified Chinese subtitles from films and TV shows. Similar to this idea, Brysbaert et al. (2011b) showed that frequency measures based on old books and non-fiction books in English are not very representative for the type of language read by university students and thus accounted for less variances than subtitle frequency measures in English. Future studies should further validate this explanation by testing whether corpus accessibility or corpus size play a more important role in determining the predictive power of the frequency measure derived from that corpus.

Replicating previous findings with virtual studies

Given the large number of words being normed for their lexical decision performance in our data, we tested the robustness of some previous lexical decision findings in the literature of compound word processing. In order to examine the validity and generalizability of the data in the current megastudy, we conducted virtual replication studies based on three published experiments discussed in the Introduction. The present megastudy contained normative data for more than 90 % of the original stimuli in these experiments. It is worth noting that the study by Peng et al. (1999) was conducted in simplified characters, whereas the Leong and Cheng (2003) study was conducted in traditional characters. By running virtual studies in these studies, we could test if the findings based on simplified characters extend to our traditional character-based normative data. Table 4 summarizes the comparison of the results of the original and virtual studies, based on our current normative data. Figures 1, 2 and 3 show the patterns of findings in the original and virtual studies.

As mentioned in the above "Character and word frequency" section, Peng et al. (1999, Experiment 1) examined whether compound words and their characters are represented at the same level in Chinese mental lexicon. They only obtained weak support for this by reporting a marginal interaction between word frequency and cumulative character frequency on lexical decision RTs in subject analyses (p = .056). However, this effect was not significant in item analyses. The item analyses of our normative data (see Table 4) reinforce the absence of this interaction in RT: F(1,72) = .30, MSE = 1,942, p = .59, η^2_p = .004, error rate: F(1,72) = .01, MSE = 38.85, p = .93, η^2_p < .001, and zRT: F(1,72) = .71, MSE = .08, p = .40, η^2_p = .01 (see Fig. 1). We also found a significant main effect of word frequency in RT: F(1,72) = 11.71, MSE = 1,942, p < .01, η^2_p = .14 and in zRT: F(1,72) = 9.76, MSE = .08, p < .01, η^2_p = .12, which was consistent with Peng et al.'s findings.

As described in the above "Semantic transparency" section, Peng et al. (1999, Experiment 2) manipulated cumulative character frequency and semantic transparency to investigate whether compound words and their characters are represented at the same level in Chinese mental lexicon. They found an interaction between semantic transparency and cumulative character frequency on lexical decision RTs, with the effect of cumulative character frequency being significantly positive for transparent words (46 ms, p < .01), but marginally negative for opaque words (−24 ms, p = .09). This result provided evidence that compound words and their characters are represented at the same level in Chinese mental lexicon. In our virtual study, our item analyses replicated this interaction effect in RT: F(1,73) = 5.66, MSE = 1,882, p < .05, η^2_p = .07 and in zRT: F(1,73) = 5.65, MSE = .08, p < .05, η^2_p = .07 (see Fig. 2). Similar to Peng et al., the interaction did not occur in error rate in item analyses, F(1,73) = 1.23, MSE = 27.55, p = .27, η^2_p = .02. Follow-up analyses showed that the effect of cumulative character frequency was significant and positive (37 ± 25 ms) for transparent words [RT: t(38) = 3.05, p < .01, Cohen's d = .96; zRT: t(38) = 2.57, p < .05, Cohen's d = .81], but numerically negative (−10 ± 33 ms) for opaque words [RT: t(35) = .64, p = .53; zRT: t(35) = .95, p = .35]. Even though the overall RT was slower in the current normative data than in Peng et al.'s experiments, which could be attributed to sampling difference (e.g., overall, the stimuli might not be as familiar to our current participants as to those in the original study). However, the RT patterns of the two studies were closely matched. Moreover, both RT and zRT findings were highly similar, indicating that the pattern of results remained the same after taking into account the differences in overall RT and variability between participants.

While the replication of the findings of Peng et al.'s (1999) Experiment 2 might suggest that the compound

Table 4 Visual lexical decision findings in original and virtual studies based on the same set of stimuli that are common in both the current megastudy and the respective original study

Original study	Condition and effect (Example in parentheses)	Number of stimuli (Original/ Replication)	RT in original study (ms)	RT in virtual study (ms)	Error rate in original study (%)	Error Rate in Virtual Study (in %)
Peng et al. (1999, Experiment 1)	Low WF-Low CCF (慈悲 [mercy])	20/20	569	641	5.5	5.7
	Low WF-High CCF (死板 [rigid])	20/17	569	631	5.0	5.1
	High WF-Low CCF (介紹 [introduction])	20/19	536	612	2.1	3.5
	High WF-High CCF (青春 [youth])	20/20	516	591	1.4	3.2
	WF × CCF Interaction		20 (ns)	11 (ns)	0.2 (ns)	-0.3 (ns)
Peng et al. (1999, Experiment 2)	Low ST-Low CCF (麻煩 [trouble])	20/19	518	613	2.9	3.2
	Low ST-High CCF (推測 [speculate])	20/18	542	623	5.9	5.7
	High ST-Low CCF (駕駛 [driving])	20/20	554	636	4.1	4.1
	High ST-High CCF (培植 [cultivate])	20/20	508	599	2.1	3.9
	ST × CCF Interaction		71**	47*	5.0 (ns)	2.7 (ns)
Leong & Cheng (2003, Experiment 1 and Experiment 2)	PI (words with the first PI character) (彈簧 [spring])	9/8	1151 (1038)	642	5.3	4.1
	PC (words with the first PC character) (現在 [present])	9/9	1195 (1059)	649	9.0	5.7
	PI (words with the second PI character) (躲藏 [hide])	9/8	1068 (967)	573	3.4	2.3
	PC (words with the second PC character) (寶貴 [valuable])	9/9	1203 (1100)	630	6.1	6.3
	Main effect of PC		90** (77**)	32*	N/A	0.4 (ns)
	Simple effect for PC vs. PI (on the first character)		44 (ns) (21 (ns))	8 (ns)	N/A	0.6 (ns)
	Simple effect for PC vs. PI (on the second character)		135** (133**)	57**	N/A	0.1 (ns)

See the main text for explanation of the findings

RT reaction time, *WF* word frequency, *CCF* cumulative character frequency, *ST* semantic transparency (High-ST = transparent; Low-ST = opaque), *PC* words with a phonological consistent character (control), *PI* words with a phonological inconsistent character, *ns* nonsignificant

$^{**}p < .01$ (two-tailed), $^{*}p < .05$ (two-tailed). All significance levels are based on item analyses

words and their constituent characters are likely represented at the same level in Chinese mental lexicon, the significant cumulative character frequency × semantic transparency interaction is in contrast to the null cumulative character frequency × word frequency interaction. Hence, the issues regarding the character and word representations in Chinese mental lexicon are clearly not straightforward and should be further investigated in future studies (e.g., by exploring other higher-order interactions between character- and word-level lexical variables).

As stated in the above "Phonological consistency" section, Leong and Cheng (2003, Experiment 1) examined the effect of phonological consistency on lexical decision performance (i.e., words with phonologically inconsistent characters were responded faster than those with phonologically consistent characters). We compared the findings of our virtual study with their overall findings (i.e., after collapsing across two groups of participants with different reading comprehension abilities), given that the overall pattern remained the same in their two participant groups. They used the same set of stimuli in both Experiments 1 and 2, which shared very similar experimental paradigms. (The RT results of their Experiment 2 are presented in the parentheses next to those of their Experiment 1 in Table 4. Given that both Experiments showed similar patterns of results, only Experiment 1's data of the original study were plotted in Fig. 3.) They did not report error rate as a function of conditions in Experiment 2 or any statistical analyses for the error rate in either experiment. Leong and Cheng obtained a significant main effect of phonological consistency across two experiments. Subsequent analyses showed that the effect was significant only when the second, but not the first, character was phonologically consistent (vs. inconsistent). The phonological consistency effect suggested an involvement of phonology at the character level in compound word recognition. The results of our virtual study replicated this pattern. The main effect of phonological consistency was significant in RT: $F(3,30) = 3.12$, $MSE = 3{,}101$, $p < .05$, $\eta^2_p = .24$, and marginally so in zRT: $F(3,30) = 2.73$, $MSE = .14$, $p = .06$, $\eta^2_p = .22$. The effect in error rate was not significant, $F(3,30) = .97$, $MSE = 28.55$, $p = .42$, $\eta^2_p = .09$. Follow-up analyses showed that relative to words with a phonologically inconsistent second character, those with a phonologically consistent second character yielded significantly 57 ± 55 ms slower responses ($p < .05$ in both RT and zRT). However, no such difference was seen when words with a phonologically inconsistent first character were compared with those with a phonologically consistent first character (8 ± 55 ms, $p = .78$ and .64 in RT and zRT, respectively). The overall RT was faster in the current normative data than in Leong and Cheng. This could be attributed to the possibility that relative to our megastudy in which participants responded to more than 900 word and nonword trials within a session, far fewer word and nonwords trials ($N = 72$) were presented to participants in Leong and Cheng's study. That is, participants' overall faster RTs in our megastudy might be due to the general practice effect as they are familiar with the task requirement. Hence, even given this potential general

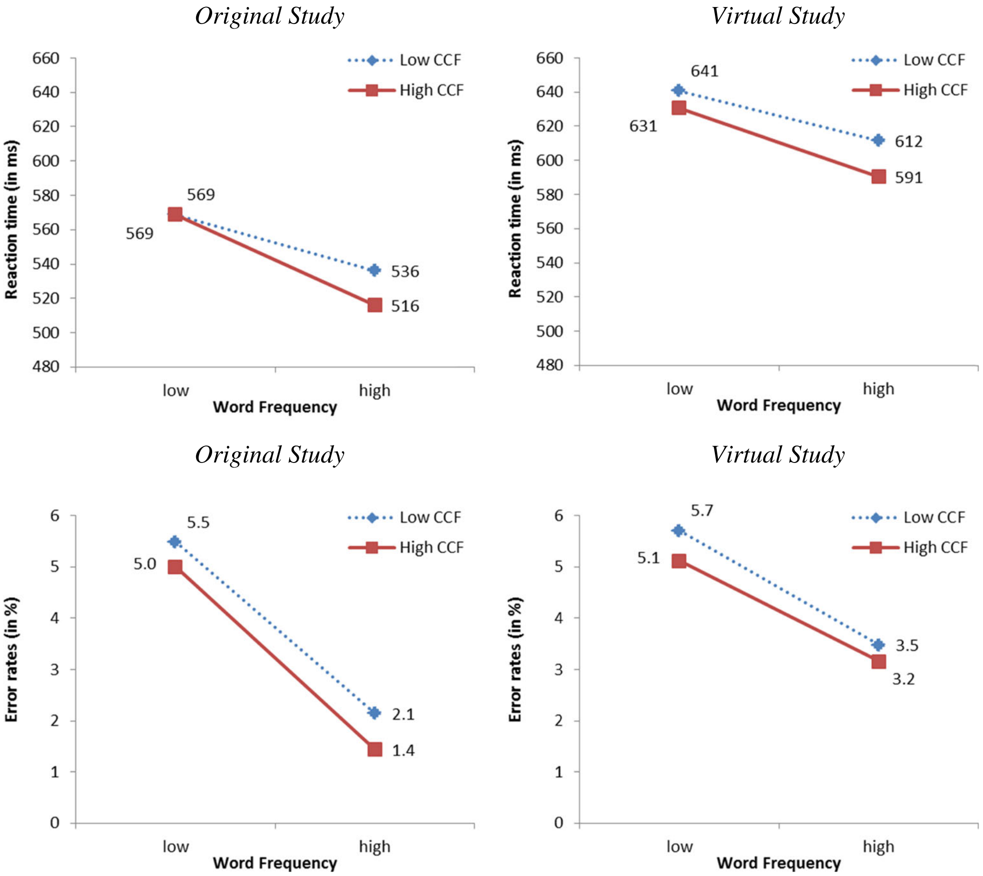

Fig. 1 Comparison between original and virtual studies in Peng et al. (1999, Experiment 1) in reaction time (upper panel) and error rate (lower panel). *CCF* cumulative character frequency

practice effect, we found the RT patterns of the original and virtual studies quite similar. Both RT and zRT findings were highly similar, suggesting that the pattern of results remained the same after taking into account the differences in overall RT and variability between participants. Overall, these indicate a potential role of phonology at the character level in the recognition of compound words.

Overall, the patterns of original and virtual studies are similar, regardless of whether the stimuli of the original studies were in traditional (Leong & Cheng, 2003) or simplified Chinese characters (Peng et al., 1999). This tentatively suggests that the results based on the current normative data, which is based on traditional characters, might be generalized to simplified characters. Nevertheless, this should be further validated in future research by running more virtual studies of previous Chinese compound word lexical decision studies.

Item-level regression analyses

For item-level multiple regression analyses, mean zRT and accuracy rate were the dependent measures. The values of the lexical variables as predictors were all mean-centered to minimize any problem of multicollinearity in regression analyses (Jaccard & Turrisi, 2003). Following previous megastudies, we entered the lexical variables in the following order: the number of strokes for the first and second characters (Step 1), the frequency of the first and second characters and word frequency (Step 2), and semantic transparency of the first and second characters (Step 3). The character and word frequencies were estimated by Cai and Brysbaert's (2010) contextual diversity subtitle measure in these analyses because this measure accounted for the largest proportion of variance among all lexical decision behavioral measures (see Tables 2 and 3). Hence, the sample size was 18,983 in these analyses. There was no multicollinearity problem, as indicated by low variance-inflation factors (all <1.17). Tables 5 and 6 report the descriptive statistics of and inter-correlational matrices among these lexical variables and the three behavioral measures (RT, zRT, and accuracy). The betas (standardized regression coefficients) of these lexical variables when they were first entered in the regression models and the R^2 change (i.e., change in proportion of variance being accounted for) at each step are reported in Table 7.

The number of strokes is often used as an index of visual complexity (e.g., Xing, Shu, & Li, 2004), with the larger

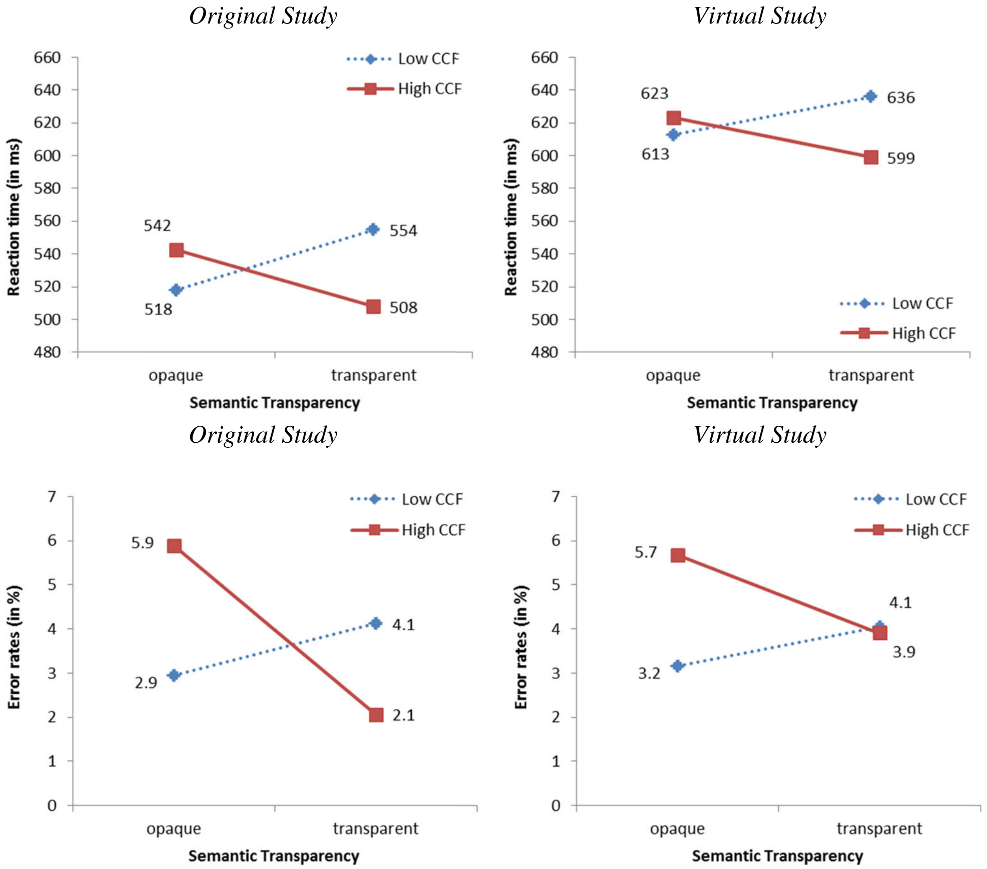

Fig. 2 Comparison between original and virtual studies in Peng et al. (1999, Experiment 2) in reaction time (upper panel) and error rate (lower panel). *CCF* cumulative character frequency

number of strokes that a character has being regarded as visually more complex (e.g., 人 [human] vs. 鬱 [depressed]). Previous studies showed that increasing the number of strokes may slow down character recognition (e.g., Leong, Cheng, & Mulcahy, 1987; see also Sze et al., 2015, for a confirmation of this effect using megastudy data). However, to the best of our knowledge, no published studies have specifically examined the effect of the first and second characters' stroke number on compound word lexical decisions. In our normative data, we obtained the effect of the number of strokes; lexical decision was slower as the number of strokes in both characters increased. This was in line with previous findings of single-character lexical decision studies (e.g., Leong et al.).

We found the effects of character and word frequency, as estimated by Cai and Brysbaert's (2010) contextual diversity subtitle measure; lexical decision was faster and more accurate when both characters were higher in frequency and when the word was higher in frequency, although the beta values for the effect of the second character's frequency on accuracy did not approach significance. Based on the beta values, the first character's frequency was a slightly stronger predictor than the second character's frequency. Moreover, in line with some previous studies (e.g., Myers et al., 2004a; Su, 1998, but see Myers et al., 2004b), we obtained the effect of semantic transparency; lexical decision was faster and more accurate when a compound word was more semantically transparent to both first and second characters. Similar to character frequency, the semantic transparency effect was also slightly stronger for the first character than the second character.

Based on the R^2 changes in each step, character and word frequency, as estimated by Cai and Brysbaert's (2010) contextual diversity subtitle measure, provided the strongest predictive power on lexical decision performance, followed by the stroke number, and semantic transparency accounted for the least variance in our analyses. Overall, these item-level regression analyses on our normative data provide a useful set of empirical benchmarks for the Chinese compound word processing literature.

Conclusion and future directions

Using a megastudy approach, we developed a database of lexical decision RTs and accuracies for more than 25,000 traditional Chinese two-character compound words. To demonstrate its potential applied value, we conducted three analyses. First, by comparing the proportion of variance being accounted for by six word frequency measures, we found that Cai and Brysbaert's (2010) contextual diversity subtitle frequency was

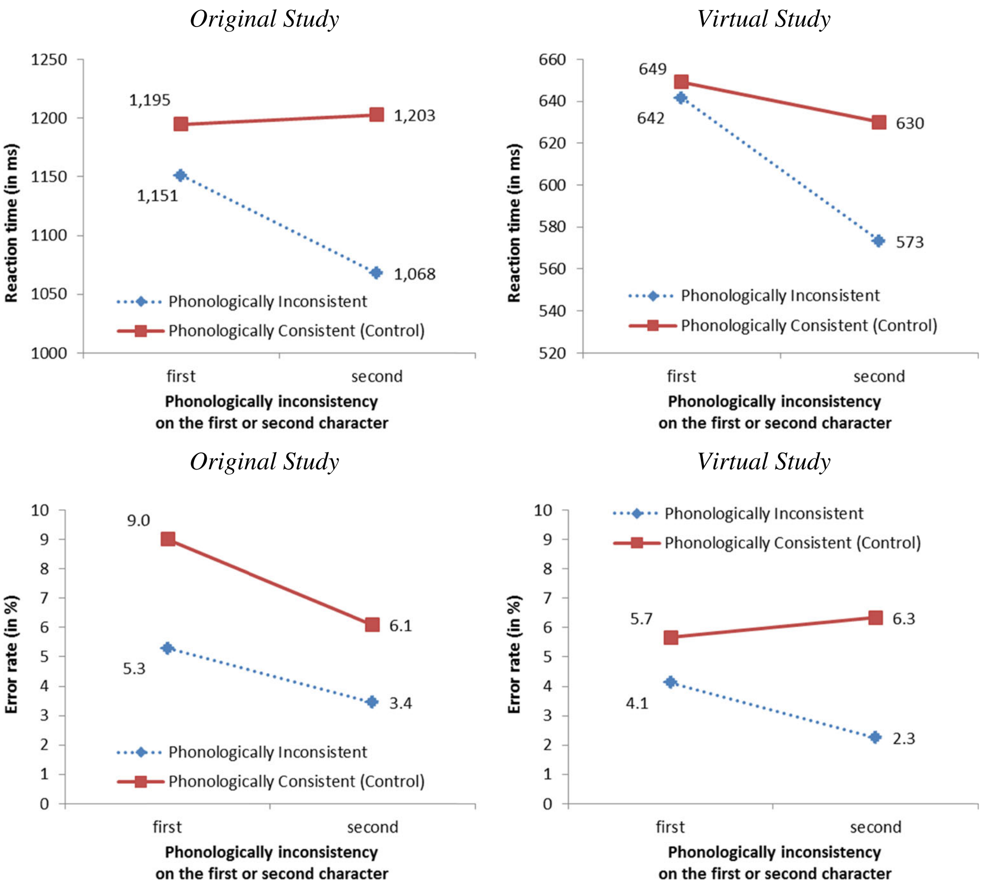

Fig. 3 Comparison between original and virtual studies in Leong and Cheng (2003, Experiment 1) in reaction time (upper panel) and error rate (lower panel)

overall the best predictor for lexical decision performance. Second, we ran virtual studies and successfully replicated the findings reported in three experiments, in which the original stimuli were mostly (>90 %) available in our normative data. Third, in our item-level regression analyses, we obtained the standard lexical effects, such as word frequency and semantic transparency effects, and further found that word frequency was the strongest predictor of lexical decision performance.

In the .xlsx file available online, we make the item-level data available to the research community that allows researchers to search for the normed lexical decision RT, zRT, and accuracy rate for words and descriptive statistics of lexical variables (Table 1) for characters and words. This database serves as a critical resource of lexical characteristics and behavioral measures for future research. For example, researchers could conduct virtual studies to test the replicability of previous studies and performing item-level regression analyses to explore the higher-order interactions among lexical variables (e.g., character frequency × word frequency × semantic transparency interaction), which would be helpful to address some theoretical issues in Chinese compound word processing, such as whether compound words and their characters are represented at the same level in the Chinese mental lexicon. Apart from these, there are some other future directions for the use of this database.

First, researchers can set up search filters to retrieve information for a specific set of words and observe whether a particular effect (e.g., character frequency) would be limited to certain levels of other lexical variables (e.g., word frequency). This would serve as the first step in developing follow-up experiments to further test the joint influence of given variables.

Second, researchers can test the generalizability of some novel effects reported in the literature by referring to the normative data, which are based on a larger word pool in the database. For example, by performing analyses on normative data of a much larger word pool in Balota et al.'s (2007) English Lexicon Project, Kang, Yap, Tse, and Kurby (2011) found that the semantic size effect (i.e., faster lexical decision RTs to words when their referents are physically large, e.g., *elephant*, than when they are small, e.g., *needle*) might not generalize beyond the original stimulus set used in earlier studies (see also Juhasz, Yap, Dicke, Taylor, & Gullick, 2011, for another example).

Third, researchers can add new variables, such as headedness of compound words (e.g., e.g., 海報 [poster] is a right-headed word, but 神經 [nerve] is not) and within-word family size of the characters (i.e., the number of words that a

Table 5 Descriptive statistics of the lexical variables and three behavioral measures (RT, zRT, and accuracy) involved in item-level regression analyses

	Mean	*SD*	Range
Accuracy	.93	.07	.70–1.00
RT	646.18	63.26	482.13–952.33
zRT	-.065	.386	−1.04–1.75
No. of strokes (first character)	10.64	4.44	1–30
No. of strokes (second character)	10.63	4.49	1–33
Log frequency (first character)	3.35	.51	.00–3.80
Log frequency (second character)	3.41	.48	.00–3.80
Log frequency (word)	1.63	.80	.00–3.80
Semantic transparency (first character)	.0019	.49	−2.14–1.61
Semantic transparency (second character)	-.0006	.52	−2.24–1.81

The character and word frequencies were estimated by Cai and Brysbaert's (2010) contextual diversity subtitle measure

RT reaction time

character can occur in the first vs. second positions) to our dataset to make it more comprehensive. For example, Yarkoni, Balota, and Yap (2008) validated their Levenshtein distance measure (a new measure of orthographic similarity) by using the normative data of Balota et al.'s (2007) English Lexicon Project, and this measure was subsequently added to the English Lexicon Project website (http://elexicon.wustl.edu).

Fourth, previous research has often focused on linear relationships between lexical variables and RTs, but a few studies have reported potential non-linear effects of some variables. For instance, New, Ferrand, Pallier, and Brysbaert (2006) found a U-shaped effect of word length on English lexical decision (facilitative for words of 3–5 letters, null for words of 6–8 letters, and inhibitory for words of 9–13 letters), contradicting the view that longer words always take more time to process. By conducting item-level multiple regression analyses on lexical decision performance based on a larger pool of words, the linear and/or nonlinear of effect of targeted variables would be captured more easily, which could not be observed in factorial-design experiments where variables are dichotomized.

Fifth, while the virtual studies based on our normative data did replicate two experiments (Peng et al., 1999), which were originally conducted in simplified characters, the overall RT difference between original and virtual studies might potentially cloud the interpretation. Using the current normative data, more virtual studies should be conducted to further test the extent to which empirical effects in simplified character processing generalize to traditional character processing. Indeed, it will be particularly interesting to conduct a parallel megastudy that involves stimuli in simplified characters and participants who are familiar with this script (e.g., young adults in mainland China). Given that very few studies have examined the difference in processing traditional versus simplified Chinese characters (see,

Table 6 Inter-correlational matrices among these non-centered lexical variables and the three behavioral measures (RT, zRT, and accuracy) involved in item-level regression analyses

		1	2	3	4	5	6	7	8	9
1	Accuracy									
2	RT	−.65**								
3	zRT	−.66**	.98**							
4	No. of strokes (first character)	.01	.08**	.09**						
5	No. of strokes (second character)	.005	.09**	.09**	.02**					
6	Log frequency (first character)	.12**	−.23**	−.24**	−.27**	−.02**				
7	Log frequency (second character)	.11**	−.22**	−.23**	−.01	−.27**	.15**			
8	Log frequency (word)	.42**	−.57**	−.58**	−.06**	−.07**	.26**	.24**		
9	Semantic transparency (first character)	.07**	−.03**	−.03**	.21**	−.03**	−.24**	−.01	−.03**	
10	Semantic transparency (second character)	.02**	.04**	.04**	.03**	.22**	−.07**	−.27**	−.14**	.07**

$^{**}p < .01$ (two-tailed). The character and word frequencies were estimated by Cai and Brysbaert's (2010) contextual diversity subtitle measure

RT reaction time

Table 7 Standardized regression coefficients (beta values) on zRT and accuracy rate

	zRT	Accuracy
Step 1		
No. of strokes (first character)	.083**	.011
No. of strokes (second character)	.085**	.004
R^2 change	*.015***	*.0001*
Step 2		
Log frequency (first character)	−.081**	.025**
Log frequency (second character)	−.078**	.013
Log frequency (word)	−.538**	.417**
R^2 change	*.340***	*.181***
Step 3		
Semantic transparency (first character)	−.069**	.078**
Semantic transparency (second character)	−.064**	.077**
R^2 change	*.009***	*.012***

**$p < .01$ (two-tailed). The character and word frequencies were estimated by Cai and Brysbaert's (2010) contextual diversity subtitle measure

RT reaction time

e.g., Chung & Leung, 2008; Lam, 2003, for exceptions although they did not include lexical decision task as their behavioral measures), this parallel megastudy would allow the effect of lexical variables on lexical decision performance to be directly compared between traditional and simplified characters in Chinese.

Sixth, it is possible to explore the link between the visual recognition of characters (Sze et al., 2014) and the visual recognition of two-character compound words (the current megastudy). However, as mentioned in the last point, the differences in stimuli (simplified characters in Sze et al. vs. traditional characters in the current megastudy) and population (Mandarin speakers in Sze et al. vs. Cantonese speakers in the current megastudy) might make it difficult to directly compare the data between two megastudies. Hence, it is important to develop a single-character lexical decision megastudy in traditional characters to explore the association between single- and two-character visual word recognition in Chinese.

Seventh, by comparing the data in this database with those in other Lexicon Projects (e.g., English in Balota et al., 2007 and French in Ferrand et al., 2010), future research could test cross-linguistic differences in compound word processing and help identify language-specific versus language-universal processes in lexical processing. For example, because the meaning is often less well specified for Chinese characters than for English words (Hoosain, 1992), it is possible that participants rely more on word-level information, rather than character-/constituent-level information, in processing Chinese words, as compared to English words.

Eighth, the large sample size involved in the current megastudy could provide sufficient number of observations that allows an investigation of the intraindividual stability of lexical processing at different points in time (e.g., each participant performed in three experimental sessions, separated by a 24-hour to a 1-week interval). Using the English Lexicon Project trial-level data, Yap, Balota, Sibley, and Ratcliff (2012) found that the participants were quite consistent in their performance across two experimental sessions in both mean RTs and RT distributional parameters (e.g., ex-Gaussian estimates, see Balota & Yap, 2011). Also, this intraindividual consistency was observed in the effect of lexical variables such as word frequency. In exploring the individual difference markers in lexical decision performance, Yap et al. showed that participants with a higher vocabulary age generally produced faster RT and more accurate word recognition performance and attenuated sensitivity to lexical variables. All these questions can potentially be explored in the current normative dataset.

Finally, the behavioral measures in the current megastudy were based on lexical decision performance. It is important to run analogous megastudy using naming RT and accuracy as behavioral measures, similar to other Lexicon Projects (e.g., English in Balota et al., 2007). This would allow a comparison of the effects of lexical variables on participants' lexical decision and naming performance, which in turn demonstrates whether the lexical effects obtained in the current normative data would be specific to lexical decision or task-general phenomena.

In summary, the current megastudy represents another important addition to the currently existing Lexicon Projects (e.g., English, Dutch, French, and Malay). Given that there are more native speakers of Chinese than any other language and two-character compound words are the most common type of word encountered in Chinese reading, a two-character Chinese compound word lexical decision database will significantly contribute to research in psycholinguistics and other research domains using Chinese word stimuli.

Acknowledgments This research was supported by General Research Fund (#451113) awarded by the Hong Kong Research Grants Council. We thank Marc Brysbaert and an anonymous reviewer for their comments on an earlier version of this manuscript and Woody On-Yiu Wong for compiling an Excel script for extracting Google character and word frequency counts.

References

Adelman, J. S., Johnson, R. L., McCormick, S. F., McKague, M., Kinoshita, S., Bowers, J. S., … Davis, C. J. (2014). A behavioral database for masked form priming. *Behavior Research Methods, 46,* 1052–1067.

Baayen, R. H. (2004). Statistics in psycholinguistics: A critique of some current gold standards. *Mental Lexicon Working Papers, 1,* 1–47.

Balota, D. A., & Yap, M. J. (2011). Moving beyond the mean in studies of mental chronometry: The power of response time distributional analyses. *Psychological Science, 20,* 160–166.

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 volume 1: Models and methods, orthography and phonology* (pp. 90–115). New York, NY: Psychology Press.

Blair, I. V., Urland, G. R., & Ma, J. E. (2002). Using Internet search engines to estimate word frequency. *Behavior Research Methods, Instruments, & Computers, 34,* 286–290.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology, 58,* 412–424.

Brysbaert, M., Keuleers, E., & New, B. (2011). Assessing the usefulness of Google Books' word frequencies for psycholinguistic research on word processing. *Frontiers in Language Sciences, 2,* 27. doi:10.3389/fpsyg.2011.00027

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

Chang, Y.-N., Hsu, C.-H., Tsai, J.-L., Chen, C.-L., & Lee, C.-Y. (2016). A psycholinguistic database for traditional Chinese character naming. *Behavior Research Methods, 48,* 112–122.

Chung, F. H.-K., & Leung, M.-T. (2008). Data analysis of Chinese characters in primary school corpora of Hong Kong and mainland China: Preliminary theoretical interpretations. *Clinical Linguistics & Phonetics, 22,* 379–389.

Cortese, M. J., Khanna, M. M., & Hacker, S. (2010). Recognition memory for 2,578 monosyllabic words. *Memory, 18,* 595–609.

Da, J. (2004). A corpus-based study of character and bigram frequencies in Chinese e-texts and its implications for Chinese language instruction. In P. Zhang, T. Xie, & J. Xu (Eds.), *Proceedings of the 4th International Conference on New Technologies in Teaching and Learning Chinese: The studies on the theory and methodology of the digitized Chinese teaching to foreigners* (pp. 501-511). Beijing: The Tsinghua University Press.

Dressler, W. U. (2006). Compound types. In G. Libben & G. Jarema (Eds.), *The representation and processing of compound words* (pp. 23–44). New York: Oxford University Press.

Fang, 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.

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.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., … Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods, 42*, 488–496.

Forster, K. I. (2000). The potential for experimenter bias effects in word recognition experiments. *Memory & Cognition, 28,* 1109–1115.

Frisson, S., Niswander-Klement, E., & Pollatsek, A. (2008). The role of semantic transparency in 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).

Hoosain, R. (1992). Psychological reality of the word in Chinese. In H. C. Chen & O. J. L. Tzeng (Eds.), *Language processing in Chinese* (pp. 111–130). Amsterdam: Elsevier Science.

Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C-S., … Buchanan, E. (2013). The Semantic Priming Project. *Behavior Research Methods, 45*, 1099-1114.

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

Jaccard, J., & Turrisi, R. (2003). *Interaction effects in multiple regression*. Newbury Park: Sage.

Juhasz, B. J., Yap, M. J., Dicke, J., Taylor, S. C., & Gullick, M. M. (2011). Tangible words are recognized faster: The grounding of meaning in sensory and perceptual systems. *Quarterly Journal of Experimental Psychology, 64,* 1683–1691.

Kang, S. H. K., Yap, M. J., Tse, C.-S., & Kurby, C. A. (2011). Semantic size does not matter: "Bigger" words are not recognised faster. *Quarterly Journal of Experimental Psychology, 64,* 1041–1047.

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono- and disyllabic words and nonwords. *Frontiers in Psychology, 1,* 174. doi:10.3389/fpsyg.2010.00174

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods, 44,* 287–304.

Kuperman, V. (2015). Virtual experiments in megastudies: A case study of language and emotion. *Quarterly Journal of Experimental Psychology, 68,* 1693–1710.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t tests and ANOVAs. *Frontiers in Psychology, 4,* 863. doi:10.3389/fpsyg.2013.00863

Lam, A. S. L. (2003). Biscriptal reading in Chinese. In H. S. R. Kao, C. K. Leong, & D. G. Gao (Eds.), *Cognitive and neuroscience studies of the Chinese language* (pp. 247–262). Hong Kong, China: Hong Kong University 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.

Leong, C.-K., Cheng, P.-W., & Mulcahy, R. (1987). Automatic processing of morphemic orthography by mature readers. *Language and Speech, 30,* 181–196.

Liu, P. D., & McBride-Chang, C. (2010). Morphological processing of Chinese compounds from a grammatical view. *Applied Psycholinguistics, 31,* 605–617.

Liu, Y., Shu, H., & Li, P. (2007). Word naming and psycholinguistic norms: Chinese. *Behavior Research Methods, 39,* 192–198.

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.

Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin, 113,* 181–190.

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: 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. (2004a). *The nature of transparency effects in Chinese compound processing*. Poster presented at the Fourth International Conference on the Mental Lexicon, Windsor, Canada.

New, B., Ferrand, L., Pallier, C., & Brysbaert, M. (2006). Re-examining word length effects in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review, 13,* 45–52.

Packard, J. L. (2000). *The morphology of Chinese: A linguistic and cognitive approach*. Cambridge, UK: Cambridge University Press.

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). NJ: Lawrence Erlbaum Associates.

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

Schneider, W., Eschman, A., & Zuccolotto, A. (2001). *E-prime user's guide*. Pittsburgh, PA: Psychology Software Tools.

Shaoul, C., Sun, C., & Ma, J. (2016) TüCoSiC: The Tübingen Corpus of Simplified Chinese. Downloaded from http://shaoul.org/TuCoSiC/

Su, Y.-C. (1998). The representation of compounds and phrases in the mental lexicon: Evidence from Chinese. *Web Journal of Modern Language Linguistics, 1*, http://wjmll.ncl.ac.uk/issue04-05/su.htm

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. *Quarterly Journal of Experimental Psychology, 68,* 1541–1570.

Taft, M. (1994). Interactive-activation as a framework for understanding morphological processing. *Language and Cognitive Processes, 9,* 271–294.

Taft, M. (2004). Morphological decomposition and the reverse base frequency effect. *Quarterly Journal of Experimental Psychology, 57,* 745–765.

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* (Vol. 1). 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). NJ: Lawrence Erlbaum Associates.

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.

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.

Yap, M. J., Balota, D. A., Sibley, D. E., & Ratcliff, R. (2012). Individual differences in visual word recognition: Insights from the English Lexicon Project. *Journal of Experimental Psychology: Human Perception and Performance, 38,* 53–79.

Yap, M. J., Rickard Liow, S. J., Jalil, S. B., & Faizal, S. S. B. (2010). The Malay Lexicon Project: A database of lexical statistics for 9,592 words. *Behavior Research Methods, 42,* 992–1003.

Yarkoni, T., Balota, D. A., & Yap, M. J. (2008). Beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review, 15,* 971–979.

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

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

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