Estimating the Effect of Word Predictability on Eye Movements in Chinese Reading using Latent Semantic Analysis and Transitional Probability

Hsueh-Cheng Wang and Marc Pomplun
Department of Computer Science, University of Massachusetts, Boston, USA

MingLei Chen and Hwawei Ko
Graduate Institute of Learning and Instruction, National Central University, Taiwan
and
Keith Rayner
Department of Psychology, University of California, San Diego, USA

Correspondence to:
Hsueh-Cheng Wang
Department of Computer Science
University of Massachusetts at Boston
Email: hchengwang@gmail.com

Abstract

Latent semantic analysis (LSA) and transitional probability (TP), two computational methods used to reflect lexical semantic representation from large text corpora, were employed to examine the effects of word predictability on Chinese reading. Subjects' eye movements were monitored and the influence of word complexity (number of strokes), word frequency, and word predictability on different eye movement measures (first fixation duration, gaze duration, and total time) were examined. We found influences of TP on first fixation duration and gaze duration and of LSA on total time. The results suggest that TP reflects an early stage of lexical processing while LSA reflects a later stage.

It is well-known that eye movements provide an indication of language processing because they are affected by lexical variables such as word frequency (Rayner, 1998). Kliegl, Grabner, Rolfs, and Engbert (2004) used a statistical control approach to examine the effect of frequency (and other variables) on all words in a sentence corpus. Repeated-measures multiple regression analysis (Lorch & Myers, 1990) was employed to remove systematic variance between subjects and to test the significance for the coefficients of variables. Kliegl et al. found that word frequency, word length, and word predictability all affect eye movement measures during reading.

Rayner and Well (1996; see also Ehrlich & Rayner, 1981) also found that the predictability of target words (as determined by a cloze task) has a strong influence on eye movements during reading. In their experiment, subjects fixated low-predictable target words longer than they did either high- or medium-predictable target words; they also skipped high-predictable words more often than they did either medium- or low-predictable target words. Subsequently, Rayner, Ashby, Pollatsek, and Reichle (2004) examined the interaction of predictability and word frequency and found that the data pattern only mildly departed from additivity with predictability effects that were only slightly larger for low-frequency than for high-frequency words. The pattern of data for skipping words was different as predictability affected only the probability of skipping for high-frequency target words.

In the E-Z Reader model (Pollatsek, Reichle, & Rayner, 2006; Reichle, Pollatsek, Fisher, & Rayner, 1998; Reichle, Rayner, & Pollatsek, 1999; 2003), word predictability within a given sentence context is considered in both first stage processing (i.e., L_1, including identification of orthographic form and a familiarity check) and second stage processing (i.e., L_2, including identification of phonological/semantic form and completion of lexical access). The model also maintains that the predictability effect is stronger in L_2 than in L_1.

The predictability effect in Chinese reading was also observed in a study by Rayner, Li, Juhasz, and Yan (2006). They found that Chinese readers, like readers of English, exploit target word predictability during reading. The results were highly similar to those of Rayner and Well (1996) with English readers: Chinese readers fixated for a shorter duration on high- and medium-predictable target words than on low-predictable target words. They were also more likely to fixate on low-predictable target words than on high- or medium-predictable target words.

Subsequently, the E-Z Reader model was extended to Chinese reading (Rayner, Li, & Pollatsek, 2007). In both the English and Chinese versions of the model, L_1 and L_2 are functions of both the frequency of the word in the language and its predictability from the prior text. As with the work mentioned above, estimates of word predictability in the Chinese model were derived from a modified cloze task procedure (Taylor, 1953) in which subjects are asked to guess word n from the prior sentence context. Typically, experiments are set up using target words that differ substantially in cloze value, often with probabilities of .70 to .90 for high-predictable words and less than .10 for low-predictable words.

Another approach to estimating eye movement behavior during reading, transitional probability (TP), was introduced by McDonald and Shillcock (2003a). They found that transitional probabilities between words have a measurable influence on fixation durations and suggested that the processing system is able to draw upon statistical information in order to rapidly estimate the lexical probabilities of upcoming words. McDonald and Shillcock (2003b) also demonstrated that TP is predictive of first fixation duration (the duration of the first fixation on a word independent of whether it is the only fixation on a word or the first of multiple fixations on it) and gaze duration (the sum of all fixation durations prior to moving to another word). The results indicated that TP might reflect low-level predictability, which influences 'early' processing measures such as first fixation duration, instead of high-level predictability, which influences 'late' processing measures. However, Demberg and Keller (2008) argued that forward TP influences not only 'early' processing measures such as first fixation duration, but also 'late' processing measures such as total time. This finding is somewhat inconsistent with the results of McDonald and Shillcock (2003b). Moreover, Frisson, Rayner, and Pickering (2005) concluded that the effects of TP are part of regular predictability effects. Accordingly, predictability measures estimated by cloze tasks often capture both low-level and high(er)-level predictability, although TP does not provide evidence for a separate processing stage. In addition, Frisson et al. argued that TP effects may not be truly independent of the traditionally considered predictability effects because TP is often correlated with frequency.

There are other computational methods that have been utilized to approximate predictability and its effect on eye movements in alphabetical languages. For instance, surprisal, a measure of syntactic complexity, was examined by Boston, Hale, Kliegl, Patil, and Vasishth (2008). Surprisal of the n^{th} word is defined by Equation 1, where the prefix probability α is the total probability for the grammatical analyses of the prefix string (see Boston et al., 2008).

$$\text{surprisal}(n) = \log_2\left(\frac{\alpha_{n-1}}{\alpha_n}\right) \tag{1}$$

Boston et al. (2008) showed that surprisal had an effect on both “early” and “late” eye movement measures. Demberg and Keller (2008) also found that surprisal can predict first fixation duration, (first pass) gaze duration, and total time.

Another computational method is conditional co-occurrence probability (CCP), a simple statistical representation of the relatedness of the current word to its context, based on word co-occurrence patterns in data taken from the Internet. It was used to predict eye movements by Ong and Kliegl (2008); they reported that CCP is correlated to frequency but that it cannot replace predictability as a predictor of fixation durations. In addition, Latent Semantic Analysis (LSA; described below; Landauer & Dumais, 1997) was used by Pynte, New, and Kennedy (2008), who reported that both single fixation duration and gaze duration effects on content words were evident using LSA.

The objective of the present study was to estimate word predictability, via the use of TP and LSA, and to further investigate predictability effects in Chinese reading. Word complexity, word frequency, and word predictability were taken into account to examine various eye movement measures: first fixation duration, gaze duration, and total time (the sum of all fixations on a word including regressions). The visual complexity of Chinese words might not be accurately represented by word length (number of characters) as in English (number of letters). In Chinese, a character is quite different from an English letter, both visually and linguistically (Rayner et al., 2007). For instance, Chinese characters have radicals and these radicals might influence visual complexity. The simplest measure to represent visual complexity of Chinese characters is number of strokes. However, the complexity of Chinese words is not well defined for several reasons. First, Chinese words are composed of one, two, three, or more characters so that the number of strokes and word length are confounded. Second, number of strokes and word frequency are correlated. Although number of strokes might not be suitable as an analogue to word length in English reading, we attempted to estimate word complexity using word length (number of characters) and average number of strokes (the average number of strokes per character).

Another goal of the present study was to examine the influence of word predictability on early and late stages of lexical processing. Taking advantage of the computational methods TP and LSA, we utilized repeated-measures multiple regression analysis to examine the main factors (word complexity, word frequency, and word predictability) in Chinese reading. LSA might be suitable for predicting eye movement patterns not only for Chinese reading but also for alphabetical languages (such as English).

Latent Semantic Analysis

LSA is a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text (Landauer & Dumais, 1997). To construct an LSA computation, a term-to-document co-occurrence matrix is first established from a corpus which embodies mutual constraints of semantic similarity of words. To solve these constraints, a linear algebra method, singular value decomposition (SVD), is used and a dimension reduction of the original matrix is performed. The meaning of each word or passage is then represented as a vector in *semantic space*. The semantic space of Chinese was established from the Academia Sinica Balanced Corpus (ASBC; Academia Sinica, 1998) which contains approximately 5 million words. Chinese words defined in ASBC were selected as term tokens in our semantic space. Documents[1] were segmented into words according to the segmentation standard[2] by Huang, Chen, Chen, and Chang (1997). The total number of segmented documents was 40463 and their length was 219 characters on average (standard deviation of 66 characters). A "stop list" was used to exclude words that occur in every document and words that do not contain much information. Words which could serve different functions were combined. For representatives of words in the corpus, we selected words that occurred more than 3 times, resulting in 49021 words being used in the term list. A term-to-document co-occurrence matrix was then established, and global and local weighting were performed. The purpose of global weighting is to reduce the importance of words that occur in every document and therefore do not help to differentiate meaning. Local weighting is aimed at diminishing the influence of words that are extremely frequent in one document and do not carry substantial meaning.

The computation of local and global weighting is described in Equations 2 and 3. In Equation 3, tf_{ij} is type frequency of type *i* in document *j*, and gf_i is the total number of times that type *i* appears in the entire collection of *n* documents (Dumais, 1991).

$$\textit{Local weight} = \log(\textit{type frequency} + 1) \tag{2}$$

$$\textit{Global weight} = \sum_j \frac{p_{ij}\log_2(p_{ij})}{\log_2 n} \textit{ where } p_{ij} = \frac{tf_{ij}}{gf_i} \tag{3}$$

For large datasets, empirical testing shows that the optimal choice for the number of dimensions ranges between 100 and 300 (Berry et al., 1999; Jessup & Martin, 2001; Lizza & Sartoretto, 2001). Table 1 shows the mean and standard deviation of cosine values of randomly selected word pairs from 50, 100, 150, 200, 250, and 300 dimensional semantic spaces and from the English semantic space (Landauer, 2007). In the present study, 300 dimensions were used for evaluation and further analysis of word predictability.

Insert Table 1 about here

Although LSA has been very successful at simulating a wide range of psycholinguistic phenomena, from judgments of semantic similarity to word categorization to discourse comprehension and judgments of essay quality (see Jones & Mewhort, 2007, for a review), it has not yet been tested on word predictability. The current study verified the credibility of differentiating between high and low predictability based on LSA. The predictable/unpredictable target words in Rayner et al. (2004), determined by a cloze task, were examined. An LSA tool was employed (http://lsa.colorado.edu/) and the setting of semantic space was *General Reading (300 factors*). Table 2 shows the LSA cosine value of two target words (*horse* and *camel*) in two contexts (*Most cowboys know how to ride a …* and *In the desert, many Arabs ride a …*)[3]. The results indicate that LSA cosine values of predictable target words were significantly higher ($t=2.97$, $df = 62$, $p<.01$) than those of unpredictable ones.

Insert Table 2 about here

Similarly, the credibility of LSA Chinese semantic space was tested. The word association norms of 600 homographs (Hue, 1996) were used to evaluate the semantic space of 300 dimensions in the Chinese language. A total of 300 subjects were randomly divided into three equally sized groups, and for each group 200 out of the 600 homographs were selected as stimuli. Subjects were asked to write down the first word that came to their mind for each of the stimulus homographs. We selected the word pairs that were written down by at least 10 subjects, which resulted in 436 out of 14,464 word pairs being selected. We (Chen, Wang, and Ko, submitted) found that the frequencies of the target word and the word associated by the subjects were significantly correlated to LSA cosine value ($r = .16$, $p < .001$).

Method

Subjects. Twelve undergraduate students from the National Chung Cheng

University (Taiwan) participated. All had normal or corrected to normal vision, and were native speakers of Mandarin. Each participant received 100 Taiwan dollars for participation in a half-hour session.

Apparatus. Subjects sat 65 cm from an LCD Monitor on which 7-line texts were presented in their entirety for them to read. At this distance, one character space equaled 1 degree of visual angle. Eye movements were recorded by an SR Eyelink-II head-mounted eye-tracker. A chin rest was provided to minimize head movements. The sampling rate was 250 Hz. Although viewing was binocular, eye movements were recorded from the right eye only.

Materials. Sixteen expository texts were used in this study. On average, the passages were 180 characters long. Each text was about 7 lines long with a maximum of 27 Chinese characters per line. Each text was followed by two yes-or-no questions. One of them was about the general idea of the passage and the other asked about a detailed fact described in the passage. The purpose of the comprehension questions was to promote the processing of text content. The predictor variables for the target words included (1) total number of strokes, (2) word length, (3) average number of strokes, (4) word frequency, (5) TP (including forward and backward TP), and (6) LSA. The number of strokes of Chinese characters was defined according to a Chinese dictionary published by the Ministry of Education, Taiwan. The total strokes of a target word were calculated as the sum across all the characters in the word. The mean of total strokes was 19.40 (standard deviation: 7.88). Word length was defined as the number of characters, and the mean and standard deviation of word length were 1.91 and 0.48. The average number of strokes of a target word was computed as the average across all characters in the word, and the mean and standard deviation were 10.13 and 3.48. Word frequency was measured by the natural logarithm of occurrence in ASBC. The mean and standard deviation of the frequency measure were 5.79 and 2.06 (range: 1.39 - 11.03). The forward transitional probability (fTP) and backward transitional probability (bTP) of the target word *n* were calculated by simple ratios of joint and marginal frequencies of its preceding word *n-1* and its following word *n+1*, respectively, as shown in Equations 4 and 5. The range of fTP was from 0 to 0.8 and its average was 0.00973, while bTP ranged from 0 to 0.75 with an average of 0.0074.

$$fTP = P(n-1|n) = f(n-1, n) / f(n-1) \tag{4}$$

$$bTP = P(n|n+1) = f(n, n+1) / f(n+1) \tag{5}$$

Most function words were excluded from the LSA calculation because LSA gives lower weights or even neglects function words that do not carry substantial meaning. It is also known that the decisions of where and when to move the eyes depend strongly on the previous fixation location (Engbert, Longtin, & Kliegl, 2002; Rayner, 1998) and function words are often skipped. Figure 1 shows an example of how the LSA cosine values of target words and their previous content[4] words were calculated. The average and standard deviation are 0.31 and 0.19, respectively.

Insert Figure 1 about here

The correlations between predictors are shown in Table 3 based on 5,324 word cases[5] which were included in a regression analysis (described below). Not surprisingly, we found that word frequency was inversely correlated to total number of strokes, word length, and average number of strokes. The correlation between word length and ln(Freq) is similar to corresponding values for alphabetic languages. Frequency was somewhat correlated to TP. As mentioned above, TP effects may not be truly independent of the traditionally computed predictability effects, and this was also found in our Chinese sample. However, LSA was only weakly correlated to word frequency and total number of strokes, word length, and average number of strokes. Because the correlation between average number of strokes and ln(Freq) is lower than the one between total number of strokes and ln(Freq) , we did not include total number of strokes as a predictor in the analysis.

Insert Table 3 about here

Procedure. After subjects read the instructions, a standard 9-point grid calibration (and validation) was completed. Subjects were instructed to read the text for comprehension. At the start of each trial, a drift calibration screen appeared, and subjects were instructed to look at the calibration dot that appeared in exactly the same position as the first character of the text. When subjects passed the drift correction, the entire text (double-spaced) appeared on the screen. Prior to the texts used for data analysis, subjects read two texts (with four questions) for practice (these texts were not included in the analysis). Given that the texts were double spaced, there was no difficulty determining which lines subjects were fixating on. They read the text at their own pace, indicating they had finished reading by pressing a button on a control pad. After the subject had pressed this button, the text disappeared and the drift correction screen appeared again. Then the first yes-or-no question appeared on the screen, and after the subject pressed the yes or no button, the second question appeared. After subjects finished the comprehension tests, the next text appeared. In some cases, calibration and validation were performed once again to increase gaze-tracking accuracy. Each subject read sixteen texts presented in random order.

Results and Discussion

The data analysis procedure used in this study was the repeated-measures multiple regression analysis suggested by Lorch and Myers (1990)[6]. Each regression analysis was performed separately for each subject. Subsequently, a one-sample t-test on the resulting regression coefficients was performed for each predictor variable to determine if the mean coefficients were significantly different from zero. The raw eye movement data[7] were processed using SR Research EyeLink Data Viewer to compute eye fixations and to remove blinks.

The first word of each text and the cases in which LSA or TP were not available in our computation were excluded from analysis. There were 11,311 word cases (which were mainly content words) available in our eye movement corpus, and 3,212 word cases were skipped[8] (i.e., did not receive any fixations). The skipped word cases were excluded, leaving 8,099 cases. In addition, only "first pass" fixations (which reflect the first left-to-right sweep of the eye over each sentence, see Boston et al., 2008) were examined, leaving 5,324 out of 8,099 word cases included in the regression analysis[9]. Table 4 shows the resulting regression coefficients. We used all predictors to obtain equations for predicting first fixation duration, gaze duration, and total time (Equations 6, 7, and 8, respectively). When the means of the natural logarithm of word frequency (5.79), word length (1.91), average number of strokes (10.13), fTP (0.0097), bTP (0.0074), and LSA (0.30) were applied to Equations 6, 7, and 8, the predicted first fixation duration (FFD), gaze duration (GD), and total time (TT) were 211, 239, and 438 ms, respectively.

$$\mathrm{FFD}_{\mathrm{pred}} = 211 + (-1.41)\, \mathit{ln(Freq)} + (-1.24)\, \mathrm{WordLength} + 0.76\, \mathrm{avg}\mathit{Strokes} + (-10.22)\, \mathit{LSA} + (-60.29)\, \mathrm{f}\mathit{TP} + (8.50)\, \mathrm{b}\mathit{TP} \tag{6}$$

$$\mathrm{GD}_{\mathrm{pred}} = 203 + (-3.47)\, \mathit{ln(Freq)} + (21.47)\, \mathrm{WordLength} + 1.63\, \mathrm{avg}\mathit{Strokes} + (-3.95)\, \mathit{LSA} + (-39.17)\, \mathrm{f}\mathit{TP} + (7.67)\, \mathrm{b}\mathit{TP} \tag{7}$$

$$\mathrm{TT}_{\mathrm{pred}} = 247 + (-6.78)\, \mathit{ln(Freq)} + (109.39)\, \mathrm{WordLength} + 4.98\, \mathrm{avg}\mathit{Strokes} + (-104.13)\, \mathit{LSA} + (104.97)\, \mathrm{f}\mathit{TP} + (1.97)\, \mathrm{b}\mathit{TP} \tag{8}$$

Insert Table 4 about here

Word Frequency. The word frequency effect was significant for first fixation duration and gaze duration, and was marginal for total time, $p=.07$ (see Table 4). This demonstrates that word frequency affects both early and late stages of lexical processing in Chinese reading. Not surprisingly, we found that the effect size for total time was greater than for gaze duration and first fixation duration with regard to the unstandardized regression coefficients. These results are consistent with those by Kliegl et al. (2004), in which a German sentence corpus was analyzed using repeated-measures multiple regression analysis.

Word Length and Complexity. In English, it has been found that word length influences gaze duration and total time, but not first fixation duration. Much of this effect is due to the fact that as words get longer, the probability that readers refixate it before moving on increases (Rayner, 1998). It is interesting that, in the current study, first fixation duration was not influenced by word length but by average number of strokes. In addition, we expected to find that average number of strokes would only affect early processing. However, the results indicated that there was a significant influence on first fixation duration, gaze duration, and total time. Despite these effects, we are unable to conclude that word complexity (estimated by average number of strokes) influences both low- and high-level lexical processing in Chinese reading because of the correlations among frequency, word length, average number of strokes, and total number of strokes.

Word Predictability. The results were similar to McDonald et al. (2003a), as there was a significant forward TP effect on first fixation duration ($p<.01$), a tendency toward an effect on gaze duration ($p=.06$), but not on total time ($p=.28$). This finding indicates that lower-level lexical processing of Chinese, like English, was influenced by forward TP because of its effect on 'early' processing measures of eye movements such as first fixation duration. However, we were not able to find any influence from backward TP, either on first fixation duration, gaze duration, or total time. The results were inconsistent with the results by McDonald et al. (2003b) who suggested that backward TP has an effect on first fixation duration, gaze duration, and total time, and low backward TP words were fixated longer than high ones[10]. LSA had a significant effect on only total time ($p<.01$), and not on first fixation duration ($p=0.27$) or gaze duration ($p=0.80$). The results thus indicate that LSA estimates higher-level lexical processing because LSA influenced 'late' processing measures of eye movements such as total time. However, the result is inconsistent with results reported by Pynte et al. (2008), who claimed that LSA has an effect on both single fixation duration and gaze duration.

In summary, since the primary objective of the present study was to investigate the effect of predictability, especially from LSA, on Chinese reading, we did not attempt to examine the inter-correlation between word complexity, word frequency, and TP. Nevertheless, we did find clear evidence that LSA predicts late-stage eye movement measures.

Conclusions

This study employed TP and LSA to predict eye fixation times and to investigate the word predictability effect on early and late stage lexical processing during Chinese reading. Although the inter-correlation of word complexity, word frequency, and TP on Chinese reading was somewhat unclear, we found that LSA can estimate higher-level word predictability effects when word complexity and word frequency effects are taken into account. It appears that TP and LSA can be used as complementary tools for deriving word predictability ratings. Local information is retrieved by TP which considers only two consecutive words, while global information is utilized by LSA to bring out latent semantic relationships among words even if they have never co-occurred in the same document (Jones et al, 2007). Word order is considered by TP for either the forward or backward direction, but not by LSA. For usability, TP can only be used on words that consecutively co-occur in a corpus, whereas LSA can compute the cosine value of any word pair included in its semantic space.

In addition to documenting the potential usefulness of both TP and LSA in the context of eye movement data, the present results also replicate prior research on

Chinese demonstrating that word predictability (Rayner et al., 2005) and word frequency (Yan et al., 2006) influence how long readers fixate on words during reading. This provides further evidence for the psychological reality of words during Chinese reading (Bai, Yan, Liversedge, Zang, & Rayner, 2008; Rayner et al., 2005; Yan et al., 2006). Chinese has intersecting levels of structure such as words, characters, and radicals, and some have argued for the priority of characters over words (Chen, Song, Lau, Wong, & Tang, 2003). Although we do not deny the importance of characters, the present results are consistent with the view that words are important in reading Chinese. The present results also document that word complexity has an influence on fixation times during Chinese reading.

In summary, the results suggest that TP reflects lower level lexical processing while LSA estimates higher level lexical processing in Chinese reading because TP influenced earlier processing measures of eye movements while LSA influenced late processing measures. However, our research, like earlier work, indicates that computational alternatives to predictability, such as TP (McDonald et al., 2003), LSA (Pynte et al., 2008), CCP (Ong et al., 2008), and surprisal (Boston et al., 2008), while providing some interesting perspectives, cannot entirely replace the standard predictability measure computed via cloze tasks.

References

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

Bai, X., Yan, G., Liversedge, S.P., Zang, C., & Rayner, K. (2008). Reading spaced and unspaced Chinese text: Evidence from eye movements. *Journal of Experimental Psychology: Human Perception and Performance*, 34, 1277-1288.

Berry, M.W., Drmac, Z., & Jessup, E. (1999). Matrices, vector spaces, and information retrieval. *SIAM Review*, *41*, 335–362.

Boston, M. F., Hale, J., Kliegl, R., Patil, U. & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, 2(1):1, 1-12

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.

Chen, M. L., Wang, H. C., & Ko H. W. (Submitted). Using Latent Semantic Analysis to create a Chinese Semantic Space and the validation of psychological reality, *Journal of Chinese Psychology*.

Choi, F.Y.Y., Wiemer-Hastings, P., & Moore, J., 2001. Latent semantic analysis for text segmentation. *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, p. 109-117.

Demberg, V. & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*. 109, 193-210.

Dumais, S. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, and Computers, 23*, 229–236.

Engbert, R., Longtin, A., & Kliegl, R. (2002). A dynamical model of saccade generation in reading based on spatially distributed lexical processing. *Vision Research. 42*, 621-636.

Ehrlich, S.F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior, 20*, 641-655.

Frisson, S., Rayner K., & Pickering, M. J. (2005). Effects of contextual predictability and transitional probability of eye movements during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 862-877.

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(2),* 47-62.

Hue, C. W., Chen, Y. J., & Chang, S. H. (1996). Word association for 600 Chinese homographs. *Chinese Journal of Psychology, 38*, 67-169.

Jessup, E., & Martin, J. (2001). Taking a new look at the latent semantic analysis approach to information retrieval. In M. W. Berry (Ed.), *Computational information retrieval* (pp. 121–144). Philadelphia: SIAM.

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

Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, 16, 262-284.

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.

Lizza, M., & Sartoretto, F. (2001). Acomparative analysis of LSI strategies. In M.W. Berry (Ed.), *Computational information retrieval* (pp. 171–181). Philadelphia: SIAM.

McDonald, S. A., & Shillcock, R. C. (2003a). Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science, 14*, 648–652.

McDonald, S. A., & Shillcock, R. C. (2003b). Low-level predictive inference in reading: The influence of transitional probabilities on eye movements. *Vision Research, 43*, 1735–1751.

Ministry of Education, R.O.C., *Chinese Dictionary* (1998). [Online]. Available: http://140.111.34.46/dict/ [Accessed: June 17, 2007]

Noortgate, W. V. D, & Onghena, P. (2006). Analysing repeated measures data in cognitive research: A comment on regression coefficient analyses. *European Journal of Cognitive Psychology, 18*, 937-952.

Ong, J. K. Y. & Kliegl, R. (2008). Conditional co-occurrence probability acts like frequency in predicting fixation durations. *Journal of Eye Movement Research*, 2(1):3, 1-7

Pynte, J., New, B. & Kennedy, A. (2008). A multiple regression analysis of syntactic and semantic influences in reading normal text. *Journal of Eye Movement Research*, 2(1):4, 1-11

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*, 372-422.

Rayner, K., Ashby, J., Pollatsek, A., & Reichle, E. D. (2004). The effects of frequency and predictability on eye fixations in reading: Implications for the E-Z Reader model. *Journal of Experimental Psychology: Human Perception and Performance, 30,* 720–732.

Rayner, K., Li, X., Juhasz, J. B., & Yan, G.. (2005). The effects of word predictability on the eye movements of Chinese readers. *Psychonomic Bulletin & Review, 12,* 1089–1093.

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.

Rayner, K., & Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin & Review, 3*, 504–509.

Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review*, 105, 125-157.

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The E-Z Reader model of eye movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences, 26,* 445–476.

Reichle, E. D., Rayner, K., & Pollatsek, A. (1999). Eye movement control in reading: Accounting for initial fixation locations and refixations within the E-Z Reader model. *Vision Research*, 39, 4403-4411.

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.

Footnotes

[1] Typically, paragraphs are used as documents in LSA computations because a paragraph often represents a main idea in English. However, ASBC does not provide paragraph information so that articles in ASBC vary substantially in their length (964 characters on average with a standard deviation of 2355 characters). To reduce the difference in length between articles and to find main ideas in the documents, articles of more than 200 characters were segmented into several documents based on the period symbol (“。”) which often represents the end of a main idea.

[2] ASBC is a representative public corpus of traditional Chinese and is widely used in research on the Chinese language. Word segmentation of materials was performed by a segmentation program provided by the CKIP group, Academia Sinica, Taiwan.(http://ckip.iis.sinica.edu.tw/CKIP/engversion/index.htm) We adjusted the segmentation of materials (less than 5 words per passage, mainly the proper nouns that cannot be recognized by the program). The segmentation was agreed upon by three native speakers of Chinese.

[3] The complete table is shown in Appendix A.

[4] We manually excluded most function words from analysis when we calculated the LSA score for each target word. Since readers might have different definitions for function words, there might be a small amount of remaining function words in the analysis according to such definitions.

[5] The correlations between predictors from 5,324, 8,099 or 11,311 word cases (described below) were highly similar to those in Table 3.

[6] Noortgate and Onghena (2006) suggested that repeated-measures multiple regression analysis (rmMRA), ANOVA, and hierarchical linear model (HLM, the precursor to Linear Mixed Models, LMM) give the same results for balanced models with positive variance estimates. Although Noortgate et al. (2006) recommended HLM over rmMRA, we found that the results from rmMRA and HLM did not make a significant difference in this study.

[7] In addition to the 12 subjects whose data were analyzed, a few other subjects were run but excluded because it was not possible to accurately identify which lines they were fixating.

[8] We found that predictable words – those in the top third of both LSA and TP values - had a higher skipping rate (39.39%) than unpredictable words – those in the bottom third of LSA and TP values (25.54%), $t(11) = 16.87$, $p<0.01$. The skipping rates of high, medium and low frequency words, categorized by equal-sized frequency intervals, were 40.61%, 25.58%, and 19.31%, respectively, with significant pairwise differences between all categories, all $ts(11)>7.78$, $ps<0.01$.

[9] The regression analysis using 8,099 word cases showed very similar results to those in Table 4, except that, for the 8,099 cases, TP has a significant effect on gaze duration ($t=2.94$, $p<.05$) and frequency has a stronger effect on first fixation duration ($t=3.31$, $p<.01$).

[10] This study also analyzed the “return sweep” word cases which were the remaining cases when the first-pass ones were excluded. There were 2,775 such cases. However, we could not find any influence from backward TP on first fixation duration, gaze duration, or total time.

Acknowledgments

The data reported in this article were collected when the first author was at the National Central University, Taiwan. Parts of the data were presented at the 14th European Conference on Eye Movements (ECEM 2007). The project was partially supported by Grant NSC095-2917-I-008-001 from the National Science Council, Taiwan. Subsequent analyses were carried out at the University of Massachusetts, Boston. Thanks to Jinmian Yang, Reinhold Kliegl, and Richard Shillcock for helpful comments on an earlier version of the article. Preparation of the article was partially supported by Grant HD26765 from the National Institute of Health. Correspondence should be addressed to Hsueh-Cheng Wang (hchengwang@gmail.com).

Table 1. Cosine values of randomly selected word pairs

Dimensions	50	100	150	200	250	300	English[1]
Mean	.108	.068	.051	.040	.029	.027	.03
Std	.157	.117	.094	.081	.069	.063	.08

Note: [1]The mean and standard deviation is suggested in Landauer (2007).

Table 2. Re-analysis of materials in Rayner et al. (2004) using LSA

Context preceding the target word	Target word	Predictability	Freq	LSA
Most cowboys know how to ride a	horse	P	H	0.62
Most cowboys know how to ride a	camel	U	L	0.1
In the desert, many Arabs ride a	horse	U	H	0.29
In the desert, many Arabs ride a	camel	P	L	0.64

Table 3. Pair-wise correlation between predictor variables (based on mainly content words)

Variable	ln(Freq)	TotalStrokes	WordLength	avgStrokes	LSA	fTP	bTP
ln(Freq)	—	-.391	-.466	-.198	.128	.260	.160
TotalStrokes		—	.613	.784	-.085	-.175	-.157
WordLength			—	.045	-.107	-.213	-.149
avgStrokes				—	-.040	-.087	-.103
LSA					—	.247	.008
fTP						—	.042
bTP							—

Note: Ln(Freq) is the natural logarithm of word frequency; TotalStrokes is the sum of the number of strokes of all characters in the word; WordLength is the number of characters, avgStrokes is the average of the number of strokes of all characters in the word; LSA is the cosine value between the target word and its preceding content word; fTP is the forward transitional probability of the target word; bTP is the backward transitional probability of the target word.

Table 4. Estimates of mean unstandardized regression coefficients and standard errors in the regression equation

	First Fixation Duration			Gaze Duration			Total Time		
Variable	Mean	Std. D	Std. E	Mean	Std. D	Std. E	Mean	Std. D	Std. E
Constant	211**	49.19	14.20	203**	69.32	20.01	247**	189.89	54.82
ln(Freq)	-1.41*	1.63	0.47	-3.47**	2.89	0.83	-6.78	11.68	3.37
WordLength	-1.24	8.81	2.54	21.47**	23.58	6.81	109.39**	45.96	13.27
avgStrokes	0.76*	1.01	0.29	1.63**	1.34	0.39	4.98**	3.36	0.97
LSA	10.22	29.92	8.64	-3.95	54.56	15.75	-104.13**	76.21	22.00
fTP	-60.29**	56.69	16.37	-39.17	64.73	18.69	104.97	322.23	93.02
bTP	8.50	88.71	25.61	7.67	126.09	39.40	1.97	90.32	26.07

Notes: Ln(Freq) is the natural logarithm of word frequency; WordLength is the number of characters, avgStrokes is the average of the number of strokes of all characters in the word; LSA is the cosine value between the target word and its previous content word; fTP is the forward transitional probability of the target word; bTP is the backward transitional probability of the target word. Estimates are based on a mean of 443 words per participant (skipped words are excluded). The constant coefficient is smaller for GD than for FFD because there is an influence of WordLength on GD but not on FFD. As shown in equation (6), (7), the predicted GD is longer than FFD when the average WordLength (1.9) is applied. All values are in milliseconds. * $p < 0.05$, ** $p < 0.01$

Figure Captions

Figure 1. The LSA cosine value of target words and their preceding content words

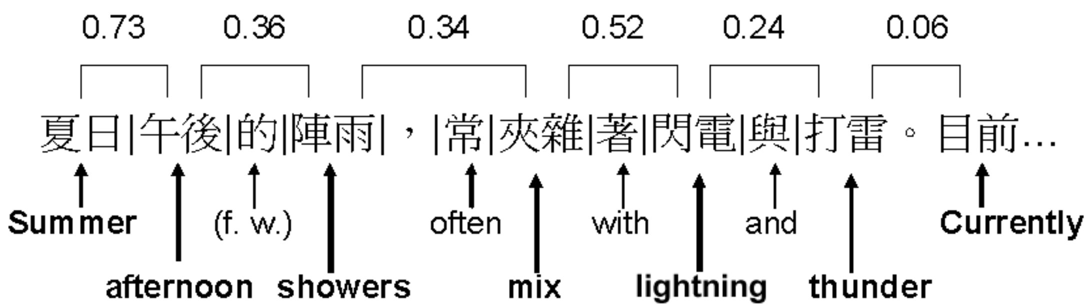

Figure 1

Appendix A

LSA cosine values of the materials in Rayner et al. (2004)

Context before target word	Target Word	Pred	Freq	LSA
Most cowboys know how to ride a	horse	P	H	0.62
Most cowboys know how to ride a	camel	U	L	0.1
In the desert, many Arabs ride a	horse	U	H	0.29
In the desert, many Arabs ride a	camel	P	L	0.64
Before warming the milk, the babysitter took the infant's	bottle	P	H	0.42
Before warming the milk, the babysitter took the infant's	diaper	U	L	0.25
To prevent a mess, the caregiver checked the baby's	bottle	U	H	0.08
To prevent a mess, the caregiver checked the baby's	diaper	P	L	0.6
June Cleaver always serves meat and	potatoes	P	H	0.34
June Cleaver always serves meat and	carrots	U	L	0.4
Bugs Bunny eats lots of	potatoes	U	H	0
Bugs Bunny eats lots of	carrots	P	L	0.05
He scraped the cold food from his dinner	plate	P	H	0.17
He scraped the cold food from his dinner	spoon	U	L	0.19
John stirred the hot soup with the broken	plate	U	H	0.15
John stirred the hot soup with the broken	spoon	P	L	0.35
The cup slipped out of Jim’s hand and hit the	floor	P	H	0
The cup slipped out of Jim’s hand and hit the	dryer	U	L	0.01
Bob folded his clean clothes on the warm	floor	U	H	0.03
Bob folded his clean clothes on the warm	dryer	P	L	0.23
The teacher kept the class quiet while she read a short	story	P	H	0.13
The teacher kept the class quiet while she read a short	diary	U	L	0.15
After writing down her secret thoughts, Sally hides her	story	U	H	0
After writing down her secret thoughts, Sally hides her	diary	P	L	0.48
Wanting children, the newlyweds moved into their first	house	P	H	0.02
Wanting children, the newlyweds moved into their first	igloo	U	L	0.09
The traditional Eskimo family lived in the	house	U	H	0.06
The traditional Eskimo family lived in the	igloo	P	L	0.34
Joey’s mother was horrified by his pet	snake	P	H	0
Joey’s mother was horrified by his pet	shark	U	L	0.02
The man was in dangerous waters when attacked by the	snake	U	H	0.06
The man was in dangerous waters when attacked by the	shark	P	L	0.14
The friends were not talking because they had a	fight	P	H	0.05

The friends were not talking because they had a	brawl	U	L	0.04
John got involved in a bar room	fight	U	H	0
John got involved in a bar room	brawl	P	L	0.03
Jenny left her jacket at work and had to return to the	office	P	H	0.04
Jenny left her jacket at work and had to return to the	locker	U	L	0.14
Ed kept gym clothes in his	office	U	H	0.1
Ed kept gym clothes in his	locker	P	L	0.11
We watched the opening night performance at the	theater	P	L	0.16
We watched the opening night performance at the	circus	U	H	0.09
We love to watch the clowns at the	theater	U	L	0.09
We love to watch the clowns at the	circus	P	H	0.12
The sailor stopped at the deserted	island	P	H	0
The sailor stopped at the deserted	casino	U	L	0
The gambler visited the	island	U	H	0.12
The gambler visited the	casino	P	L	0.12
The camera crew finished filming the	movie	P	H	0.51
The camera crew finished filming the	diver	U	L	0.1
After exploring an underwater cave, the	movie	U	H	0
After exploring an underwater cave, the	diver	P	L	0.11
While away at war, Fred mailed his mother a	letter	P	H	0.05
While away at war, Fred mailed his mother a	compass	U	L	0.02
The lost hiker carefully checked his	letter	P	H	0
The lost hiker carefully checked his	compass	U	L	0.06
He planned to refinish the hardwood	floor	P	H	0.04
He planned to refinish the hardwood	shelf	U	L	0.01
The librarian returned the books to the appropriate	floor	U	H	0.05
The librarian returned the books to the appropriate	shelf	P	L	0.23
After cleaning her teeth, Dr. Sam wiped Mary’s	mouth	P	H	0.35
After cleaning her teeth, Dr. Sam wiped Mary’s	cheek	U	L	0.17
She kissed her old friend on the	mouth	U	H	0
She kissed her old friend on the	cheek	P	L	0.18