Behav Res
DOI 10.3758/s13428-013-0360-z

Predicting raters' transparency judgments of English and Chinese morphological constituents using latent semantic analysis

Hsueh-Cheng Wang · Li-Chuan Hsu · Yi-Min Tien · Marc Pomplun

Abstract The morphological constituents of English compounds (e.g., "butter" and "fly" for "butterfly") and two-character Chinese compounds may differ in meaning from the whole word. Subjective differences and ambiguity of transparency make judgments difficult, and a computational alternative based on a general model might be a way to average across subjective differences. In the present study, we propose two approaches based on latent semantic analysis (Landauer & Dumais in Psychological Review 104:211–240, 1997): Model 1 compares the semantic similarity between a compound word and each of its constituents, and Model 2 derives the dominant meaning of a constituent from a clustering analysis of morphological family members (e.g., "butterfingers" or "buttermilk" for "butter"). The proposed models successfully predicted participants' transparency ratings, and we recommend that experimenters use Model 1 for English compounds and Model 2 for Chinese compounds, on the basis of differences in raters' morphological processing in the different writing systems. The dominance of lexical meaning, semantic transparency, and the average similarity between all pairs within a morphological family are provided, and practical applications for future studies are discussed.

Keywords Semantic transparency · Latent semantic analysis · Chinese · Compound words · Morphological family · Semantic consistency · Clustering

H.-C. Wang (✉) · M. Pomplun
Department of Computer Science, University of Massachusetts Boston, Boston, MA, USA
e-mail: hchengwang@gmail.com

L.-C. Hsu
School of Medicine and Graduate Institute of Neural and Cognitive Sciences, China Medical University, Taichung, Taiwan

Y.-M. Tien
Department of Psychology and Chung Shan Medical University Hospital, Chung Shan Medical University, Taichung, Taiwan

Published online: 20 June 2013

English compounds and semantic transparency

A compound word is a word composed of at least two free morphological constituents that refer to a new concept. Compound words with two transparent constituents are defined as TT (transparent–transparent; see Frisson, Niswander-Klement, & Pollatsek, 2008; Libben, Gibson, Yoon, & Sandra, 2003; Pollatsek & Hyönä, 2005) when the whole-word meaning can be grasped through its individual constituents, such as "cookbook." Compound words are regarded as being semantically opaque (opaque–opaque, OO), when the word's meaning cannot be fully derived from its constituents —for example, "cocktail." Some compound words are considered partially opaque (opaque–transparent, OT, or transparent–opaque, TO) when the primary meaning of only one of the constituents is related to the meaning of the compound, such as in "butterfly" or "staircase," respectively. Several models have attempted to explain the access mechanisms of compound words from the mental lexicon (see Frisson et al., 2008, for a review). The *whole-word model* (Butterworth, 1983) proposes that a compound is accessed as a whole, so that the transparency of the constituents does not influence the processing of a word. The *morphological decomposition model* (Taft, 1981) suggests that readers decompose a compound into its constituents, followed by access to the constituents' meanings, and then construct the whole-word meaning on the basis of the individual constituents. The *parallel dual-route (process) model* (Baayen, Dijkstra, & Schreuder, 1997) suggests that a whole-word lookup route and a decomposition route compete with each other, implying that semantic transparency possibly plays a role in deciding which route will be used.

The effect of the transparency of the constituents in compound words during reading, however, is not as robust as the frequency effect, which has been found to influence gaze fixation time on each of the constituents (see Rayner, 2009, for a review). Pollatsek and Hyönä (2005) manipulated the

frequency (i.e., their occurrence in print) of constituents and the transparency of Finnish compound words, and they found longer gaze durations (the sum of all fixations made on a word prior to a saccade to another word, see Rayner, 1998, 2009) on low-frequency first constituents of either transparent or opaque compounds, as compared to high-frequency first constituents, but the eye-movement measures did not differ between transparent and opaque constituents. They concluded that the identification of both transparent and opaque compound words does not rely on constructing the meaning from the components. In a similar experiment, Frisson et al. (2008) used three types of opaque compound words—OT, TO, and OO—with matched TT words and found, consistent with Pollatsek and Hyönä, no significant difference in eye-movement measures due to this transparency manipulation. However, they did find longer gaze durations on opaque than on transparent compounds when the compounds were presented with a space between the constituents. They therefore suggested that the meaning of English compound words is not constructed from its parts but from the whole word, unless readers are forced to process the first and second constituents separately. However, inconsistent results were found in a study by Juhasz (2007), who manipulated the frequency of the constituents of transparent and opaque compound words and demonstrated that opaque compounds received longer gaze duration. She suggested that the decomposition of compound words occurs for both transparent and opaque compounds.

It is also known that morphological family affects semantic transparency (see Feldman, Basnight-Brown, & Pastizzo, 2006). A constituent may be part of one or many morphological family members; for example, the family of the constituent "butter" consists of "butterfly," "buttercup," "butterfingers," "buttermilk," "butterscotch," and "butterfat," among others. Within a morphological family, individual family members may vary in semantic transparency: For example, the meaning of "butter" is context-sensitive, so that it is more transparent in the meaning of "buttermilk" than in the meaning of "butterfly." Schreuder and Baayen (1997) suggest that upon reading a word, its family members become coactivated, which leads to a larger global activation in the mental lexicon. Several studies have focused on the effect of semantic transparency on morphological facilitation (Feldman & Soltano, 1999; Feldman, Soltano, Pastizzo, & Francis, 2004). The general finding is that a more concrete constituent with a larger family size (the number of morphological family members) is processed faster and more accurately in lexical-decision tasks.

Although the processing of compound words shares many common characteristics across languages, many differences have also been found, so that it is unclear whether the results found in alphabetical languages could be applied, for example, to the processing of Chinese.

Chinese compounds and semantic transparency

Approximately 74 % of all words in the Chinese language are made up of two characters (Zhou & Marslen-Wilson, 1995), with some words consisting of only one character, and others consisting of three or more characters. A Chinese character is a writing unit that has a single syllable and one or more meanings. Most Chinese characters are approximately equal to single morphemes, and therefore the majority of Chinese words can be considered bimorphemic compound words (referred to as *compounds*). Chinese compounds, similar to English ones, differ in how the meanings of the first and second characters relate to the meaning of the word. Some Chinese compounds are semantically transparent—that is, both characters are transparently related to the meaning of the whole word. Other words are fully opaque—that is, the meaning of neither constituent is related to the meaning of the compound—or partially opaque. Table 1 lists some examples of transparent, opaque, and partially opaque Chinese words.

Similar to morphological families in English, a Chinese character—for example, 馬 ("horse")—can be shared by its morphological family members—for example, 馬鞍 ("saddle") and 馬虎 ("careless")—and the meaning of the character and those morphological family members may not be consistent in meaning (see Mok, 2009, for a review). For example, the character 馬 ("horse") consists of morphological family members, including 馬背 ("horseback") and 馬鞍 ("saddle"), that are semantically related to "horse," but others, such as 馬虎 ("careless"), 馬桶 ("stool"), or 馬來 ("Malaysian"), are not. The position of a Chinese character within a two-character compound does not provide strong constraints on the activation of the whole-word units in general (Taft, Zhu, & Peng, 1999). However, the meaning of some compounds—for example, 領帶 ("necktie")—may differ from the meanings of compounds in which the characters are transposed—for example, 帶領 ("guide"). Sometimes a constituent—for example, 調—may have different meaning and pronunciation when it is located in the initial (meaning "adjust", pronunciation *tiáo*) or final (meaning "high or low tone/key", pronunciation *diáo*) positions.

Unlike English and other alphabetic writing systems, Chinese words are written without spaces in a sequence of characters. The concept of a word is not as clearly defined in Chinese as it is in English, which means that Chinese readers may disagree somewhat about where word boundaries are located (see Mok, 2009; Rayner, Li, & Pollatsek, 2007, for reviews). According to the segmentation standard by Huang, Chen, Chen and Chang (1997), used by the Academia Sinica Balanced Corpus (ASBC; Academia Sinica, 1998), not all characters stand on their own as one-character words.

Studies have investigated how Chinese compound words are accessed in the mental lexicon in different tasks by

Table 1 Examples of transparent, opaque, and partially opaque Chinese words

Transparency	Whole Word	First Character	Second Character
TT	球場 (“ball court”)	球 (“ball”)	場 (“court”)
OO	壽司 (“sushi”)	壽 (“age”)	司 (“to be in charge of”)
TO	智商 (“I.Q.”)	智 (“intelligent”)	商 (“commerce”)
OT	追悼(“commemorate”)	追 (“chase”)	追 (“mourn”)

manipulating frequency (see Zhou, Ye, Cheung, & Chen, 2009, for a review). In a series of experiments with varied whole-word and constituent frequencies, Chen and Chen (2006) showed that compound-word production in Chinese is not sensitive to morpheme frequency, even when all of the stimuli are semantically transparent, and they suggested that morphological encoding is only minimally involved in the production of Chinese transparent compound words. Consistent results were obtained by Janssen, Bi, and Caramazza (2008), who found that compound-word production is determined by the compound's whole-word frequency either in Chinese or in English, and not by its constituent morpheme frequency. Their results support the view that compounds are stored in their full form. However, inconsistent results were obtained for low-frequent compounds during reading (Yan, Tian, Bai, & Rayner, 2006). These authors investigated the effect of (two-character) compound-word and constituent (character) frequency on word processing during reading on eye movements, and they suggested that when a compound is frequent and has been seen quite often in print, it is accessed as a single entity in the mental lexicon of Chinese readers, whereas when it is infrequent, the compound needs to be accessed via the constituents (and hence, an effect of character frequency emerges).

In studies in which the frequency and transparency of Chinese compound words are manipulated, Hung, Tzeng, and Chen (1993) reported a reliable constituent frequency effect for fully transparent compounds (TT) but not for opaque ones (either TO, OT, or OO) in a lexical-decision task. They also obtained a significant whole-word frequency effect for both transparent and opaque compounds. They suggested that a whole-word representation exists for all types of Chinese compounds, even for fully transparent ones, but that separate morphemic representations in the mental lexicon exist only for transparent compounds. Mok (2009) reported a larger word superiority effect (WSE) for opaque compounds (either TO, OT, or OO) than for transparent ones in a modified Reicher–Wheeler paradigm, which briefly presents words and letters followed by a mask (see Mok, 2009; Reicher, 1969; Wheeler, 1970). The WSE describes more accurate recognition when a target character—for example, 态 (“appearance”)—is in the context of a compound word 态度 (“manner”), as opposed to when the same target is in a position-matched nonword control 态备 (“appearance–equipped”). They also found a larger WSE for Chinese compounds with high whole-word frequencies than for ones with low whole-word frequencies. The results imply that all types of Chinese compounds, including TT, have corresponding whole-word entries in the mental lexicon, and that the constituents of TT compounds are activated more distinctively than the ones of opaque compounds (OT, TO, and OO).

Taken together, these results show a reliable whole-word frequency effect and indicate that whole-word representations exist in the mental lexicon, even for TT compounds. Fully transparent compounds tend to be accessed via constituents (1) when the compounds are low-frequent during reading (Yan et al., 2006), (2) in lexical-decision tasks (Hung et al., 1993), or (3) in a modified Reicher–Wheeler paradigm (Mok, 2009). Although inconsistent results have been found in different tasks, studying semantic transparency is clearly important for understanding morphological processing in Chinese.

Estimating semantic transparency

Transparency rating of English compounds

Transparency ratings are the most common method to obtain transparency information. For instance, Pollatsek and Hyönä (2005) selected 80 compound words, 40 of which they assumed to be semantically transparent, and the other 40 to be opaque. They asked eight subjects to rate these words regarding their transparency using a 7-point scale (with 1 for *totally transparent* and 7 for *totally opaque*), and the ratings were clearly lower for the supposedly transparent sets than for the supposedly opaque ones. Frisson et al. (2008) asked 40 participants to rate transparency in terms of the appropriate categories, either OT, TO, OO, or TT, for each compound. Frisson et al. found good agreement between the subjects' choices and the predefined classifications.

Transparency ratings of Chinese compounds

In Mok (2009), semantic transparency judgments were made in two passes, one by an experimenter and by five trained participants' analysis on the basis of dictionary definition,

and the other by 30 naïve participants. A 6-point scale rating, in which 1 was *opaque* and 6 *transparent*, was used for both passes. In general, a constituent was classified as transparent if the rating was greater than 3.5, and as opaque if the rating was less than 3.5.

Subjective differences and ambiguity of transparency

Unfortunately, estimates of semantic transparency are often subjective and vary strongly across raters. Mok (2009) pointed out that response biases in transparency judgments may be due to (1) a subject's understanding of the meaning of stimuli being different from the dictionary definition, (2) the meaning of a compound not being dissociated from its constituents, or (3) a subject not clearly knowing the meaning of the presented materials. Furthermore, subjective differences may also be caused by the instructions for the transparency judgments—for example, by dictionary definitions or by other means—which may have great influence on the ratings, since the concept of a word is not clearly defined for Chinese readers. For a constituent with multiple, inconsistent meanings, raters may make subjective decisions leading to inconsistent results.

Sometimes even the meaning of transparent compounds cannot be unambiguously determined from the meanings of their constituents (see Frisson et al., 2008). Inhoff, Starr, Solomon, and Placke (2008) indicated that a semantic relationship often exists between an opaque lexeme and its compound; for example, even though "jailbird" typically refers to a person rather than an animal, it can convey useful semantic information, such as being caged or wishing to fly free. This subjectivity and variability also occurs in the characters of Chinese compounds. A general model may be a way to average across subjective differences.

Models to predict transparency using LSA

In this study, we propose models using latent semantic analysis (LSA) for predicting raters' transparency judgments. LSA is a method to determine the semantic similarity of words and sets of words by statistical computations applied to a text corpus (Landauer & Dumais, 1997; Landauer, McNamara, Dennis, & Kintsch, 2007). Typically, the terms are words, and a term-to-document co-occurrence matrix is established from a corpus. Then, a mathematical method, singular value decomposition (SVD), is used to reduce the dimensions of the original matrix (see Martin & Berry, 2007). The meaning of each term is represented as a *vector* in *semantic space*. One can compute the semantic similarity values for any two terms in a given language using the LSA cosine value, which ranges from −1 to 1, but rarely goes below 0 because the matrix is made up of word counts, which are strictly positive. Since only very small negative numbers occur in the left singular matrix, the dot product (and the cosine) tends to have a lower bound close to zero. In the semantic space for "general reading up to 1st year college" (abbreviated as SP-E) with the 300 dimensions used in the present study, randomly chosen pairs of words have a mean of .03 and a standard deviation of approximately .08 (see Landauer et al., 2007). An LSA website is freely available (http://lsa.colorado.edu/, accessed September, 2010; see Dennis, 2007).

LSA has been used to investigate morphological decomposition; for example, Rastle, Davis, Marslen-Wilson, and Tyler (2000) investigated morphologically complex words with semantically transparent embedded stems (e.g., "depart" vs. "departure") and opaque embedded stems (e.g., "apart" vs. "apartment"). Similarly, Diependaele, Duñabeitia, Morris, and Keuleers (2011) used LSA to estimate transparency between full words and constituent-embedded stems, which yielded "viewer" versus "view" as being highly transparent and "corner" versus "corn" as highly opaque.

Since the LSA-based method may be able to estimate the transparency of English compounds, it could possibly be applied to Chinese two-character words in a similar manner. Following the principle of creating semantic spaces (Quesada, 2007), our previous studies (M. L. Chen, Wang, & Ko, 2009; Wang, Pomplun, Ko, Chen, & Rayner, 2010) built an LSA semantic space of Chinese (abbreviated as SP-C) from ASBC, which contains approximately 5 million words (or 7.6 million characters). The texts in ASBC were collected from different topics. Word segmentation was performed manually according to the standard by Huang et al. (1997). For representatives of words in the corpus, words that occurred less than four times among the 5 million words were excluded in SP-C. Most of the excluded words were proper names and technical nouns. A 49,021 × 40,463 term-to-document co-occurrence matrix was then established. SP-C has been shown to successfully estimate word predictability (see Wang et al., 2010) and word association (see Chen et al., 2009) in Chinese.

However, LSA was merely developed as a tool in morphological-decomposition studies to validate human transparency judgments, instead of being a model based on a theoretical foundation that can be strongly predictive and correlated with transparency judgments. Furthermore, LSA has not yet been tested for two-constituent compound words in English or Chinese, which raises the question of whether the cross-linguistic comparison could be made in the same manner. Therefore, we adopted the idea of comparing the meanings of a compound and of each of its constituents in Model 1. We also proposed a Model 2 based on the theoretical foundation of morphological family members, and this model may explain how a rater accomplishes semantic

judgment tasks. Furthermore, it was necessary to evaluate the discrimination performance of the proposed models for human transparency ratings. We evaluated how LSA estimates transparency using the English compound materials in Frisson et al. (2008), the Chinese compounds in Mok (2009), and transparency rating conducted for the present study. The objectives of building general models are to develop a research method to average across subjective differences and to allow a linkage between the theoretical foundation and technical understanding of LSA for transparency judgments.

Model 1: Whole word versus each of its constituents

One proposed idea of modeling how transparency judgment is done by raters is to compute the LSA cosine values between a compound word and each of its constituents. The parameters of Model 1 include the *semantic space*, the number of *dimensions*, and the *comparison type*. For example, using SP-E, 300 dimensions, and "term-to-term" comparison, the LSA cosine value between "staircase" and "stair" is .57, whereas the one between "staircase" and "case" is .07. Since the constituent "stair" and the compound word "staircase" result in a clearly higher cosine value, "stair" is considered semantically transparent, whereas "case" is considered opaque.

To accomplish the comparisons of Model 1, the compound words are required to have their constituents occur in the semantic space on their own, which becomes a constraint of Model 1. The term-to-document matrix of SP-C uses the unit of words, which may be single- or multicharacter words. Within the 49,021 words available in SP-C, 31,637 are two-character words, and for 3,921 out of these two-character words, either the first or the second character is not a standalone word occurring more than three times in the corpus. That is, 12 % of the compound-word cases were unavailable in the Model 1 computations.

Model 2: Whole word versus dominant meaning of each of its constituents

A possible solution for the constraint in Model 1 is proposed in Model 2, which assumes that a rater accesses the *dominant meaning* of a constituent from its morphological family. Chinese compounds sharing common constituent morphemes were consistently found to facilitate each other (Zhou, Marslen-Wilson, Taft, & Shu, 1999). It is also known in English that morphologically related words in the mental lexicon are linked in meaning, and a morpheme shared by more morphological family members allows for more rapid activation of a word's meaning, and therefore faster responses in word recognition tasks (see Bybee, 1988; Feldman et al., 2006). Therefore, Model 2 takes the polysemy of a constituent into account and may be well-adapted to morphological processing models. Furthermore, Model 2 overcomes the limitation of Model 1 that some characters do not exist as one-character words, which is especially useful for Chinese.

The first step of Model 2 is to obtain the dominant meaning of the constituent from its morphological family members that a rater would possibly activate. We computed the LSA cosine values of the pairs in the morphological family members as a distance function. Using a hierarchical clustering algorithm and a given threshold, we classified these morphological family members into *semantic clusters*. Subsequently, a cluster—with the largest family size, or the highest sum of word frequencies (occurrence in corpus) when multiple clusters obtained the same family size—was considered the dominant meaning. An example of the transparency of the constituent "butter" in "butterfly" is determined as follows: The morphological family members "butter," "butterfly," "buttercup," "butterfingers," "buttermilk," "butterscotch," and "butterfat" are activated; the LSA cosine values among them are shown in Table 2. The semantic similarities, although in high-dimensional space, can be visualized by multidimensional scaling (MDS) in two dimensions, as is shown in Fig. 1a. According to the distance measure by LSA and the agglomerative hierarchical clustering algorithm (implemented in MATLAB; The MathWorks, Inc., Natick, MA), "butter," "buttercup," "buttermilk," "butterscotch," and "butterfat" are in one cluster, and "butterfly" and "butterfingers" are in their own clusters. We applied "term-to-document" comparison (i.e., each word in the "document" is represented as a vector and weighted according to its frequency in the corpus, and the "document" is the vector addition of the weighted vectors) between the compound (e.g., "butterfly") and the dominant meaning cluster (e.g., the string "butter buttercup buttermilk butterscotch butterfat") in order to compute the LSA cosine value (in this example, .04). Similarly, the LSA cosine values between the Chinese character 馬 ("horse") and its morphological family members are presented in Table 3,

Table 2 Latent semantic analysis (LSA) cosine values among "butter," "butterfly," "buttercup," "butterfingers," "buttermilk," "butterscotch," and "butterfat"

	butter	-fly	-cup	-fingers	-milk	-scotch	-fat
butter (2085)	1						
butterfly (630)	.04	1					
buttercup (24)	.09	.09	1				
butterfingers (3)	0	–.1	–.1	1			
buttermilk (13)	.44	−0	.12	.01	1		
butterscotch (7)	.45	.05	−0	.02	.35	1	
butterfat (39)	.12	−0	.04	0	.11	.16	1

The frequency for each word in the British National Corpus (BNC, per 100 million words) is shown in parentheses

and the MDS result is illustrated in Fig. 1b. 馬 ("horse"), 馬背 ("horseback"), 馬鞍 ("saddle"), and 馬車 ("carriage") are grouped in one cluster, 馬來 ("Malaysian") and 馬國 ("Malaysia") form another one, whereas 馬虎 ("careless"), 馬桶 ("stool"), and 馬腳 ("a clue of") are in their own clusters.

The parameters in Model 2 include *semantic space*, *number of dimensions*, *comparison type*, *morphological family definition*, *threshold and distance function* of the clustering algorithm, and *dominant meaning definition* (by family size or by frequency). The definition of morphological family for a constituent is considered *position-specific*; for example, the morphological family of "butter" in "butterfly" are words starting with "butter" as a constituent. The selection of a threshold in a clustering algorithm is related to the distance function as well as to the LSA values in a given semantic space. A low threshold may generate too many clusters, whereas a high threshold may group unrelated members in one cluster.

The access of the dominant meaning of a constituent may be different between English and Chinese compounds. The English writing system contains clear word boundaries, and the constituents in English compounds are usually standalone words. However, the concept of a word is not clearly defined in Chinese, and it is possible that Chinese readers derive the meaning of a character implicitly from its morphological family. Therefore, due to the cross-linguistic difference, the settings of the model parameters may differ between the English and Chinese languages.

Model evaluation

We evaluated Models 1 and 2 on the basis of three data sets: English compounds using the materials of Frisson et al. (2008), Chinese compounds in Mok (2009), and a transparency-rating procedure conducted for the present study. To represent semantic similarity, we report the descriptive statistics and distributions of the LSA cosine values computed by Model 1 (*M1*) and by Model 2 (*M2*) as the major measures, and the average LSA cosine values between all pairs within a morphological family (*Co*, henceforth referred to as *consistency*) as a supportive measure. A nonparametric test was performed using Mann–Whitney *U* tests. Since the models attempted to map the LSA cosine values (a continuous variable) onto dichotomous transparency results (either O or T), we performed a receiver operating characteristic (ROC; Green & Swets, 1966) analysis, which plots the hit rate on the *y*-axis as a function of the false-alarm rate on the *x*-axis. The discriminatory ability of a model is reflected in the area under the ROC curve (AUC), which typically ranges from .5 (chance level) to 1 (perfect performance).

English compounds in Frisson et al. (2008)

Method

Stimuli For the English compounds, we reanalyzed the materials in Frisson et al. (2008), which included 10 OO, 14 OT, 10 TO, and 34 TT compounds (i.e., 44 opaque and 92 transparent constituents). The proportions of subjects' choices agreeing with the predefined classification was 65 % for OO, 71 % for OT, 65 % for TO, and 86 % for TT. Moreover, the proportions of subjects classifying at least one of the constituents as opaque for the predefined opaque words were very high: 95 % for OO, 93 % for OT, and 95 % for TO.

Analysis For the computations in Models 1 and 2, 40 opaque and 84 transparent constituents were available using SP-E and 300 dimensions. The "term-to-term" comparison was

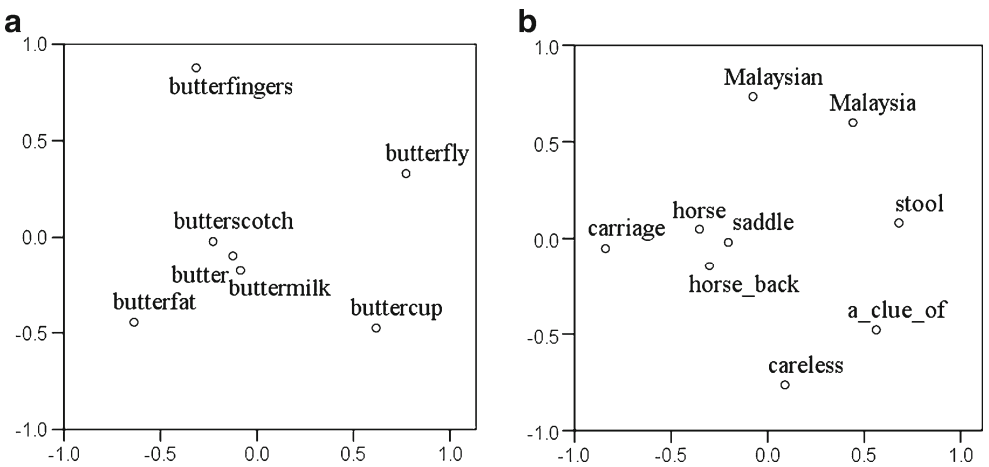

Fig. 1 Multidimensional scaling (MDS) results for examples of semantic relationships: (**a**) "butter" and its morphological family, and (**b**) 馬 (horse) and its morphological family. The *x*- and *y*-axes represent Dimensions 1 and 2, respectively, of the abstract, two-dimensional Euclidean output spaces of the MDS algorithms

Table 3 Latent semantic analysis (LSA) cosine values between the character 馬 ("horse") and its morphological family members

	馬	馬背	馬鞍	馬車	馬虎	馬桶	馬腳	馬來	馬國
馬 (horse, 342)	1								
馬背 (horse back, 14)	.83	1							
馬鞍 (saddle, 4)	.74	.74	1						
馬車 (carriage, 37)	.17	.07	.03	1					
馬虎 (careless, 13)	–.02	–.04	–.01	–.04	1				
馬桶 (stool, 23)	–.05	–.04	–.03	.01	.10	1			
馬腳 (a clue of, 4)	.00	.04	–.05	.01	.13	.02	1		
馬來 (Malaysian, 11)	.08	.06	.04	.04	.00	–.09	–.03	1	
馬國 (Malaysia, 12)	.03	–.01	–.03	.03	.01	–.04	.02	.15	1

The meaning and frequency (per 5 million words) in ASBC for each word are shown in parentheses

used in Model 1 and "term-to-document" was used in Model 2. In Model 2, we defined morphological family as the compounds sharing the same constituents in the same position. For the agglomerative hierarchical clustering algorithm, a distance function and a threshold were used to decide which morphological family members should be combined. The distance function between pairs of morphological family members was set to one minus the absolute value of the LSA cosine value, and the threshold was set to .8. The dominant meaning was defined as the cluster with the highest sum of frequencies. The details are reported in Appendix Table 5.

Results and discussions

Descriptive statistics We found that the M1 results for the transparent constituents (mean = .29, standard deviation = .21) were significantly higher than those for opaque ones (mean = .07, standard deviation = .09), $U = 430.50$, $N_1 = 40$, $N_2 = 84$, $p < .001$. The results were consistent in M2: Higher cosine values were obtained for transparent constituents (mean = .31, standard deviation = .26) than for opaque constituents (mean = .05, standard deviation = .16), $U = 564.00$, $N_1 = 40$, $N_2 = 84$, $p < .001$. Co was found to be higher for transparent constituents (mean = .21, standard deviation = .14) than for opaque constituents (mean = .17, standard deviation = .13), $U = 1{,}442$, $N_1 = 44$, $N_2 = 92$, $p < .01$. The distributions of LSA cosine values of transparent and opaque constituents computed by Models 1 and 2 are shown in Fig. 2a and b.

ROC Figure 3a illustrates the ROC curves for Models 1 and 2, and the AUCs are .87 and .83, respectively. An "optimal" cutoff point in the ROC curve, defined as the shortest Euclidian distance to the point (0, 1) (perfect performance; i.e., false-alarm rate of 0 and hit rate of 1), can be used to find a LSA cosine value that performs a good separation between opaque and transparent constituents. The optimal cutoff point of Model 1 is at a hit rate of .74 and a false-alarm rate of .10 when the threshold of the LSA cosine value is set to .135. The optimal cutoff point for Model 2 is at a hit rate of

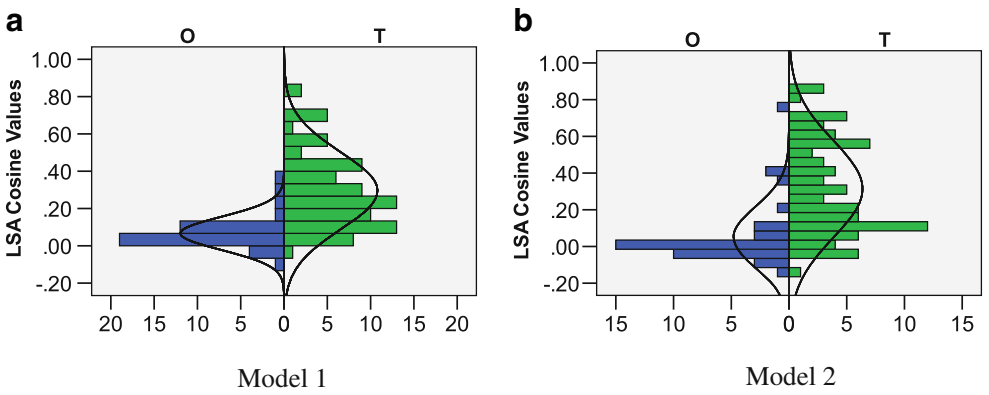

Fig. 2 Distributions of LSA cosine values of opaque (O) and transparent (T) constituents computed by (**a**) Model 1 and (**b**) Model 2 for the materials in Frisson et al. (2008)

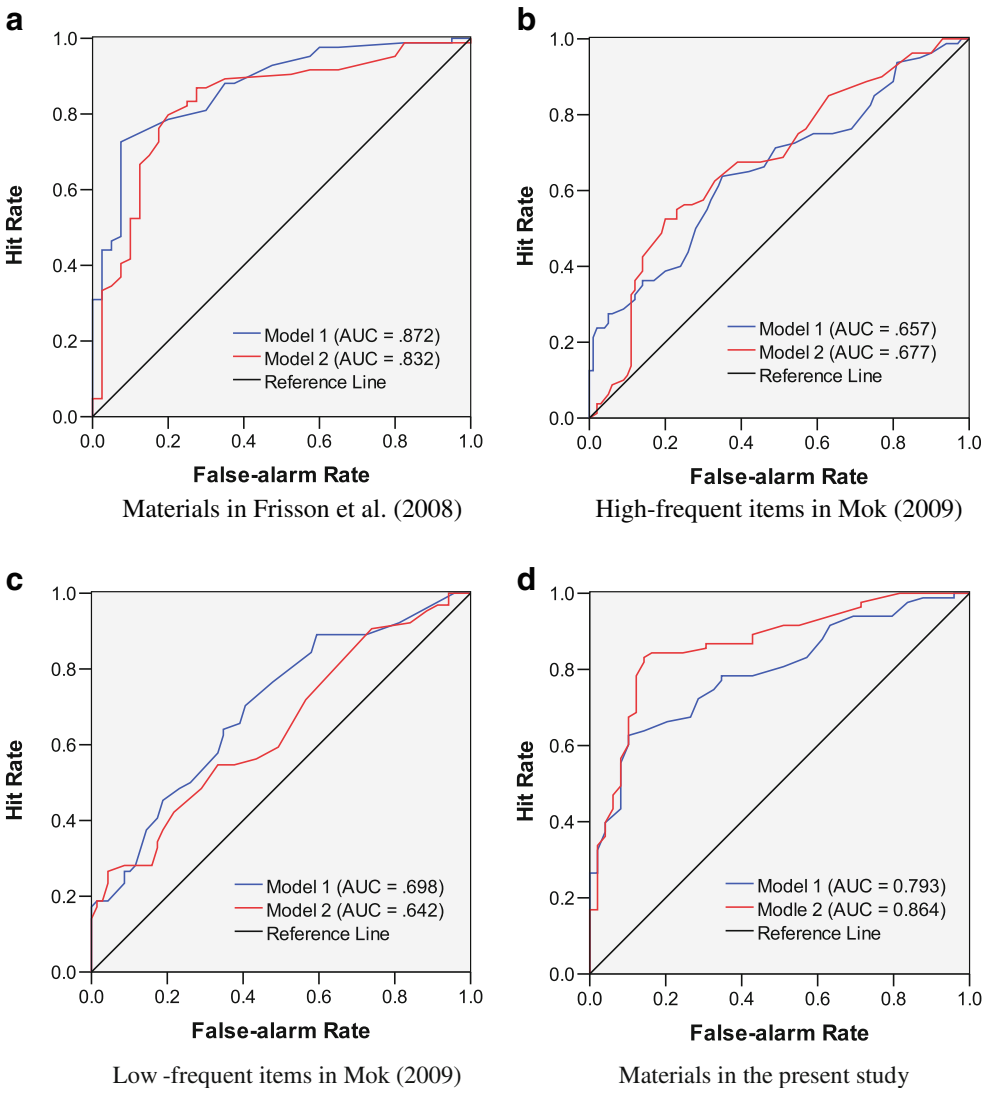

Fig. 3 Receiver operating characteristic analysis of Models 1 and 2. AUC, area under the curve

.80 and a false-alarm rate of .20, for which the threshold of the LSA cosine value is .085.

Model 1 showed a higher hit rate than does Model 2 for all false-alarm rates except those between .2 and .4. Both models perform good prediction of human transparency judgments, and Model 1 has slightly better performance than Model 2. The overall results suggest that LSA successfully captures the transparency conditions in the materials of Frisson et al. (2008).

Chinese compounds in Mok (2009)

The total of 190 compounds were divided into two sub-data-sets: high-frequency items and low-frequency items. Each sub-data-set contained half of the compounds. The agreement between two passes (one by dictionary definition and the other by subject rating) was high (Cohen's kappa = .83), and low-frequency items (Cohen's kappa = .87) obtained higher agreement than did high-frequency items (Cohen's kappa = .78). The compounds in Mok (2009) were presented in simplified script, and they were converted into traditional script in SP-C. The details are listed in Appendix Table 6.

Method

Stimuli for high-frequency items The final classification included 21 TT, 21 TO, 22 OT, and 31 OO compounds, resulting in 85 transparent and 105 opaque constituents. Due to the different segmentation standards applied to the Mok (2009) study in SP-C, the compound 幻灯 ("slideshow," as 幻燈 in traditional script) was converted into 幻燈機

("slideshow machine") and 幻燈片 ("slides"), 忘年 ("old age") was converted into 忘年之交 ("an old friend"), 开交 ("to conclude [impossible] to finish," as 開交 in traditional script) was converted into 不可開交 ("to conclude impossible to finish"). Model 1 evaluated 80 out of 85 transparent and 100 out of 105 opaque constituents. Model 2 overcame the limitation of nonstandalone characters in Model 1, resulting in 85 out of 85 transparent and 103 out of 105 opaque constituents being available for its evaluation.

Stimuli for low-frequency items These were 23 TT, 24 TO, 26 OT, and 22 OO compounds, which contained 96 transparent and 94 opaque items. There were minor differences in usage between simplified and traditional scripts, and 穷蛋 ("pauper," as 窮蛋 in traditional script) was converted into 窮光蛋 ("pauper"). Five out of the 95 compounds (站队, 逃奔, 环打, 洋灰, and 白事) are not used in traditional script. The limitation of SP-C is that only words occurring at least four times in the ASBC corpus were included, resulting in 20 out of the 95 words being excluded from the computation. In total, 64 out of the 96 transparent and 69 out of the 94 opaque items were available in Model 1, and 64 out of the 96 transparent items and 71 out of the 94 opaque items were available in Model 2.

Analysis The model parameters were set as follows: SP-C was used for the computations in Models 1 and 2. The "term-to-term" comparison was used in Model 1 and the "term-to-

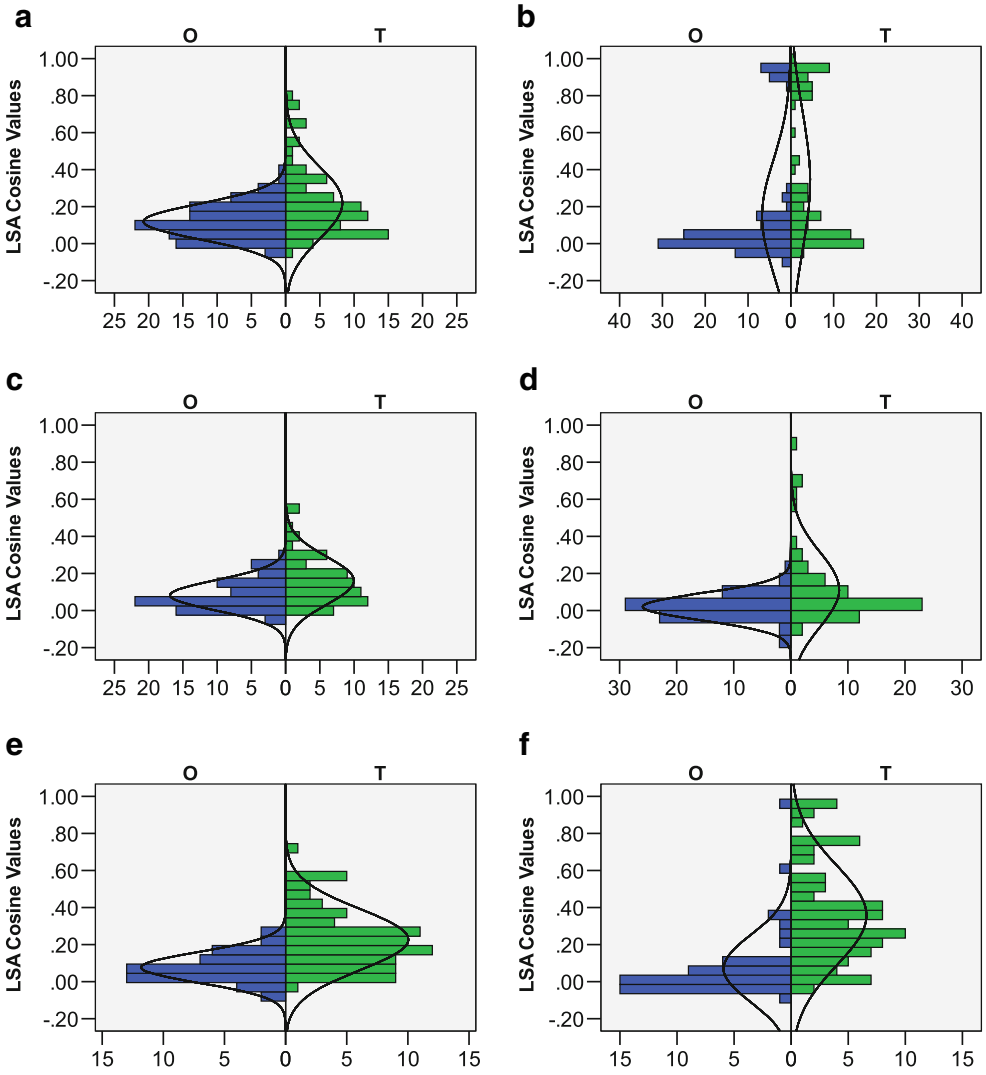

Fig. 4 Distributions of latent semantic analysis (LSA) cosine values of opaque (O) and transparent (T) constituents of (**a**) high-frequency items by Model 1, (**b**) high-frequency items by Model 2, (**c**) low-frequency items by Model 1, and (**d**) low-frequency items by Model 2 in the materials of Mok (2009). Panels **e** and **f** are for the materials in the present study, predicted by Models 1 and 2, respectively

Table 4 Constituent frequency, family size, and consistency when a cutoff point of .4 of the latent semantic analysis (LSA) cosine was selected, for the high-frequency items in Mok (2009)

	Constituent Frequency		Family Size		Consistency	
	Mean	Std	Mean	Std	Mean	Std
T response to T	389	698	24.64	19.72	.14	.09
T response to O	1,245	313	27.95	19.97	.09	.03
O response to T	63	98	14.58	19.42	.10	.05
O response to O	1,180	2,425	37.53	31.50	.08	.03

document" was used in Model 2, except that 幻燈機 幻燈片 was based on the "document-to-term" comparison in Model 1 and the "document-to-document" comparison in Model 2. For the clustering algorithm in Model 2, the threshold setting was .5, and the dominant meaning was defined as the cluster with the largest family and the highest sum of frequencies if the family sizes of multiple clusters were the same.

Results and discussions

Descriptive statistics for high-frequent items We found that the M1 results for the transparent constituents (mean = .22, standard deviation = .19) were significantly higher than those for the opaque words (mean = .12, standard deviation = .10), U = 2,741.50, N_1 = 80, N_2 = 100, $p < .001$. Again, for transparent constituents (mean = .34, standard deviation = .37), M2 obtained higher LSA cosine values than those for opaque constituents (mean = .15, standard deviation = .30), U = 2,779.50, N_1 = 85, N_2 = 103, $p < .001$. For Co of high-frequency items, consistent with the English constituents, transparent constituents (mean = .12, standard deviation = .07) yielded higher values than did opaque constituents (mean = .09, standard deviation = .04), U = 3,047.5, N_1 = 83, N_2 = 104, $p < .01$.

ROC for high-frequency items Figure 3b illustrates the ROC curves for Models 1 and 2, and the AUCs are .66 and .68, respectively. These ROC curves show that Model 2 generated more false alarm cases in the beginning of the curve than Model 1, which were caused by the "O responses to T" cases. Nevertheless, Model 2 overall slightly outperformed Model 1. The distributions of LSA cosine values of transparent and opaque constituents computed by Models 1 and 2 are shown in Fig. 4a and b.

Misclassified item analysis for high-frequency items Since a few opaque items have high LSA cosine values, and also a few transparent items have low LSA values in Fig. 4b, we selected a cutoff point of .4 of the LSA cosine value and summarize the constituent frequency, family size, and consistency in Table 4. A one-way analysis of variance (ANOVA) for the four transparency conditions (T responses to T, T responses to O, O responses to T, and O responses to O) indicated a significant overall effect of condition on the consistency, $F(3, 183) = 11.94$, $p < .001$. Post-hoc tests using Bonferroni-adjusted p values revealed that consistency was higher in the condition of T responses to T than in each of the other conditions (all ps < .01). None of the other comparisons were significant (all ps > .37). A one-way ANOVA revealed a marginal difference in frequency across the four conditions, $F(3, 186) = 2.41$, $p = .07$. The frequency of O responses to T (mean: 63) was numerically lower than in the other conditions, but the difference did not reach significance in the Bonferroni-corrected post-hoc tests (all ps > .30). These results suggest that when the consistency of constituents was high, the prediction for transparent constituents tends to be more accurate. Constituent frequency might be a possible reason for the misclassification of Model 2 in high-frequency items (see Fig. 4b), since Model 2 defines the dominant meaning as the cluster with higher family size and frequency. We suggest that the polysemy and frequency of constituents may affect model predictions.

Descriptive statistics for low-frequency items We found that transparent constituents (mean = .16, standard deviation = .13) obtained higher M1 values than did opaque ones (mean = .08, standard deviation = .08), U = 1,335, N_1 = 66, N_2 = 103, $p < .001$, and transparent constituents (mean = .11, standard deviation = .20) obtained higher M2 values than did opaque ones (mean = .02, standard deviation = .07), U = 1,335, N_1 = 66, N_2 = 103, $p < .01$. For Co, transparent constituents (mean = .10, standard deviation = .04) showed marginally higher values than did opaque ones (mean = .09, standard deviation = .03), U = 3,689, N_1 = 93, N_2 = 94, $p = .06$.

ROC for low-frequency items The AUCs of the ROC curves for Models 1 and 2 were .70 and .64, respectively (see Fig. 3c), which may be caused by subjects accessing the meaning of a constituent via its standalone form, rather than via its morphological family. The distributions of LSA cosine values of transparent and opaque constituents computed by Models 1 and 2 are shown in Fig. 4c and d.

Discussion In general, the models showed less predictive power for the Mok (2009) materials than for the Frisson et al. (2008) materials. Nevertheless, both Models 1 and 2 demonstrated considerable discrimination performance and may represent how experimenters performed transparency

judgments by dictionary definition or how naïve raters activated meanings from morphological family members. Model 1 represents the meaning of a constituent when it is stand-alone, which may be closer to its dictionary definition than it is in Model 2. On the other hand, Model 2 represents the dominant meaning (defined as higher family size and frequency) derived from its morphological family members, and Model 2 may be closer to a subject's rating than is Model 1 when the stimuli are high-frequency.

Transparency rating in the present study

We found that model prediction may be affected by the procedures for transparency rating—for example, the two passes (one by dictionary definition and the other by subject rating) in the study by Mok (2009). It is also important to notice that the transparency rating in Mok is somewhat different from the one in Frisson et al. (2008), who asked participants to categorize a word as being OT, TO, OO, or TT. Following Frisson et al. (2008), we performed a rating study, which required subjects to respond either T or O for each constituent when the compounds were presented.

Method

Stimuli We selected 80 bisyllabic words (two characters in written form) in traditional script, given that SP-C is built on a traditional script corpus. Seventy-seven out of the 80 words were compound words, and the others were composed of foreign-word translations based on pronunciation. The materials were selected from those used by Lee (1995), Lee (2007) and Tsai (1994). Totals of 83 out of 89 transparent and 49 out of 53 opaque constituents were available for evaluating Model 1. All 89 transparent and 53 opaque constituents were available to Model 2, since Model 2 overcomes the constraint of Model 1 that some constituents are stand-alone and therefore unavailable in SP-C. The detailed results are shown in Appendix Table 7.

Subjects Those compounds were rated by 11 students who completed a college degree in Taiwan. All of the subjects were native speakers of Chinese (traditional script).

Analysis The measure of human rating of each constituent was calculated as the probability with which subjects responded T to the constituent—for example, .91 for 10 out of 11 participants responding T. The characters with probabilities greater than or equal to .6 were categorized as transparent, whereas the ones with probabilities less than or equal to .4 were considered as opaque. The means and standard deviations of the human ratings were .85 and .13, respectively, for transparent characters, and .11 and .11 for opaque characters. All other settings for Models 1 and 2 were identical to those in the analysis of the Mok (2009) compounds.

Results and discussions

Descriptive statistics For M1, transparent constituents (mean = .23, standard deviation = .16) were higher than opaque constituents (mean = .08, standard deviation = .08), $U = 841$, $p < .001$. Similarly, transparent constituents (mean = .36, standard deviation = .27) obtained higher M2 than opaque ones (mean = .07, standard deviation = .18), $U = 655.5$, $p < .001$. The results also reveal that Co is higher for transparent constituents (mean = .15, standard deviation = .10) than for opaque ones (mean = .09, standard deviation = .04), $U = 1{,}268.5$, $p < .001$.

Correlations The Spearman rank correlations (a nonparametric test) between the human rating probabilities and Model 1 and between human rating probabilities and Model 2 were .50 and .58, respectively, ps < .001.

ROC Figure 3d illustrates the results of the ROC analysis, and the AUCs for Models 1 and 2 were .80 and .86, respectively. Both Models 1 and 2 are predictive to the results of human transparent judgments, and Model 2 obtained a numerically higher AUC and correlation than did Model 1. As we mentioned above, the concept of a word is not as clearly defined in Chinese as in English, and Chinese readers might learn the polysemy of characters implicitly from polymorphemic words. Therefore, Model 2 may in general be a better approach than Model 1 to predict transparency ratings for the constituents of Chinese compounds.

General discussion

The most important outcome of the present study is the proposed computational approach of using LSA to estimate semantic transparency and consistency measures, which may benefit psycholinguistic studies as a research method.

A research method for psycholinguistic studies

The effect of semantic transparency and consistency in Chinese reading The results, such as semantic transparency and consistency estimations, could be adapted to further Chinese reading research. Some unpublished studies have addressed the semantic transparency of two-character Chinese words (Lee, 1995; Lee, 2007). C. Y. Lee found opposite results by

manipulating word frequency, character frequency, and word transparency in a lexical-decision task. She found that the response times of opaque words were shorter than those of transparent words, and the character frequency effects were only significant in transparent words. These results suggest that opaque words tend to be stored as a whole unit, whereas the constituents of transparent words are represented separately in the mental lexicon. Moreover, P. J. Lee used eye-tracking to investigate how word frequency, word transparency, and character frequency influence eye-movement measures during reading. The study revealed shorter first-fixation and gaze durations for opaque words than for transparent words. Character frequency effects were found in transparent words (also known as *compositional words*) in first-pass measures; low-frequency characters were fixated longer. These eye-movement results were consistent with the findings of C. Y. Lee, but contrasted with the results of the English compound-word study by Frisson et al. (2008) that eye-movement measures did not differ between opaque words and their transparent controls. Since these studies were from non-peer-reviewed work and details of the works are restricted, we suggest applying the proposed transparency and consistency measures in further studies to examine the relationship between the continuous cosine values and genuinely continuous data such as fixation durations or lexical decision latencies, in order to address the open questions about semantic transparency.

Transparency of semantic radicals of Chinese characters In Chinese character orthography, the most common structure is a semantic–phonetic compound character that is composed of a semantic radical on the left and a phonetic radical on the right (see Yan, Zhou, Shu, & Kliegl, 2012). Generally, a semantic radical represents the meaning of the whole character, whereas the phonetic radical provides roughly the pronunciation. Some radicals can be standalone characters, but others cannot. Similar to the constituents of English and Chinese compound words, the meaning of a semantic radical may or may not be semantically related to the whole character. Radical semantic transparency refers to how a semantic radical semantically relates to the meaning of its semantic–phonetic compound character. For example, the semantic radical 馬 ("horse") in the character 驢 ("donkey) is considered semantically transparent, whereas the semantic radical 氵 ("water") in the character 法 ("law") is opaque. Furthermore, the meaning of a semantic radical might differ from its meaning when it is a standalone character; for example, 貝 ("shell") was used as currency in ancient China, and therefore many characters with this radical are related to "money"—for instance, 賺 ("earn money") or 賒 ("loan money")—and are not semantically related to "shell" in contemporary Chinese.

The semantic transparency of radicals has been found to affect sublexical processing (see Yan et al., 2012, for a review). Adult readers were found to be able to process characters with transparent radicals more efficiently than those with opaque radicals in semantic-categorization and lexical-decision tasks (Chen & Weekes, 2004; Hsiao, Shillcock, & Lavidor, 2007). Shu and Anderson (1997) instructed 292 Chinese children in the first, third, or fifth grade to produce a two-character word from four candidate characters with different semantic radicals but the same phonetic radicals. They found that children performed better when the semantic radicals of the target characters were transparent. Using the proposed Model 1 in this context would be problematic, since nearly 50 % of all semantic radicals are not one-character standalone words. We therefore suggest adopting Model 2 to compute the semantic transparency and consistency of semantic radicals.

Language development The transparency and consistency measures computed by the proposed methods can also be beneficial for educational purposes. Chen, Hao, Geva, Zhu, and Shu (2009) suggested that Chinese children's abilities of vocabulary acquisition and character reading are related to how well they can construct a new compound word from familiar morphemes. The results of the present study might provide useful guidance for designing teaching materials—for example, to first teach children characters with high consistency or a larger morphological family size in order to teach general rules to construct new compounds, and then teach opaque compounds at the whole-word level.

Implications for raters' morphological processing Many issues remain when setting our model parameters, such as: Should constituent position be included in the morphological family? Should the size of morphological families be limited? For the constituents with low consistency, which meaning is activated by a rater? What is the optimal threshold for a given semantic space? Recent work by Crepaldi, Rastle, Davis, and Lupker (2012) has demonstrated evidence for position-independent representations of the constituent morphemes of compounds. Furthermore, it seems possible that position-independent constituents would improve the measurement of transparency in Model 2 by adding extra compounds with common meanings. For example, Model 2 is not presently able to compute a cosine measure for the constituent "nail" in "nailbrush," but presumably this would be possible if the model were to consider the "nail" in "fingernail" or "toenail," in which "nail" occurs in a different position. To address these issues, more empirical work will be required. From the results in the present study, it appears that the way in which transparency judgments are carried out affects the model parameter settings. We imply that the

meanings activated by human raters during transparency judgments may be individually different, and that each rater might have a different threshold for the "cutoff" of opacity.

Limitations and future work

Limitation of Chinese corpus The SP-C used in the present study was built using traditional script, and it will be important to test the compatibility between the traditional and simplified scripts. From the materials in Mok (2009) using simplified script, all high-frequency items except one (due to segmentation standard) were covered in SP-C, but five out of the 95 low-frequency items were not used in traditional scripts. Furthermore, 10 % of simplified characters map to multiple (two to four) traditional characters, which increases morphological or semantic ambiguity (see Tsai, Kliegl, & Yan, 2012). This ambiguity caused by one-to-many mappings between simplified and traditional scripts could be further studied using semantic spaces based on simplified and traditional Chinese corpora.

The current limitations of the proposed method in Chinese might be the relatively small corpus size, which is due to the fact that no spaces appear between words in the Chinese writing system, and thus an automatic word segmentation algorithm is required. Hong and Huang (2006) introduced the Chinese Gigaword Corpus containing 1.1 billion Chinese characters, including 700 million traditional characters from Taiwan's Central News Agency and 400 million simplified characters from China's Xinhua News Agency (all simplified characters were converted into traditional characters). Automatic and partially manual word segmentation were carried out, and the accuracy was estimated to be above 95 %. Cai and Brysbaert (2010) published SUBTLEX-CH, based on a corpus (47 million characters) of film and television subtitles, and they suggested that SUBTLEX-CH is a good estimate of daily language exposure and captures much of the variance in word-processing efficiency. It is possible that Chinese semantic spaces could be established on the basis of those larger corpora.

Semantic similarity measure for computing transparency Other computational alternatives of semantic similarity could also be used. Since LSA requires document information—that is, a set of words that relate to the same topic, a corpus without specific document information (such as a corpus from film subtitles) may turn to the hyperspace analog to language (HAL; Burgess & Lund, 2000; Lund & Burgess, 1996). Similar to LSA, HAL is a semantic space model, but HAL moves an *n*-word window, serving as a document, along a text corpus. An alternative approach, BEAGLE (Jones & Mewhort, 2007), incorporates word order information on top of LSA. Since order information is important for some constituents, BEAGLE might be adopted to compute semantic similarities within the morphological family of a constituent (such as in the examples shown in Tables 2 and 3). Furthermore, Maki, McKinley, and Thompson (2004) provided semantic distance norms derived from WordNet (Fellbaum, 1998; see also Miller, 1990), and they found that these semantic distance measures closely resembled featural similarity and were distinct from LSA. This measure may be suitable for detecting semantic transparency with regard to the exocentric interpretation, such as *shape* information between "seahorse" and "horse," which is inaccessible for LSA (LSA cosine value for this example: .01). Therefore, we suggest that the WordNet-based measure could be integrated into semantic space models to account for a broader range of transparency interpretations.

Model parameters and performance The model parameters in the present study include (1) semantic space, (2) number of dimensions, (3) comparison type in Models 1 and 2, (4) definition of morphological family (position-specific or position-free), (5) definition of dominant meaning, and (6) distance function and threshold of the clustering algorithm. A single set of parameter values was used for each data set reported in this article, and these values were estimated arbitrarily. The optimization issues of LSA have been studied in Lifchitz, Jhean-Larose, and Denhière (2009) regarding optimal tuning of lemmatization, stop-word list, term weighting, pseudodocuments, and normalization of document vectors (see also Shaoul & Westbury, 2010). It is also possible to combine multiple models—for example, by regression analysis or AdaBoost (Freund & Schapire, 1997; see also Bishop, 2006)—to improve model performance.

In the present study, we propose a method to average across subjective differences and transparency ambiguities. Corroborating evidence from two different languages was presented by testing the stimuli used in prior compound-word studies (Frisson et al., 2008; Mok, 2009), as well as ratings obtained for the present study. The results show good prediction (AUCs > .8) for the English compounds of Frisson et al. and for the ratings of Chinese data in the present study, and reasonable predictive performance for the Chinese compounds from Mok's data. Semantic consistency was found to be higher in transparent constituents than in opaque ones in all the materials of our evaluations. The results may be beneficial to psycholinguistic studies.

Author note Portions of the data were presented at the Asia–Pacific Conference on Vision (APCV) 2010, and the first author was awarded an APCV Student Travel Grant. Part of the study was supported by Grant No. R01 EY021802 from the National Institutes of Health (NIH) to M.P. Thanks to Keith Rayner, Jinmian Yang, Marc Brysbaert, and several anonymous reviewers for helpful comments on this study.

Appendix

Table 5 Reanalysis of transparency conditions in Frisson et al. (2008) using LSA

		Initial Constituent					Final Constituent				
ST	Compounds	C1	FS	Co	M1	M2	C2	FS	Co	M1	M2
OO	honeymoon	honey	9	.12	.03	–.07	moon	8	.08	.01	–.07
	hamstring	ham	26	.06	–.01	–.07	string	5	.30	.02	.02
	blockbuster	block	11	.13	.01	.00	buster	3	.04	.01	.01
	pocketbook	pocket	5	.31	.38	.22	book	10	.20	.04	–.02
	cocktail	cock	16	.08	.04	–.03	tail	8	.05	–.04	.08
	network	net	4	.07	.01	–.03	work	30	.08	.08	–.06
	jackpot	jack	13	.07	.1	.08	pot	11	.08	.01	.01
	deadline	dead	8	.11	–.01	–.05	line	60	.08	.13	.11
	pineapple	pine	6	.25	.11	.03	apple	6	.11	.28	.41
	flapjack	flap	4	.46	.09	.01	jack	15	.07	0	–.04
OT	godchild	god	17	.15	N/A	N/A	child	6	.11	N/A	N/A
	buckwheat	buck	15	.11	.14	.01	wheat	3	.05	.37	.33
	crowbar	crow	13	.12	.01	.06	bar	9	.15	.23	.25
	ragweed	rag	11	.13	0	.01	weed	4	.12	.41	.33
	restroom	rest	37	.11	.11	.12	room	31	.24	.17	.12
	ladybug	lady	5	.26	.04	–.06	bug	3	.26	.44	.51
	peanut	pea	6	.19	.11	.74	nut	7	.14	.13	.61
	sandalwood	sandal	3	.13	.02	–.01	wood	32	.07	.09	.09
	horseradish	horse	15	.31	.06	.02	radish	3	.23	.09	.11
	horseplay	horse	15	.31	.06	.01	play	7	.15	.04	–.03
	trenchcoat	trench	3	.28	N/A	N/A	coat	5	.17	N/A	N/A
	dragonfly	dragon	4	.29	.12	.09	fly	10	.14	.43	.57
	peppermint	pepper	6	.19	.26	.40	mint	4	.22	.33	.54
	butterfly	butter	9	.09	.04	–.04	fly	10	.14	.27	.41
TO	chatterbox	chatter	6	.11	.06	.02	box	11	.11	–.07	–.12
	lumberjack	lumber	3	.18	.17	.16	jack	15	.07	.05	–.01
	nightmare	night	17	.18	.32	.20	mare	2	.79	.11	.01
	staircase	stair	6	.34	.57	.69	case	8	.20	.07	.00
	litterbug	litter	4	.21	N/A	N/A	bug	9	.09	N/A	N/A
	gingersnap	ginger	3	.16	N/A	N/A	snap	8	.14	N/A	N/A
	sideburns	side	26	.17	.05	–.05	burns	3	.07	.01	–.01
	warhead	war	68	.09	.1	.08	head	20	.09	–.02	–.02
	doughnut	dough	3	.39	.31	.53	nut	7	.14	.13	.35
	heirloom	heir	4	.33	.17	.07	loom	3	.14	.13	–.06
TT	moonlight	moon	7	.10	.46	.48	light	24	.14	.25	.23
	toothache	tooth	8	.32	.85	.85	ache	4	.11	.14	.09
	dishwasher	dish	9	.14	.52	.67	washer	2	.44	.44	.80
	barnyard	barn	11	.07	.71	.63	yard	14	.12	.34	.14
	cookbook	cook	12	.21	.16	.11	book	10	.20	.34	.33
	farmland	farm	13	.39	.55	.57	land	77	.08	.67	.40
	haystack	hay	13	.06	.43	.60	stack	6	.14	.2	.38
	rainfall	rain	17	.29	.68	.66	fall	12	.08	.17	.40
	honeybee	honey	9	.12	.67	.73	bee	3	.57	.82	.84
	paintbrush	paint	20	.15	.22	.21	brush	4	.21	.33	.35

Table 5 (continued)

		Initial Constituent					Final Constituent				
ST	Compounds	C1	FS	Co	M1	M2	C2	FS	Co	M1	M2
	snowball	snow	14	.51	.65	.62	ball	10	.24	.03	–.02
	gingerbread	ginger	3	.16	.2	.39	bread	2	.32	.32	.30
	nailbrush	nail	4	.57	N/A	N/A	brush	4	.21	N/A	N/A
	teacup	tea	47	.12	.21	.09	cup	4	.25	.13	.00
	songbook	song	2	.89	.1	.10	book	10	.20	.33	.28
	firewood	fire	23	.14	.31	.26	wood	32	.07	.53	.54
	pillbox	pill	11	.07	–.03	–.04	box	11	.11	.11	.12
	toothpaste	tooth	8	.32	.41	.55	paste	6	.21	.42	.72
	meatball	meat	4	.17	.22	.21	ball	10	.24	.01	–.05
	gunshot	gun	15	.13	.13	.08	shot	5	.10	.22	.29
	raincoat	rain	17	.29	.39	.37	coat	5	.17	.36	.45
	flowerpot	flower	8	.27	.12	.10	pot	11	.08	.04	–.14
	clothesline	clothes	4	.32	.15	.01	line	60	.08	.26	.18
	rattlesnake	rattle	4	.34	.39	.71	snake	5	.42	.58	.57
	lumberyard	lumber	3	.18	.18	.13	yard	14	.12	.23	.10
	rainstorm	rain	17	.29	.57	.54	storm	5	.23	.42	.47
	headache	head	37	.10	.19	.16	ache	4	.11	.25	.87
	woodshed	wood	23	.12	.1	.05	shed	2	.04	.27	.20
	sandcastle	sand	22	.12	N/A	N/A	castle	2	.14	N/A	N/A
	mousetrap	mouse	4	.23	.03	.02	trap	4	.08	.19	.15
	hairbrush	hair	16	.26	.11	.11	brush	4	.21	.06	.04
	chessboard	chess	2	.59	.59	.68	board	23	.12	.07	.05
	pocketknife	pocket	5	.31	.09	–.05	knife	4	.28	.23	.18
	mailbag	mail	7	.39	.72	.70	bag	4	.25	.2	.21

For the present and the following appendix, ST is semantic transparency. FS is number of morphological family size, Co is semantic consistency, M1 and M2 are the LSA cosine values calculated by Models 1 and 2, respectively

Table 6 Re-analysis of the materials in Mok (2009)

		Compound			Initial Constituent					Final Constituent				
WF	ST	Word	SP-C	WF	FreqS	FS	Co	M1	M2	FreqS	FS	Co	M1	M2
H	TT	奴隶	奴隸	52	6	6	.13	.26	.96	8	3	.08	.06	.96
		座谈	座談	95	965	9	.08	–.03	–.05	934	32	.07	.14	.07
		统治	統治	299	20	17	.11	.38	.20	200	23	.08	.29	.18
		冬季	冬季	132	117	11	.16	.66	.78	316	16	.16	.29	.85
		校长	校長	769	1,375	28	.21	.66	.75	1,220	87	.07	.03	.12
		纪录	紀錄	366	0	11	.06	N/A	.95	32	20	.09	.06	.15
		晚饭	晚飯	56	180	19	.10	.33	.32	280	16	.39	.77	.78
		形状	形狀	113	82	20	.12	.40	.14	42	27	.09	.17	.28
		震荡	震盪	76	32	16	.10	.02	.02	7	12	.07	.04	.97
		优良	優良	154	19	26	.09	.15	.31	25	14	.05	.06	.11
		真理	真理	191	1,809	32	.10	.18	.93	100	65	.06	.14	.03
		装饰	裝飾	80	203	29	.08	.11	.39	22	18	.14	.26	.26
		解剖	解剖	47	139	38	.07	.09	.00	0	1	N/A	N/A	.98
		保护	保護	893	74	42	.08	.19	.04	52	22	.09	.36	.01

Table 6 (continued)

		Compound			Initial Constituent					Final Constituent				
WF	ST	Word	SP-C	WF	FreqS	FS	Co	M1	M2	FreqS	FS	Co	M1	M2
		包含	包含	345	117	24	.06	.03	.23	189	6	.14	.21	.93
		反抗	反抗	100	444	51	.08	.38	.01	119	5	.18	.00	.16
		收获	收穫	153	359	52	.06	.17	–.02	0	1	N/A	N/A	.93
		文艺	文藝	83	685	62	.07	.12	.46	17	25	.09	.18	.10
		光线	光線	120	355	47	.09	.78	.81	364	67	.08	.06	–.02
		对待	對待	129	13,944	47	.06	.33	.03	294	19	.10	.21	.03
		电动	電動	123	166	57	.09	.04	–.05	474	92	.08	.17	.30
H	OO	矛盾	矛盾	282	26	3	.05	.07	.93	19	4	.03	.04	.93
		伙计	夥計	4	29	4	.12	.12	.06	184	22	.09	.02	–.05
		封建	封建	31	184	14	.05	.07	.01	415	32	.13	.33	.17
		东西	東西	1,726	344	50	.07	.13	.04	382	23	.06	.13	–.07
		营养	營養	160	26	18	.07	.01	–.08	268	31	.10	.20	.09
		影响	影響	1291	46	22	.17	.02	.00	127	10	.11	.01	–.02
		条件	條件	1,044	1,724	13	.07	.07	.03	1,651	27	.09	.06	–.02
		田径	田徑	94	148	12	.12	–.02	–.03	8	15	.07	–.05	–.03
		抽象	抽像	117	216	28	.07	N/A	N/A	80	22	.09	N/A	N/A
		积极	積極	919	47	17	.07	.11	.90	1,195	9	.12	.26	–.02
		客观	客觀	214	47	27	.08	.07	–.02	164	28	.09	.18	.85
		简直	簡直	223	17	27	.09	.20	–.02	323	14	.11	.42	.02
		客气	客氣	125	47	27	.08	.15	.10	491	96	.08	.26	–.02
		先生	先生	2,423	2,465	31	.06	.08	.07	619	99	.06	.07	–.01
		走狗	走狗	8	1,955	47	.11	–.03	–.03	387	10	.17	.06	.05
		方针	方針	74	216	25	.06	.21	–.01	129	8	.05	.02	.01
		难道	難道	291	1,430	44	.09	.19	.03	733	99	.07	.26	.15
		感冒	感冒	70	126	32	.13	.09	.04	86	5	.06	.22	.13
		神经	神經	184	350	65	.09	.04	–.03	1,232	33	.07	.09	–.06
		活该	活該	11	627	23	.06	.09	.00	3,380	3	.09	.05	.04
		马虎	馬虎	13	342	38	.07	–.02	–.05	197	7	.21	.02	.00
		本领	本領	53	3,462	34	.08	.06	.03	165	19	.08	.21	.05
		体会	體會	240	177	40	.09	.08	.00	14,066	77	.08	.27	.00
		行李	行李	77	345	55	.07	.28	.13	620	5	.04	.00	–.04
		手段	手段	237	1476	40	.09	.08	.02	1,174	24	.08	.09	.08
		分寸	分寸	17	1,100	99	.06	.30	–.02	69	4	.02	.04	–.03
		花生	花生	28	623	72	.10	.17	.13	619	99	.06	.09	.01
		出息	出息	9	863	102	.07	.13	.01	17	25	.08	.00	–.07
		发明	發明	166	162	81	.06	.08	.00	162	59	.06	.04	.05
		老婆	老婆	113	1,062	99	.08	.22	.04	7	11	.14	.20	.30
		无聊	無聊	143	2,519	109	.08	.17	.06	88	4	.25	.31	.91
H	TO	态度	態度	1,067	7	3	.07	.01	.91	679	71	.08	–.04	–.10
		摔跤	摔跤	8	79	10	.15	.19	.14	12	2	.03	.03	.00
		或者	或者	1,074	8,317	6	.18	.18	.06	7,221	36	.06	.14	–.01
		晓得	曉得	197	5	4	.08	.17	.85	5,969	71	.07	.22	.01
		熊猫	熊貓	240	156	9	.23	.05	–.02	147	7	.12	.24	.16
		坏蛋	壞蛋	12	339	9	.12	.13	.06	118	11	.11	.12	.05
		介绍	介紹	627	0	9	.06	N/A	.93	0	1	N/A	N/A	.93
		徒弟	徒弟	21	27	9	.06	.03	.01	30	16	.12	.11	.10
		似乎	似乎	1,078	278	4	.41	.46	.17	41	9	.14	.10	.18

Table 6 (continued)

		Compound			Initial Constituent					Final Constituent				
WF	ST	Word	SP-C	WF	FreqS	FS	Co	M1	M2	FreqS	FS	Co	M1	M2
		尾巴	尾巴	116	62	12	.21	.66	.82	15	15	.07	.19	.97
		责备	責備	33	37	10	.11	.16	.07	69	29	.08	.02	–.02
		场合	場合	146	1,615	11	.12	.09	.03	100	45	.08	.13	–.01
		突然	突然	627	46	13	.10	.36	.06	99	87	.10	.13	.25
		字眼	字眼	54	1,060	25	.18	.27	.24	635	37	.08	.10	.00
		如果	如果	5,336	3,142	22	.08	.11	.91	35	23	.08	.06	.06
		领袖	領袖	216	165	32	.06	.05	.14	6	7	.09	.01	.94
		奇迹	奇跡	23	62	26	.08	.15	–.04	13	19	.10	.08	.07
		指头	指頭	22	472	38	.08	.19	.01	834	141	.07	.27	.04
		主席	主席	412	67	71	.06	.03	–.02	49	13	.10	.18	.24
		正经	正經	15	2,573	52	.06	.06	.00	1,232	32	.07	.02	.06
		平等	平等	295	88	63	.07	.20	.20	8,070	27	.08	.15	.05
H	OT	昆虫	昆蟲	137	0	2	.00	N/A	.97	87	13	.27	.51	.83
		仔细	仔細	374	8	3	.14	.20	.90	143	9	.15	.24	.90
		漆黑	漆黑	27	30	5	.12	.00	–.07	415	17	.11	.25	–.02
		跟前	跟前	19	3,539	12	.07	.16	.08	2,944	27	.08	.10	.15
		陆续	陸續	352	47	13	.08	.24	.90	39	14	.10	.15	.01
		幻灯	幻燈機 幻燈片	38	8	9	.11	.10	.05	142	17	.09	.00	.00
		卫星	衛星	274	0	10	.06	N/A	.02	117	38	.18	.30	.62
		汽车	汽車	470	6	7	.06	.31	.97	985	60	.20	.74	.83
		轮船	輪船	29	65	17	.06	.12	.06	468	31	.16	.53	.46
		肥皂	肥皂	38	53	10	.13	.17	.09	0	2	.43	N/A	.96
		太阳	太陽	429	2,893	27	.07	.13	.04	43	24	.06	.11	.06
		目标	目標	1,227	34	12	.09	–.01	.06	21	21	.10	.04	.03
		吃苦	吃苦	30	2,636	28	.15	.21	.15	310	21	.15	.36	.07
		麻烦	麻煩	164	18	14	.06	.10	.88	54	4	.19	.24	.88
		以及	以及	2,511	13,172	27	.12	.27	.03	13,758	29	.09	.23	.08
		方便	方便	621	216	25	.06	.10	.14	2,723	14	.11	.16	.83
		投降	投降	53	127	42	.06	.00	–.06	67	8	.10	.11	–.01
		毛病	毛病	106	297	20	.07	.13	.08	358	25	.18	.34	.25
		血汗	血汗	22	207	31	.12	.06	.00	61	8	.15	.07	.01
		清早	清早	13	153	77	.08	.12	.07	694	19	.08	.18	.07
		老师	老師	2,871	1,062	99	.08	.04	.01	129	39	.08	.53	.80
		打听	打聽	52	1,448	94	.07	.11	.00	1,833	18	.12	.18	.01
L	TT	雇用	雇用	0	13	1	N/A	N/A	N/A	3,045	102	.07	N/A	N/A
		站队	站隊	0	740	12	.08	N/A	N/A	284	45	.11	N/A	N/A
		侵扰	侵擾	11	16	11	.12	.07	.01	11	11	.10	.10	.16
		触犯	觸犯	24	35	19	.05	.02	–.09	168	16	.12	.39	.16
		查询	查詢	242	109	30	.10	.19	–.02	18	10	.12	.08	.01
		试卷	試卷	7	283	26	.07	.17	.25	83	7	.08	.18	–.01
		丰盛	豐盛	38	32	17	.11	.20	.16	70	12	.09	.28	.02
		逃奔	逃奔	0	90	24	.11	N/A	N/A	45	5	.13	N/A	N/A
		环打	環打	0	104	22	.07	N/A	N/A	1,448	42	.11	N/A	N/A
		远程	遠程	26	760	41	.07	.16	.08	72	41	.07	–.01	.00
		尽兴	盡興	28	177	17	.09	.19	.01	19	28	.06	.12	.00
		短枪	短槍	0	383	27	.07	N/A	N/A	136	12	.18	N/A	N/A
		见闻	見聞	31	1,682	22	.08	.05	.00	131	15	.09	.07	.06

Table 6 (continued)

		Compound			Initial Constituent					Final Constituent				
WF	ST	Word	SP-C	WF	FreqS	FS	Co	M1	M2	FreqS	FS	Co	M1	M2
		衣裤	衣褲	8	57	17	.15	.19	.14	4	5	.22	.13	.01
		古稀	古稀	0	263	46	.11	N/A	N/A	27	4	.06	N/A	N/A
		引证	引證	5	92	33	.06	.08	.00	21	30	.07	.06	.05
		后援	後援	12	7,752	55	.06	.11	–.07	9	11	.09	–.02	.73
		石料	石料	0	78	41	.10	N/A	N/A	30	37	.08	N/A	N/A
		火攻	火攻	0	238	42	.09	N/A	N/A	77	13	.16	N/A	N/A
		分身	分身	3	1,100	99	.06	N/A	N/A	1,398	78	.07	N/A	N/A
		中级	中級	37	12,231	121	.05	.11	.26	478	31	.09	.13	–.02
		公议	公議	0	59	98	.06	N/A	N/A	28	22	.14	N/A	N/A
		天际	天際	22	5,038	85	.07	.15	.37	291	19	.06	.18	.01
L	OO	袖珍	袖珍	4	6	4	.11	–.04	.13	4	6	.07	–.02	.02
		洗尘	洗塵	4	255	21	.12	.01	.03	18	20	.11	.06	–.02
		忘年	忘年之交	5	330	8	.16	.07	.07	10,127	61	.10	.15	.03
		端午	端午	5	207	16	.08	.06	.01	17	16	.19	–.01	.06
		造化	造化	10	181	30	.07	.12	.00	47	69	.06	.28	–.06
		垂青	垂青	4	10	11	.07	.06	–.02	80	14	.04	.14	–.16
		烧卖	燒賣	0	167	17	.10	N/A	N/A	557	15	.12	N/A	N/A
		便当	便當	118	2,723	16	.08	.05	–.01	2,486	50	.10	.07	.01
		粉刺	粉刺	13	27	11	.13	.18	.25	73	7	.06	.07	.06
		乌贼	烏賊	4	6	14	.07	.01	.05	56	7	.13	.07	.04
		麻将	麻將	21	18	14	.06	.15	–.10	7,858	22	.09	.04	.05
		利落	俐落	25	0	1	N/A	N/A	N/A	117	50	.08	.24	.02
		物色	物色	11	254	21	.06	.00	–.03	174	79	.12	.04	.01
		耳光	耳光	13	119	14	.08	.14	.17	355	74	.08	.02	–.03
		云雨	雲雨	4	137	17	.09	.19	.08	297	24	.17	.12	.05
		木耳	木耳	2	80	47	.07	N/A	N/A	119	18	.17	N/A	N/A
		百合	百合	26	404	11	.05	N/A	.00	100	45	.08	.06	.03
		红颜	紅顏	3	481	45	.06	N/A	N/A	13	4	.06	N/A	N/A
		风流	風流	19	434	59	.08	.15	.10	106	55	.08	.21	.13
		花甲	花甲	1	623	72	.10	N/A	N/A	95	11	.05	N/A	N/A
		开交	不可開交	12	1,023	114	.06	.16	.06	135	24	.07	–.04	.11
		发指	髮指	2	26	4	.18	N/A	N/A	472	15	.10	N/A	N/A
L	TO	蝇头	蠅頭小利	2	5	1	N/A	N/A	N/A	834	141	.07	N/A	N/A
		赖皮	賴皮	6	40	7	.10	.11	.07	102	35	.08	.09	.00
		辈份	輩份	6	49	5	.14	.32	.10	824	24	.06	.04	–.11
		辞令	辭令	4	21	11	.08	.06	.08	1,814	24	.07	.08	.07
		登基	登基	27	38	21	.07	.23	–.02	11	16	.06	.26	.13
		使节	使節	10	4,645	14	.08	.06	.01	193	29	.06	.00	–.04
		紧凑	緊湊	18	109	21	.09	.05	–.05	41	3	.01	.00	–.04
		溜达	溜達	0	22	6	.11	N/A	N/A	748	38	.06	N/A	N/A
		拜堂	拜堂	3	111	14	.08	N/A	N/A	78	26	.08	N/A	N/A
		问津	問津	8	2,103	24	.09	.00	.01	4	10	.04	.02	.11
		饱和	飽和	35	51	8	.07	.01	.09	13,585	35	.06	.12	.04
		穷蛋	窮光蛋	7	143	5	.20	.38	.29	118	11	.11	.17	.11
		牙床	牙床	5	58	12	.31	.31	.56	254	16	.08	.07	.03
		眉宇	眉宇	5	53	11	.15	.14	.06	0	7	.05	N/A	–.06
		败北	敗北	14	123	14	.15	.57	.72	379	17	.10	–.01	.01

Table 6 (continued)

		Compound			Initial Constituent					Final Constituent				
WF	ST	Word	SP-C	WF	FreqS	FS	Co	M1	M2	FreqS	FS	Co	M1	M2
		私房	私房	1	33	27	.06	N/A	N/A	178	34	.08	N/A	N/A
		劳顿	勞頓	1	16	18	.17	N/A	N/A	144	24	.09	N/A	N/A
		悲切	悲切	0	42	16	.19	N/A	N/A	78	18	.11	N/A	N/A
		书香	書香	33	1,600	38	.17	.55	.61	100	24	.11	.11	.04
		名声	名聲	33	2,263	53	.07	.06	.01	480	72	.12	.03	–.04
		安顿	安頓	20	48	50	.06	.12	–.03	144	24	.09	.03	–.03
		同窗	同窗	5	950	47	.07	.02	.31	219	10	.08	.03	–.03
		外快	外快	9	1,781	91	.06	.12	–.01	1,192	18	.10	.17	.12
		水母	水母	11	1,445	98	.11	.06	.06	86	18	.09	.06	–.15
L	OT	盲肠	盲腸	5	26	9	.14	–.05	–.06	13	8	.15	.18	–.02
		径自	逕自	11	38	4	.20	.04	–.02	1,580	18	.09	.15	.10
		仓促	倉促	16	9	8	.06	.03	.06	22	9	.08	.00	.06
		午夜	午夜	52	17	10	.15	.13	.12	374	24	.12	.31	.10
		掌故	掌故	8	46	12	.07	–.02	–.05	684	16	.08	.07	.05
		寻常	尋常	28	120	13	.12	.23	–.03	2,022	22	.13	.16	.07
		笔挺	筆挺	3	380	21	.12	N/A	N/A	109	2	.08	N/A	N/A
		格言	格言	11	57	12	.06	.13	–.01	1,568	44	.07	.24	.05
		软片	軟片	43	105	13	.05	.02	–.05	987	53	.09	.05	.07
		处女	處女	13	913	21	.10	.00	.02	790	44	.10	.43	.03
		脚本	腳本	20	442	19	.11	.02	.00	3,462	40	.06	.09	–.01
		云雀	雲雀	3	137	17	.09	N/A	N/A	6	5	.20	N/A	N/A
		比邻	比鄰	11	2,475	26	.07	.03	–.02	18	8	.10	.16	–.01
		马步	馬步	2	342	38	.07	N/A	NA	567	30	.08	N/A	NA
		放晴	放晴	4	1,033	54	.07	.09	–.01	33	7	.06	.29	.25
		洋灰	洋灰	0	12	16	.07	N/A	N/A	73	6	.07	N/A	N/A
		起诉	起訴	33	1,158	48	.06	.02	.02	8	13	.16	.33	.89
		可取	可取	12	8,508	46	.08	.06	–.03	361	42	.07	.08	.02
		海报	海報	141	625	76	.16	.02	–.03	335	47	.07	.22	–.04
		当今	當今	96	2,486	47	.08	.09	.01	471	8	.24	.24	.13
		正法	正法	1	2,573	52	.06	N/A	N/A	754	86	.07	N/A	N/A
		成语	成語	50	1,284	65	.06	.24	.00	108	54	.09	.22	.05
		气质	氣質	153	491	36	.08	.06	–.01	120	48	.08	.17	.14
		风尘	風塵	10	434	59	.08	.20	.06	18	20	.11	.07	–.04
		白事	白事	0	596	53	.07	N/A	N/A	4,008	106	.06	N/A	N/A
		打滚	打滾	26	1,448	94	.07	.25	.16	41	5	.14	.32	.19

For the present and the following appendix, WF represents high (H) or low (ST) frequent items. SP-C is the traditional-script conversion of compounds in SP-C. WF and FreqS are the counts of whole words and constituents in ASBC (5 million words), in which FreqS is counted for one-character words and does not include occurrences in other multi-character words

Table 7 Eighty compounds selected in the present study

		Compound		Glossary			Initial Constituent						Final Constituent					
ST		Word	WF	Word	Initial	Final	Rate	FreqS	FS	Co	M1	M2	Rate	FreqS	FS	Co	M1	M2
T	T	筆鋒	5	writing style	pen	cutting edge	.64	509	12	.14	.25	.35	.64	118	3	.61	.03	.12
T	T	預約	67	reserve	in advance	appointment	.64	179	52	.11	.05	.19	1.00	602	22	.09	.06	–.02
T	T	夜市	68	night market	night	market	.82	516	14	.19	.05	.08	.91	715	22	.09	.16	.23
T	T	末期	70	last phase	last	phase	.91	128	7	.13	.24	.75	.64	939	39	.10	.04	.29
T	T	見面	204	meet	see	face	.91	1,464	12	.10	.31	.37	.64	2,838	65	.08	.36	.11
T	T	實話	54	truth	true	word	.82	1,287	30	.09	.18	.15	.64	1,466	23	.13	.22	.24
T	T	補救	49	redeem	mend	save	1.00	139	12	.09	.20	.34	.91	296	10	.12	.00	–.02
T	T	毆打	43	beat up	beat up	hit	.91	5	1	N/A	.40	.93	1.00	1,441	10	.12	.15	.13
T	T	歉意	24	apology	apology	opinion/ wish	1.00	52	2	.07	.07	.96	.73	1,910	61	.09	.17	.25
T	T	窮苦	20	poverty-stricken	poor	bitter	1.00	154	3	.37	.19	.32	.82	578	16	.19	.18	.27
T	T	羞怯	14	shy	shame/ shy	fear/ timid	1.00	76	4	.11	.05	.07	1.00	15	3	.03	N/A	.97
T	T	乾燥	62	dry	dry	dry	1.00	297	11	.12	.37	.54	1.00	38	2	.05	–.05	.96
T	T	開門	61	open a door	open	door	.64	2,248	144	.07	.28	.16	1.00	836	41	.08	.47	.41
T	T	起點	59	starting point	start	point	.73	3,046	29	.07	.21	.27	.64	1,745	54	.09	.06	.16
T	T	擴散	41	spread	expend	loose/ leisurely	.91	72	8	.13	.01	.03	1.00	217	10	.07	.10	.00
T	T	帳篷	38	tent	tent/ account	fluffy	.64	57	8	.10	.07	.04	.91	21	2	.08	N/A	.97
T	T	跑步	35	run	run	step	1.00	573	13	.19	.43	.50	.73	575	19	.11	.19	.25
T	T	邪惡	29	evil	evil	wickedness	1.00	28	2	.20	.20	.90	1.00	172	8	.15	.14	.40
T	T	進來	190	come in	move forward	come	.91	1,691	60	.09	.56	.26	1.00	8,247	68	.12	.37	.42
T	T	球場	167	ball court	ball	court	.82	1,141	15	.38	.56	.70	.82	710	39	.09	.39	.37
T	T	地形	142	terrain	earth	shape	.82	4,773	46	.10	.17	.53	.91	1,888	24	.09	.13	.14
T	T	產量	114	production	produce	quantity	.91	644	16	.11	.58	.40	.82	1,022	45	.10	.13	.13
T	T	貝殼	59	shell	shellfish	shell	.91	280	6	.07	.37	.14	.91	203	4	.19	.01	–.01
T	T	射箭	45	shoot an arrow	shoot	arrow	.82	284	8	.16	.26	.65	.82	167	3	.15	.28	.75
T	T	吊橋	20	drawbridge	hang	bridge	.64	49	1	N/A	.03	.00	.91	167	8	.27	.42	.46
T	T	診療	8	diagnosis	diagnose	medical care	1.00	30	3	.43	N/A	.57	.73	62	3	.28	N/A	.39
T	T	笑臉	33	smiling face	smile	face	1.00	694	11	.23	.47	.46	.64	375	5	.17	.56	.50
T	T	鳴叫	16	cry of insects or birds	cry of insects or birds	(of animals) to call	1.00	81	6	.29	.26	.84	1.00	1,098	8	.22	.12	.23
T	T	花季	10	flowering season	flower	season	.73	1,901	35	.14	.28	.31	.73	227	11	.21	.18	.28
T	T	旱災	10	drought	drought	disaster	1.00	24	1	N/A	.51	.75	.91	181	7	.23	.18	.34
T	T	陰涼	5	shady and cool	moon/ hidden/ negative	cool	.73	137	18	.13	.13	.04	1.00	118	6	.24	.24	.27
T	O	智商	15	IQ	intelligence	commerce	.91	177	11	.11	.04	.01	.09	270	27	.08	–.01	.01
T	O	啟迪	15	inspire	open	progress	.82	124	11	.07	.04	.42	.09	109	3	.02	N/A	.02
T	O	字母	36	letter	character	mother	.82	1,553	18	.18	.50	.65	.09	1,159	13	.09	.06	–.02
T	O	質地	30	texture	quality / substance	earth	.91	359	10	.08	.28	.00	.00	4,773	68	.10	.15	.09
T	O	婉約	7	graceful	gentle	appointment	.82	21	2	.04	N/A	.20	.09	602	22	.09	.07	.05
T	O	同樣	742	equal	same	appearance/ type	1.00	2,330	35	.08	.31	.35	.27	2,646	22	.11	.29	.37

Table 7 (continued)

ST		Compound		Glossary			Initial Constituent						Final Constituent					
		Word	WF	Word	Initial	Final	Rate	FreqS	FS	Co	M1	M2	Rate	FreqS	FS	Co	M1	M2
T	O	水面	66	water surface	water	surface	.82	2,919	61	.14	.55	.56	.36	2,838	65	.08	.24	.29
O	T	垂死	21	almost dead	hanging	dead	.09	126	3	.06	.08	–.03	1.00	668	22	.13	.28	.29
O	T	追悼	11	commemorate	chase	mourn	.36	224	23	.10	.01	–.02	.91	4	2	.42	N/A	.77
O	T	宿命	26	predestination	lodge	life	.27	40	3	.05	.08	–.05	.91	650	26	.10	.32	.34
O	T	會計	97	accounting	meet/be good at/ association	calculation	.18	4,776	21	.11	.08	.00	.73	616	16	.12	.05	.03
O	T	主意	79	idea	master	meaning	.18	1,508	45	.07	.24	.05	.82	1,910	61	.09	.24	.21
O	T	開水	39	boiled water	open	water	.18	2,248	144	.07	.13	.05	.91	2,919	56	.13	.26	.23
O	T	松鼠	23	squirrel	pine	mouse	.18	130	4	.07	.15	.59	1.00	287	6	.13	.19	.33
O	T	充飢	8	alleviate one's hunger	fill/ full	hungry	.36	352	11	.14	.14	.05	.64	38	6	.22	.15	.71
O	T	腹案	11	a plan in one's mind	belly	plan/ case/ file/ table	.27	95	5	.17	.03	.03	.82	313	30	.16	.09	.13
O	O	和尚	169	monk	peace	yet	.00	5,193	14	.08	.03	–.02	.00	133	6	.09	.11	.08
O	O	便當	118	lunch box	convenient	think/ equal/ appropriate	.18	969	10	.13	.05	–.02	.00	1,739	14	.12	.07	.01
O	O	道地	33	genuine	way/ speak	earth / ground	.00	2,287	36	.12	.12	.08	.09	4,773	68	.10	.10	.06
O	O	打點	21	get ready	hit	point	.09	1,441	63	.09	–.08	.34	.00	1,745	54	.09	–.01	–.03
O	O	杯葛	27	boycott	cup	a Chinese surname	.00	202	4	.27	.02	.00	.09	45	4	.07	.02	.06
O	O	沙拉[a]	22	salad	sand	pull	.00	367	15	.14	.02	.03	.00	627	5	.03	.02	–.05
O	O	員外	13	rich landlord	member	outside	.00	444	5	.09	.00	.00	.00	1,624	26	.12	.04	–.03
O	O	滑稽	13	funny	slippery/ smooth	check/ examine	.09	193	10	.10	.16	.09	.18	12	5	.05	N/A	–.06
O	O	休克	11	shock	rest	subdue/ gram	.09	284	9	.11	–.02	.20	.00	492	15	.07	.06	.03
O	O	張羅	11	prepare	spread/ sheet	net/ gross	.18	1,026	18	.05	.05	.02	.27	410	6	.06	–.02	–.06
O	O	摩登[a]	10	modern	rub/ touch	mount/ publish	.09	293	2	.06	N/A	–.09	.09	155	4	.09	.00	.03
O	O	行頭	10	costume/ wardrobe	go/carry out/ conduct	head	.18	1,952	32	.08	.14	.09	.09	2,234	81	.09	.07	.10
O	O	郎中	7	physician	young man	middle/ inside	.36	234	2	.17	–.08	–.02	.00	5,339	32	.08	.14	.11
O	O	壽司[a]	5	sushi	age	in charge of	.00	76	3	.04	.06	.00	.00	159	3	.07	.01	–.05
O	O	將就	5	put up with	general	nearby/ engage	.09	1,427	5	.19	.03	–.05	.18	5,516	5	.19	.15	.02
O	O	凱旋	16	triumphant return	victorious/ name	revolve/ rotate	.18	73	4	.03	*N*/A	.98	.09	121	7	.08	.08	.00
T	.	無能	59	incompetent	without	capability/ energy	.73	1,191	69	.11	.13	.20	.55	3,396	29	.08	.13	.11
T	.	進度	155	progress	advance	degree	.64	1,691	60	.09	.11	.35	.55	1,202	53	.09	.04	.12
T	.	好處	305	benefit	good	place	.91	5,553	40	.11	.22	.23	.45	942	26	.11	.04	.20
T	.	美學	84	aesthetics	beauty	learn/ knowledge	1.00	1,730	40	.08	.22	.17	.45	4,470	51	.10	.02	.07
T	.	住所	25	residence	living	place/ that which	1.00	1,025	13	.20	.31	.25	.45	2,587	13	.08	.15	.10
T	.	旅社	18	hotel	traveling	organization	.73	311	13	.18	.10	.14	.55	373	10	.09	–.06	–.03
O	.	調皮	26	mischievous	mix/ adjust	skin/ fur/ naughty	.00	391	28	.08	.07	–.04	.45	521	18	.09	.14	–.04
O	.	生氣	283	angry/ vital	produce/ birth	gas/ air/ anger	.00	5,444	40	.09	.26	.13	.45	2,276	57	.09	.45	.37
O	.	姑息	5	over-tolerate	aunt	rest	.09	298	4	.08	.16	.25	.45	387	14	.11	–.02	–.11
.	T	部長	196	minister	part/ military unit	person in charge/ grow	.55	1,663	26	.12	.10	.02	.64	3,374	66	.08	.01	.28
.	T	呆帳	23	bed debt	slow-witted/ stay	account	.45	83	3	.15	.01	–.01	.82	57	5	.15	.26	.78

Table 7 (continued)

ST	Compound		Glossary			Initial Constituent					Final Constituent				
	Word	WF	Word	Initial	Final	Rate	FreqS	FS	Co	M1	M2	Rate	FreqS	FS	Co	M1	M2
. T	茶杯	31	tea cup/ cup	tea	cup	.55	193	17	.51	.87	.87	.91	202	3	.38	.74	.78
. T	飯碗	20	rice bowl/ job	rice	bowl	.55	313	12	.30	.29	.37	.64	61	3	.21	.25	.42
. T	理想	450	ideal	reason/ logic/ manage	think	.45	1,078	17	.12	.20	.53	.73	2,695	18	.13	.20	.40
. O	夥計	4	buddy/ waiter	partner/ plenty	calculation/ idea	.55	57	2	.05	.12	.06	.27	616	16	.12	.02	.00
. O	沈著	11	composed/ calm	sink/ heavy	a move/ step	.45	26	14	.09	.09	.43	.00	4,398	26	.10	.18	.08
. .	狼狽	11	embarrassed	wolf	a kind of wolf	.55	149	1	N/A	.04	.00	.55	4	1	N/A	N/A	.97

Rate indicates the proportions of responses from 11 raters who responded "T." [a] Bisyllabic words (two characters in written form)
Period in ST indicates that the Rate of the constituent is greater than 0.4 and less than 0.6.

References

Academia Sinica (1998). Academia Sinica Balanced Corpus (Version 3) [Electronic database]. Taipei, Taiwan.

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. doi:10.1006/jmla.1997.2509

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York, NY: Springer.

Burgess, C., & Lund, K. (2000). The dynamics of meaning in memory. In E. Dietrich & A. B. Markman (Eds.), *Cognitive dynamics: Conceptual and representational change in humans and machines* (pp. 117–156). Mahwah, NJ: Erlbaum.

Butterworth, B. (1983). Lexical representation. In B. Butterworth (Ed.), *Language production (Vol. II): Development, writing and other language processes* (pp. 257–294). London, UK: Academic Press.

Bybee, J. L. (1988). Morphology as lexical organization. In M. Hammond & M. Noonan (Eds.), *Theoretical morphology: Approaches in modern linguistics* (pp. 119–141). London, UK: Academic Press.

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS One, 5*, e10729. doi:10.1371/journal.pone.0010729

Chen, T. M., & Chen, J. Y. (2006). Morphological encoding in the production of compound words in Mandarin Chinese. *Journal of Memory and Language, 54*, 491–514.

Chen, X., Hao, M., Geva, E., Zhu, J., & Shu, H. (2009). The role of compound awareness in Chinese children's vocabulary acquisition and character reading. *Reading and Writing, 22*(5), 615–631.

Chen, M. L., Wang, H. C., & Ko, H. W. (2009). The construction and validation of Chinese semantic space by using latent semantic analysis [In Chinese]. *Chinese Journal of Psychology, 51*, 415–435.

Chen, M. J., & Weekes, B. S. (2004). Effects of semantic radicals on Chinese character categorization and character decision. *Chinese Journal of Psychology, 46*, 179–195.

Crepaldi, D., Rastle, K., Davis, C. J., & Lupker, S. J. (2012). Seeing stems everywhere: Position-independent identification of stem morphemes. *Journal of Experimental Psychology. Human Perception and Performance, 39*, 510–525.

Dennis, S. (2007). How to use the LSA website. In T. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 57–70). Mahwah, NJ: Erlbaum.

Diependaele, K., Duñabeitia, J. A., Morris, J., & Keuleers, E. (2011). Fast morphological effects in first and second language word recognition. *Journal of Memory and Language, 64*, 344–358. doi:10.1016/j.jml.2011.01.003

Feldman, L. B., Basnight-Brown, D., & Pastizzo, M. J. (2006). Semantic influences on morphological facilitation, concreteness and family size. *Mental Lexicon, 1*, 59–84.

Feldman, L. B., & Soltano, E. G. (1999). Morphological priming: The role of prime duration, semantic transparency, and affix position. *Brain and Language, 68*, 33–39.

Feldman, L. B., Soltano, E. G., Pastizzo, M. J., & Francis, S. E. (2004). What do graded effects of semantic transparency reveal about morphological processing? *Brain and Language, 90*, 17–30.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences, 55*, 119–139.

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.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psycho physics*. New York, NY: Wiley.

Hong, J.-F., & Huang C.-R. (2006, November). *Using Chinese Gigaword Corpus and Chinese Word Sketch in linguistic research*. Article presented at the 20th Pacific Asia Conference on Language, Information and Computation (PACLIC-20). Wu-Han, China.

Hsiao, J. H., Shillcock, R., & Lavidor, M. (2007). A TMS examination of semantic radical combinability effects in Chinese character recognition. *Brain Research, 1078*, 159–167.

Huang, C. R., Chen, K. J., Chen, F. Y., & Chang, L. L. (1997). Segmentation standard for Chinese natural language processing. *Computational Linguistics and Chinese Language, 2*, 47–62.

Hung, D. L., Tzeng, O. J. L., & Chen, S. Z. (1993). Activation effects of morphology in Chinese lexical processing [In Chinese]. *World of Chinese Language, 69*, 1–7.

Inhoff, A. W., Starr, M. S., Solomon, M., & Placke, L. (2008). Eye movements during the reading of compound words and the influence of lexeme meaning. *Memory & Cognition, 36*, 675–687. doi:10.3758/MC.36.3.675

Janssen, N., Bi, Y. C., & Caramazza, A. (2008). A tale of two frequencies: Determining the speed of lexical access for Mandarin Chinese and English compounds. *Language & Cognitive Processes, 23*, 1191–1223.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review, 114*, 1–37. doi:10.1037/0033-295X.114.1.1

Juhasz, B. J. (2007). The influence of semantic transparency on eye movements during English compound word recognition. In R. van Gompel, W. Murray, & M. Fischer (Eds.), *Eye movements: A window on mind and brain* (pp. 373–389). Oxford, UK: Elsevier.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211–240. doi:10.1037/0033-295X.104.2.211

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2007). *Handbook of latent semantic analysis*. Mahwah, NJ: Erlbaum.

Lee, C. Y. (1995). *The representation of semantically transparent and opaque words in mental lexicon* [In Chinese]. Unpublished master's thesis, National Chung Cheng University, Chia-Yi, Taiwan.

Lee, P. J. (2007). *The representation of semantically transparent and opaque words in mental lexicon: Evidence from eye movements* [In Chinese]. Unpublished master's thesis, National Chung Cheng University, Taipei, Taiwan.

Libben, G., Gibson, M., Yoon, Y. B., & Sandra, D. (2003). Compound fracture: The role of semantic transparency and morphological headedness. *Brain and Language, 84*, 50–64.

Lifchitz, A., Jhean-Larose, S., & Denhière, G. (2009). Effect of tuned parameters on an LSA multiple choice questions answering model. *Behavior Research Methods, 41*, 1201–1209. doi:10.3758/BRM.41.4.1201

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers, 28*, 203–208. doi:10.3758/BF03204766

Maki, W. S., McKinley, L. N., & Thompson, A. G. (2004). Semantic distance norms computed from an electronic dictionary (WordNet). *Behavior Research Methods, Instruments, & Computers, 36*, 421–431. doi:10.3758/BF03195590

Martin, D. I., & Berry, M. W. (2007). Mathematical foundations behind latent semantic analysis. In T. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 35–55). Mahwah, NJ: Erlbaum.

Miller, G. A. (Ed.). (1990). WordNet: An on-line lexical database [Special issue]. *International Journal of Lexicography, 3*(4).

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

Pollatsek, A., & Hyönä, J. (2005). The role of semantic transparency in the processing of Finnish compound words. *Language and Cognitive Processing, 20*, 261–290.

Quesada, J. (2007). Creating your own LSA spaces. In T. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 71–88). Mahwah, NJ: Erlbaum.

Rastle, K., Davis, M. H., Marslen-Wilson, W. D., & Tyler, L. K. (2000). Morphological and semantic effects in visual word recognition: A time-course study. *Language & Cognitive Processes, 15*, 507–537. doi:10.1080/01690960050119689

Rayner, K. (1998). Eye movement in reading and information processing: 20 years of research. *Psychological Bulletin, 24*, 372–422. doi:10.1037/0033-2909.124.3.372

Rayner, K. (2009). The 35th Sir Frederick Bartlett Lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology, 62*, 1457–1506.

Rayner, K., Li, X., & Pollatsek, A. (2007). Extending the E-Z Reader model of eye movement control to Chinese readers. *Cognitive Science, 31*, 1021–1033.

Reicher, G. M. (1969). Perceptual recognition as a function of meaningfulness of stimulus material. *Journal of Experimental Psychology, 81*, 275–280.

Schreuder, R., & Baayen, R. H. (1997). How complex simplex words can be. *Journal of Memory and Language, 37*, 118–139.

Shaoul, C., & Westbury, C. (2010). Exploring lexical co-occurrence space using HiDEx. *Behavior Research Methods, 42*, 393–413. doi:10.3758/BRM.42.2.393

Shu, H., & Anderson, R. C. (1997). Role of radical awareness in the character and word acquisition of Chinese children. *Reading Research Quarterly, 32*, 78–89.

Taft, M. (1981). Prefix stripping revisited. *Journal of Verbal Learning and Verbal Behavior, 20*, 289–297.

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

Tsai, C.-H. (1994). *Effects of semantic transparency on the recognition of Chinese two-character words: Evidence for a dual-process model* [In Chinese]. Unpublished master's thesis, National Chung Cheng University, Chia-Yi, Taiwan.

Tsai, J.-L., Kliegl, R., & Yan, M. (2012). Parafoveal semantic information extraction in traditional Chinese reading. *Acta Psychologica, 141*, 17–23.

Wang, H.-C., Pomplun, M., Ko, H. W., Chen, M. L., & Rayner, K. (2010). Estimating the effect of word predictability on eye movements in Chinese reading using latent semantic analysis and transitional probability. *Quarterly Journal of Experimental Psychology, 63*, 1374–1386.

Wheeler, D. D. (1970). Processes in word recognition. *Cognitive Psychology, 1*, 59–85.

Yan, G., Tian, H., Bai, X., & Rayner, K. (2006). The effect of word and character frequency on the eye movements of Chinese readers. *British Journal of Psychology, 97*, 259–268.

Yan, M., Zhou, W., Shu, H., & Kliegl, R. (2012). Lexical and sublexical semantic preview benefits in Chinese reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 1069–1075.

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

Zhou, X., Marslen-Wilson, W., Taft, M., & Shu, H. (1999). Morphology, orthography, and phonology in reading Chinese compound words. *Language & Cognitive Processes, 14*, 525–565.

Zhou, X., Ye, Z., Cheung, H., & Chen, H.-C. (2009). Processing the Chinese language: An introduction. *Language & Cognitive Processes, 24*, 929–946.