Estimating Semantic Transparency of Constituents of English Compounds and Two-Character Chinese Words using Latent Semantic Analysis

Hsueh-Cheng Wang (hchengwang@gmail.com)
Li-Chuan Hsu (lchsu@mail.cmu.edu.tw)
Yi-Min Tien (tien@mercury.csmu.edu.tw)
Marc Pomplun (marc@cs.umb.edu)

Department of Computer Science, University of Massachusetts at Boston,
100 Morrissey Boulevard, Boston, MA 02125 USA

Abstract

The constituents of English compounds (e.g., butter and fly for butterfly) and two-character Chinese words may differ in meaning from the whole word. Furthermore, the meanings of the words containing the same constituent (e.g., butter in “butterfingers”, or “buttermilk”) may or may not be consistent. Estimating semantic transparency of a constituent is usually difficult and subjective because of these uncertainties and ambiguities. It is rather unexplored why a constituent is considered transparent/opaque by raters, and how its polysemy correlates to its transparency. We propose a computational method for predicting semantic transparency based on Latent Semantic Analysis. We computed the primary meaning of a constituent by a clustering analysis and compared it to the whole-word meaning. The proposed method successfully predicted participants’ transparency ratings, and may explain the cognitive processes in raters when classifying semantic transparency of English compounds and two-character Chinese words.

Keywords: compound words; semantic transparency; latent semantic analysis; Chinese; clustering.

Introduction

A compound word is a word composed of at least two free lexemes that refer to a new concept. Compound words with two constituents are defined as semantically transparent (transparent-transparent, referred to as TT, see Frisson, Niswander-Klement, & Pollatsek, 2008) when the whole word meaning can be grasped through its individual constituents, such as *cookbook.* Compound words are regarded as semantically opaque (opaque-opaque, OO), when their meaning cannot be fully derived from its constituents, e.g., *cocktail*. Some compound words are considered partially opaque (opaque-transparent, OT, or transparent-opaque, TO) when the primary meaning of one of the constituents is related to the meaning of the compound, such as *butterfly* or *staircase*, respectively.

Typically, transparency ratings are the most common method to obtain transparency information. Transparency rating experiments have used target words that differed substantially in their estimated transparency by researchers or a group of participants. 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 participants to rate these words regarding their transparency using a 7-point scale (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. Similarly, Frisson et al. (2008) asked 40 participants to rate transparency in terms of appropriate categories (e.g., opaque-transparent (OT), transparent-opaque (TO), opaque-opaque (OO), and transparent-transparent (TT)), and there was good agreement between the participants’ choices and the predefined classification by Frisson et al. (2008). The proportion of participants’ choices agreeing with the predefined classification was 65% for OO, 71% for OT, 65% for TO, and 86% for TT. Moreover, the proportion of participants classifying at least one of the constituents as opaque for the predefined opaque words was very high: 95% for OO, 93% for OT and 95% for TO. Inhoff et al. (2008) selected “headed” and “tailed” compound words, i.e., compound words whose meaning was primarily defined by their first or second constituents, respectively. They had 13 participants rate 390 compound words using an 11-point scale ranging from 0 to 10, where 0 indicated that the meaning of the compound was solely associated with the meaning of the first constituent, while 10 denoted that the meaning of the compound was solely associated with the one of the second constituent. Compounds with mean ratings below 4 (mean: 3.34) or above 6 (mean: 7.18) were considered to be headed or tailed, respectively. It is important to notice that the definition of headed and tailed compound words might not equal the TO and OT conditions discussed above. For example, the second constituent of a headed compound may be opaque or transparent, as long as its meaning is less closely related to the compound than the meaning of the first constituent is.

Two-character Chinese words, similar to English compound words, differ in how the meanings of the first and second characters relate to the meaning of the word. Some two-character Chinese words are semantically transparent, i.e., both characters are transparently related to the meaning of the whole word. Other words are fully opaque, i.e., 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.

According to the estimation of Zhou and Marslen-Wilson (1995), 74% of Chinese words are made up of two characters, although some words consist of only one character and some consist of three or more characters. A Chinese character is a writing unit which has a single syllable and meaning(s). It is approximately equal to a morpheme in most cases. However, 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 might somewhat disagree where word boundaries are located (see Rayner, Li, & Pollatsek, 2007, for a review). 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 constitute one-character words. Furthermore, a Chinese character might be shared by many words, but the meaning of the character and those words may not be consistent.

Table 1. Examples of transparent, opaque, and partially opaque Chinese words. The meaning of the whole word and the primary meanings of 1st and 2nd characters are shown in parentheses.

	Whole Word	1st Character	2nd Character
TT	球場 (ball court)	球 (ball)	場 (court)
OO	壽司 (sushi)	壽 (age)	司 (in charge of)
TO	智商 (I.Q.)	智 (Intelligent)	商 (commerce)
OT	開水(boiled water)	開 (open)	水 (water)

Early studies of the morphological processing of Chinese polymorphemic words asked how compound words are represented in the mental lexicon and how their lexical processing in visual or auditory word recognition is performed. Recent studies investigated semantic composition (Mok, 2009) and frequency effects (see Zhou, Ye, Cheung, & Chen, 2009, for a review). In Mok (2009), the experimenter pre-defined semantic transparency on a 6-point scale, where 1 is opaque and 6 is transparent. A constituent was classified transparent if the rating was equal to or greater than 4, and opaque otherwise. Five participants were then provided the 6-point scale again for each constituent. Constituents with an average rating greater than 3.5 were classified as transparent, and the others were categorized as opaque.

There are also a few unpublished studies attempting to estimate semantic transparency of Chinese two-character words by researchers or human raters such as Tsai (1994), Lee, C. Y. (1995), and Lee, P. J. (2007). For example, a five-point scale ranging from 1 to 5 was used in the study by Lee (2007), and words were considered transparent when the average score was below 2 while opaque when the average score was greater than 4. Tsai (1994) categorized opaque words into OT and TO conditions, but Lee (1995) and Lee (2007) generalized OT, TO, or OO conditions as opaque words (referred to as idiomatic words).

Unfortunately, estimates of semantic transparency are often subjective and vary across raters, and sometimes even the meaning of transparent compounds cannot be unambiguously determined from the meanings of their constituents (see Frisson et al., 2008). Inhoff et al. (2008) pointed out 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 topic was also studied in the literature on conceptual combination (e.g., Wisniewski, 1996; Costello & Keane, 2000), which indicated that one part of a compound has an *exocentric interpretation* such as *shape* (“seahorse” is a fish whose head is the *shape* of a horse’s head) or the head concept (in the “seahorse” case the diagnostic predicate being shape). Participants might be able to interpret constituents being defined as opaque to a meaning related to the compound according to some kinds of relation (e.g., shape) or the polysemy of the constituent and compound. This subjectivity and variability also occurs in characters of Chinese two-character words. Therefore, a computational model may be a way to average across subjective differences of estimating semantic transparency.

Predicting Transparency using Latent Semantic Analysis

This study proposes a computational method for estimating transparency using Latent Semantic Analysis (LSA). LSA is a method to represent the meaning of words by statistical computations applied to a text corpus (Landauer & Dumais, 1997; Landauer, McNamara, Dennis, & Kintsch, 2007). Typically, 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 similiarity values for any two terms in a given language using the LSA cosine value, which ranges between -1 and 1, but rarely goes below 0. Randomly chosen pairs of words have a mean of 0.03 and a standard deviation of approximately 0.08 (see Landauer et al., 2007). An LSA web site is freely available (http://lsa.colorado.edu/, accessed September, 2010; see Dennis, 2007).

LSA has been successful at simulating judgments of semantic similarity, word categorization, discourse comprehension, essay quality (see Landauer & Dumais, 1997; Landauer et al., 2007; Jones & Mewhort, 2007, for a review). LSA has also 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”). Furthermore, Diependaele, Dunabeitia, Morris and Keuleers

(2011) used LSA to estimate transparency between full words and constituent-embedded stems, which yields "viewer" vs. "view" as being highly transparent and "corner" vs. "corn" as highly opaque.

One possible method (abbreviated as C2W) of measuring semantic transparency is, similar to Diependaele et al. (2011), to compute the LSA cosine values between the compound word and each of its constituents. For example, the LSA cosine value between "staircase" and "stair" is 0.57 while the one between "staircase" and "case" is 0.07. Since the constituent "stair" and the compound word "staircase" result in a clearly higher cosine value, "stair" is considered semantically transparent, while "case" is considered opaque. However, this computation for English words may or may not reflect how a Chinese rater classifies a constituent as transparent/opaque for two-character Chinese words.

One possible solution is to access the primary meaning of a constituent. The first step of our proposed idea is to find words containing a constituent that a rater possibly activates. Since a constituent may have several meanings, the *primary meaning* of the constituent is computed by a hierarchical clustering algorithm. Since LSA cosine values rarely go below 0 in high-dimensional spaces, we use one minus the absolute value of the LSA cosine as distance function and a given threshold. The selection of this threshold is discussed below in the Reanalysis of Previous Data and General Discussion sections. Since word frequency is important for word recognition and reading (see Rayner et al., 2007), the cluster with the highest sum of word frequency is considered the primary meaning. For example, the transparency of constituent "butter" in "butterfly" is determined as follows. Using the text corpus "general reading up to 1st year college," the LSA cosine values among "butter", "butterfly", "buttercup", "butterfingers", "buttermilk", "butterscotch", "butterfat", and "butterwick" are shown in Table 2. Based on the LSA cosine values, semantic relationships can be visualized by multi-dimensional scaling (MDS) as presented in Figure 1. The results of the cluster analysis are demonstrated in Figure 2. The group of "butter", "buttercup", "buttermilk", "butterscotch", "butterfat", and "butterwick" are considered primary meaning for their highest sum of frequency. According to a threshold 0.9, "butterfly" and "butterfingers" are clustered individually. We applied "document to term" comparison to compute the LSA cosine value between the primary meaning cluster (i. e., a string of "butter buttercup buttermilk butterscotch butterfat butterwick") and "butterfly", and 0.04 is obtained. This approach is abbreviated as M2W.

The M2W approach takes the polysemy of a constituent into account and works even when a constituent is not a stand-alone word. M2W is especially useful for the Chinese language in which many characters do not exist as one-character words in the corpus (described below).

Since the LSA-based method may be able to estimate transparency of English compounds, it could possibly be applied to Chinese two-character words in a similar manner.

Table 2. The LSA cosine values among "butter", "butterfly", "buttercup", "butterfingers", "buttermilk", "butterscotch", "butterfat", and "butterwick".

butter	-fly	-cup	-fingers	-milk	-scotch	-fat	-wick
1							
0.04	1						
0.09	0.09	1					
0	-0.1	-0.1	1				
0.44	-0	0.12	0.01	1			
0.45	0.05	-0	0.02	0.35	1		
0.12	-0	0.04	0	0.11	0.16	1	
-0	0.01	0.12	-0	0.09	0.03	0.04	1

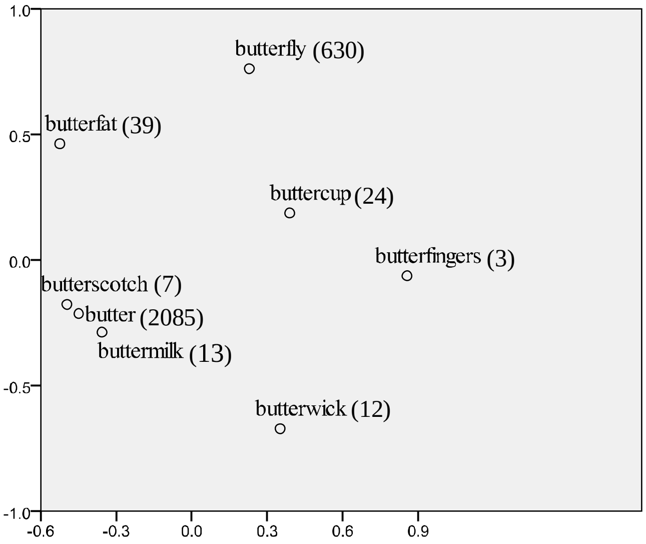

Figure 1: The MDS result for an example of semantic relationships for "butter" and words containing "butter". The frequency for each word in British National Corpus (BNC) is shown in parentheses. The x and y axis represent dimensions 1 and 2, respectively, of the abstract, two-dimensional Euclidean output space of the MDS algorithm.

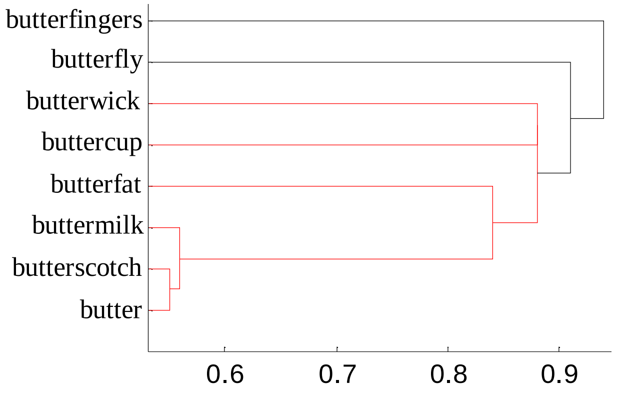

Figure 2. The results of hierarchical clustering of the example word “butter”.

Following the principle of creating semantic spaces (Quesada, 2007), our previous studies (Wang et al., 2010; Chen, Wang, & Ko, 2009) built an LSA semantic space of Chinese (abbreviated as SP-C) from ASBC which contains approximately 5 million words (or 7.6 million characters). Texts in ASBC were collected from different topic areas and classified using five criteria: genre, style, mode, topic, and source, in order to make ASBC a representative sample of modern Chinese language. 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 3 times per 5 million were excluded. A 49021 x 40463 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 in Chinese language (see Chen, Wang, & Ko, 2009).

The term-to-document co-occurrence matrix of SP-C was established using the unit of words, which may be one or more Chinese characters. The C2W approach requires Chinese two-character words to have their constituent characters appear as single-character words more than 3 times in the corpus. Within the 49021 words available in SP-C, 31,637 are two-character words. For 3,921 out of these 31,637 two-character words, either the first or second characters are unavailable due to the frequency restriction. Nevertheless, the M2W approach can still compute the primary meaning of characters despite this characteristic of Chinese. It is even possible that a Chinese reader does not have a single-character representation in his or her mental lexicon for non-stand-alone characters. The polysemy of a character might be involved and the primary meaning might be obtained during lexical access.

Table 2 shows examples of the polysemy of character “馬” (horse). The whole-word meanings of words such as “馬背” (horse back) and “馬鞍” (saddle) are close to “馬” (horse), while the ones of words, e.g., “馬虎” (careless) and “馬桶” (stool) are not. The character “馬” in the word “馬來” (Malaysian, pronunciation: ma-lai) and “馬國” (Malaysia) refers to the abbreviation of Malaysia/Malaysian because of its pronunciation. Figure 3 demonstrates the clustering of character “馬” - the meaning of “馬” is “horse” in words 1 to 4, and is related to “Malaysia” in words 8 and 9. Since the sum of frequency in ASBC for the group of words 1 to 4 is the highest, the group of words 1 to 4 is considered the primary meaning of character “馬”.

It is necessary to verify the proposed computational method by comparing its results with human transparency ratings. We evaluated how LSA estimates transparency of English compounds using the materials of Frisson et al. (2008). The evaluation for two-character Chinese words was conducted by re-analyzing the materials of Tsai (1994) and Lee (2007).

Table 2. A list of character “馬” as one-character word and the two-character words beginning with character “馬”. C2 Meaning is the primary meaning of the second character. WFreq is whole-word frequency in ASBC.

	Word	Whole-Word Meaning	C2 Meaning	WFreq
1	馬	horse		342
2	馬背	horse back	back	14
3	馬鞍	saddle	saddle	4
4	馬車	carriage	car	37
5	馬虎	careless	tiger	13
6	馬桶	stool	tub	23
7	馬腳	a clue of	foot	4
8	馬來	Malaysian	come	11
9	馬國	Malaysia	country	12

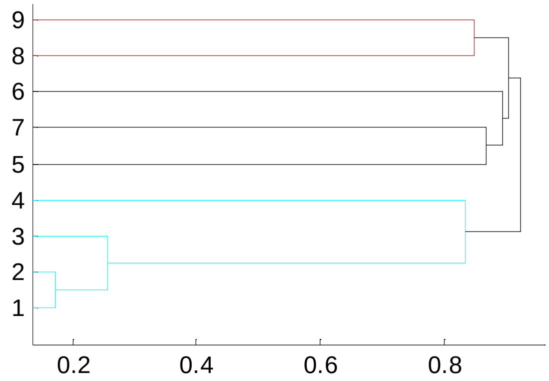

Figure 3. The results of hierarchical clustering of example “馬” with numbers referring to Table 2.

Reanalysis of Previous Data

Ten opaque-opaque, 14 opaque-transparent, and 10 transparent-opaque compounds defined in Frisson et al. (2008) were estimated by our classifiers using C2W and M2W. A receiver operating characteristic (ROC) analysis was performed and the area under the curve (AUC) was used as measurement. Figure 4 illustrates the ROC curves for C2W and M2W (threshold = 0.1, 0.8, and 1), and the AUCs are 0.82, 0.74, 0.82, and 0.75, respectively. A threshold too low may generate too many groups, while a threshold too high only produces one group and therefore causes more false alarm cases. We found that when a compound is high-frequent and its constituent is opaque and low-frequent, the primary meaning of the constituent might be taken over by the compound and therefore the constituent is incorrectly considered transparent. We suggest that C2W could be used when the constituent is low-frequent, and an item-level human judgment should be performed for further analysis.

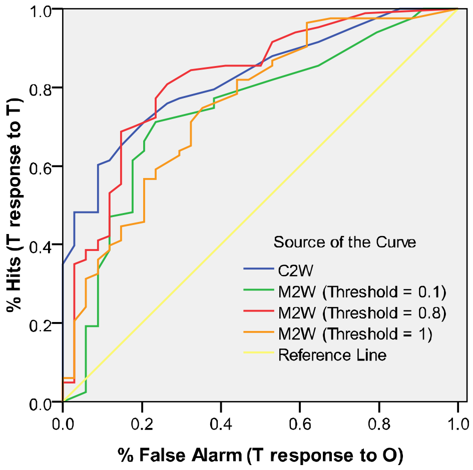

Figure 4: ROC curves for C2W and M2W (threshold = 0.1, 0.8, and 1) for English readers.

For two-character Chinese words, we pre-defined 160 characters selected from the materials of Tsai (1994), Lee, C. Y. (1995), and Lee, P. J. (2007) as either transparent (T) or opaque (O), then those characters were rated by eleven students who completed a college degree in Taiwan participated. All participants were native speakers of Chinese (traditional script). Participants were presented with the two-character word, and asked to respond either "T" or "O" for each constituent. The measure of human rating of each constituent was calculated as the probability with which participants responded "T" to the constituent, e.g., 0.91 for 10 out 11 participants responding "T." The means and standard deviations of human rating (Human), C2W, and M2W are shown in Table 3, where there are 99 transparent and 50 opaque constituents available for C2W. Figure 5 illustrates the results of the ROC analysis, and the AUCs of human rating, C2W, and M2W are 0.99, 0.76, and 0.85, respectively. The Spearman rank correlations (a non-parametric test) between Human and C2W and between Human and M2W are 0.48 and 0.53, respectively. These results suggest that M2W not only overcomes the constraint that C2W is unable to compute transparency when constituents are unavailable in SP-C, but also outperforms C2W in ROC and correlation analyses. As 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. We suggest that M2W may be a better approach than C2W for predicting transparency of constituents of two-character Chinese words.

Table 3. The means and standard deviations (in parentheses) of human rating (Human), C2W, and M2W.

	Human	C2W	M2W
T	0.79 (0.19)	0.22 (0.17)	0.35 (0.27)
O	0.13 (0.14)	0.08 (0.06)	0.07 (0.10)

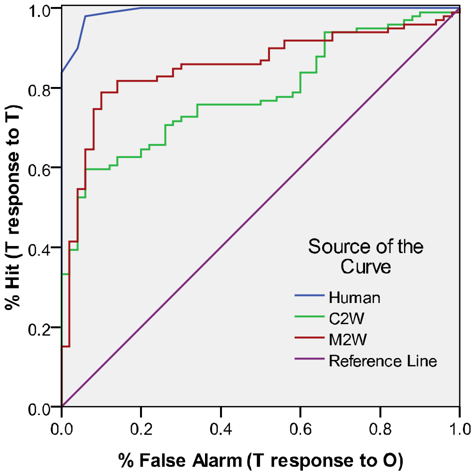

Figure 5: ROC curves for human rating (Human), C2W and M2W for Chinese readers.

General Discussion

The most important outcome of the current study is its proposed computational method of using LSA to estimate semantic transparency, which may reflect the polysemy of constituents and how raters access meanings. Corroborating evidence from two different languages was presented by testing the method with English compounds used in prior compound word study (including Frisson et al., 2008) and two-character Chinese words in the transparency judgment in this study.

The results could be adapted to further Chinese reading research using eye movements. For example, it is still being debated how Chinese words are accessed by readers. Yan, Tian, Bai, and Rayner (2006) investigated the effect of word and character frequency on word processing, and they suggested that when a two-character word 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 word needs to be accessed via its characters (and hence an effect of character frequency emerges). However, some studies have argued for the priority of characters over words (e.g., Chen, Song, Lau, Wong, & Tang, 2003). Therefore, it is still unclear how opaque and transparent words are processed during natural reading. It would be valuable to address these issues using semantic transparency and eye-movement analysis.

The current limitations of the proposed method in Chinese might be the relatively small corpus size. Cai and Brysbaert (2010) published SUBTLEX-CH based on a larger 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 a Chinese LSA semantic space could be established based on this larger corpus as long as the corpus provides enough information in terms of "documents", i.e., a set of words that relate to the same topic in a document. It is important to notice that the size of a corpus is not its only criterion of being representative, but the selection of texts covering different varieties in a corpus should also be taken into account. Furthermore, there are traditional and simplified scripts of Chinese, and it is important to test whether the semantic space built by traditional Chinese is compatible with simplified Chinese.

In addition to the selection of the threshold, since it is related to the distance function of the clustering algorithm and the LSA values, we suggest that an optimization test should be performed for each semantic space. We imply that a threshold might be involved in the transparency judgments by human raters and that each participant might have a different threshold for the "cut-off" of opacity. It should be clear that the use of the proposed computational method is not intended to replace the standard measures that are based on human raters, but that it offers a different perspective and an opportunity to examine the lexical processing for estimating semantic transparency.

Acknowledgments

Portions of the data were presented at the Asia-Pacific Conference on Vision (APCV) 2010. Thanks to Keith Rayner, Jinmian Yang, and Marc Brysbaert for helpful comments on this study.

References

Academia Sinica. (1998). *Academia Sinica balanced corpus* (Version 3) [Electronic database]. Taipei, Taiwan.

Cai, Q. & Brysbaert, M. (2010). SUBTLEX-CH: Chinese Word Frequencies Based on Film Subtitles. PLoS ONE 5(6): e10729. doi: 10.1371/journal.phone.0010729.

Chen, H.-C., Song,H., Lau, W. Y.,Wong, K. F. E., & Tang, S. L. (2003). Developmental characteristics of eye movements during reading. In C. McBride-Chang & H. C. Chen (Eds.), *Reading development in Chinese children* (pp. 157–169).Westport, CT: Praeger.

Dennis, S. (2007). How to use the LSA website. In T. Landauer, D. McNamara, S. Dennis & W. Kintsch Eds. *Handbook of Latent Semantic Analysis*. Erlbaum, 57-70.

Diependaele, K., Dunabeitia, J. A., Morris, J., & Keuleers, E. (2011). Fast morphological effects in first and second language word recognition. *Journal of Memory and Language*, 64(4), 344-358.

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.

Inhoff, A. W., Starr, M. S., Solomon, M. P., & Lars, P. (2008). Eye movements during the reading of compound words and the influence of lexeme meaning. *Memory & Cognition*, 36(3), 675-687.

Jones, M. N. & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, 114, 1-37.

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.

Landauer, T. K., McNamara, D. S., Dennis S., & Kintsch W. (2007). *Handbook of Latent Semantic Analysis*, Lawrence Erlbaum Associates.

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

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

Martin, D. I. & Berry, M. W. (2007). Mathemetical fundations behind latent semantic analysis. In T. Landauer, D. McNamara, S. Dennis & W. Kintsch Eds. *Handbook of Latent Semantic Analysis*. Erlbaum, 35-55.

Mok, L. W. (2009). Word-superiority effect as a function of semantic transparency of Chinese bimorphemic compound words. *Language and Cognitive Processing*, 24 (7/8), 1039-1081

Pollatsek, A. & Hyönä, J. (2005). The role of semantic transparency in the processing of Finnish compound words. *Language and Cognitive Processing,* 20 (1/2), 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*. Erlbaum, 71-88.

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 and Cognitive Processes*, 15(4/5), 507-537.

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.

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

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.

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.

Zhou, X., Ye, Z., Cheung, H., Chen, H.-C. (2009). Processing the Chinese language. *Language and Cognitive Processing*, 24 (7/8), 929-946.