COMPUTATIONAL MODELING OF EYE MOVEMENTS – FROM READING TO SCENE VIEWING

A Dissertation

by

HSUEH-CHENG WANG

Submitted to the Office of Graduate Studies
University of Massachusetts Boston,
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2012

Computer Science Program

COMPUTATIONAL MODELING OF EYE MOVEMENTS – FROM READING TO SCENE VIEWING

A Dissertation Presented
by
HSUEH-CHENG WANG

Approved as to style and content by:

Marc Pomplun, Professor
Chairperson of Committee

Dan Simovici, Professor
Member

Jun Suzuki, Associate Professor
Member

Mohinish Shukla, Assistant Professor
Member

Dan Simovici, Graduate Program Director
Computer Science Program

Peter Fejer, Chairperson
Computer Science Program

ABSTRACT

COMPUTATIONAL MODELING OF EYE MOVEMENTS – FROM READING TO SCENE VIEWING

December 2012

Hsueh-Cheng Wang, B.A., National Taiwan University, Taiwan
M.S., National Taiwan University, Taiwan
Ph.D., University of Massachusetts Boston

Directed by Professor Marc Pomplun

My dissertation focuses on developing computational models of eye movements for understanding how cognitive processes (e.g., visual information processing, word recognition, attention, and oculomotor control) can work together to perform a complex everyday task. In a theoretical framework, many biologically-inspired computational methods were used and found psychologically plausible to predict human behaviors and simulate human cognition. For eye movements in reading, I proposed models of visual encoding, word identification, and semantic

integration in contexts. Using singular value decomposition (SVD), I was able to predict the most important strokes for Chinese character recognition (Wang, Angele, Schotter, Yang, Simovici, Pomplun, & Rayner, under revision). Furthermore, I used a vector space model (latent semantic analysis, LSA) to explain how readers rate the semantic transparency of English and Chinese compound words. A linear regression model was then developed to estimate contextual predictability during reading (Wang, Chen, Ko, Pomplun, & Rayner, 2010), and a connectionist model was used to represent the activations of concepts in working memory (Plummer, Wang, Tzeng, Pomplun, & Rayner, 2012).

My interests in reading and vision studies provided interdisciplinary research opportunities, which I pursued by applying methods and concepts from reading research to the viewing of real-world scenes. Regarding eye movements in natural scene viewing, I studied when and where we fixate, resulting in a model for gaze transition using LSA (Wang, Hwang, & Pomplun, 2010; Hwang, Wang, & Pomplun, 2011). The final part of my dissertation focuses on studying how texts attract attention in natural scene viewing (Wang & Pomplun, 2012) compared to attraction by saliency and edge density. I have also developed a model of this effect of texts on visual attention that includes an automatic text detector (Wang, Lu, Lim, & Pomplun, 2012).

The results of my doctoral thesis will broaden our understanding of low-level

and higher-level cognitive processing as well as cultural differences during reading and real-world scene viewing. The findings should eventually lead to practical applications, e.g., contribute to the development of more effective automatic text detectors, or making a great difference to visually challenged people's lives by assisting them in reading and scene viewing.

ACKNOWLEDGMENTS

I would like to thank my advisor Dr. Marc Pomplun, who has always been open and supportive during my Ph.D. studies. He guided me in the right research directions and helped me improve my writing skills. He is my role model, and I will always appreciate what he said: “all that matters to me now is that your career works out well.” I am grateful that God gave me such a great mentor, who is also my good friend to share life experiences.

I would like to thank my thesis committee, Dr. Dan Simovici, Dr. Jun Suzuki and Dr. Mohinish Shukla, for their commitment and for their insightful suggestions on my dissertation work.

I would like to thank my lab members Dr. Alex Hwang and Dr. Chia-Chien Wu; my collaborators at UCSD, Dr. Keith Rayner, Dr. Liz Schotter, Dr. Bernhard Angele, Dr. Jinmian Yang, and Patrick Plummer; my collaborators in Singapore, Dr. Joo-Hwee Lim and Dr. Shijian Lu; my collaborators in Taiwan, Dr. Yi-Min Tien and Dr. Li-Chuan Hsu; and finally all the faculty and staff members at UMass Boston.

Last but not least, I would like to thank my parents and all members of the Boston Taiwanese Christian Church for their unconditional love and support.

STATEMENT ABOUT COLLABORATIVE WORK

In a highly interdisciplinary field such as the modeling of human eye movements during cognitive tasks, collaboration with other researchers from various departments is important. Much of the work reported in this thesis was carried out collaboratively, which I believe led to stronger, more insightful results and a better learning experience for all researchers involved. However, for all of the work included in this thesis, I developed the underlying concepts and methods. Furthermore, I was in charge of conducting all studies and am the lead author of the resulting journal or conference proceedings publications, with the exception of two articles. In these two studies in which I am the second author (Hwang, Wang, & Pomplun, 2011; Plummer, Wang, Tzeng, Pomplun, & Rayner, 2012), my contribution and commitment to the work were significant in the former and equaled that of the lead author in the latter. In the following paragraphs, I will describe my contributions to all studies reported in this thesis.

Chapter 2 is now under minor revision for publication in the *Journal of Research in Reading, special issue on Chinese reading* as Wang, H. C., Schotter, E., Angele, B., Yang, J. M., Simovici, D., Pomplun, M., & Rayner, K. "Using Singular Value Decomposition to Investigate Degraded Chinese Character Recognition: Evidence from Eye Movements during Reading." The idea of applying SVD to Chinese characters was first inspired in the CS724 course at UMass Boston by me, Dr. Marc Pomplun, and Dr. Dan Simovici. The actual stroke removal study was initiated and carried out when I visited the Rayner Eyetracking Lab at UCSD. All authors contributed significantly in this project. This study was awarded a Honorable Mention at the 4th Doctoral Research Symposium, Department of Computer Science, UMass

Boston, USA, 2012. Parts of the data were presented at the 16th European Conference on Eye Movements (ECEM 2011). I was invited as a Guest Speaker to present this study at the China International Conference on Eye Movements (CICEM 2012). Preparation of the article was supported by NIH grants EY021802 to Marc Pomplun, HD026765 to Keith Rayner and a fellowship to Elizabeth R. Schotter from DC000041.

Chapter 3 is now under review for publication in the journal *Behavior Research Methods*. Parts of the data were presented at the Asia-Pacific Conference on Vision (APCV), Taipei, 2010, and I was awarded a Student Traveling Grant (5 awards out of approximately 200 submissions). Partial data were published as Wang, H. C., Tien, Y. M., Hsu, L. C., & Pomplun, M. (2012), "Estimating Semantic Transparency of Constituents of English Compounds and Two-Character Chinese Words using Latent Semantic Analysis" in the *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci),* Sapporo, 2012.

Chapter 4 was published as 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“ in the *Quarterly Journal of Experimental Psychology*, 64, 7, pages 1374-1386. 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. Preparation of the article was partially supported by Grant HD26765 from the National Institute of Health. Thanks to Jinmian Yang, Reinhold Kliegl, and Richard

Shillcock for helpful comments on an earlier version of the article.

Chapter 5 was published as Plummer, P., Wang, H. C., Tzeng, Y. T., Pomplun, M., & Rayner K. (2012), "A Connectionist Model of Concept Activation during Reading using Latent Semantic Analysis and LandScape Model" in the *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci 2012)*, Sapporo, Japan, and the study was awarded a Student Travel Award (21 awards out of 798 contributions). The idea of applying LSA to the LandScape Model was inspired by me and Dr. Yuhtsuen Tzeng in 2010, and later carried out by all authors when I visited the Rayner Eyetracking Lab at UCSD. Preparation of the article was supported by Grant R01EY021802 from the National Eye Institute, USA, to Marc Pomplun, Grant T32 DC00041 from the Center for Research in Language to Keith Rayner, and Grant 97-2410-H-194-090-MY2 from the National Science Council, Taiwan, to Yuhtsuen Tzeng.

Chapter 6 was published as Wang, H.-C., Hwang, A. D. & Pomplun, M. (2010), "Object Frequency and Predictability Effects on Eye Fixation Durations in Real-World Scene Viewing" in the *Journal of Eye Movement Research*, 3(3):3, pages 1-10. It was supported in part by NIH grant R15EY017988 to Marc Pomplun.

Chapter 7 was published as Hwang, A. D., Wang, H. C., and Pomplun, M. (2011), "Semantic Guidance of Eye Movements in Real-world Scenes" in *Vision Research*, 51, pages 1192-1205. Part of the data in Experiment 1 was shared in both Chapters 6 and 7, but distinct aspects of the data were analyzed to address different questions. This research was supported by Grant Number R15EY017988 from the National Eye Institute to Marc Pomplun.

Chapter 8 was published as Wang, H. C. & Pomplun, M. (2012), "The Attraction of Visual Attention to Texts in Real-World Scenes" in the *Journal of Vision*, 12(6):26, pages 1–17. Parts of the data were presented at the European Conference on Visual Perception (ECVP 2010) and the Asia Pacific Conference of Vision (APCV 2011). Thanks to Melissa Võ, Gregory Zelinsky, and an anonymous reviewer for their helpful comments on earlier versions of the article. Preparation of the article was supported by Grant R01 EY021802 from the National Institutes of Health (NIH) to Marc Pomplun.

Finally, Chapter 9 was published as Wang, H. C., Lu, S. J., Lim, J. H., & Pomplun, M. (2012), "Visual Attention is Attracted by Text Features Even in Scenes without Text" in the *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci 2012)*, Sapporo, Japan. This study was initiated in 2011 when I worked as a summer intern in the Computer Vision and Image Understanding (CVIU) Department, Institute for Infocomm Research (I^2R) in Singapore. Subsequent analyses were carried out at the University of Massachusetts, Boston. Preparation of the article was supported by Grant R01EY021802 from the National Eye Institute to Marc Pomplun.

TABLE OF CONTENTS

CHAPTER Page

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

LIST OF TABLES

Table Page

Table Page

CHAPTER 1

INTRODUCTION

In our everyday life, visual information is essential for our interaction with the environment, and sometimes even for our survival. For example, we *read* the newspaper, books, and web pages for retrieving, learning, and comprehending information and ideas. We continually shift our gaze to *inspect scenes* to understand the real world, or search for an object, e.g., look for a key. We *pay attention* to traffic signs or displays showing directions to a hospital or grocery store.

Basic Characteristics of Eye Movements

We inspect our environment by means of very quick eye movements, called *saccades*. Between these saccades, our eyes remain relatively still during *fixations* to acquire visual information. Saccadic eye movements are very fast so that only a blur would be perceived during their execution. However, this does not actually happen, because our perception is essentially “turned off” during saccades, an effect known as “saccadic suppression” (Martin, 1974), which implies that useful information is acquired only during fixations. We make saccades very frequently – approximately 3 to 4 saccades per second - because of our biological *acuity* limitations. When we look ahead, our visual field can be categorized into three regions: *foveal*, *parafoveal*, and *peripheral*, according to the eccentricity from the fixation location. Acuity is very good in the foveal region (the central 2 degrees), decreasing in the parafoveal region (2 to 5 degrees), and poor in the peripheral region (outside 5 degrees). Therefore, we have to move our eyes and allocate the foveal region to where we want to see clearly (Jacobs, 1986; Rayner & Morrison, 1981). The saccade lengths and fixation durations vary depending on different visual stimuli, tasks, and other factors and typically range from 1 to 4 degrees and around 200 to 400 milliseconds, respectively (see Rayner, 1998).

Why is Studying Eye Movements Important?

Tracking an observer's eye movements is a technique that has been widely used to provide important insights into how we process visual information in various tasks, e.g., reading, visual search, scene perception, and natural behaviors, such as music reading, drawing, walking, and driving. (see Rayner, 1998; 2009; Findlay, 2004; Henderson, 2003; Henderson & Ferreira, 2004; Land, 2006, for reviews).

Each task involves very complex processes in our visual system as well as in higher cognitive functions, and these processes rarely occur isolated. For example, reading is a complicated cognitive process of decoding symbols for comprehension, in which the type of the text (e.g., the language in which it is written) strongly interacts with the reader’s prior knowledge. Similarly, scene viewing involves dynamic processes. A viewer may direct his or her attention to an important location or search for important (salient) visual inputs using low-level visual processing. In addition to visual attention, a viewer recognizes objects, understands the scene or interprets a social situation using higher-level processing, and these processes may coordinate with each other.

Interdisciplinary Approaches

Scientists in different fields, such as artificial intelligence, linguistics, and psychophysics, are devoted to a common goal: to understand the nature of human vision and its relation to cognition and mind. Therefore, it is important to bring together investigations from diverse disciplines and perspectives. *Computational models* are theoretical frameworks for understanding how human cognitive processes (e.g., visual information processing, word/ object recognition, attention, and oculomotor control) can work together to perform a complex everyday task, reflecting the constraints and strengths of human cognitive mechanisms. Therefore, computational models are particularly well suited for interdisciplinary approaches underlying human vision across stimuli and tasks. This dissertation is aimed at developing computational models that can be applied to reading and scene viewing.

Models of Eye Movements in Reading

Eye movements provide an indication of language processing because they are sensitive to how people process words with different lexical characteristics such as *word frequency* (the normative frequency of occurrence in a text corpus). Specifically, eye fixation times are longer on low-frequency words than on high-frequency words (Rayner, 1998). Furthermore, predictability of target words in the context of a (beginning) sentence has been found to have a strong influence on eye movements (Rayner & Well, 1996; see also Ehrlich & Rayner, 1981). In Rayner and Well’s (1996) experiment, subjects fixated unpredictable target words longer than either highly or moderately predictable target words; highly predictable words were also skipped more often than moderately predictable or unpredictable target words. Typically, researchers use a number of word-based measures of eye-movement behaviors to reflect different stages of visual and lexical processing, including *first fixation duration* (FFD, 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), *gaze duration* (GD, the sum of all fixation durations prior to moving to another word), and *total time* (TT, the duration sum of all fixations on a word including regressions; see Reichle, 2003, for a review).

Some computational models of eye movements in reading were proposed. In the E-Z Reader model (Reichle, Pollatsek, Fisher, & Rayner, 1998; Reichle, Rayner, & Pollatsek, 1999; 2003), word frequency and predictability within a given sentence context influence how we identify the orthographic, phonological, and semantic form of a word. Kliegl, Grabner, Rolfs, and Engbert (2004) used a statistical control approach to examine the effect of lexical 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.

Models of Eye Movements in Scene Viewing

Real-World Scene Viewing

Studies in reading and scene viewing have been found to share some common mechanisms, for example, that our visual system recognizes visual stimuli (words in a sentence or objects in an image) through a hierarchical process that includes a part-based stage beginning with independent feature detection (e.g., Biederman, 1987; Hubel & Wiesel, 1962, 1963; McClelland & Rumelhart, 1981; Selfridge, 1959; Taft, Zhu, & Peng, 1999). Subsequently, these features are combined into higher-level, more meaningful components for the semantic activation of the overall concept (i.e., the word or object), or for further sentence comprehension or scene understanding.

However, there are many fundamental differences between reading and scene viewing, e.g., unlike the left-to-right sweep over each sentence in English reading, there is no regular direction of visual scanning in scene viewing. Traditionally, eye-movement studies during reading and scene viewing have been addressing different issues, and the data have been analyzed using separate paradigms and methods.

Models of Visual Attention

Visual attention has been a great interest for vision scientists. Studies have indicated that visual attention and eye movements are tightly coupled during scene inspection or search tasks (see Findlay, 2004), including the factors *low-level visual saliency* (e.g., Itti, Koch & Niebur, 1998; Bruce & Tsotsos, 2006; Itti & Koch, 2001; Parkhurst, Law, & Niebur, 2002), and *top-down control* (e.g., Hwang, Higgins, & Pomplun, 2009; Peters & Itti, 2007; Pomplun, 2006; Zelinsky, 2008). It is also important to consider the relative contributions of *objects* and low-level features. Elazary and Itti (2008) used the LabelMe image dataset (Russell, Torralba, Murphy & Freeman, 2008) to examine the relation between objects and low-level saliency, as computed by Itti et al.'s model, and they found that salient locations tend to fall within "interesting objects" defined by objects people choose to label. Their finding was later refined by Nuthmann and Henderson (2010), who showed that viewers tend to fixate close to the center of objects and emphasized the importance of objects in memorization and preference tasks. Einhäuser, Spain and Perona (2008) further investigated whether observers attend to interesting objects by asking them to name objects they saw in artistic evaluation, analysis of content, and search tasks. Einhäuser et al. (2008) found that saliency combined with object positions determines which objects are named frequently. They concluded that both low-level saliency and objects need to be integrated in order to capture attention. Attentional capture could be driven by some particular classes of objects, which attract eye fixations independently of their low-level visual saliency. There may be specific features of texts, similar to faces, that attract attention but differ from those features that are typically associated with visual saliency. For instance, Cerf, Harel, Einhäuser, and Koch (2007) showed that a model combining low-level saliency and face detection achieved better estimation of fixation locations than low-level saliency alone. Similarly, Judd, Ehinger, Durand, and Torralba (2009) added object detectors for faces (Viola & Jones, 2004) and persons (Felzenszwalb, McAllester, & Ramanan, 2008) to their model and obtained better prediction of human fixations. Cerf, et al. (2009) refined the standard saliency model by adding a channel indicating regions of faces, texts, and cell phones, and demonstrated that the enhancement of the model significantly improved its ability to predict eye fixations in natural images.

Computational and Experimental Methods

In the theoretical framework of computational models, many computational methods are found psychologically plausible to predict human behaviors observed from experimental methods. During reading, for example, *surprisal,* a measure of syntactic complexity, was found to predict first fixation duration, (first pass) gaze duration, and total time (see Boston, Hale, Kliegl, Patil, & Vasishth, 2008; Demberg & Keller, 2008). 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, was used to predict eye movements (Ong & Kliegl, 2008). For scene viewing, studies have been using saliency models to predict the distribution of human fixations (Bruce & Tsotsos, 2006;

Harel, Koch, & Perona, 2007; Einhäuser, Spain, & Perona, 2008; see also Zhao & Koch, 2011).

It is important to determine how predictive a computational model is to experimental data. Among many metrics, receiver operating characteristic (ROC, Green & Swets, 1966) has been widely applied to many binary classifier problems, especially in eye movements in scene viewing. ROC is a signal detection measure, which is used to evaluate the discriminatory performance for target and non-target distributions. As proposed by Tatler, Baddeley, and Gilchrist (2005), pixels are selected and labeled either as fixated (target) or non-fixated (non-target) locations according to a subject's eye-movement data. A threshold is systematically moved, and the hit rate and false-alarm rate of discrimination between target and non-target are changed. A fixated location with a salience value above a given threshold is a true positive (hit), whereas a non-fixated pixel with salience value greater than the same threshold is a false positive (false alarm). A curve is then plotted indicating the false alarm rate as a function of the hit rate, and the area under the curve (AUC) is used to represent the discriminatory performance. If a model can completely separate the distributions of targets and non-targets, the AUC will be 1 (perfect performance), whereas the AUC will be 0.5 for chance level (random performance, no predictive power).

This dissertation specifically focuses on how *singular value decomposition (SVD)* and *Latent Semantic Analysis (LSA)* can be predictive and correlated with the results from experimental data.

Singular Value Decomposition

SVD (Strang, 1993) is a powerful linear algebra factorization technique for a rectangular matrix. Similar to Principal Component Analysis (PCA), SVD is a dimension reduction method in linear algebra to retain the least redundant components contained in a matrix (see Elden, 2007). These dimension reduction methods have been extensively used for pattern recognition in many vision problems. In the face recognition domain, when faces are correctly aligned and scaled, a reduced set of orthogonal functions can be used to reconstruct faces (Craw & Cameron, 1991; Turk & Pentland, 1991). For natural scene images, it can also been used to obtain just enough visual input for scene gist recognition (see Torralba & Oliva, 2003, for a review). As shown in Figure 1, any m x n matrix A can be decomposed into a product of three matrices,

$$\mathbf{A} = \mathbf{U} \Sigma \mathbf{V}^{\mathrm{T}}$$

$$= \underbrace{\begin{bmatrix} | & | & | \\ \mathbf{u}_1 \cdots & \mathbf{u}_r \cdots & \mathbf{u}_m \\ | & | & | \end{bmatrix}}_{m\times m} \underbrace{\begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix}}_{m\times n} \underbrace{\begin{bmatrix} - & \mathbf{v}_1 & - \\ & \vdots & \\ - & \mathbf{v}_r & - \end{bmatrix}}_{n\times n},$$

Figure 1: Singular Value Decomposition

where U is an m x m matrix of orthonormal columns, V^T is an n x n matrix of orthonormal rows, and ∑ is a non-negative m x n matrix with singular values σ_1, …, σ_r. Most software packages for numerical calculations such as MATLAB (The MathWorks, Inc, Natick, MA) contain the computation of SVD.

Latent Semantic Analysis

SVD was first introduced to the field of information retrieval (see Dumais, Furnas, Landauer, Deerwester, & Harshman, 1988; Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990), and later widely used in psycholinguistic studies, known as LSA. LSA is a technique that applies SVD to a text corpus in order to represent the meaning of words by statistical computations (Landauer & Dumais, 1997; Landauer, McNamara, Dennis, & Kintsch, 2007; Martin & Berry, 2007). 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. Typically, terms are words, and a term-to-document co-occurrence matrix is established from a corpus. Furthermore, global and local weighting are 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 1 and 2. In Equation 2, tf_{ij} is the term frequency of term i in document j, and gf_i is the total number of times that term i appears in the entire collection of n documents (Dumais, 1991).

$$\textit{Local weight = log (term frequency + 1)} \quad (1)$$

$$Global\ weight = \sum_j \frac{p_{ij} \log_2(p_{ij})}{\log_2 n} \quad where\ p_{ij} = \frac{tf_{ij}}{gf_i} \quad (2)$$

For large datasets, empirical testing shows that the optimal choice for the number of dimensions ranges between 100 and 300 (Berry, Drmac, & Jessup, 1999; Jessup & Martin, 2001; Lizza & Sartoretto, 2001).

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 between -1 and 1, but the dot product (and the cosine) tends to have a lower bound close to zero.

Objectives and Organizations

With the theoretical foundation and technical understanding of what SVD and LSA are accomplishing and how this is achieved, we will be able to establish computational models to answer the fundamental questions in reading and scene viewing, i.e., what strokes are more important in a Chinese character than others? What is the semantic relationship between two words or between two objects? How does the context in a sentence give us a hint for an upcoming word? Does the semantic relationship of two objects in a scene affect where and how long we fixate? Do the characteristics and features of a word or an object influence our visual attention?

The first part of this dissertation, Chapters 2 to 5, employs SVD and LSA to address specific questions about eye movements in reading. We will start in Chapter 2 with how SVD predicts the important strokes of Chinese characters and allows us to investigate the mechanism of visual encoding of degraded character recognition. In

Chapter 3, LSA is used to estimate semantic transparency of English and Chinese compound words and how the compound words are processed by English and Chinese speakers. Chapters 4 and 5 address the semantic integration in the sentence context for an upcoming target word using LSA and a connectionist model.

The second part, Chapters 6 to 9, investigates eye movements in scene viewing. Adopting the paradigm and analysis methods from reading studies, Chapter 6 investigates the processing time of objects during scene viewing. Chapter 7 applies the linguistic-based LSA technique to visual search and scene viewing experiments and demonstrates how semantics influence viewers' eye fixations. Chapter 8 investigates a special class of objects - “texts” in real-world scenes - and reveals the factors influencing the attentional capture of texts. Chapter 9 extends the findings from Chapter 8 by adding an automatic text detector to a visual attention model.

Finally, the conclusions regarding the general findings derived from the computational models based on SVD and LSA, the implications of linguistic and visual processing for words, objects, and scene texts, cross-linguistic investigations, and practical applications, are given in Chapter 10.

CHAPTER 2

USING SINGULAR VALUE DECOMPOSITION TO INVESTIGATE DEGRADED CHINESE CHARACTER RECOGNITION: EVIDENCE FROM EYE MOVEMENTS DURING READING

As mentioned above, since the dimension reduction methods have been extensively applied to solve many problems in vision research, SVD may be adopted to determine what information is important for successful word recognition. It is known that not all letters are of equal importance to the word recognition process. When letters of words are replaced by other letters, changes to *initial* letters are more disruptive than changes to medial or final letters (Rayner & Kaiser, 1975; Rayner, White, Johnson, & Liversedge, 2006). Furthermore, *exterior* letters are more important than word internal letters (Jordan, Patching, & Thomas, 2003; Rayner et al., 2006). These data suggest that in alphabetic languages the position of the letters within a word is very important and some letter positions are more important than others. What about component features of a very different orthography, like Chinese, which does not use letters?

Similar results were found in a previous study on stroke removal during Chinese reading, which indicates that removing initial strokes from Chinese characters makes them harder to read than removing final or internal ones (Yan, Bai, Zang, Bian, Cui, Qi, Rayner, & Liversedge, 2012). In contrast to alphabetic languages, the orthographic units of a Chinese character are written within a single box, not in a linear order. These units, *strokes,* are simple features (e.g., dots, lines, or curves) or combinations of simple features that vary in complexity (Zhang, Wang, Zhang, & Zhang, 2002). For example, the strokes of the character 场 in its writing order are 一 , 丨 , ㇀ , ㇋, ㇒ , and ㇒, where the fourth stroke ㇋ is composed of 5 features. Strokes are written in a defined order that generally follows the order of left to right, top to bottom, and exterior to interior. Furthermore, each stroke's contribution to the character is less clear in Chinese. While in alphabetic orthographies, one or more lines represent a letter, which in turn represents a sound (although alphabetic languages differ in the degree to which letters correspond directly to sounds), in Chinese, each stroke or few strokes does not necessarily correspond to a sound or other unit of representation, but rather the configuration of all the strokes together contributes to the character's overall meaning.

Not surprisingly, without all the strokes of a character intact, subjects have difficulty identifying or remembering the character, depending on the extent and nature of stroke removal. Tseng, Chang, and Wang (1965) removed different portions of the strokes of characters ranging from 10% to 60%, using different methods: from the beginning or from the ending of the canonically written stroke order, or strokes that did not greatly change the character's visual configuration. Subjects were required to fill in the strokes that had been removed and performed better when the ending strokes were removed than when beginning strokes were removed and best when the basic configuration of the character was retained.

Yan et al. (2012) showed similar results to Tseng et al. for subjects reading sentences in which the characters had 15%, 30%, or 50% of their strokes removed either at the beginning or the end of the writing order, or strokes that did not contribute to the character's configuration. They found that when 30% or 50% of strokes were removed reading was disrupted. Furthermore, in those conditions, characters in which the strokes that do not contribute to the configuration were removed were easiest to read, characters with ending strokes removed were more difficult to read, and characters with beginning strokes removed were the most difficult to read. Taken together, these data suggest that the strokes of a character that are written first are more important than the strokes that are written last or the ones that do not contribute to character configuration. The results are generally consistent with the findings of alphabetical languages that beginning or exterior letters seem to be more important than ending or interior ones.

However these studies with Chinese raise the question of whether there is something privileged about the first-written strokes or whether another aspect of the strokes at the beginning of the writing order is what causes them to be more important for character identification. To test this, we turned to *SVD* to investigate whether the contribution of these strokes to the configuration of the character drives their importance for identification. In this study, we used four oriented filters to decompose each character into simple segments that roughly map on to features such as oriented lines of vertical, horizontal, and diagonal orientations, as presented in Figure 2. Representing characters by their oriented line vectors may be well-adapted to a biologically plausible model (such as the feature level in Taft et al., 1999) of character identification for early visual processing. In order to combine these features into higher-level, more meaningful components that contribute to activation of the overall concept, we used SVD as a possible method to determine which segments contribute the most information to Chinese characters. The actual input into SVD is a matrix containing 40x40 grayscale pixels of the original character, which provide only low-level visual information (such as features, orientations, spatial structure) but not higher-level linguistic information (such as writing order or the composition of radicals).

SVD is particularly well suited to decompose and summarize Chinese characters because the orientations of strokes are typically vertical, horizontal, diagonal, or combinations thereof. Thus, linear algebra and SVD are easily adapted to Chinese characters. The least redundant components determined by SVD were considered the most important ones for recognition in this study. As we mentioned above, strokes may be simple features or combinations of simple features that vary in length and complexity. In contrast to previous studies, the SVD method of stroke deletion avoids the problem of stroke complexity by decomposing strokes into segments, thus reducing bias toward longer or more complex strokes seeming more important or disruptive if deleted.

In the present study we investigated whether the most important strokes in a character are, in fact, those that are the first to be written, as suggested by Tseng et al. (1965) and Yan et al. (2012). Furthermore we investigated whether the most important strokes can be identified by SVD. Importantly, if SVD identifies that the most disruptive strokes to delete are the same as those identified by Tseng et al. (1965) and Yan et al. (2012), it would suggest that there is something about the characters in those positions that are more informative (less redundant) than other characters, and that aspect must be part of the visual configuration of the character because visual configuration is the only input to the SVD algorithm. We had three hypotheses: (1) when the most important segments are removed, reading would be most disrupted, (2) when the least important segments are removed, reading would be least disrupted, and (3) the most important strokes would be those that retained the character configuration and the least important to be those that were less related to character configuration.

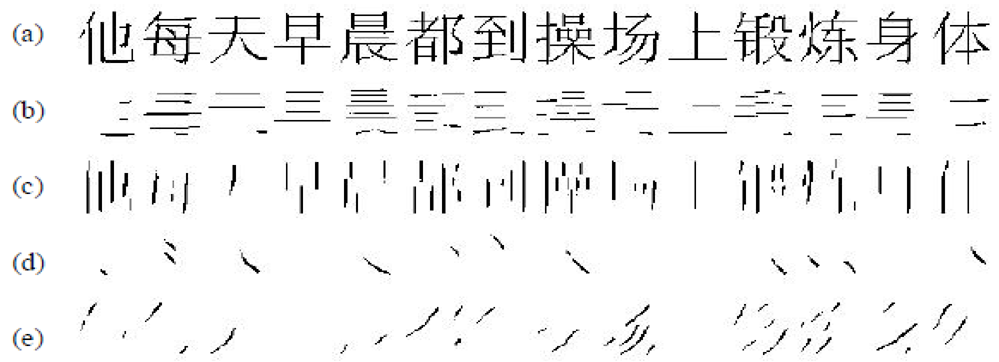


Figure 2: Decomposition of characters. (a) all segments, (b) horizontal segments, (c) vertical segments, (d) diagonal (top-left to bottom-right) segments, and (e) diagonal (top-right to bottom-left) segments.

Similar to Yan et al. (2012), we had subjects read sentences with characters that had elements deleted. In contrast to Yan et al., we deleted segments, as opposed to strokes and only used one degree of removal: 30%. In the present study, we removed: (1) the least important segments according to SVD, (2) the most important segments according to SVD, or (3) randomly selected segments. We also included a control condition in which no segments were deleted.

To use SVD to delete segments, we first determined the important components identified by SVD as most (highest ranked) or least (lowest ranked) important (see Figure 3). Using the same principles, we then reconstructed the characters using only a subset of the components. As demonstrated in Figure 4, reconstruction of characters using some components from SVD results in some pixels from a given segment being retained, but not all. To create the characters that were ultimately used in the experiment, we retained a segment when the SVD method reconstructed 35% of the original segment.

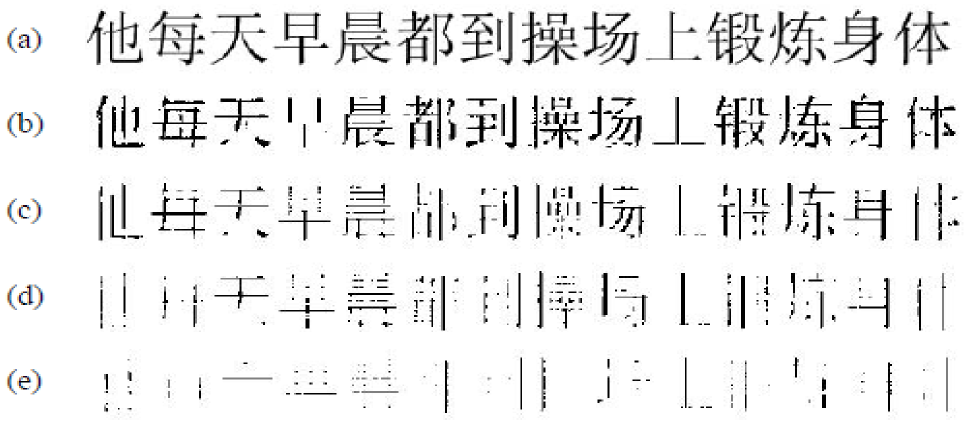


Figure 3: SVD reduction for Chinese characters. Sentence (a) is the original sentence without any reduction. Sentences (b) to (e) are the sentences after removing the least important 20%, 40%, 60%, and 80% information as determined by SVD.

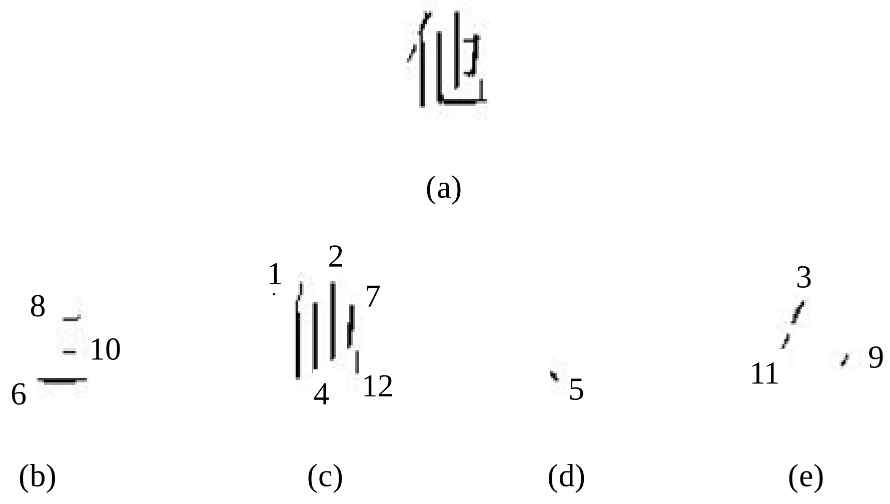


Figure 4: Determining the importance of segments by SVD. (a) Original character, (b) horizontal segments, (c) vertical segment, (d) to (e) diagonal segments. Numbers 1 to 12 represent from the most (highest ranked) to least (lowest ranked) segments identified by SVD.

Method

Subjects. Sixteen students at the University of Massachusetts at Boston were recruited for the experiment. All subjects were native speakers of Chinese between the ages of 19 and 30 years old with normal or corrected-to-normal vision. Each participant received 10 dollars for participation in a half-hour session.

Materials. Sentences were taken from the fifty sentences in Yan et al. (2012), two sentences were used as practice stimuli and 48 were used in the experiment. Each sentence contained 14 characters for which character frequency ranged from 21 to 38,075 per million (mean: 3343), frequencies were taken from the China Newspaper Database (see Yan et al., 2012). The 48 sentences for the experiment were presented in four experimental conditions: (1) all segments retained, (2) the least important 30% of segments removed, (3) the most important 30% of segments removed, and (4) 30% of segments randomly selected to be removed. The degree of deletion was set to be as close to 30% of total number of pixels as possible. This method, deleting characters based on the number of pixels, does not consider spatial information such as orientation or length of segments. For example, two segments with the same length but different thicknesses, would not be considered equivalent sizes. Stimuli were counter-balanced so that each subject read 12 sentences in each condition, and across all subjects, each sentence was seen equally often in each condition. Each sentence was followed by a true/false comprehension question. Examples of these stimuli are shown in Figure 5.

Apparatus. Eye movements were recorded using an SR Research EyeLink 1000 system with a sampling frequency of 1000 Hz run in desktop mode. After calibration, the average calibration error was 0.5˚. Stimuli were presented on a 19-inch Dell P992 CRT monitor with a refresh rate of 85 Hz and a screen resolution of 1024×768 pixels.

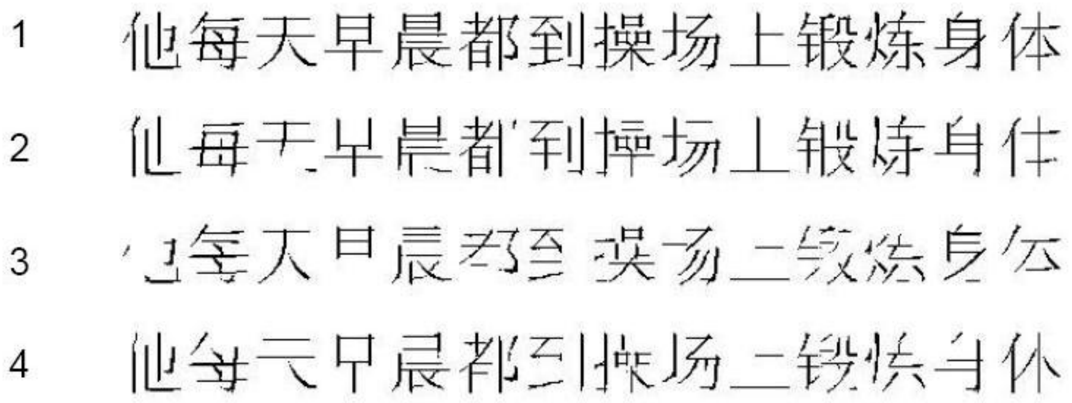


Figure 5: Four conditions of reading stimuli. Sentence 1 represents the all retained condition, sentence 2 represents the least important 30% segments removed condition, sentence 3 represents the most important 30% segments removed condition, and sentence 4 represents the 30% randomly selected segments condition.

Procedure. Subjects were instructed to read the degraded sentences and try to understand them as accurately as possible. At the beginning of the experiment, a standard 3-point calibration and validation of the gaze recording were completed. Following two practice sentences, subjects viewed 48 sentences in random order. Each sentence was followed by a true or false question. At the start of each sentence, a calibration box appeared at the position of the first character, and once a fixation was detected inside the box, the sentence appeared. If the fixation did not trigger the sentence to appear, the experimenter recalibrated the tracker and continued the experiment. Subjects read the sentence at their own pace, and pressed a button to indicate they had finished. Subsequently, the stimulus disappeared and the question appeared, and subjects responded yes or no by pressing the corresponding button.

Results

We examined a number of *global reading measures,* which index reading efficiency (Rayner, 1998, 2009). These measures include *comprehension accuracy, total sentence reading time, mean fixation duration, number of progressive saccades, number of regressive saccades*, and *mean forward saccade length*. In general, reading efficiency is positively related to comprehension accuracy and mean forward saccade length and negatively related to total sentence reading time, mean fixation duration, number of progressive saccades and number of regressive saccades. Trials with a total sentence reading time of less than 1000 ms were considered as accidentally terminated (see means and standard deviations in Table 1), and trials with total reading time above four standard deviations from the mean were considered outliers. These trials were excluded from analysis, which removed 5.8% of the data. Means and standard deviations of global reading measures are shown in Table 2.1. Repeated measures one-way ANOVAs by-subjects (F_1) and by-items (F_2) for four experimental conditions was performed, and then post-hoc tests using SPSS corrected Bonferroni adjusted p-values were carried out for paired comparisons between conditions.

Comprehension Accuracy. Comprehension accuracy was high in all conditions: 90% in the all retained condition, 89% in the least important removed condition, 80% in the most important removed condition, and 85% in the randomly removed condition. One-way ANOVAs revealed an overall difference in accuracy across the four conditions, $F_1(3; 45) = 3.37$, $p < .05$, $F_2(3; 141) = 4.61$, $p < .01$. Post-hoc tests indicated that texts in the least important removed condition were comprehended better than in the most important removed condition by subjects and by items (both *p*s < .05). The comparison between the all retained condition and the most important removed condition was significant by items ($p < .05$) but not by subjects (p = .21). None of the other comparisons were significant (all *p*s > .21). These results indicate that participants read and comprehended the sentences well, but that removing the segments identified as the most important via SVD caused some comprehension difficulty.

Table 1: Means and standard deviations of global reading measures (ComAcc, Comprehension Accuracy; TSR, Total Sentence Reading Time; MFD, Mean Fixation Duration; NPS, Number of Progressive Saccades; NRS, Number of Regressive Saccades; MSL, Mean Saccade Length) across experimental conditions. Standard deviations are shown in parentheses.

	Removal condition			
	All Retained	Most Important	Least Important	Randomly selected
ComAcc	0.90 (0.08)	0.80 (0.13)	0.89 (0.08)	0.85 (0.08)
TSR (s)	3.98 (1.30)	9.90 (4.12)	4.29 (1.18)	5.73 (1.59)
MFD (ms)	209 (40)	254 (36)	219 (34)	226 (36)
NPS	10.00 (3.29)	19.76 (7.81)	10.52 (2.59)	13.09 (3.43)
NRS	4.96 (2.09)	12.98 (5.82)	5.27 (1.98)	7.23 (2.44)
MSL (degrees)	2.20 (0.59)	1.76 (0.43)	1.93 (0.48)	1.87 (0.49)

Total sentence reading time. One-way ANOVAs revealed significant differences across the conditions (see Figure 6; $F_1(3; 45) = 31.02$, $p < .001$, $F_2(3; 141) = 54.21$, $p < .001$). Post-hoc tests showed that subjects read slowest in the most important removed condition compared to each of the other conditions (all *p*s < .01). Subjects read slower in the randomly removed condition than in the all retained condition (both *ps* < .001) and the least important removed conditions by-subjects (*p* < .001) and marginally by-items (*p* = .06). There was no significant difference between the least important removed condition and the all retained conditions (both *ps* > .12). These results indicate that reading fluency was reduced not only in the most important removed condition, but also the randomly removed condition. In contrast, readers read the sentences as efficiently in the least important removed condition as in the all retained condition.

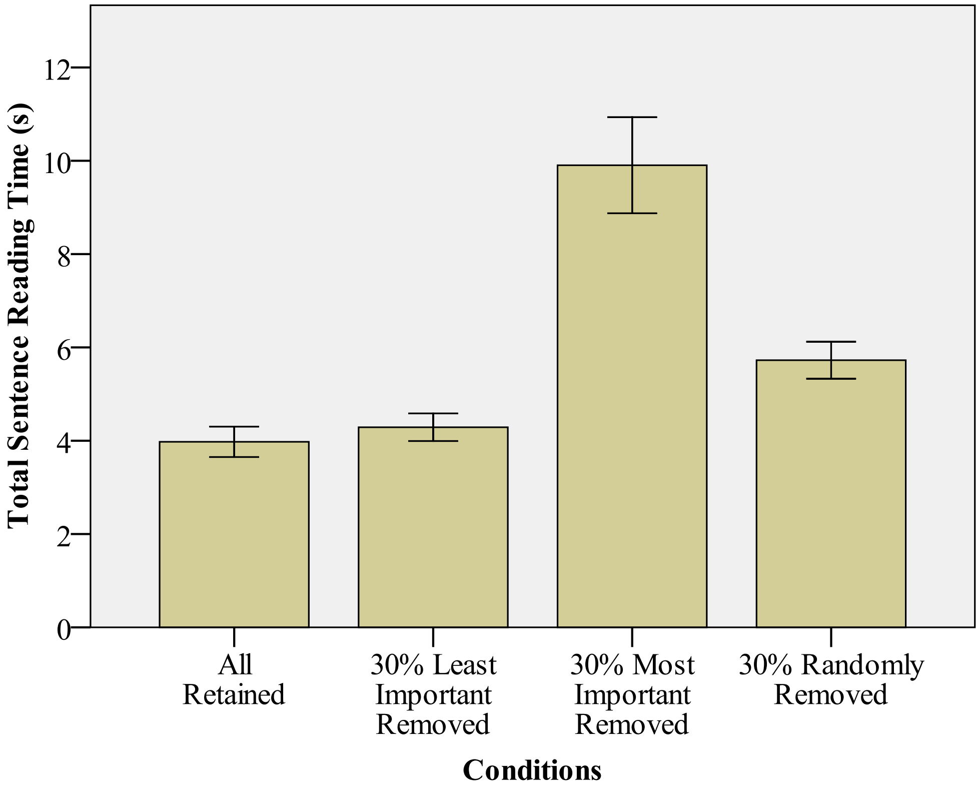


Figure 6: Total sentence reading time as a function of condition. Error bars are based on standard errors.

Mean fixation duration. One-way ANOVAs revealed a significant overall effect of condition on mean fixation duration ($F_1(3; 45) = 47.20$, $p < .001$, $F_2(3; 141) = 25.98$, $p < .001$). Post-hoc tests revealed that the mean fixation duration in the most important removed condition were significantly longer than each of the other conditions (all *p*s < .001). Mean fixation duration in the randomly removed condition was longer than that in the all retained condition ($p < .01$). Mean fixation duration in the least important removed condition was longer than in the all retained condition in the by-items analysis ($p < .05$) but not in the by-subjects analysis (p = .09). None of the other comparisons were significant (all *p*s > .09).

Number of saccades. Similar to the total sentence reading time analyses, One-way ANOVAs revealed a significant overall effect of condition on the number of progressive saccades ($F_1(3; 45) = 28.20$, $p < .001$, $F_2(3; 141) = 48.71$, $p < .001$), see Figure 7. Post-hoc tests revealed that there were more progressive saccades in the most important removed condition compared to each of the other conditions (all *p*s < .01). More progressive saccades were produced in the randomly removed condition than in the all retained and the least important removed conditions (both *ps* < .05). There was no significant difference between the least important removed and the all retained conditions ($p > .13$).

One-way ANOVAs revealed the overall effect of condition on the number of regressive saccades to be significant ($F_1(3; 45) = 29.66$, $p < .001$, $F_2(3; 141) = 51.41$, $p < .001$). Post-hoc tests showed that there were more regressive saccades in the most important removed condition compared to each of the other conditions (all *p*s < .01). More regressive saccades were found in the randomly removed condition than in the all retained and the least important removed conditions (both *ps* < .05). There was no significant difference between the least important removed and the all retained conditions ($p > .52$).

Mean forward saccade length. One-way ANOVAs revealed a significant overall effect of condition ($F_1(3; 45) = 29.62$, $p < .001$, $F_2(3; 141) = 7.12$, $p < .001$). Post-hoc tests of both by-subjects and by-items analyses indicated that forward saccade length in the all retained condition was longer than in the least important removed, the most important removed, and randomly removed conditions in by-subjects analysis (all ps < .001), and in by-item analysis (p = .06, p < .01, and p < .05, respectively). Forward saccade length in the least important removed condition was significantly longer than in the most important removed condition in by-subjects analysis (p < .05), but not in by-items analysis (p > .24). Forward saccade length in the randomly removed condition was marginally longer than in the most important removed condition in the by-subjects analysis (p = .07), but not in the by-items analysis (p > .50).

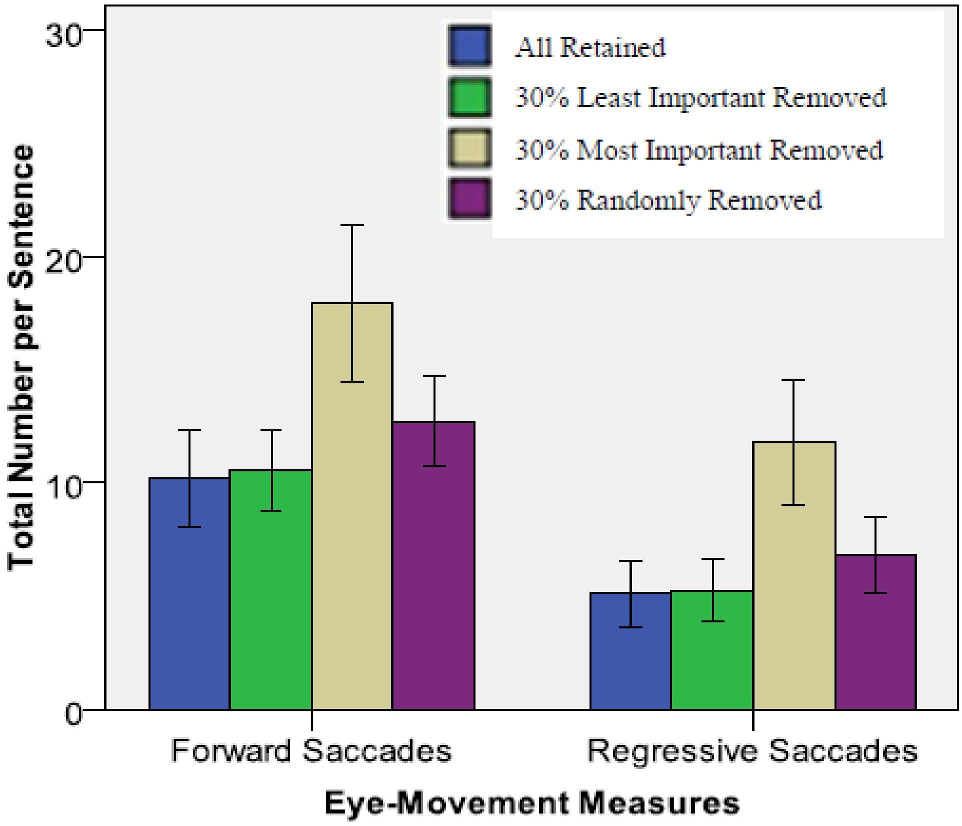


Figure 7: Number of forward and regressive saccades per sentence. Error bars are based on standard errors.

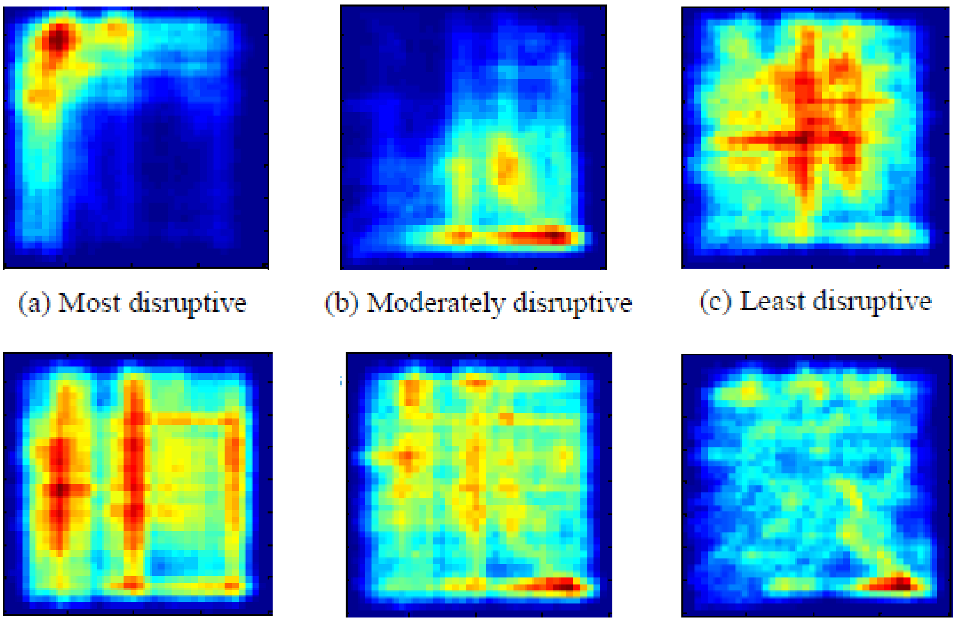
Taken together, these results suggest that reading was most disrupted when the most important segments were removed, moderately disrupted when randomly selected segments were removed, and least disrupted (equivalent to reading with all strokes retained) when the least important strokes were removed. Our data are similar to the data reported by Yan et al. (2012), in that some conditions (beginning strokes removed in their study and most important segments removed in our study) were more disruptive than others (ending strokes removed in their study and randomly selected strokes in our study). Furthermore, both studies found that removing certain character elements (those that did not contribute to the configuration in their study and the least important segments in our study) did not alter reading compared to reading intact characters.

Additional Analyses

The overall results indicate that the mathematical method SVD captured the most informative segments of Chinese characters. However, it is possible that SVD identifies the same strokes that were deleted in the Yan et al. study. Alternatively, SVD may identify a different set of segments in the characters, but if so, what is it about those segments that are most informative and impair reading most when deleted?

Distribution of Degradation Position

Since stroke order is correlated to stroke position (e.g., beginning strokes tend to be located at the top left position of characters) we compared the distribution of deleted elements from the different conditions in Yan et al. (2012) and the conditions used in the current study. As shown in Figure 8, the distribution of the positions where strokes were removed in the Yan et al. (2012) study is different from the distribution in the current study. As expected for the Yan et al. materials, the top left positions tended to be removed in the beginning stroke removal condition, the bottom right positions tended to be removed in the ending strokes removal condition and character internal positions tended to be removed in the configuration retaining condition. In contrast, while SVD identified the most important segments as those located on the left side of the character, they are more widely distributed than those in the Yan et al. study. Similarly, SVD identified the least informative segments as those in the bottom right location of the character, again, with a wider distribution that in the Yan et al. study. Lastly, the configuration retaining condition (from the Yan et al. study) and the randomly selected condition (from the current study) are similarly centered around the middle of the character, but again, with a wider distribution with the SVD method. Thus, it seems that the beginning strokes/segments (i.e., those located in the upper and left-side positions of the character) tend to be more important for character identification than those that are located in the bottom and right-side positions.



(a) Most disruptive (b) Moderately disruptive (c) Least disruptive

(d) Most disruptive (e) Moderately disruptive (f) Least disruptive

Figure 8: (a) to (c): Distributions of removed strokes (30%) in Yan et al. (2012) which were the most, moderately, and the least disruptive. (a) Beginning strokes, (b) ending strokes, and (c) configuration retaining strokes. (d) to (f): Distributions of removed segments (30%) in this study which were the most, moderately, and the least disruptive. (d) Most important segments removed, (e) random segments removed, and (f) least important segments removed.

Because these locations were also identified by SVD, which has no information about the order of writing strokes, writing order may not be the cause of these strokes' importance, but rather be correlated with it, and character configuration may be the important factor instead. However, the configuration of a Chinese character is not well-defined. In the following analyses, we propose a computational method for representing character configuration and measuring the degradation percentage of characters in Yan et al. (2012) and this study.

Measuring Character Configuration using Contour

In object recognition, contour is important for successful recognition of degraded objects; observers are more accurate at identifying degraded characters when vertices are retained than when the midsections of lines are retained (see Biederman, 1987, for a review). We extracted vertices from each segment and simplified the contour of the character using *convex hull*–the shape formed by a "tight rubber band" that surrounds all the vertices, shown in Figure 9.

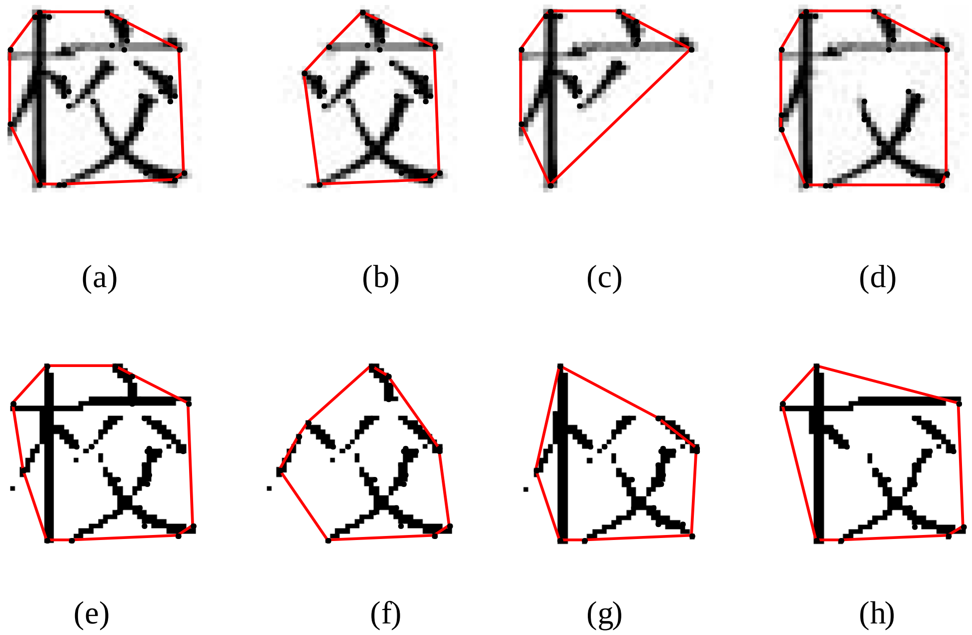


Figure 9: Defining the contours of characters using convex hull. (a) to (d) are sample characters in Yan et al. (2012). (a) Character without removal, (b) the most disruptive removal (beginning strokes), (c) moderately disruptive removal (ending strokes), and (d) the least disruptive removal (configuration retaining). (e) to (h) are sample characters in this study. (e) All retained, (f) the most disruptive removal (most important segments), (g) moderately disruptive removal (randomly selected segments), and (h) the least disruptive removal (least important segments).

Two similarity measures were computed between the original character and each of the degraded characters for the Yan et al. stimuli and our own stimuli: the *proportion of overlapping vertices* and the *proportion of overlapping perimeters*. Overlapping vertices are the number of matching vertices, and overlapping perimeters are the sum of length of matching edges. As shown in Table 2, the results indicate that the least disruptive conditions yielded the highest similarity with the original characters, while the most disruptive conditions yielded the lowest similarity. A one-way ANOVA for the degree of disruption (most, moderately, and least disruptive conditions; 48 sentences each condition) was performed, and then post-hoc tests using SPSS corrected Bonferroni adjusted p-values were conducted for paired comparisons between conditions. In the analysis for the Yan et al. (2012) stimuli, the results showed significant main effects of the proportion of overlapping vertices and perimeters (F(2; 94) = 589.89, p < .001 and F(2; 94) = 390.51, p < .001), respectively. Post-hoc tests revealed that the proportion of overlapping vertices and perimeters of most disruptive conditions was significantly lower than the ones of moderately and least disruptive conditions (all *p*s < .001), and the ones of the moderately disruptive conditions was significantly lower than the ones of the least disruptive conditions (all *p*s < .001). In the current study, there was also an overall effect across conditions for the proportion of overlapping vertices (F(2; 94) = 52.72, p < .001) and perimeters (F(2; 94) = 68.60, p < .001). Post-hoc tests revealed that the moderately and least disruptive conditions contained characters with higher proportions of overlapping vertices and perimeters than the most disruptive conditions (all *p*s < .001). However, while the least disruptive conditions had slightly more overlapping vertices and perimeters than the moderately disruptive conditions, this difference was not statistically significant (both *ps* > .34). These results indicate that the least disruptive conditions are those that retained most of the contours of the original characters. Thus, degrading the contour of the original character is likely what made the beginning stroke condition (Yan et al. study) and the most important removed condition (the present study) lead to the poorest reading efficiency. The non-significant results between least and moderately disruptive conditions in the current study may imply that contour is not the only factor that influences the degree of disruption produced by deleting strokes; degradation position may also be important.

General Discussion

This study investigated how readers recognized Chinese characters that were degraded using SVD to identify the most important and least important segments. We found that reading was most impaired when subjects read sentences with the most important segments removed. Reading was not impaired when the least important segments were removed and reading was moderately impaired when randomly selected segments were removed. Our data suggest that SVD is a powerful tool in determining what the most informative segments of Chinese characters are.

Table 2: The mean and standard deviation of similarity measures of contour using convex hull between undegraded and degraded characters in the Yan et al. (2012) and the current study.

	Yan et al. (2012)	
	Proportion of Overlapping Vertices	Proportion of Overlapping Perimeters
Most Disruptive	0.59 (0.04)	0.46 (0.05)
Moderately Disruptive	0.73 (0.04)	0.52 (0.05)
Least Disruptive	0.86 (0.04)	0.75 (0.07)
	Current study	
Most Disruptive	0.71 (0.05)	0.47 (0.07)
Moderately Disruptive	0.78 (0.05)	0.61 (0.07)
Least Disruptive	0.80 (0.04)	0.63 (0.07)

Comparisons between spatial distributions of deleted strokes between the most disruptive and least disruptive conditions for the present study and the Yan et al. (2012) study revealed that, in both studies, the most important strokes/segments for Chinese character identification are the ones that retained the character configuration.

When these elements are removed, the contour of the character changes most dramatically and, consequently, reading is most impaired. Conversely, the least important strokes/segments are those that, when deleted, do not greatly change the contour of the character and therefore cause no reading impairment. We suggest that the convex hull might be a useful tool for measuring the configuration of a character.

In addition to their contribution to character configuration, there may be other reasons why some strokes are more important than others. The most important strokes in both studies tended to be located on the left side of the character and the least important tended to be located in the bottom right portion of the character. It is clear that the importance of these elements cannot be due to the order in which the strokes are written since SVD has no information about writing order. Rather, the left hand strokes may be important for several possible reasons. First, semantic radicals tend to be located on the left or top side of Chinese characters, and there is much evidence from a range of paradigms to suggest that reading a complex character involves the processing of its component radicals (see Taft et al., 1999; Zhou, Ye, Cheung, & Chen, 2009). Therefore, when the strokes/segments that contribute to these radicals are missing, reading is more impaired than when other strokes/segments are missing. Alternatively, radicals that are on the top and left hand side of the character tend to have fewer strokes than radicals on the bottom or right hand side. Therefore, deleting these strokes would delete a greater proportion of the character, leading to more impaired reading. Future research should investigate these possible contributions of the left and top strokes/segments to the identification of Chinese characters.

In short, the present study demonstrates that SVD can identify the most and least informative strokes of Chinese characters. When these strokes are deleted, reading is impaired more or less, respectively. These data are similar to the data reported by Yan et al. (2012), using a different method. Both methods identify left strokes/segments as being the most informative for Chinese character identification.

CHAPTER 3

PREDICTING RATERS' TRANSPARENCY JUDGMENTS OF ENGLISH AND CHINESE MORPHOLOGICAL CONSTITUENTS USING LATENT SEMANTIC ANALYSIS MODELS

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. Sometimes, the whole word meaning can be *transparent*, i.e., grasped through its individual constituents, such as in *cookbook*, but sometimes *opaque*, i.e., cannot be fully derived from its constituents, e.g., *cocktail*. The judgments of semantic transparency are often subjective and vary strongly across raters, and a general model may be a way to average across subjective differences.

LSA, the SVD-based application in linguistic studies, has been 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). Therefore, LSA may be a solution to the problem of estimating semantic transparency of a compound word.

The current study proposes two models based on Latent Semantic Analysis (Landauer & Dumais, 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 based on a clustering analysis of morphological family members (e.g., “butterfingers” or “buttermilk” for “butter”). The proposed models account for polysemy of constituents and successfully predicted participants’ transparency ratings. The models may explain the morphological processing when raters classify semantic transparency of English and Chinese compounds.

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 Libben, Gibson, Yoon, & Sandra, 2003; Pollatsek & Hyönä, 2005; Frisson, Niswander-Klement, & Pollatsek, 2008). In contrast, some compounds with two opaque constituents are regarded as semantically opaque (opaque-opaque, OO). Other 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. There have been several models explaining 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 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 based on 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 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 constituent (see Rayner 2009, for a review). Pollatsek and Hyönä (2005) manipulated the frequency (i.e., their occurrence in print) of constituents and transparency of Finnish compound words, and they found longer gaze duration (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 and found, consistent with Pollatsek and Hyönä (2005), no significant difference in eye-movement measures due to this transparency manipulation. However, they found longer gaze duration on opaque compounds than on transparent ones 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 constituents of transparent and opaque compound words and demonstrated that opaque compounds receive 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 size and semantic concreteness affect semantic transparency (see Feldman, Basnight-Brown, & Pastizzo, 2006). A constituent may consist of one or many *morphological family* members; for example, the constituent "butter" consists of "butterfly", "buttercup", "butterfingers", "buttermilk", "butterscotch", "butterfat", "butterwick", among others. Within a morphological family, individual family members may vary in semantic transparency, e.g., the meaning of "butter" is context-sensitive and more transparent in the meaning of "buttermilk" than in the meaning of "butterfly". In other words, semantic similarity among morphological members varies (referred to as *concreteness*, see Feldman et al., 2006). Schreuder and Baayen (1997) suggest that upon reading a word, its family members become co-activated, which leads to a larger global activation in the mental lexicon. There have been several studies focusing 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 larger family size (the number of morphological family members) is processed faster and more accurately in lexical decision tasks.

While there are many common characteristics across languages, there are also many differences so that it is unclear if 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 which 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, 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 3 lists some examples of transparent, opaque, and partially opaque Chinese words.

Table 3: 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	開水(boiled water)	開 (open)	水 (water)

Similar to morphological families in English, a Chinese character, e.g., “馬” (horse), can be shared by its morphological family members, e.g., “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” (horse back) and “馬鞍” (saddle), which 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 whole-word units in general (Taft, Zhu, & Peng, 1999). However, the meaning of some compounds, e.g., “領帶” (necktie), may differ from the ones in which the characters are transposed, e.g., “帶領” (guide). Sometimes a constituent, e.g., “調”, may have different meaning and pronunciation when it locates in the initial (meaning: adjust, pronunciation: tiáo) or final positions (meaning: high or low tone/ key, pronunciation: diáo).

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; Mok, 2009, 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 manipulating frequency (see Zhou, Ye, Cheung, & Chen, 2009, for a review). In a series of experiments with varied whole-word and constituent frequency, Chen T. M. and Chen J. Y. (2006) showed that compound word production in Chinese is not sensitive to morpheme frequency even when all 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).

For studies manipulating frequency and transparency of Chinese compound words, Hung, Tzeng, and Chen (1993) reported a reliable constituent frequency effect for fully transparent compounds (TT) but not 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 suggest that there exists a whole-word representation for all types of Chinese compounds, even fully transparent ones, but separate morphemic representations in the mental lexicon only exist 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). WSE describes a more accurate recognition when a target character, e.g., “态” (appearance), is in the context of a compound word “态度” (manner), as opposed to the same target being in a position-matched non-word control “态备” (appearance; equipped). They also found a larger WSE for Chinese compounds with high whole-word frequency than for ones with low whole-word frequency. The results imply that all types of Chinese compounds, including TT, have corresponding whole-word entries in the mental lexicon, and 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. 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 (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 appropriate categories (e.g., opaque-transparent (OT), transparent-opaque (TO), opaque-opaque (OO), and transparent-transparent (TT), and there was good agreement between the subjects' choices and the predefined classification by Frisson et al. (2008). The proportion of subjects' choices agreeing with the predefined classification was 65% for OO, 71% for OT, 65% for TO, and 86% for TT. Moreover, the proportion of subjects 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.

Transparency Rating of Chinese Compounds

In Mok (2009), the semantic transparency judgments were made in two passes, one by an experimenter and five trained participants' analysis based on dictionary definition, and the other by 30 naïve participants. A 6-point scale rating, where 1 is opaque and 6 is 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 smaller than 3.5. There were 190 compounds, half of which were categorized as high frequent and the other half as low frequent. The agreement between two passes was high (Cohen's kappa = 0.83), and low-frequent items (Cohen's kappa = 0.87) obtained higher agreement than high-frequent items (Cohen's kappa = 0.78). The final classification of the 190 compounds used in the stimuli included 21 TT, 21 TO, 22 OT, and 31 OO in high-frequent items, and 23 TT, 24 TO, 26 OT, and 22 OO in low-frequent items. In total, there were 85 transparent and 105 opaque constituents in high-frequent items and 96 transparent and 94 opaque items in low-frequent items.

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 dictionary definition, (2) the meaning of a compound is not dissociated from its constituents, or (3) a subject not clearly knowing the meaning of the presented materials. Furthermore, the subjective difference may also be caused by the instructions for transparency judgments, e.g., by dictionary definition or not, 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 meanings and low concreteness, 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 et al. (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 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 kind of relation (e.g., shape) or the polysemy of the compound. This subjectivity and variability also occurs in characters of Chinese compounds. A general model may be a way to average across subjective differences.

Models to Predict Transparency using LSA

This study proposes models using Latent Semantic Analysis (LSA) for reflecting raters' transparency judgments based on their mental lexicons. In the semantic space "general reading up to 1st year college" (abbreviated as SP-E) and 300 dimensions used in the current study, 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 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”). Similarly, 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.

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. Following the principle of creating semantic spaces (Quesada, 2007), our previous studies (Wang, Pomplun, Ko, Chen, & Rayner, 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 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 4 times per 5 million were excluded in SP-C, and most of the excluded words were proper names and technical nouns. 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 (see Chen, et al., 2009).

However, this idea of comparing the meaning of a compound and its constituent was used as a tool to validate human transparency judgments, instead of a theoretical foundation and technical understanding of what LSA is accomplishing and why. Therefore, a model based on the theoretical foundation that can be strongly predictive and correlated with transparency judgments is needed. Specifically, how is the word meaning of compounds represented in the mental lexicon by LSA, and how do human raters access the meanings of a compound and its constituents, in transparency judgment? Furthermore, LSA has not yet been tested for two-constituent compound words in English or Chinese, which raises the question whether the cross-linguistic comparison could be made in the same manner. Therefore, we adopted the idea of comparing the meaning of a compound and each of its constituents to 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 is 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) and the Chinese compounds in Mok (2009) and in the present study. The objectives of building general models are 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 vs. Each of its Constituents

One proposed idea of modeling how transparency judgment is done by raters is to compute the LSA cosine values between the compound word and each of its constituents. The parameters of Model 1 include: the *semantic space,* number of *dimensions,* and *comparison type*. For example, using SP-E, 300 dimensions, and “term-to-term” comparison, 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.

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 multi-character 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 second characters are not stand-alone words occurring more than 3 times in the corpus. That is, 12% of the compound word cases were unavailable in the Model 1 computations.

Model 2: Whole Word vs. 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 a 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.

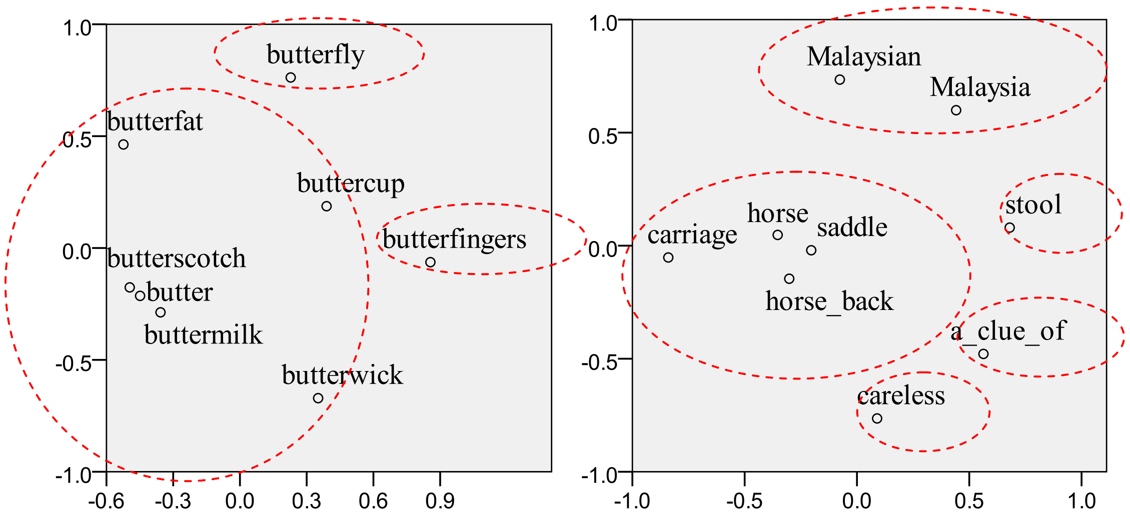
The first step of Model 2 is to obtain the semantic concreteness and dominant meaning of the constituent from its morphological family members that a rater would possibly activate. We computed the LSA cosine values of the pairs in the morphological family members as a distance function. Using a hierarchical clustering algorithm and a given threshold, we classified these morphological family members into *semantic clusters*. Subsequently, a cluster with largest family size or highest sum of word frequency was considered the dominant meaning. An example of the transparency of constituent "butter" in "butterfly" is determined as follows. The morphological family members "butter", "butterfly", "buttercup", "butterfingers", "buttermilk", "butterscotch", "butterfat", and "butterwick" are activated; the LSA cosine values among them are shown in Table 4. The semantic similarities, although in high-dimensional space, can be visualized by multi-dimensional scaling (MDS) in two dimensions, as shown in Figure 10a. 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”, “butterfat”, and “butterwick” are in one cluster, and “butterfly” and “butterfingers” are in their own clusters. We applied “term-to-document” comparison between the compound (e.g., “butterfly”) and the dominant meaning cluster (e.g., the string “butter buttercup buttermilk butterscotch butterfat butterwick”) to compute the LSA cosine value (in this example, 0.04). Similarly, the LSA cosine values between the Chinese character “馬” (horse) and its morphological family members are presented in Table 5, and the MDS result is illustrated in Figure 10b. “馬” (horse), “馬背” (horse back), “馬鞍” (saddle), and “馬車” (carriage) are grouped in one cluster, “馬來” (Malaysian) and “馬國” (Malaysia) form another one, while “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*, e.g., 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 the LSA values in a given semantic space. A low threshold may generate too many clusters, while a high threshold may group unrelated members in one cluster.

Table 4: The LSA cosine values among “butter”, “butterfly”, “buttercup”, “butterfingers”, “buttermilk”, “butterscotch”, “butterfat”, and “butterwick”. The frequency for each word in the British National Corpus (BNC) is shown in parentheses.

	butter	-fly	-cup	-fingers	-milk	-scotch	-fat	-wick
butter (2085)	1							
butterfly (630)	0.04	1						
buttercup (24)	0.09	0.09	1					
butterfingers (3)	0	-0.1	-0.1	1				
buttermilk (13)	0.44	-0	0.12	0.01	1			
butterscotch (7)	0.45	0.05	-0	0.02	0.35	1		
butterfat (39)	0.12	-0	0.04	0	0.11	0.16	1	
butterwick (12)	-0	0.01	0.12	-0	0.09	0.03	0.04	1



(a) (b)

Figure 10: The MDS result for an example of semantic relationships for (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 space of the MDS algorithm.

The access of the dominant meaning of a constituent may be different between English and Chinese compounds. There are clear word boundaries in the English writing system, and constituents in English compounds are usually stand-alone 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 model parameters may differ between the English and Chinese languages.

Table 5: The LSA cosine values between the character “馬” (horse) and its morphological family members. The meaning and frequency in ASBC for each word are shown in parentheses.

	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)	0.83	1							
馬鞍(saddle, 4)	0.74	0.74	1						
馬車(carriage, 37)	0.17	0.07	0.03	1					
馬虎(careless, 13)	-0.02	-0.04	-0.01	-0.04	1				
馬桶(stool, 23)	-0.05	-0.04	-0.03	0.01	0.10	1			
馬腳(a clue of, 4)	0.00	0.04	-0.05	0.01	0.13	0.02	1		
馬來(Malaysian, 11)	0.08	0.06	0.04	0.04	0.00	-0.09	-0.03	1	
馬國 (Malaysia, 12)	0.03	-0.01	-0.03	0.03	0.01	-0.04	0.02	0.15	1

Model Evaluation

We evaluated how Models 1 and 2 estimate transparency of English compounds using the materials of Frisson et al. (2008) and of Chinese compounds using Mok (2009). We also selected 160 constituents and performed transparency rating. The descriptive statistics and distribution of LSA cosine values of opaque and transparent constituents are reported, and a non-parametric test is performed using

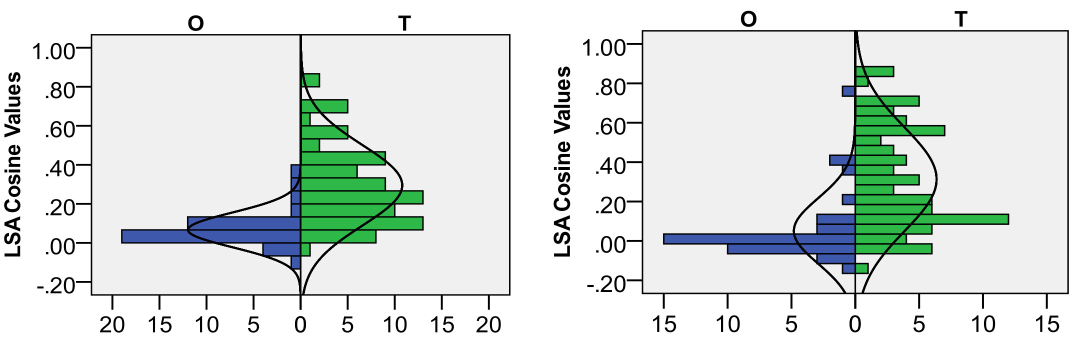
Mann-Whitney U tests. Since the models attempt to map the LSA cosine values (a continuous variable) on dichotomous transparency results (either O or T), we perform a ROC analysis

English Compounds

For English compounds, we re-analyzed 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). 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 used in Model 1 and "term-to-document" was used in Model 2. For 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 are 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 0.8. The dominant meaning was defined as the cluster with the highest sum of frequencies. Concreteness was found to be higher for transparent constituents (mean: 0.21, standard deviation: 0.14) than for opaque constituents (mean: 0.17, standard deviation: 0.13), U = 1442, N_1 = 44, N_2 = 92, $p <$ .01.

The distribution of LSA cosine values of transparent and opaque constituents computed by Models 1 and 2 is shown in Figures 11a and 11b. For Model 1, we found that the cosine values of transparent constituents (mean: 0.29, standard deviation: 0.21) were significantly higher than those of opaque ones (mean: 0.07, standard deviation: 0.09), U = 430.50, N1 = 40, N2 = 84, p < .001. In Model 2, the results were consistent in that higher cosine values were obtained for transparent constituents (mean: 0.31, standard deviation: 0.26) than for opaque constituents (mean: 0.05, standard deviation: 0.16), U = 564.00, N1 = 40, N2 = 84, p < .001. Figure 12a illustrates the ROC curves for Models 1 and 2, and the AUCs are 0.87 and 0.83, respectively. An ‘optimal’ cut-off 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 cut-off point of Model 1 is at a hit rate of 0.74 and a false-alarm rate of 0.10 when the threshold of the LSA cosine value is set to 0.135. The optimal cut-off point for Model 2 is at a hit rate of 0.80 and a false-alarm rate of 0.20, where the threshold of the LSA cosine value is

0.085. Model 1 showed a higher hit rate than Model 2 for all false-alarm rates except those between 0.2 and 0.4. 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).



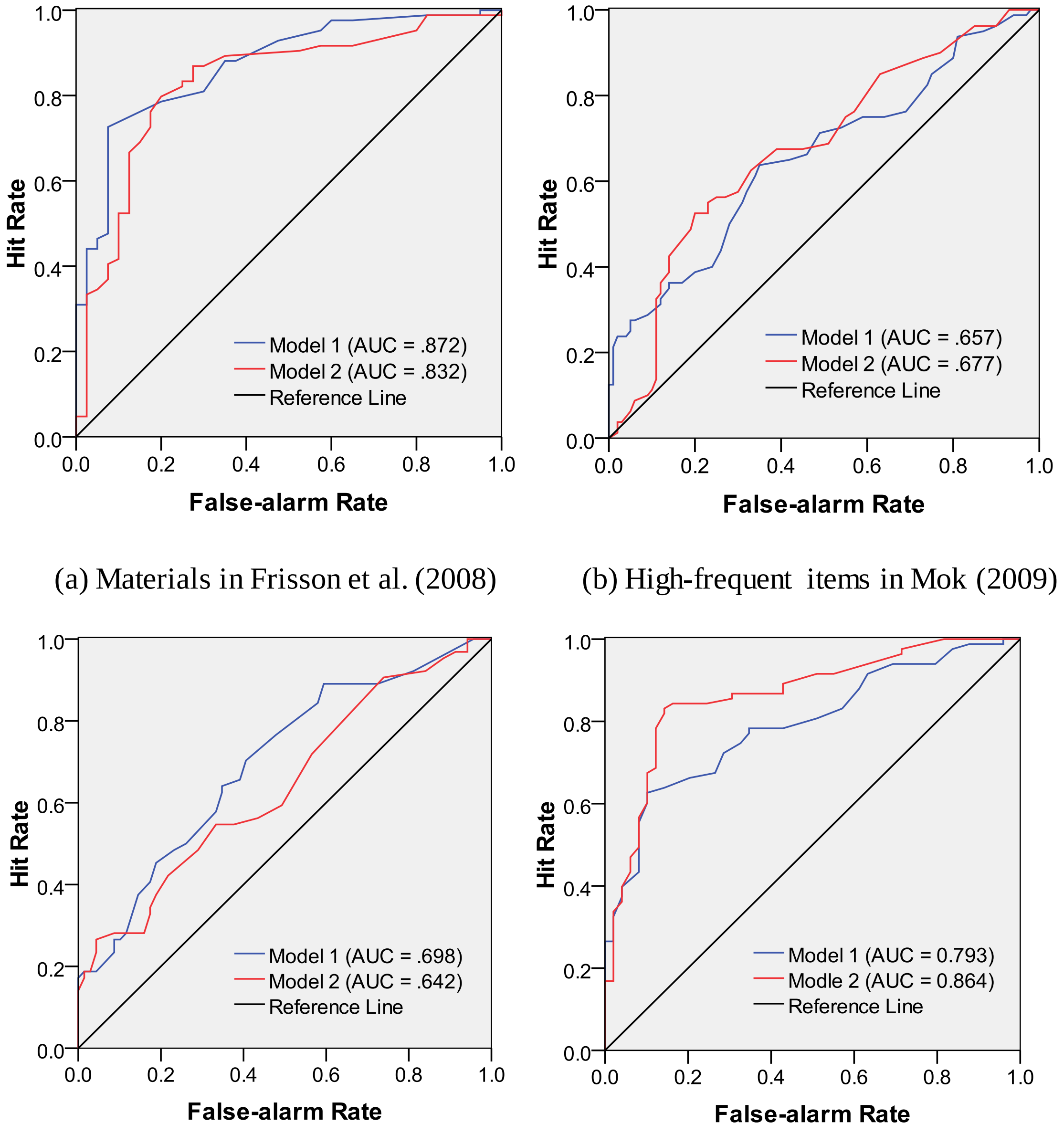
(a) Model 1

(b) Model 2

Figure 11: The 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).

Chinese Compounds

The compounds in Mok (2009) were presented in simplified script, and they were converted into traditional script in SP-C. 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) 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). For low-frequent 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 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 4 times in the ASBC corpus are included, resulting in 20 out of 95 words being excluded from computation.



(a) Materials in Frisson et al. (2008)

(b) High-frequent items in Mok (2009)

(c) Low-frequent items in Mok (2009)

(d) Materials in the present study

Figure 12: ROC analysis of Model 1 and Model 2.

In addition to the material in Mok (2009), in the present study we selected 80 compounds in traditional script, and constituents of those compounds were rated by eleven students who completed a college degree in Taiwan. All participants were native speakers of Chinese (traditional script). Participants were presented with compound words, 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 of 11 participants responding "T." The characters with probabilities greater than or equal to 0.6 were categorized as transparent, while the ones with probabilities less than or equal to 0.4 were considered as opaque. The means and standard deviations of the human ratings were 0.85 and 0.13, respectively, for transparent characters and 0.11 and 0.11 for opaque characters.

Materials in Mok (2009)

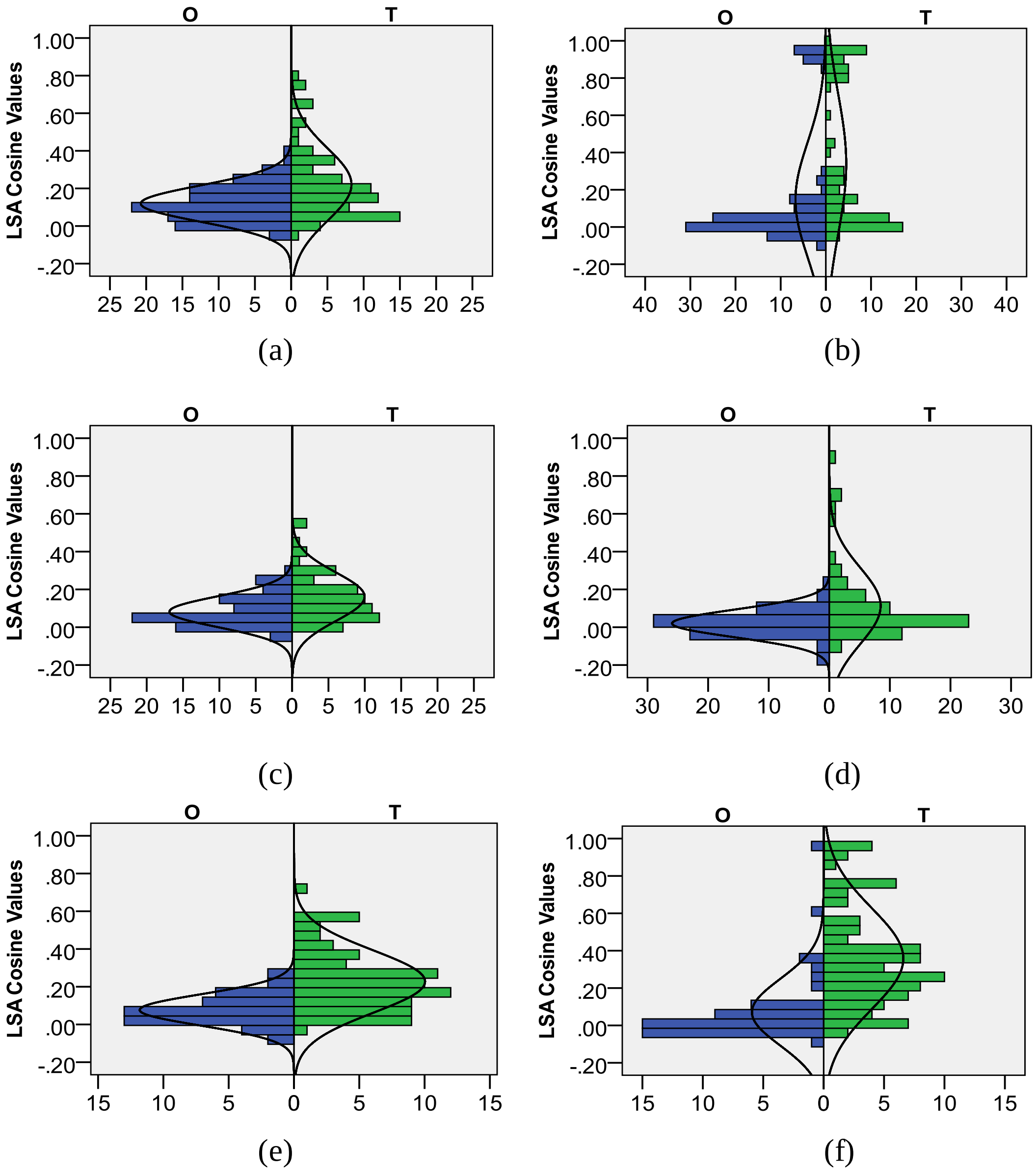
The model parameters were set as follows: SP-C was used for the computations in Models 1 and 2. The "term-to-term" comparison was used in Model 1 and the "term-to-document" was used in Model 2 except that "幻燈機 幻燈片" used "document-to-term" in Model 1 and "document-to-document" in Model 2. For the clustering algorithm in Model 2, the threshold setting was 0.5, 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.

For concreteness of high-frequent items, consistent to English constituents, transparent constituents (mean: 0.12, standard deviation: 0.07) yielded higher values than opaque constituents (mean: 0.09, standard deviation: 0.04), U = 3047.5, N_1 = 83, N_2 = 104, *p* < .01. For the concreteness in low-frequent items, transparent constituents (mean: 0.10, standard deviation: 0.04) showed marginally higher values than opaque ones (mean: 0.09, standard deviation: 0.03), U = 3689, N_1 = 93, N_2 = 94, *p* = .06.

High-frequent Items. The evaluation of Model 1included 80 out of 85 transparent and 100 out of 105 opaque constituents, and the LSA scores of transparent constituents (mean: 0.22, standard deviation: 0.19) were significantly higher than the ones of opaque words (mean: 0.12, standard deviation: 0.10), U = 2741.50, N_1 = 80, N_2 = 100, *p* < .001. Model 2 overcame the limitation of non-stand-alone characters in Model 1, resulting in 85 out of 85 transparent and 103 out of 105 opaque constituents being available for its evaluation. Again, transparent constituents (mean: 0.34, standard deviation: 0.37) obtained higher LSA cosine values than opaque constituents (mean: 0.15, standard deviation: 0.30), U = 2779.50, N_1 = 85, N_2 = 103, $p < .001$. The distribution of LSA cosine values of transparent and opaque constituents computed by Models 1 and 2 are shown in Figures 13a and 13b. In the evaluation of Model 2 with high-frequent items, there were a few opaque items with high LSA cosine values and also a few transparent items with low LSA values. It is possible that the polysemy and frequency of constituents somehow affects model predictions.

To further explore why some opaque constituents obtained high and some transparent constituents obtained low LSA cosine value in Figure 13b, we selected a cut-off point of 0.4 of the LSA cosine value and summarized the constituent frequency, family size, and concreteness in Table 6. A one-way 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 concreteness F(3; 183) = 11.94, $p < .001$. Post-hoc tests using Bonferroni adjusted p-values revealed that concreteness in the condition of T responses to T was higher compared to 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 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 concreteness 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 frequent items (see Figure 13b), since Model 2 defines dominant meaning as the cluster with higher family size and frequency.

Figure 12b illustrates the ROC curves for Models 1 and 2, and the AUCs are 0.66 and 0.68, 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.



(a) (b) (c) (d) (e) (f)

Figure 13: The distributions of LSA cosine values of opaque (O) and transparent (T) constituents of (a) high-frequent items by Model 1, (b) high-frequent items by Model 2, (c) low-frequent items by Model 1, and (d) low-frequent items by Model 2 in the materials in Mok (2009). (e) and (f) are the materials in the present study, which were predicted by Models 1 and 2, respectively.

Table 6. The constituent frequency, family size, and concreteness when a cut-off point of 0.4 of the LSA cosine was selected in the high-frequent items in Mok (2009).

	Constituent Frequency		Family Size		Concreteness	
	Mean	Std	Mean	Std	Mean	Std
T response to T	389	698	24.64	19.72	0.14	0.09
T response to O	1245	313	27.95	19.97	0.09	0.03
O response to T	63	98	14.58	19.42	0.10	0.05
O response to O	1180	2425	37.53	31.50	0.08	0.03

Low-frequent Items. For the evaluation of Model 1, there were 64 out of 96 transparent and 69 out of 94 opaque items available, and transparent constituents (mean: 0.16, standard deviation: 0.13) obtained higher LSA cosine values than opaque ones (mean: 0.08, standard deviation: 0.08), U = 1335, N_1 = 66, N_2 = 103, $p <$ .001. For Model 2, 64 out of 96 transparent items and 71 out of 94 opaque items were available, and transparent constituents (mean: 0.11, standard deviation: 0.20) obtained higher LSA cosine values than opaque ones (mean: 0.02, standard deviation: 0.07), U = 1335, N_1 = 66, N_2 = 103, $p <$ .01. The distribution of LSA cosine values of transparent and opaque constituents computed by Model 1 and 2 are shown in Figures 13c and 13d. The agreement between the two passes (one by dictionary definition and the other by subject rating) in the Mok (2009) study was higher in low-frequent items than in high-frequent items. It is possible that subjects access the compound meaning by constituent in low-frequent items, and it is therefore easier to dissociate the meanings of constituents from the compounds than in high-frequent items. The AUCs of the ROC curves for Models 1 and 2 were 0.70 and 0.64, respectively (see Figure 12c), which may be caused by subjects accessing the meaning of a constituent via its stand-alone form, rather than via its morphological family.

In general, the models showed less predictive power for the Mok (2009) materials than for the Frisson (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 Model 1 when the stimuli are high-frequent. Since the materials of Mok (2009) were selected by both dictionary definition and subject rating, the efficiency of Model 2 may be underestimated in the materials in Mok (2009). To test this hypothesis, we tested a dataset whose transparency judgments were performed in the present study, and our expectation according to the hypothesis was that Model 2 would outperform Model 1.

Transparency Rating in the Present Study

There were 83 out of 89 transparent and 49 out of 53 opaque constituents 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 LSA cosine values of transparent constituents were higher than opaque constituents for both models, U = 841 and U = 655.5, respectively, *ps* < .001. Figure 12d illustrates the results of the ROC analysis, and the AUCs for Models 1 and 2 were 0.80 and 0.86, respectively. The Spearman rank correlations (a non-parametric test) between human rating probability and Model 1 and between human rating probability and Model 2 were 0.50 and 0.58, respectively. Both Models 1 and 2 are predictive to the results of human transparent judgments, and Model 2 numerically obtained higher AUC and correlation than Model 1. 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. Therefore, Model 2 may in general be a better approach than Model 1 to predict transparency ratings for constituents of Chinese compounds.

General Discussion

The most important outcome of the current study is the proposed computational model 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 stimuli used in prior compound word studies (Frisson et al., 2008; Mok, 2009) as well as in the present study. The proposed models demonstrate considerable performance at distinguishing transparent and opaque constituents. Semantic concreteness is found to be higher in transparent constituents than in opaque ones in all the materials of our evaluations.

Dominance of Lexeme Meaning

In a related study, Inhoff, Star, Solomon, and Placke (2008) manipulated the lexeme frequency (high vs. low) and dominance (“headed” vs. “tailed”) of compound words. They found that compound words were parsed into constituent lexemes and that lexeme dominance influenced compound recognition. For example, in the HL condition (high-frequency first lexeme and low-frequency second lexeme), first fixation duration for headed compounds was shorter than for tailed compounds. 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 subjects 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.

Semantic Transparency and Concreteness in Reading

The results, such as semantic transparency and concreteness estimations, could be adapted to further Chinese reading research. There are some unpublished studies addressing semantic transparency of two-character Chinese words (Lee, C. Y., 1995; Lee, P. J., 2007). Lee, C. Y. (1995) found opposite results by manipulating word frequency, character frequency, and word transparency in a lexical decision task. She found that RTs 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 while the constituents of transparent words are represented separately in the mental lexicon. Moreover, Lee, P. J. (2007) used eye tracking to investigate how word frequency, word transparency, and character frequency influence eye-movement measures during reading. The study found shorter first fixation and gaze duration for opaque words as compared to 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 findings by Lee (1995), 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 are from non-peer reviewed work and details of the works are restricted, we suggest applying the proposed transparency and concreteness measures in further studies using different tasks or paradigms to address the open questions about semantic transparency.

Educational Applications

The transparency and concreteness measures 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 were 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, e.g., to first teach children characters with high concreteness or larger morphological family size as general rules to construct new compounds, and then teach opaque compounds at the whole-word level.

Transparency of Semantic Radicals of Chinese Characters

In Chinese character orthography, the most common structure is a semantic-phonetic compound character which 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 while the phonetic radical provides roughly the pronunciation. Some radicals can be stand-alone characters but others cannot. Similar to 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, while 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 stand-alone character, e.g., "貝" (shell) was used as currency in ancient China, and therefore many characters with this radical are related to "money", e.g., "賺" (earn money), "賒" (loan money), which are not semantically related to "shell" in contemporary Chinese.

The semantic transparency of radicals has been found to affect sub-lexical 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 (Hsiao, Shillcock, & Lavidor, 2007; Chen & Weekes, 2004). 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 stand-alone words. We therefore suggest adopting Model 2 to compute the semantic transparency and concreteness of semantic radicals.

Limitations and Future Work

The SP-C used in the current study was built using traditional script, and it is important to test the compatibility between the traditional and simplified scripts. From the materials in Mok (2009) using simplified script, all high-frequent items except one (due to segmentation standard) were covered in SP-C, but 5 out of 95 low-frequent items were not used in traditional scripts. Furthermore, 10% of simplified characters map to multiple (two to four) traditional characters, which increases the morphological or semantic ambiguity (see Tsai, Kliegl, & Yan, 2012). The 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 corpuses.

The current limitations of the proposed method in Chinese might be the relatively small corpus size, which is due to the fact that there are no spaces between words in the Chinese writing system, and 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 is 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 based on those larger corpuses, although the corpuses mentioned above are currently limitedly accessible.

There are other computational alternatives of semantic similarity. Since LSA requires document information, i.e., 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, 1999), 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 by LSA (LSA cosine value for the example: 0.01). Therefore, we suggest that the WordNet-based measure could be integrated into semantic space models to account for a broader range of transparency interpretations.

Estimation of 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 clustering algorithm. A single set of parameter values were used for each dataset reported in this paper, 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). For the parameters (4) to (6), there are many issues, such as: should constituent position be included in the morphological family? Should the size of morphological families be limited? For the constituents with low concreteness, which meaning is activated by a rater? What is the optimal threshold for a given semantic space? To address these issues, more empirical work is required. From the results in the present study, it appears that the way in which transparency judgments were carried out affects model parameter settings. We imply that the meaning activated by a human rater during transparency judgments may be individually different, and each rater might have a different threshold for the "cut-off" of opacity. The present study proposes general models to average across the subjective differences and transparency ambiguity, and these models may capture some basic aspects of a rater's morphological processing during transparency judgments.

CHAPTER 4

ESTIMATING THE EFFECT OF WORD PREDICTABILITY ON EYE MOVEMENTS IN CHINESE READING USING LATENT SEMANTIC ANALYSIS AND TRANSITIONAL PROBABILITY

The preceding Chapter 3 has demonstrated that LSA-based models can explain the word identification process for both English and Chinese speakers. In the present chapter, we will be using LSA to address on higher-level linguistic processing, the *predictability* effect, between a target word and its prior context during sentence processing.

The predictability of target words (as determined by raters in a cloze task) has been found to strongly influence eye movements during reading (Rayner & Well, 1996; see also Ehrlich & Rayner, 1981). 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.

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 3, where the prefix probability α is the total probability for the grammatical analyses of the prefix string (see Boston et al., 2008).

$$\text{surprisal(n)} = \log_2\left(\frac{\alpha_{n-1}}{\alpha_n}\right) \qquad (3)$$

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.

Although LSA has been very successful at simulating a wide range of psycholinguistic phenomena, 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 7 shows the LSA cosine value of two target words (*bottle* and *diaper*) in two contexts (*Before warming the milk, the babysitter took the infant's …* and *To prevent a mess, the caregiver checked the baby's …*). 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.

Table 7: Re-analysis of materials in Rayner et al. (2004) using LSA

Context preceding the target word	Target word	Predictability	Freq	LSA
Before warming the milk, the babysitter took the infant's	bottle	P	H	.42
Before warming the milk, the babysitter took the infant's	diaper	U	L	.25
To prevent a mess, the caregiver checked the baby's	bottle	U	H	.08
To prevent a mess, the caregiver checked the baby's	diaper	P	L	.60

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.

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

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

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 14 shows an example of how the LSA cosine values of target words and their previous content4 words were calculated. The average and standard deviation are 0.31 and 0.19, respectively.

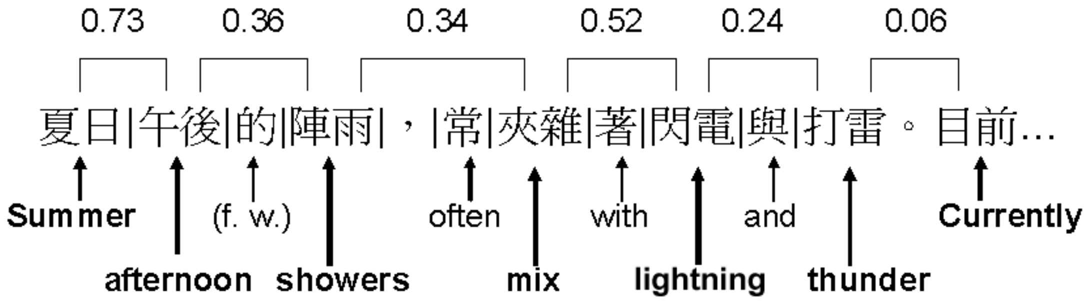


Figure 14: The LSA cosine value of target words and their preceding content words

The correlations between predictors are shown in Table 8 based on 5,324 word cases 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.

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.

Table 8: Pair-wise correlation between predictor variables (based on mainly content words). 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.

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							—

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). 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 (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. Table 9 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.

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

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

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

Table 9: Estimates of mean unstandardized regression coefficients and standard errors in the regression equation. 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

	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

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 9). 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.

CHAPTER 5

MODELING CONCEPT ACTIVATION IN WORKING MEMORY DURING ONLINE SENTENCE PROCESSING

Using LSA and TP, we were able to estimate the results obtained by the cloze task and eye movements during reading. However, these computational methods did not explain how the semantic representation of each content word in a sentence is activated in working memory. In this chapter, we propose a connectionist model (Landscape model, see van den Broek, 2010) and LSA to determine the predictability of a word and its corresponding semantic representation associated in a neural network. LSA is used to establish connections between words and simulate the long-term semantic associations among concepts. This model may provide a means of investigating how language comprehension is affected by the activation of concepts in working memory.

The predictability of a given word can, to a large extent, be conceptualized as the degree to which the semantic concept represented by the word is associated with the preceding context. By treating incoming lexical items as semantic concepts that interactively influence working memory processes, prior context for a word can be represented as inputs which influence the activation of associated concepts and have the potential to facilitate or inhibit the processing of upcoming words. As a result, the higher the activation of a concept when it is encountered, the more processing of the concept is facilitated. Importantly, individuals can allocate their processing attention to only a finite number of linguistic items at a given moment. Thus, any model of language processing and working memory must set limits to the number of lexical-semantic concepts that can be simultaneously active and exert an appreciable influence on the processing of upcoming lexical inputs.

A Connectionist Model for Sentence Reading

This study proposes a computational model to monitor the activations of concepts in working memory. The computation of concept activation is derived from a connectionist model (the Landscape model, see van den Broek, 2010). The current model does not have distributed semantic representations; rather, words are represented as localized semantic "concepts" with weighted connections to a network of additional concepts. The semantic connections among concepts in the simulation are computed using LSA cosine values based on the default 300 dimension semantic space, “general reading up to 1st year college”, available at the LSA@CU Boulder website (http://lsa.colorado.edu/). LSA represents word meaning and computes associations by applying a linear algebra method, singular value decomposition (SVD), to a large corpus of text (see Landauer & Dumais, 1997).

The Landscape model is a connectionist approach to instantiating comprehension using psychologically plausible algorithms that can potentially be used to model several aspects of text comprehension (see van den Broek, 2010; Tzeng, van den Broek, Kendeou, & Lee, 2005). The architecture of the conventional Landscape model assumes that as a reader proceeds through a text in reading cycles (with each cycle roughly corresponding to the reading of a new sentence), concepts fluctuate in activation as a function of four sources of information: the current processing cycle, the preceding cycle, the current episodic text representation, and reader’s background knowledge. With the reading of each cycle, particular concepts are activated and added as nodes to the episodic memory representation of the text. If a concept is already part of the text representation and is reactivated, its trace is strengthened. Furthermore, co-activation of concepts leads to the establishment (or strengthening) of connections between those concepts. The resulting network representation influences subsequent activation patterns. This phenomenon is called the *cohort effect*. These cyclical and dynamically fluctuating activations lead to the gradual emergence of an episodic memory representation and discourse model of the text in which textual propositions and inferences are connected via semantic relations (such as causal and referential links). Thus, the model captures the fluctuations of concepts during reading (Linderholm, Virtue, Tzeng, & van den Broek, 2004), as well as readers' memory representation of text (Tzeng, 2007). As such, this model has prescribed mechanisms that can link the iterative and reciprocal relations between fluctuations of activations and the episodic text representation. However, there are necessary differences with regard to how readers generate and update active discourse representations for the comprehension of an individual sentence, compared to the processing of a longer narrative or expository text. For the comprehension of an individual sentence, a reader must primarily rely on establishing connections between relevant concepts in working memory and pre-existing long-term semantic representations. For a longer text, on the other hand, readers are often able to take advantage of more extensive and detailed context and presumably a more enriched discourse model. Thus, the current computational approach adapts the Landscape Model to a connectionist framework more suitable for capturing sentence reading.

Moreover, the current model utilizes LSA in order to represent pre-existing connections between semantic representations stored in long-term memory (i.e., background or world knowledge).

In the current model, as with the Landscape model, text inputs are represented by an *input matrix* and each is indexed as a *Mention* (concepts being read from the text). The conventional Landscape model also defines other sources of activation including *Referential* (for building referential coherence), *Causal*, and *Enabling* (for the causal explanation of the current statement), but those activations are as of yet, not implemented here. The input matrix for example sentence: "The knight uses his sword to fight the dragon" is shown in Table 10.

Table 10: Input matrix for the *Knight* example.

cycle	knight	Use	sword	fight	Dragon
1	1	0	0	0	0
2	0	1	0	0	0
3	0	0	1	0	0
4	0	0	0	1	0
5	0	0	0	0	1

Initially, the sentence is segmented into component concepts: “knight”, “use”, “sword”, “fight”, and “dragon”; as, currently, only content words are considered as concepts. The model assumes that each word is processed sequentially. In each cycle, the concept of *Mention* receives 1 unit of activation. In addition to the sequential activation of concepts, the influence of semantic knowledge and pre-existing lexical associations between concepts is established using LSA corpus-learned associations. Table 11 presents the *connection matrix* for the example sentence. The values are always between -1 and 1, but are rarely below 0 because of LSA’s high-dimensional space.

Table 11: Connection matrix for the example.

.	knight	use	sword	fight	dragon
knight	1	.01	.64	.15	.28
use	.01	1	.03	-.02	.06
sword	.64	.03	1	.20	.40
fight	.15	-.20	.20	1	.13
dragon	.28	.06	.40	.13	1

The activation values for each concept are represented in an m x n *activation matrix*, where m represents the number of concepts in the sentence and n represents the number of *cycles*. Each column in the matrix thus represents the status of each concept. The *activation matrix* takes each column of the *input matrix* as raw input and processes it row by row. In our model, the activation during the current reading cycle is defined by Equation (9):

$$A_i^{cycle} = \sum_{j=1} \delta A_j^{cycle-1} \sigma(S_{ij}) + \sum_{j=i} input_i^{cycle} \sigma(S_{ij}) \qquad (9)$$

A_i^{cycle} is the activation of concept *i* during the current cycle. Starting from the summation (Σ) term in Equation (9), for all activated concepts in the previous reading cycle, each activation value is multiplied by a transformation function σ of connection strength (S_{ij}) and by the cohort activation parameter δ. S_{ij} is the strength of the relation from concept *j* to *i*. For the current cycle, $input_i^{cycle}$ is the activation of concept *i* in the *input matrix*. The sum of the net inputs for these *m* concepts is multiplied by the transformation function σ of connection strength (S_{ij}).

The conventional Landscape model uses a sigmoid function as the transformation function σ to control the possible linear growth of spreading of activation and limit the effect of cohort activation to those strongly related to the concept. Since S_{ij} is usually between 0 and 1, a linear function with absolute value is used in this study. The value of the cohort activation parameter, δ, directly determines the amount of cohort activation and can be used to mimic individual differences in spreading of activation. Our model assumes that for any concept, its cohort activation can never exceed its input activation. For this reason the model will take the larger of the input and cohort activation values, and *Mention* is the maximum activation a concept can receive. Furthermore, there is a system parameter *Activation Threshold*; any activation below that threshold is set to zero.

The working memory constraint is implemented by a parameter *WMC (Working Memory Capacity)*. When the actual sum of activation exceeds the value of WMC, the activation of each concept is scaled down using Equation (10):

$$A_i^{cycle} = A_{i,Actual}^{cycle} \cdot \frac{WMC}{\sum_{i=1}^{m} A_{i,Actual}^{cycle}} \tag{10}$$

For the example sentence, the activation matrix is shown in Figure 15. For the 1st cycle the activation of “knight” is 1, from the *Mention* input. There is no cohort effect for the first reading cycle since no previous cycle exists. The activations for “use”, “sword”, “fight”, and “dragon” are established by multiplying their connections, .01, .64, .15, and .28 respectively, and the input of “knight” (1). The activation of “use” does not reach the threshold (set to 0.1) and as a result receives an activation of 0. For the 2nd cycle when “use” is being processed, the activation of each concept is calculated according to Equation (9). Figure 1 illustrates that the activation of “dragon” increases from cycle 1 to cycle 4 because of relatively strong connections to “knight”, “use”, “sword”, and “fight.” Conversely, the activation of “use” decreases from cycle 2 to 5 because its connections to “sword”, “fight”, and “dragon” are relatively weak (less than .06).

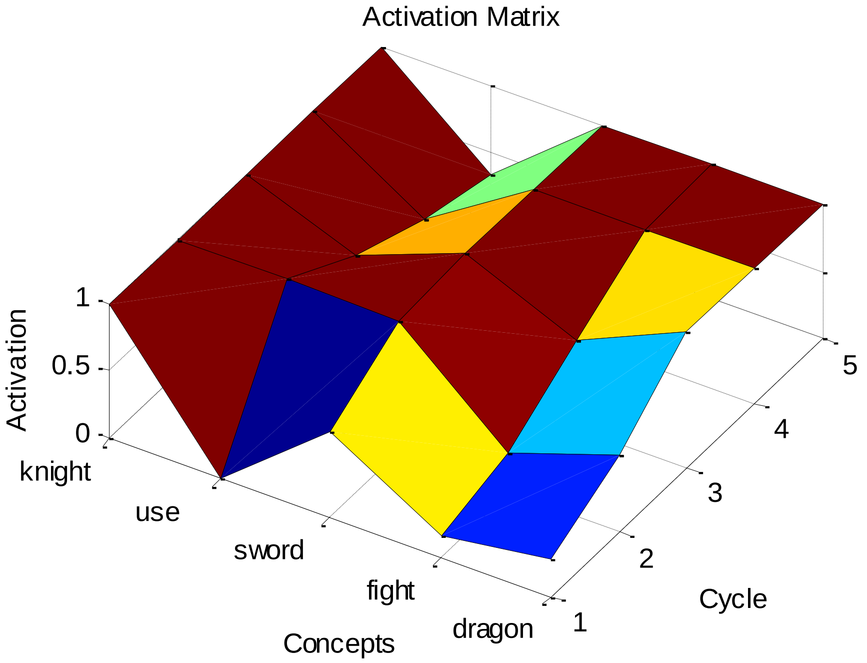


Figure 15: The “landscape” of the activation matrix for the Knight example.

The conventional Landscape model updates the connection strengths in its episodic memory using a learning algorithm in order to adjust active discourse representations for the comprehension of a longer text. In this study, we assume that the background knowledge (represented by in the connection matrix) is not altered during sentence reading.

In summary, by assuming (a) the words in a sentence are read and processed sequentially, and (b) long-term memory representations (i.e., background knowledge) are not altered during sentence reading, we propose a computational model of sentence reading which takes advantage of an existing discourse comprehension model that is designed to take into account contextual effects. The proposed model allows us to examine several factors that affect sentence comprehension; namely, (1) semantic concept activation in working memory, (2) background knowledge, and (3) working memory capacity. To assess the model's ability to investigate linguistic processing we will compare its performance to the conventional cloze task.

Reanalysis of Previous Data

The key objective for this implementation is to disambiguate high from low semantic constraint in sentence contexts. Another objective of this implementation is to demonstrate that the Landscape (LS) model surpasses previously utilized methods as an alternative to the cloze task. In order to demonstrate that the proposed computational model is a more suitable metric of contextual constraint than previous computational measures, (i.e., Wang et al., 2010), we re-analyzed the materials from Gollan, Slattery, Goldenberg, Van Assche, Duyck, and Rayner (2011), in which predictable or unpredictable target word conditions were confirmed by a norming cloze task. We estimated predictability of a target word by (1) the previous content word, (2) all words in prior context, and (3) the estimates of the proposed connectionist model in this study. The question being: can our model outperform other predictors in differentiating high- and low-constraint contexts and produce higher by-item correlation to cloze values.

Participants. Twenty undergraduate students at the University of California, San Diego, participated. All participants were native speakers of English.

Materials. There were 90 target words; all target words were embedded in either a high-constraint (HC) or low-constraint (LC) sentence. For example, "the hockey player moved on the ice on his ______" (S1) was considered HC while "The little girl was very happy when she unwrapped her brand new ______" (S2) was LC for the target "skates". For the cloze task target words in HC context were generated 87% of the time, whereas the ones in LC context were generated less than 3% of the time.

Procedure. Participants were presented with the sentences up to the target words, and asked to provide one-word continuations for each sentence.

Analysis. The first estimate of predictability for each target word was derived by one previous content word (PreCont) of each target, e.g., the previous content word of S1 is “ice,” while the one of S2 is “new.” The second approach computed the LSA cosine value using all words in the previous context (AllW). The final estimate was derived from Landscape model of sentence processing described above in the previous section (LS). We manually segmented the sentence into concepts and removed function words such as “a”, “the”, “in”, etc., for instance, “hockey / player / moved / ice” for S1. The parameters of our model were set as following: $\delta = .7$, Mention = 1, Activation Threshold = .1, and WMC = 7. The averages and standard deviations of Cloze, PreCont, AllW, and LS for HC and LC are described in Table 12.

Table 12. The averages and standard deviations (in parentheses) of Cloze, PreCont, AllW, and LS for HC and LC conditions.

	Cloze	PreCont	AllW	LS
HC	.87 (.13)	.17 (.16)	.21 (.16)	.66 (.29)
LC	.03 (.03)	.05 (.11)	.04 (.07)	.13 (.20)

Results

As shown in Figure 16, a receiver operating characteristic (ROC) analysis demonstrates that the area under the curves (AUC) of Cloze, PreCont, AllW, and LS are 1, .70, .87, .91, respectively. The LS model obtains a higher AUC than AllW or PreCont. Furthermore, a correlation analysis demonstrates that the Pearson correlations between Cloze and PreCont, AllW, and LS are .39, .56, and .70, respectively.

The results suggest that the LS model can simulate much of the linguistic processing subjects perform when producing cloze responses (and presumably during normal reading). It is important to note that levels of activation expressed by the model are not intended to predict exact cloze probabilities, but to successfully differentiate highly constrained and unconstrained sentence contexts as well as the conventionally used cloze task. The LS model also demonstrates superiority over objective measures that utilize only the prior content word or LSA connections between content words exclusively.

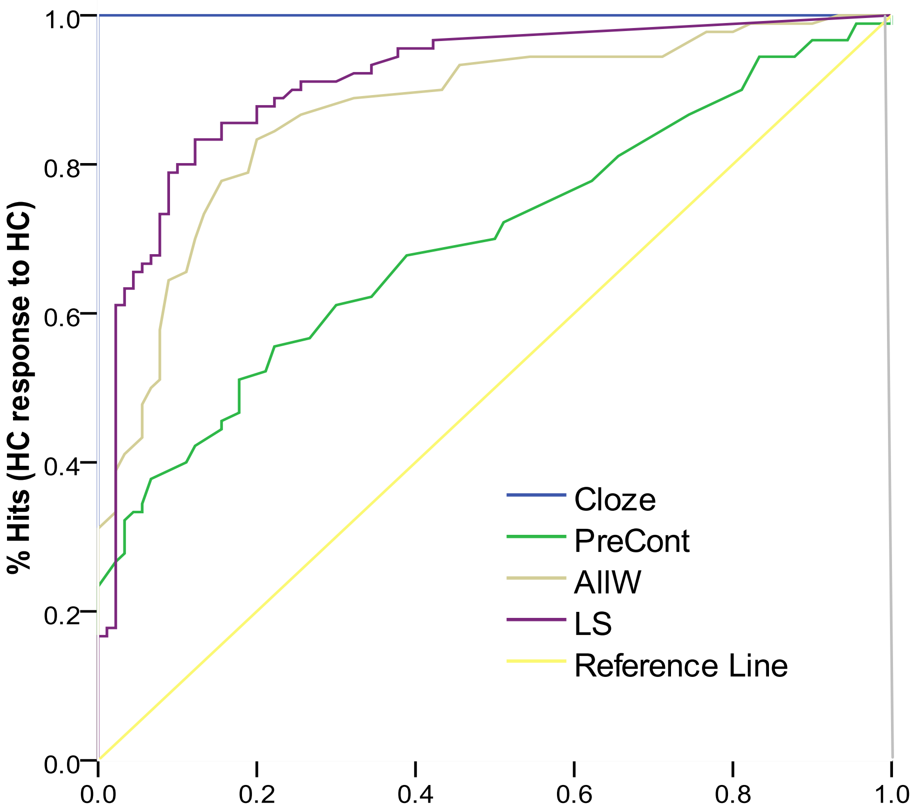


Figure 16: ROC curves for Cloze, PreCont, AllW, and LS.

Discussion

The current implementation of the model has demonstrated that it is an effective measure of contextual constraint in that it differentiates high and low-constraint sentence contexts better than previously employed alternatives to the cloze task. Furthermore, model activations for target words correlate with cloze responses more highly than previous computational methods of measuring contextual constraint. We believe this is an initial step toward the ultimate objective of representing both the fluctuating activation of lexical-semantic concepts in working memory during online sentence processing and how the processing of upcoming words can be influenced by prior context. The next logical step is to employ the LS Model as a metric of processing difficulty that can be compared to the behavioral record and used to generate predictions about eye movements during reading. It has been aptly demonstrated that reading times on words are influenced by the preceding linguistic context (Rayner, 1998; 2009). Moreover, discourse-mediated spreading activation across lexical-semantic representations has been proposed as an appreciable source of predictability effects during reading (Morris, 1994; Pynte et al., 2008; Traxler, Foss, Seely, Kaup, & Morris, 2000). Thus, modeling the process whereby linguistic inputs activate concepts in long-term memory and continuously influence working memory operations during sentence comprehension is an important endeavor in psycholinguistics.

As shown by the comparison to standard cloze responses, the current model can be used to reliably derive predictability of word *n* given the preceding context. The model generates a specific level of activation for word *n*, assuming that each word in the preceding context has been identified and associated concepts have been engaged in working memory. As demonstrated above, this predicted level of activation correlates to cloze probabilities for a target word (*n*).

Critically, when using the LS model, in many cases the level of activation for word *n* will provide a more psychologically realistic measure of word processing difficulty when compared to cloze proportions, especially in neutral or unconstrained contexts. For instance, referencing cloze scores alone, there is no distinction between words that are plausible, yet not highly-predictable, and those that are completely implausible or anomalous given the preceding sentence context. In fact, it is quite feasible for plausible target words in unconstrained sentence frames to receive cloze probabilities at or around zero; however, low cloze probabilities are not necessarily indicative of potential processing difficulty. The manner in which the cloze task is conventionally used produces binary measures (to the extent that non-target responses are ignored). In this way, the current computational model may produce a more accurate representation than cloze scores with regard to indexing online word processing difficulty. As such, in the future we will assess the LS Model's goodness-of-fit to reading times and other eye movement data.

By modifying the framework of the conventional Landscape model to reduce the size of text segments being processed during a reading cycle and situating activated concepts within limited working memory resources, we have attempted a psychologically plausible computational model of semantic effects on sentence comprehension. Crucially, the fluctuating activation of within sentence concepts is not determined merely by summing its cumulative activation across all preceding words; rather, the interactive and co-dependent influence of the prior sequence of words determines the extent to which the prior sentence context results in activation for a particular semantic concept (and its corresponding lexical representation).

The model is also a useful tool for investigating the number of semantic entities that are generally active in working memory, as well as the upper limits for the number of lexical-semantic items simultaneously activated. Computationally examination of working memory limitations during reading could provide insight into what linguistic constructions are likely to elicit processing difficulty, result in longer fixation times, and lead to inter-word regressions during sentence reading. Model outputs can also be used to inform inferences as to which concepts are likely to maintain relatively high levels of activation in working memory.

While among the most sophisticated computational frameworks in the field of cognitive science, current models of eye movement control during reading do not focus on how prior words render specific words predictable. The more well-developed models of oculomotor behavior and language comprehension represent the predictability of a given word in a sentence using only its cloze probability (Engbert, Nuthmann, Richter & Kliegl, 2005; Reichle et al., 1998; 2003). Our model successfully attempts to represent the cognitive processes that are sensitive to semantic constraint. Future implementations of the LS model will be capable of more thoroughly examining aspects of language processing and eye movement behavior. The *connection matrix* in the LS model can operationalize a variety of linguistic characteristics stored explicitly, or otherwise represented, in long-term memory. Semantically-based connection weights can be modified to accommodate mediation by lexical and sub-lexical frequency characteristics. In addition, the *connection matrix* could be modified to capture morphological, orthographic, and phonological similarity between lexical items. As of now, the LS model is a computational alternative to the cloze that is sensitive to both strong and subtle changes in contextual semantic constraint. Ultimately, the model will be expanded in an effort to achieve more comprehensive measurement of lexical-semantic predictability as it affects reading behavior.

CHAPTER 6

OBJECT FREQUENCY AND PREDICTABILITY EFFECTS ON EYE FIXATION DURATIONS IN REAL-WORLD SCENE VIEWING

As indicated by the findings from Chapter 5, LSA can be used to estimate word predictability in reading. This successful application raises the question whether the perception of objects in everyday scene viewing can be studied in a similar way.

Although the influence of semantic factors on fixation time during scene viewing has been studied (Loftus & Mackworth, 1978; Henderson & Hollingworth, 1999; Hollingworth & Henderson, 1999; Henderson & Ferreira, 2004), the semantic consistency in scene viewing studies was not equal to predictability in reading studies. The semantic consistency studies often added a “strange” object to the stimuli that seldom occurred in real-world scenes. Moreover, it has not been tested whether frequency and predictability effects apply to scene viewing in ways analogous to reading. To analyze eye movement data using object-based measures in scene images in a way analogous to word-based measures being used in reading, object segmentation is required. However, the results of automated segmentation and labeling of images are still unsatisfactory for such a purpose.

Therefore, I chose to base this study on an annotated image database, the freely available LabelMe image dataset (Russell, Torralba, Murphy & Freeman, 2008). It contains a large number of scene images that were manually segmented into annotated objects. The locations of objects are provided as coordinates of polygon corners (as shown in Figure 17) and are labeled by English words or phrases. The given locations and labels of objects make the analysis of object-based eye-movement measures possible. Furthermore, the labels of English words provide an opportunity to apply linguistic-based methods, such as LSA, to scene viewing, and compare the results between vision and language.

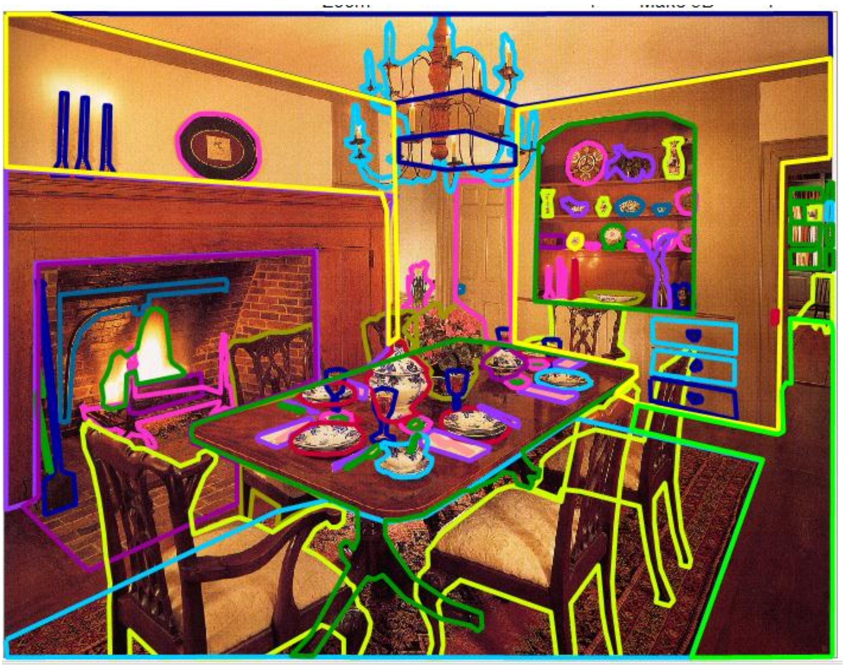


Figure 17: A dining room scene in LabelMe.

Methods

Participants. Twelve participants performed this experiment. All were students at the University of Massachusetts Boston, aged between 19 to 40 years old. Each was entitled to a $10 honorarium.

Apparatus. Eye movements were tracked and recorded using an SR Research EyeLink II system with a sampling frequency of 500 Hz. After calibration, the average error of visual angle in this system is 0.5˚. Stimuli were presented on a 19-inch Dell P992 monitor with a refresh rate of 85 Hz and a screen resolution of 1024×768 pixels. Participants' responses were entered using a game-pad.

Materials. A total of 200 images (1024×768 pixels) of real-world scenes, including landscapes, home interiors, and city scenes, were selected from the LabelMe database (http://labelme.csail.mit.edu/) as stimuli. Objects in each scene were annotated with corner coordinates of characteristic polygons defining the outline of the object shape and were labeled with English words. Each scene contained an average of 53.03 labeled objects (median = 40), covering 92.88% of the scene area.

Procedure. Following five practice trials, participants viewed the 200 scene images in random order. For each trial, they were instructed to inspect the scene and memorize it as thoroughly as possible. After a five-second presentation of each scene, an English word was shown for three seconds. Subjects had to manually report whether the object indicated by the word had been presented in the previously viewed scene. Target present and absent cases were evenly distributed among the 200 trials.

Data Analysis

Deriving Object Frequency and Predictability from Linguistic Analysis

Since the objects in the LabelMe scenes are labeled as English words or phrases, we are able to derive object frequency and predictability from a text corpus and LSA, similar to reading studies. Object frequency was computed from the British

National Corpus (BNC). We used the “basic level” of scene structure (see Oliva & Torralba, 2001) to describe each image in the materials, such as “landscape”, “bedroom”, “dining room”, “office”, “street”, or “kitchen.”. Object predictability was estimated using the LSA@CU (http://lsa.colorado.edu/) tool with the semantic space “General Reading up to 1st year college (300 factors).” Object labels and scene descriptions in single English words were considered as “terms”, while those in English phrases were considered as “documents” for LSA computation. For example, the object label “dish washer” and scene description “kitchen” used “document to term” as comparison type, and the result was 0.42. Since LSA computes the cosine value between two vectors, the highest value in LSA computation is one. The cosine value is usually close to zero for random vector pairs in a high-dimensional space. The computed predictability represents semantic consistency between an object and its embedding scene. Sometimes, the labels in the LabelMe dataset are not consistently applied to the same objects, for example, a "computer screen" in one scene could be labeled "monitor" in the next scene. Since the cosine value of the vectors representing “computer screen” and “monitor” is high (0.6), we could still get similar object-scene consistency using LSA with different synonyms of labels.

Deriving Object Frequency and Predictability from Visual Scene Analysis

Since LabelMe contains a large number of object labels, those labels can serve as a data source for computing frequency and predictability. The frequency of objects was accumulated across all available LabelMe scenes. To compute predictability based on scene data, we established a semantic space using labels in LabelMe.

There were 39,879 scene images and 303,033 annotated object labels (retrieved in February 2009). The objects with empty labels and the scene images without any object were excluded, which resulted in 39,724 scene images and 303,020 object labels. Since the labels were collected manually from the Internet, there existed different labels for identical objects (such as “car”, “SUV”, and “car occluded” for a car). We used a translating list (e.g. “car occluded” to “car”) provided by LabelMe to reduce the variability of object labels. This translation reduced the number of distinct object labels from 10,696 to 7,373.

We constructed a term-to-document matrix in which terms were defined as object labels and documents were images. Subsequently, a term-to-document matrix containing 303,020 objects was established, and local weighting was performed. Local weighting is aimed at diminishing the influence of objects that are extremely frequent in one document. Global weighting is often applied in text corpora in order to reduce the importance of terms that occur in every document and therefore do not help to differentiate meaning, such as function words (“a” or “the”). However, we did not apply global weighting because the dataset of visual scenes was quite different from a text corpus. For example, object labels do not contain function words, which occur in every document in a text corpus. The computation of local weighting is described in Equation 1 in Chapter 1.

Dimension reduction was then performed for the term-to-document matrix, and a “semantic space” was established. In this semantic space, each vector had 500 dimensions, which is within the typical range for LSA studies (Landauer et al., 2007). To compute predictability, every object label and scene image was considered as a vector in our LabelMe semantic space. We calculated the cosine value, representing semantic similarity, between the vectors of one object and its embedding scene image.

Identifying Fixated Object

The proportion of the area in the selected scene images covered by annotated object regions was 92.88%. However, this dense coverage has the disadvantage that many objects were occluded by others. In fact, 34% of all fixations in the present study were located in the intersection of two or more object regions so that it was hard to identify the actually fixated object, i.e., visible object that occluded the others. Even when we identified background objects by their labels (e.g., “wall”, “floor”, or “ceiling”), the percentage of fixations with multiple object regions was still 22%. Therefore, we estimated the depth-order of the intersecting objects based on the number of characteristic corners contributed by each object and the similarity of each object’s intersecting and non-intersecting parts in terms of their brightness (Histogram Intersection Similarity Method; Swain & Ballard, 1991).

Data Selection

There were a total of 19,767 fixated objects by all participants in our experimental data. Since the cognitive processes underlying the fixation of foreground and background objects might be different, we excluded 1,512 background fixation cases. In addition, we had to exclude 677 cases because the labels of fixated objects were not included in the LSA@CU tool, resulting in 17,578 cases. Subsequently, we categorized these cases to high/low frequency, high/low predictability, and large/small object size by selecting the top and the bottom 6,000 cases for each variable. Only cases that were categorized by all three predictors (frequency, predictability, and size) were selected. This selection resulted in 5,816 cases in the linguistic analysis and 5,951 cases in the visual scene analysis.

Object Size

In reading studies, there is a clear relationship between the probability of fixating a word and its length: As length increases, the probability of fixating a word increases (see Rayner, 1998, for a review). In scene viewing, large objects might receive more fixations than small ones. In our experiment, small objects contained an average of approximately 3,400 pixels, while large objects contained an average of approximately 90,000 pixels (see Table 13 below). The large objects might often have been larger than the observers' perceptual span, and therefore multiple fixations were required to identify the object. In this study, we analyzed large and small objects separately.

Correlations among Predictors

The correlations among predictors acquired from linguistic (represented by suffix 'L') and visual scene (marked by suffix 'V') analysis are shown in Table 13 for small objects and in Table 14 for large objects. We found that FreqL (object frequency derived from BNC) and FreqV (object frequency accumulated from LabelMe) are highly correlated in both small and large objects. Both FreqL and FreqV are weakly correlated with PredL. Moreover, both FreqL and FreqV are weakly correlated with PredV for small objects, but are somewhat correlated in large objects. This correlation might have been caused by not applying global weighting in the LabelMe semantic space. Therefore, although the influence of a highly frequent object in one scene was reduced by local weighting, highly frequent objects distributed across many scenes still received high predictability values. In addition, we found that Size was only weakly related to FreqL, FreqV, PredL, and PredV for either small objects or large objects, which means that object size did not significantly influence other predictors in such categorization.

Eye Movement Data Analysis

We separated the raw data into four groups: linguistic analysis for small objects, visual scene analysis for small objects, linguistic analysis for large objects, and visual scene analysis for large objects. The data in these four groups were submitted separately to an analysis of variance (ANOVA) with frequency (low vs. high) and predictability (low vs. high) as within-subject factors. The average area covered by objects (number of pixels), the natural logarithm of frequency, and the cosine values indicating predictability of objects in each group are shown in Table 15.

Table 13: Correlation among predictors of small objects. FreqL is the natural logarithm of frequency from British National Corpus (BNC); FreqV is the natural logarithm of frequency accumulated from LabelMe; PredL is the cosine value between the target object and its scene gist computed by LSA@CU; PredV is the cosine value between the target object and its scene computed from out LabelMe semantic space; Size is the number of pixels enclosed in the polygon of an object provided by the LabelMe dataset.

	FreqL	FreqV	PredL	PredV	Size
FreqL	—	.494	-.032	.118	-.095
FreqV		—	.068	.141	.155
PredL			—	.108	.073
PredV				—	.050
Size					—

Table 14: Correlation among predictors of large objects. Variable names are identical to Table 13.

	FreqL	FreqV	PredL	PredV	Size
FreqL	—	.551	-.013	.340	.174
FreqV		—	.049	.329	.149
PredL			—	.036	.108
PredV				—	-.033
Size					—

Table 15: The average area, frequency, and predictability of objects. Ling is linguistic analysis, and Visual is visual scene analysis; Pixels is the number of pixels enclosed in the polygon of an object provided by the LabelMe dataset.

			Frequency		Predictability	
	Size	Pixels	Low	High	Low	High
Ling	Small	3,350	6.24	9.90	0.01	0.39
Ling	Large	99,504	6.56	9.88	0.02	0.42
Visual	Small	3,495	4.21	9.67	0.30	0.66
Visual	Large	89,134	4.45	9.75	0.28	0.68

Results and Discussion

Linguistic Analysis – Small Objects

First Fixation Duration (FFD). In this study, FFD is defined as the first fixation on an object independent of whether it is the only fixation on an object or the first of multiple fixations on it. The main effect of frequency was found to be significant, F(1, 11) = 5.60, p<.05, indicating that high-frequent objects received less fixation time than low-frequent ones. There was neither a main effect of predictability, F(1, 11) = 0.03, p=.88, nor an interaction of the factors, F(1, 11) = .45, p = .52. The results of the FFD analysis in terms of mean values and their standard deviation are shown in Table 16.

In reading studies, FFD usually reflects early visual and lexical processing (including identification of orthographic form and a familiarity check), and is affected by both frequency and predictability (see Rayner, Ashby, Pollastsek, & Reichle, 2004, for a review). Consistent to reading studies, during scene viewing, we observed a frequency effect, but we failed to find predictability effect. Two factors may be responsible for this effect: First, we suggest that the high-frequent objects might be processed faster than low-frequent objects, which is similar to that high-frequent words are processed faster than low-frequent words. Second, since predictability was estimated as object-scene consistency, predictability effect are likely to only affect late stage processing, which might be reflected in total time (TT) but not FFD.

Gaze Duration (GD). In scene viewing, first-pass GD is hard to compute because there is no regular direction of visual scanning (such as the first left-to-right sweep over each sentence in English reading). In addition, due to the different size and shape of objects, it is also difficult to define a "spotlight" and know what objects in the scan path were actually processed. We computed GD using the sum of all fixation durations prior to moving to another object. We found significant main effects for frequency, F(1, 11) = 10.00, p<.01, but not for predictability, F(1, 11) = 0.18, p = .67. The interaction of frequency and predictability was not significant, F(1, 11) = .002, p = .96. The mean values and their standard deviation are shown in Table 17.

In reading studies, GD reflects higher-level lexical processing, such as the identification of a word's phonological and/or semantic forms, and lexical access. GD is also often influenced by both frequency and predictability during reading (see Rayner et al., 2004, for a review). The current frequency effect, and the fact that it is more pronounced for GD than for FFD, are consistent with findings from reading studies. We presume that the frequency effects were due to the same reasons discussed in the FFD section.

Total Time (TT). In this study, TT is computed as the duration sum of all fixations on an object including regressions. The main effects of frequency and predictability on TT were significant, F(1, 11) = 6.92, p<.05, and F(1, 11) = 5.78, p<.05, respectively, while there was no significant interaction of frequency and predictability, F(1, 11) = .36, p=.56. The mean and standard deviation of TT across factors are shown in Table 18.

In reading research, TT includes gaze duration and re-fixations and is thought to reflect information integration. In the present study, both frequency and predictability effects on TT were found. We suggest the explanation that participants re-fixate low-frequent or low-predictable objects more often than high-frequent or high-predictable objects. Both effects were consistent with results from reading studies.

Table 16: Summary of mean and standard deviation for FFD based on linguistic analysis for small objects

Freq	Pred	Mean (ms)	Std (ms)
Low	Low	265	29
	High	271	44
High	Low	260	41
	High	252	29

Table 17: Summary of mean and standard deviation for GD based on linguistic analysis for small objects

Freq	Pred	Mean (ms)	Std (ms)
Low	Low	302	36
	High	301	52
High	Low	288	40
	High	285	45

Table 18: Summary of mean and standard deviation for TT based on *linguistic analysis for small objects*

Freq	Pred	Mean (ms)	Std (ms)
Low	Low	340	45
	High	334	62
High	Low	333	49
	High	314	56

Visual Scene Analysis – Small Objects

First Fixation Duration. We failed to find significant effects on frequency, predictability, and their interaction, F(1, 11) = .01, F(1, 11) = .07, and F(1, 11) = .19, respectively. The results suggest that frequency and predictability derived from visual scene analysis might not be good predictors of FFD.

Gaze Duration. In scene viewing, again, we could not find frequency, predictability, and their interaction on GD, F(1, 11) = .02, F(1, 11) = 2.67, and F(1, 11) = .55, ps > .1. The results were similar to FFD.

Total Time. . The main effect of predictability was significant, F(1, 11) = 9.88, p < .01; low-predictable objects received more total time than high-predictable objects. There was no frequency effect, F(1, 11) = 2.36, p > .1 and no interaction, F(1,

11) = 4.59, p > .05. The results suggest that, similar to linguistic analysis, predictability tends to affect TT, which reflects information integration. The mean and standard deviation of TT across factors are shown in Table 19.

Linguistic Analysis – Large Objects

First Fixation Duration. The main effect of predictability was significant, F(1, 11) = 4.96, p<.05; surprisingly, high-predictable objects received more fixation time than low-predictable ones. This predictability effect is the inverse of that found in reading studies. There was neither a main effect of frequency, F(1, 11) = 0.01, p=.92, and no interaction of factors. The results of the FFD analysis in terms of mean values and their standard deviation are shown in Table 20.

Gaze Duration. We found significant main effects for both frequency, F(1, 11) = 13.35, p<.01, and predictability, F(1, 11) = 10.75, p < .01. High-frequent and high-predictable objects were fixated longer than low-frequent and low-predictable ones, respectively, which is the inverse of results from reading studies. The interaction was not significant. The mean values and their standard deviation are shown in Table 21.

Total Time. The main effects of frequency and predictability on TT were significant, F(1, 11) = 24.44, p<.001, and F(1, 11) = 23.14, p<.001, respectively. The direction of the effects was identical to GD and, again, the inverse of that found in reading studies. Besides, there was no significant interaction of frequency and predictability, F(1, 11) = .93, p=.36. The mean and standard deviation of TT across factors are shown in Table 22.

Table 19: Summary of mean and standard deviation for TT based on *visual scene analysis for small objects*

Freq	Pred	Mean (ms)	Std (ms)
Low	Low	362	77
	High	353	52
High	Low	421	99
	High	336	42

Table 20: Summary of mean and standard deviation for FFD based on linguistic analysis for large objects

Freq	Pred	Mean (ms)	Std (ms)
Low	Low	260	35
	High	269	37
High	Low	260	46
	High	268	34

Table 21: Summary of mean and standard deviation for GD based on *linguistic analysis for large objects*

Freq	Pred	Mean (ms)	Std (ms)
Low	Low	390	62
	High	431	71
High	Low	444	111
	High	505	109

Table 22: Summary of mean and standard deviation for GD based on *linguistic analysis for large objects*

Freq	Pred	Mean (ms)	Std (ms)
Low	Low	496	79
	High	576	102
High	Low	582	127
	High	694	120

Visual Scene Analysis – Large Objects

First Fixation Duration. Using the visual scene measure, we failed to find significant effects of frequency, predictability, or their interaction on FFD for large objects, F(1, 11) = 2.55, F(1, 11) = 1.80, and F(1, 11) = .01, respectively, ps > .1.

Gaze Duration. Similarly to FFD, there were no effects by frequency, predictability, or their interaction on GD, F(1, 11) = .38, F(1, 11) = 2.06, and F(1, 11) = .32, ps > .58.

Total Time. We found a marginal effect of predictability on TT, F(1, 11) = 4.52, p = .057, indicating a trend for high-predictable objects toward receiving greater TT than low-predictable objects. The effect was the inverse of what has been found in reading studies. There was no frequency effect, F(1, 11) = .62, p > .1 and no interaction, F(1, 11) = .03, p > .87. The mean and standard deviation of TT across factors are shown in Table 23.

Table 23: Summary of mean and standard deviation for GD based on *visual scene analysis for large objects*

Freq	Pred	Mean (ms)	Std (ms)
Low	Low	587	83
	High	636	190
High	Low	560	96
	High	617	95

General Discussion

The results for small objects suggest that FreqL has effects on FFD, GD, and TT that are similar to those found in reading studies. Although the correlation of FreqL and FreqV is high (see Table 13), the results suggest that FreqV was a better predictor of FFD, GD, and TT compared to FreqL.

It is interesting that although PredL and PredV were only weakly correlated, both PredL (object predictability derived from linguistic analysis) and PredV (object predictability computed from visual scene analysis) influenced TT. We suggest that both PredL and PredV capture object-scene consistency, and both influenced TT. The frequency effects observed in scene viewing on FFD might reflect early visual processing, GD might reflect the higher-level cognitive activities (such as semantic activation), and TT might reflect information integration as observed in reading studies. Predictability effects on TT in scene viewing might reflect late-stage semantic verification of object-scene consistency.

Object size had a substantial influence on GD and TT - larger objects were fixated longer. More importantly, small and large objects induced very different frequency and predictability effects: For large objects, frequency effects were found on GD and TT in the linguistic analysis but not in the visual scene analysis. Predictability effects were found on FFD, GD, and TT in the linguistic analysis and on TT in the visual scene analysis. Interestingly, the direction of all effects for large objects was the inverse of that found in reading studies.

We suggest that the processing of large objects might be particularly demanding and thus induce gaze behavior that is substantially different from processing both small visual objects and written words. Conceivably, the size and complexity of large objects may often not allow its inspection within a single fixation. In the current memorization task in which the participants need to quickly develop a semantic understanding of the scene, inspecting objects that are frequent and predictable (i.e., consistent with the scene gist) may be most efficient. Therefore, the inspection of large objects that are less useful in this regard may not be completed but interrupted after the initial fixation. If such effects exist, it is clear that the object-based measures for large objects in scene viewing play a different role than those for both small objects and written words.

Based on the current results, we propose that frequency and predictability, from both visual scene and linguistic analysis, and size should be taken into account to develop a computational model of fixation durations in scene viewing. As discussed above, it is important to notice that the current same-direction effects for small objects and inverse effects for large objects were observed in a brief-presentation memorizing task. We suggest that such effects may vary for different tasks (for example, long-presentation memorizing tasks, visual search, counting objects, or scene gist recognition), which we will investigate in future studies. Regarding the design of such experiments, the current data indicate that LabelMe and

LSA are useful, complementary tools for studying eye movements during scene viewing.

CHAPTER 7

SEMANTIC GUIDANCE OF EYE MOVEMENTS IN REAL-WORLD SCENES

In the previous chapter, we found that reading and scene viewing share some mechanisms, and the semantic effects estimated by LSA influence fixation durations. It is possible that semantic factors, such as meaning and semantic relations among objects, also affect where we look.

There have been studies of contextual effects in visual search, in which eye movements were constrained by contextual spatial knowledge of the scene, e.g., information about the objects likely to be found (Neider & Zelinsky, 2006; Torralba, Oliva, Castelhano & Henderson, 2006), and studies of primitive semantic effects based on co-occurrence of objects in terms of implicit learning (Chun, & Jiang, 1998; Chun & Phelps, 1999; Manginelli & Pollmann, 2008), the contextual relations investigated in those experiments depended on the spatial distribution or the consistency of scene objects.

Scene consistency has been the subject of numerous studies. One line of research has focused on objects that are not semantically consistent with the scene gist, referred to as “semantic violations”, such as an octopus in a farmyard, or a microscope in a kitchen (Biederman, Mezzanote & Rabinowitz, 1982; Bonitz & Gordon, 2008; Henderson, Weeks & Hollingworth, 1999; Joubert, Fize, Rousselet & Fabre-Thorpe, 2008; Loftus & Mackworth, 1978; Stirk & Underwood, 2007; Underwood, Humphreys & Cross, 2007). Another line of research has studied objects that are semantically consistent but located in unexpected places in the scene structure or in unusual orientations, referred to as “syntax violations”, e.g., a floating cocktail glass in a kitchen or a fire hydrant on top of a mailbox in a street scene (Becker, Pashler & Lubin, 2007; Biederman, Mezzanote & Rabinowitz, 1982; Gareze & Findlay, 2007; Vo & Henderson, 2009).

The mechanisms underlying the effects of semantic or syntax violations on eye movements are still not well understood. There is a current debate on whether semantic inconsistence guides eye movements in a similar way as visual saliency does. For example, Biederman et al (1982), Stirk et al (2007), Underwood et al (2007), Becker et al (2007), Bonitz and Gordon (2008) and Loftus and Mackworth (1978) found that inconsistent objects are often found earlier and detected more accurately, and they conclude that it might be the result of parafoveal or peripheral information processing that enables object identification. On the contrary, Henderson et al (1999) and Vo et al (2009) found no evidence for such extrafoveal analysis. Finally, Joubert et al (2008) found a mixed result of lower detection rate and faster reaction time for scene-inconsistent objects than for scene-consistent ones. Despite these varying results, there seems to be consensus that after the identification of an object, semantically or syntactically inconsistent objects draw more attention, suggesting that the estimation of semantic or syntactic relations is simultaneously processed with object identification.

It should be noted that these observed effects on eye movements are based on a single object-scene relation (semantic or syntactic violation) that rarely occurs in the real-world. The above studies thus over-simplify high-level visual perception, making it problematic to apply their findings to common cases in which the conceptual relations among all scene objects contain no semantic or syntactic violations.

Recently, there have been many efforts to understand the role of conceptual semantic influence on attention using various experimental methods. Belke et al. (2008) and Moores et al. (2003) used a visual search paradigm, in which the search target was verbally specified before a set of object drawings was displayed. By analyzing observers' response times and eye movements, these studies demonstrated that attention was preferentially attracted to those objects that were semantically similar to the target. Corresponding effects were obtained by Huettig and Altmann (2006) and Yee and Sedivy (2006) using the visual world paradigm. In their work, observers' eye-movement bias was analyzed while they were looking at multiple, well-segregated object drawings and listening to spoken object names. However, all of these studies were limited to briefly presented, simple search displays containing four to eight objects that were "intuitively" selected for their semantic relations – a scenario that drastically differs from any real-world situation. Moreover, these studies only demonstrate a generic tendency of semantic influence when activated by external, verbal stimuli.

While these previous findings point out the relevance of semantics to scene inspection and visual search, their contribution to our everyday control of visual attention is still unclear. For example, whenever we routinely inspect our real-world visual environment, is it possible that the semantic similarity among objects in the scene influences our visual scan paths (gaze transitions)? Conceivably, in order to quickly develop a semantic understanding of a given scene, observers may inspect semantically similar objects consecutively. If such effects exist, do they depend on the visual task, e.g., scene inspection or visual search, and do they vary for the same scene over time? Such *semantic guidance* has not been studied, most likely due to the difficulties of assigning eye fixations to objects in real-world scenes and due to the intricacy of defining semantic relations among objects. Moreover, a quantitative approach of assessing semantic guidance in eye-movement data is necessary.

Analyzing eye fixations on objects in scene images requires object segmentation and labeling. There have been numerous attempts to solve these problems automatically, ranging from global scene classification (Bosch , Munoz & Marti, 2007; Grossberg & Huang, 2009; Le Saux & Amato, 2004; Rasiwasia & Vasconcelos, 2008) to local region labeling (Athanasiadis, Mylonas, Avrithis & Kollias, 2007; Chen, Corso, & Wang, 2008; Li, Socher & Li 2009). However, their results are still unsatisfactory compared to human performance in terms of segmentation and descriptive labeling.

In order to estimate the effect of semantic similarities between objects purely based on visual scenes, the co-occurrence of objects in a large number of scene images and the importance of each object in the scene context - defined by its attributes such as size, location or luminance - would have to be carefully considered. For example, objects of frequent co-occurrence, close proximity, or similar shape could be considered as semantically similar. Unfortunately, analyzing a sufficient amount of scenes and computing semantic relations directly from the image data sources is impractical. It is important to notice, however, that semantic relations are formed at the conceptual rather than at the visual level and thus do not have to be derived from image databases. Consequently, any database that can generate a collection of contexts or knowledge might be used to represent the semantic similarity of objects.

For the present study, we chose the linguistics-based computational method referred to as Latent Semantic Analysis (LSA; Landauer & Dumais, 1997) to serve as a quantitative measure of semantic similarity between objects. Since annotated objects in LabelMe have descriptive text labels, their semantic similarity can be estimated by calculating cosine values for the labels of object pairs. In this study, we used the

LSA@CU text/word latent semantic analysis tool developed by the University of Colorado at Boulder for LSA computation.

Equipped with above tools, we conducted two experiments to study two everyday visual activities - scene inspection (Experiment 1) and scene search (Experiment 2). For each recorded eye fixation during scene inspection, we generated a semantic saliency map. These maps were similar to feature-wise saliency maps used in visual feature guidance analysis (e.g., Hwang et al., 2009; Peters & Itti, 2007; Pomplun, 2006; Zelinsky, 2008), but were entirely based on the semantic similarities between the currently fixated object and the other objects in the scene. Here, saliency was defined as the corresponding LSA cosine value. If observers' immediate gaze transitions between objects are guided by the objects' semantic similarities, then these saliency maps should predict the next saccade target at an above-chance level. We measured these effects of *transitional semantic guidance* using the Receiver Operator Characteristic (ROC). Similarly, in the scene search experiment, we additionally measured *target-induced semantic guidance* by computing saliency as the LSA cosine between the search target and the label of each non-target scene object, followed by an identical ROC analysis. Several control analyses were conducted to exclude confounds and ensure that actual semantic guidance was measured.

Experiment 1

Method

Participants. Ten subjects participated in Experiment 1. All of them were students at the University of Massachusetts Boston, aged between 19 to 40 years old, with normal or corrected-to-normal vision. Each participant received a $10 honorarium.

Apparatus. Eye movements were tracked and recorded using an SR Research EyeLink-II system with a sampling rate of 500Hz. After calibration, the average error of visual angle in this system is 0.5°. Stimuli were presented on a 19-inch Dell P992 monitor. Its refresh rate was 85Hz and its resolution was 1024×768 pixels. Subjects' responses were entered using a game-pad.

Materials. A total of 200 photographs (1024×768 pixels) of real-world scenes, including landscapes, home interiors, and city scenes, were selected from the LabelMe database (http://labelme.csail.mit.edu/, downloaded on March 10, 2009) as stimuli. Objects in each scene were annotated with polygon coordinates defining the outline of the object shape, and they were labeled with English words. When displayed on the screen, the photographs covered 40°×30° of visual angle. Each scene contained an average of 53.03±38.14 labeled objects (in the present work, '±' always indicates a mean value and its standard deviation), and the median object number per image was 40. On average, labeled objects covered 92.88±10.52% of the scene area.

*Procedure***. S**ubjects were instructed to inspect the scenes and memorize them for subsequent object recall tests (see Figure 18a). After the five-second presentation of each scene, an English word was shown and subjects were asked whether the object indicated by the word had been present in the previously viewed scene. Subjects had to respond within three seconds by pressing a button on the game pad. If they were unable to make the decision within that period, the trial would time out and the next trial would begin.

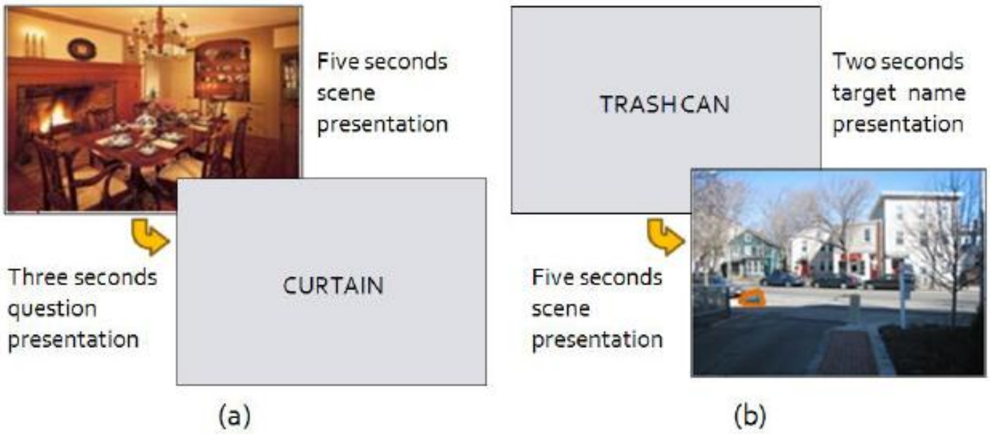

Five seconds scene presentation
Three seconds question presentation
CURTAIN
(a)
TRASHCAN
Two seconds target name presentation
Five seconds scene presentation
(b)

Figure 18: Examples of the experiment procedures. (a) Scene inspection task and (b) scene search task. The target object is marked for illustrative purpose.

Data Analysis

Table 24 shows examples of LSA cosine values for various object labels used in LabelMe scene image “Dining20” in terms of the reference object label “FORK”. This label has, for instance, a higher cosine value (greater semantic similarity) with “TABLE TOP” (0.43) than with “SHELVES” (0.09). This difference indicates that in the text corpus, “FORK” and “TABLE TOP” occur in more similar contexts than do “FORK” and “SHELVES”, which is plausible since, for example, forks are used for eating food on table tops rather than on shelves. The important feature of LSA is that it can quantify higher-level conceptual semantic similarity, regardless of any geometrical relation, functional relation or visual relation.

Since the images in the LabelMe database were annotated by many different contributors, objects are not always labeled consistently. For instance, the same object could be labeled “computer screen” in one image and “monitor” in another. In this study, we preserved the original object labels as much as possible by minimizing any kind of label modification - only spelling mistakes were corrected. Since LSA represents each term or document as a vector in semantic space, inconsistent but appropriate labels (synonyms) are mapped onto similar vectors. Therefore, the semantic similarity between synonyms is typically very high. While measuring the agreement among contributors in a meaningful way would require data beyond those offered by the LabelMe database, a study by Russell et al. (2008) suggests a high level of agreement. In their study, WordNet (see Fellbaum, 1998) was used to unify different contributors' object descriptions. They found only a small increase in the number of returned labels for several object queries before and after applying WordNet, indicating good consistency of the labels in the database.

Table 24: Sample LSA cosine values.

Label 1	Label 2	Cosine
…	…	…
FORK	TABLE TOP	0.43
FORK	PLATE	0.34
FORK	CANDLESTICKS	0.27
FORK	FIRE PLACE	0.17
FORK	SHELVES	0.09
…	…	…

To compute semantic similarity for each pair of object labels in our materials, a web-based LSA tool, LSA@CU (http://lsa.colorado.edu), developed at the University of Colorado at Boulder, was used. This tool was set to create a semantic space from "general readings up to 1st year college" and 300 factors (dimensions). Based on this space, we computed semantic similarity as the LSA cosine value, ranging between 0 and 1, for each object label compared to all other objects' labels for the same image. LSA cosine values are sometimes slightly smaller than 0 because of the high-dimensional space computation; we rounded negative values to zero. The average semantic similarity value for the pairs of labels in our materials was 0.245±0.061.

The reason for choosing LSA as our semantic similarity measure is that it is one of the fundamental and widely used approaches for estimating semantic relationships at the conceptual level based on semantic analysis among words, sentences, or documents. It would be difficult to reach a consensus about how semantic similarity should be measured, and LSA is just one possible approach that may capture relationships between words at the conceptual level. It is thus important to keep in mind that in the present work, we define "semantic similarity" to be similarity as measured by LSA.

*Constructing Semantic Saliency Maps for Gaze Prediction***.** We introduce the term “semantic saliency map” to refer to a saliency map purely based on the LSA cosine between the label of a given object (the currently fixated object during inspection or the target object during search) and the labels of other objects in the scene. The semantic saliency maps were normalized so that the total volume under them was one. In the current study, semantic saliency maps served as predictors of gaze behavior, as described in the following section.

Figure 19 shows examples of semantic saliency maps generated for a subject’s individual fixations on various objects in the scene (“Dining20”). As expected, when subjects fixate on an object labeled “PLANT IN POT” (highlighted in Figure 19b), semantically similar objects like “FLOWER IN VASE” and “DECORATIVE POT” receive high semantic activations. The LSA cosine values between “PLANT IN POT” and “FLOWER IN VASE”, and between “PLANT IN POT” and “DECORATIVE POT” are 0.53 and 0.37, respectively. In Figure 19c, where the subject is currently looking at one of the forks on the table, the semantically most similar objects are the other forks on the table (LSA cosine=1.00), which are maximally activated in the semantic saliency map. Objects that are semantically similar to “FORK”, such as

"BOWL" (LSA cosine=0.46) and "PLATE" (LSA cosine=0.34) still get higher activations compared to rather dissimilar objects like "FIREPLACE" (LSA cosine=0.19) or "CHANDELIER" (LSA cosine=0.13). Similar semantic saliency elevation between "FLAME" and "CANDLE STICKS" (LSA cosine=0.59) can be seen in Figure 19d.

*Measuring Semantic Guidance***.** Similar to bottom-up and top-down effects in visual feature guidance, we defined two kinds of hypothetical semantic effects that might guide eye movements in real-world scenes. One is *transitional semantic guidance*, which can be computed for both scene inspection and scene search, and the other is *target-induced semantic guidance*, which can only be computed for scene search and will be discussed in the context of Experiment 2.

Transitional semantic guidance affects immediate gaze transitions from one object to another. In other words, this guidance influences the choice of the next object to be inspected; our hypothesis is that there is a bias toward selecting objects that are semantically similar to the currently fixated object. Since this type of guidance is thought to influence transitions between objects, we measured it by analyzing only those eye movements that transitioned from one object to another.

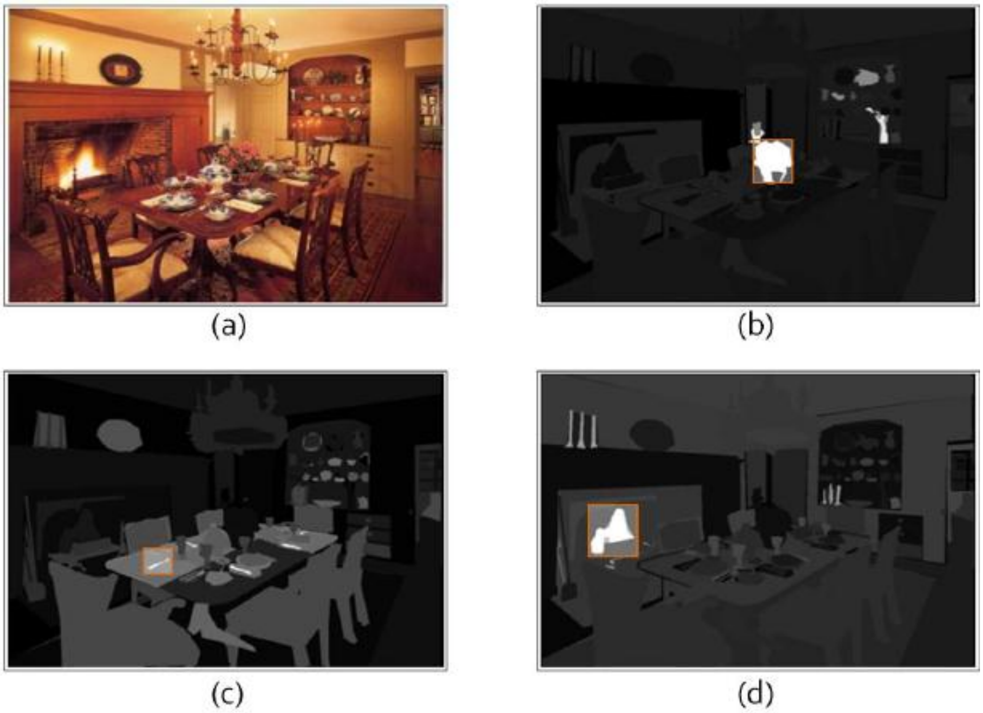


Figure 19: Examples of semantic saliency maps. The reference object (for instance, the currently fixated one) is marked with an orange square. (a) Original scene image (Dining20). (b) Semantic saliency map during gaze fixation on an object labeled as “PLANT IN POT”; greater brightness indicates higher activation. (c) Semantic saliency map when the observer fixates on an object labeled as “FORK”. (d) Semantic saliency map while fixating on an object labeled as “FLAME”. As it can clearly be seen, semantically more similar objects receive higher activation; for example, candle sticks in (d) are activated by the reference object labeled “FLAME”.

This restriction led to the exclusion of 36.2% of the saccades (23.5% within-object saccades and 12.7% saccades starting or landing outside of any marked objects) from the analysis of transitional semantic guidance in Experiment 1 (fixation sequences and durations within objects were examined in a separate study by Wang, Hwang & Pomplun, 2010). To be clear, this exclusion only affected saccades in the semantic guidance analysis, and no data were excluded from any fixation analyses.

In order to compute the transitional guidance measure, we first translated the sequences of eye fixations into sequences of inspected objects. For each gaze transition in a given scene, a semantic saliency map was generated based on the currently fixated object (see Figure 20). Subsequently, the ROC value was computed for the semantic saliency map as a predictor of the next object to be fixated by the subject. This calculation was very similar to previous studies using visual saliency maps (Hwang et al, 2009; Tatler, Baddeley & Gilchrist, 2005). All ROC values computed along scan paths, excluding successive fixations on the same object, were averaged to obtain the extent of transitional semantic guidance during the inspection of a real-world scene. If gaze transitions were exclusively guided by semantic information, making semantic saliency a perfect predictor of gaze transitions, then the average ROC value across all scenes should be close to one. If there were no semantic effects on gaze transitions at all, the average ROC value should be close to 0.5, indicating prediction at chance level.

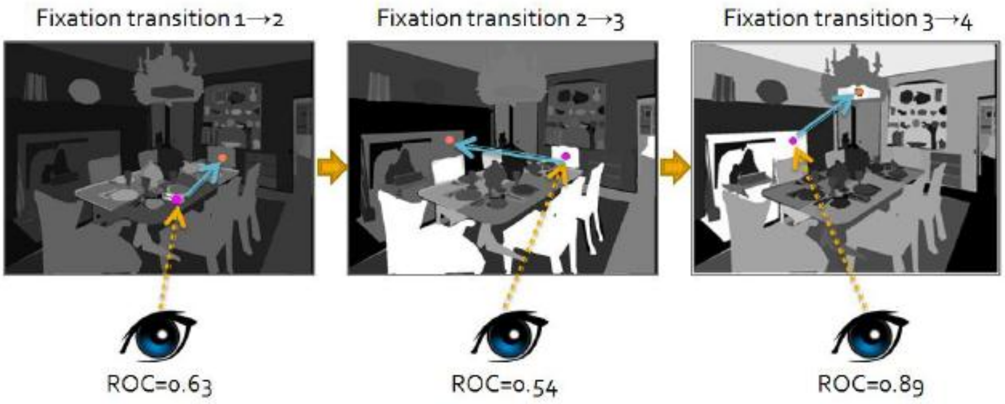


Figure 20: Examples of *transitional semantic guidance* computation. For each fixation transition, a semantic saliency map of the scene is generated based on the currently fixated object (dashed arrows). The semantic saliency of the next fixation target determines the guidance score (ROC value) of that gaze transition (solid arrows). The average of these scores across all gaze transitions during a trial is computed as the transitional semantic guidance for that trial.

Similar to the majority of studies using visual saliency maps (e.g., Bruce & Tsotsos, 2006; Hwang et al, 2009; Itti & Koch, 2001; Parkhurst et al, 2002), our semantic saliency maps for both inspection and search were static, i.e., did not account for the observer's gain in scene knowledge over time (see Najemnik & Geisler, 2005). Clearly, as with the visual saliency maps, this characteristic does not imply that, at stimulus onset, observers instantly build a complete, static semantic map that guides all of their subsequent eye movements. Observers certainly do not identify all objects in the scene at once, which would be necessary to instantly build a complete semantic map (see Torralba et al., 2006). Instead, we assume the semantic exploration of the scene to be an iterative process. For example, at the beginning of the inspection or search process, subjects may mostly identify objects that are close to the initial (central) fixation position. From this initial set, in the case of transitional semantic guidance, subjects tend to choose objects that are semantically similar to the initially fixated one. As inspection progresses, subjects identify more objects in the scene. While these dynamics are not reflected in our saccadic similarity maps, for the purpose of the current study, they are a straightforward, initial approach to investigating the existence and extent of semantic guidance.

Excluding Possible Confounds with Control Data Sets and Analyses. In order to control for possible confounds in the measurement of semantic guidance, ROC values were computed for three control data sets, namely (1) Random fixations, (2) Dissociated fixations, and (3) Greedy Model fixations. The random case, consisting of randomly positioned fixations, served as a test for correct and unbiased computation of ROC values. For example, if the normalized semantic saliency maps were biased toward greater saliency for larger objects, we may receive above-chance ROC values even for random fixations, because larger objects are likely to receive more fixations than small objects. The random data for each subject and trial were sequences of randomly selected fixation positions in the scene, simulating unbiased and unguided fixations. We used a homogeneous pseudo-random function to place fixations at random pixel coordinates (x, y) on the image. For each simulated trial, the number of gaze transitions during the inspection period was kept identical to the empirical number in a given subject's data (see Figures 21a and 21b). Any ROC values for the random case that deviate substantially from the chance level of 0.5 would indicate a bias in our ROC measure.

The dissociated case was introduced to control for likely confounds in the guidance measures: It is possible that semantically more similar objects tend to be spatially closer to each other in real-world images (proximity effect). Since amplitudes of empirical saccades during both scene inspection (5.81±4.30˚ of visual angle) and scene search (6.43±5.27˚) are significantly shorter (both ts(9)>24.125, ps<0.001) than those of random saccades (12.59±6.08˚ and 13.17±6.50˚, respectively), we might overestimate the extent of semantic guidance of actual eye movements simply because they favor transitions between spatially close objects. Furthermore, it is known that our eye fixations are biased toward the center of a presented image during experiments under laboratory conditions (Tatler, 2007), and real-world images are often biased by a tendency of photographers to put interesting objects in the center. Therefore, the empirical eye fixation distribution is unlikely to resemble the artificial, homogeneous distribution created in the random control case.

To measure the potential proximity effect on our guidance measure, we computed the ROC value for dissociated fixations and scenes, that is, we analyzed the eye fixation data measured in scene *n* against the object data from scene (*n*+1), and the eye fixation data in scene 200 against the object data from scene 1, in the randomized sequence of scenes. This technique conserved the spatial distribution statistics of the empirical eye movements while eliminating semantic guidance effects (see Figures 21a and 21c). Consequently, an ROC elevation above 0.5 in the dissociated case would indicate distribution (e.g., proximity) effects, and the ROC difference between empirical and dissociated fixations measures the actual strength of semantic guidance.

However, even the dissociated case may not provide sufficient assurance against artifacts entering the data, because it may be distorted by breaking the mapping between fixations and objects. While inspecting or searching through a scene, subjects presumably tend to fixate on objects. However, in the dissociated case, these fixations are superimposed on a different scene and do not necessarily land on objects anymore. Furthermore, successive fixations that transitioned between objects in the original scene may, in the dissociated case, land on the same object and would then be excluded from analysis.

In order to ensure that this characteristic of the dissociated case did not lead to a misinterpretation of the guidance data, we implemented the Greedy Model of gaze transitions. Following the idea of greedy, i.e., locally optimizing algorithms, this model always transitions from its current fixation location to the center of the display object with the shortest Euclidean distance from it, excluding the currently fixated object. If the semantic similarity maps predict the empirical transitions better than they predict the Greedy Model's transitions, this would further support the existence of transitional semantic guidance. For this evaluation, due to possible proximity effects (see above), it is important to only compare saccades of similar amplitudes.

When testing the greedy model, we found that if we simply started it at the screen center and let it perform a series of transitions that matched the number of transitional saccades in empirical scan paths, in most cases the model remained substantially closer to the screen center than the human gaze trajectories would. This was the case even when the model was prevented from visiting any object more than once. However, as discussed above, it is problematic to compare the guidance characteristics of scan paths with clearly distinct spatial distributions, and therefore we decided to use the empirical fixations as the starting points for all of the model's gaze transitions. To be precise, we took every object fixation from every subject and scene, recorded a transition from the current fixation to the closest object, and continued with the subject's next fixation, and so on (see Figures 21a and 21d). This approach allowed us to compute transitions that were not guided by semantic information but only by proximity, while preserving the spatial distribution of eye movements and their targeting of scene objects. Since each model transition was associated with a given subject's fixation, we compared empirical and model ROC values within subjects.

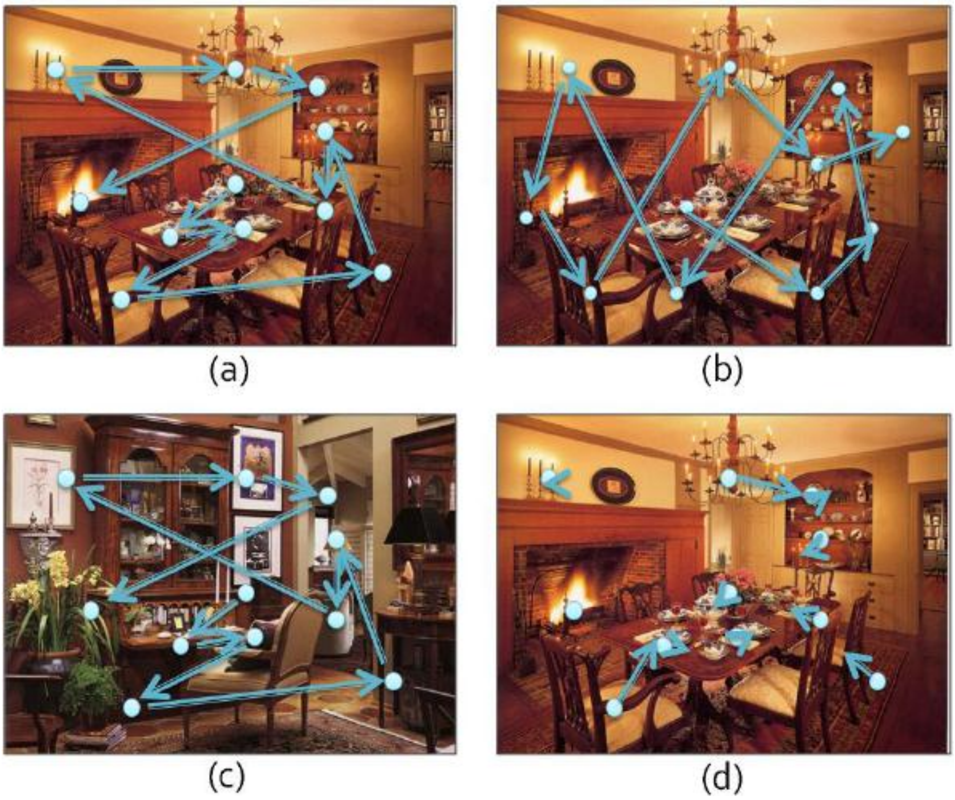


Figure 21: Examples of the three control cases. (a) Empirical eye movements in one of the scenes. (b) Random case, in which we computed the ROC value of simulated random fixations in the same scene. (c) Dissociated case, in which empirical eye fixation data were analyzed for different scenes than those in which they actually occurred. We computed the ROC value of eye movements made in scene 1 as predicted by the saliency map for scene 2, eye movements made in scene 2 as predicted by saliency in scene 3, and so on. (d) Gaze transitions produced by the Greedy Model, which uses empirical fixations as starting points and the spatially closest display objects as endpoints of its transitions.

As a final control measure, we computed transitional guidance by visual similarity, in order to rule out that transitional guidance is caused by low-level visual similarity rather than semantic similarity. Conceivably, visual and semantic similarities are positively correlated – semantically similar objects may be more likely to share visual features than do semantically dissimilar ones. For example, intuitively, different kinds of plants are semantically similar, and they are also visually similar, as green is their predominant color. Therefore, gaze guidance by both visual and semantic similarity of objects and the correlation between the two similarities have to be considered in order to get conclusive results.

Our visual similarity measure considered the following four important object characteristics: color, size, compactness, and orientation. Color similarity between two objects was measured by a simple, robust histogram matching method called Histogram Intersection Similarity Method (HISM; Swain & Ballard, 1991); Chen (2008) demonstrated the accuracy of the HISM method for estimating perceptual color similarity. Following these studies, our current color similarity measure included the three components of the DKL color model which is based on the human eye's cone receptor sensitivity regarding three wavelengths (short, medium and long) and double opponent (red-green, blue-yellow and luminance) cell responses (see

Krauskopf, Lennie & Sclar, 1990; Lennie, Derrington & Krauskopf, 1984). The computation of these features and their similarity is described in detail in Hwang et al. (2009).

Object size was measured as the number of pixels covered by the polygon that outlined the object. The compactness of an object was defined as the square of its perimeter, measured as the sum of the length of the enclosing polygon's edges, divided by the object's area. Compactness tells us whether an object is rather disc-shaped (low value) or elongated like a stick (high value). Finally, the orientation of an object was determined by computing a linear regression on all pixels belonging to that object, and taking the angle between the horizontal axis and the resulting regression line as the orientation measure. Size, compactness and orientation values were scaled to vary between 0 and 1.

The overall similarity between two objects was then computed as the product of similarity values along the four feature dimensions (color, size, compactness, and orientation), where color similarity was measured by the HISM method, size and compactness similarity were defined as one minus the absolute distance between feature values, and angular similarity was defined as the minimum angle difference between two line orientations. This type of multiplicative feature similarity was found to yield more robust results than additive techniques (e.g., Hwang et al., 2009).

It is clear that adding more visual feature dimensions to our similarity measure could still, at least slightly, improve that measure. In order to estimate the extent of such improvements, we also computed the measure with only the DKL color component, making it insensitive to size, compactness, and orientation. To assess the quality of visual similarity measurement, we computed the correlation between visual and semantic similarity across all object pairs in our study. This correlation is assumed to yield a positive coefficient for sound visual similarity measures (see below). We found a correlation of r=0.15 for the full visual similarity measure and only a very small decrease, r=0.147, for the color-only measure. This finding is in line with our previous studies (e.g., Hwang et al., 2009), showing that among commonly computed low-level visual features, color features exert by far the strongest guidance of eye movements. Moreover, this finding suggests that adding even more visual features is unlikely to drastically increase the correlation between visual and semantic similarity. It thus seems appropriate to use the full version of the current measure for estimating visual similarity guidance in the present context.

Results and Discussion

Basic Performance Measures. Subjects produced an average of 15.5±4.0 fixations per trial. Among those fixations, subjects made 11.1±1.7 gaze transitions between distinct objects, with average fixation duration of 248±29ms. The average saccade amplitude was 5.81±4.30˚. Response accuracy, measured as the percentage of correctly identified target-present and target-absent cases, was 72.0%.

Transitional Semantic Guidance and Control Cases. As described above, we computed four ROC values to study transitional semantic guidance, which were based on (1) empirical data, (2) random fixations, (3) dissociated fixations, and (4) the Greedy Model's gaze transitions. As shown in Figure 22a, the transitional semantic guidance value of simulated random fixations during scene inspection (0.508±0.123) was close to 0.5, i.e., chance level. This result indicates that the ROC computation was applied correctly and that the normalized saliency maps used for our analysis were unbiased. Moreover, the ROC value was significantly greater for the dissociated gaze-scene pairs (0.583±0.143) than for the random fixations, t(9)=17.16, p<0.001, evidencing the hypothesized proximity effect. Finally, the empirical eye movements had a significantly higher ROC value (0.646±0.127) than the random fixations, t(9)=23.28, p<0.001, and disassociated ones, t(9)=12.46, p<0.001. Consequently, we can conclude that although there is a significant proximity effect, actual transitional guidance still plays a significant role independently of proximity.

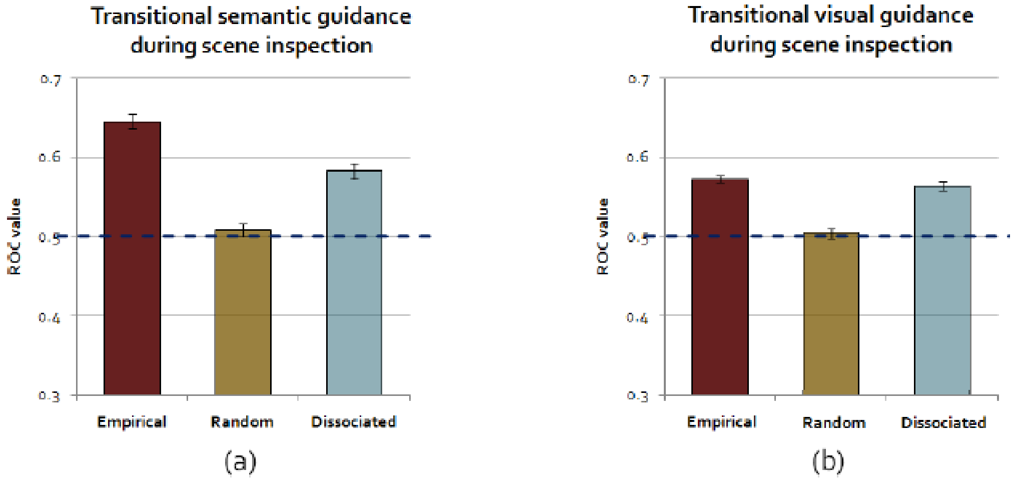


Figure 22: (a) Transitional semantic guidance and (b) transitional visual guidance during scene inspection (Experiment 1) as measured by the ROC method, with dashed lines indicating chance level and error bars representing standard error of the mean. The difference between the empirical and dissociated cases indicates the existence of semantic guidance.

As discussed above, due to proximity effects, the ROC analysis for the Greedy Model had to be performed separately for different saccade amplitude intervals. At the same time, this analysis had the added benefit of providing some insight into both the nature of the proximity effect and semantic guidance as a function of saccade amplitude. Figure 23 shows an ROC comparison of the empirical gaze transitions and those generated by the Greedy Model. Since the transitions produced by the greedy model tended to be shorter (3.43±2.75 degrees) than the subjects' transitions (5.81±4.30˚), t(9)=7.69, p<0.001, there were not many transitions larger than 10 degrees (3.2% of all transitions) to allow interval-based analysis. The cut-off point for empirical transitions was set to 18 degrees, with 1.8% of the transitions being longer. It can clearly be seen that the ROC values for the empirical transitions were consistently greater than those for the modeled ones. Comparing the average ROC values for saccade amplitudes below 10 degrees within subjects revealed a significant difference, t(9)=26.13, p<0.001, between empirical (0.667) and model data (0.586). Furthermore, the data for the Greedy Model show strong proximity effects, as evidenced by ROC values above 0.6 for object-to-object transitions shorter than 3 degrees. The ROC values decrease with longer transitions and seems to virtually disappear for transitions longer than 9 degrees. This pattern contrasts with the ROC values for the empirical transitions, which not only exceed the model's ROC for short transitions but remain at a constantly high level, even for transitions longer than 18 degrees. These results further support the view that the elevated ROC for the empirical eye movements is not an artifact caused by the arrangement of objects and their local contexts.

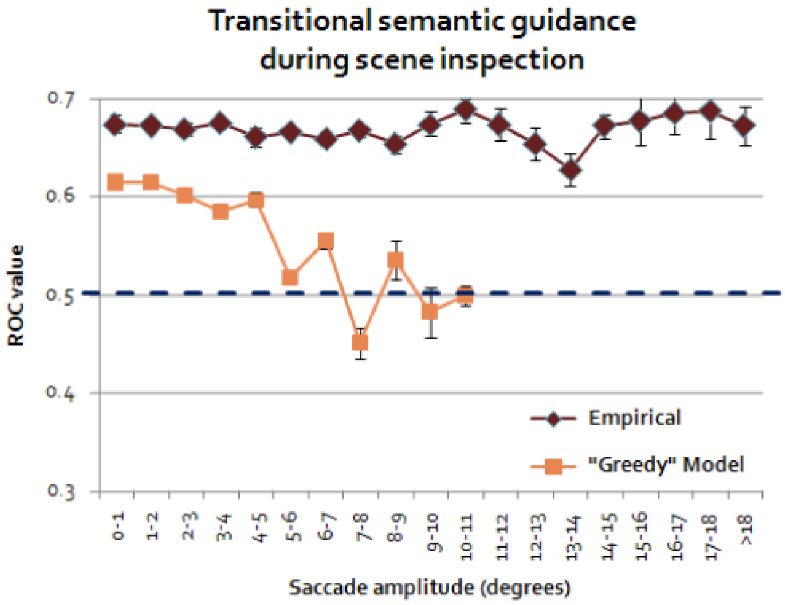


Figure 23: Comparison of transitional semantic guidance during scene inspection (Experiment 1) between empirical gaze transitions and transitions generated by the Greedy Model. Results are shown separately for different saccade amplitude (distance between transition starting point and endpoint) intervals. Note that all ROC values for saccades longer than 18 degree and 10 degree for the empirical and model data, respectively, were collapsed into one data point for each series. The dashed line indicates ROC chance level, and error bars show the standard error of the mean.

As discussed in the previous section, the final step in the guidance data analysis was to rule out the possibility that the observed guidance effects were purely caused by low-level visual similarity of objects instead of their semantic similarity. In order to quantify the influence of visual similarity on gaze movements, the correlation between visual and semantic similarity within objects was computed. For this computation, all possible pairings of objects across all 200 scenes used for the current experiments were analyzed (879,998 object pairs, average visual similarity of 0.161±0.152). As expected, the result shows a slight positive correlation between the two similarity measures, $r=0.15$, $p<0.001$. This finding suggests the possibility that the semantic guidance measured above could be an artifact resulting from strong guidance of eye movements by visual similarity and its correlation with semantic similarity. To examine this possibility, we computed ROC values based on the visual similarity of objects for the empirical, random, and dissociated cases in a manner analogous to our semantic guidance calculation. In this computation, the saliency maps that were generated for each gaze transition between distinct objects represented visual similarity, instead of semantic similarity, between the currently fixated object and all other objects in the scene.

As illustrated in Figure 22b, the random control case showed a near-chance ROC level of 0.510±0.092, and the dissociated case revealed an elevated ROC value (0.564±0.088) as compared to the random case, t(9)=7.09, p<0.001, demonstrating a visual proximity effect. However, the difference between visual similarity guidance for the empirical data (0.573±0.059) and the dissociated data did not reach statistical significance, t(9)=1.29, p>0.1. This finding indicates that semantic similarity, and not visual similarity, is the main factor underlying the current results.

Time Course of Transitional Semantic Guidance. Since we found significant semantic guidance effects, it is sensible to ask when this guidance starts and whether it is sustained throughout the trial. These temporal changes of semantic guidance during scene perception might help to understand the underlying mechanisms. We decided to examine transitional semantic guidance for each of the first nine gaze transitions after stimulus onset, which include 68.6% of all fixations, and an average value over 10th or later fixations.

As shown in Figure 24, transitional semantic guidance influences gaze movements throughout the trial, starting from the first gaze transition. However, since only saccades transitioning between different objects were included in the analysis, the present data cannot conclusively show whether this guidance is fully activated over all object in the scene at the first saccade in a trial. Nevertheless, the data suggest that semantic saliency guides the attentional selection of visual objects in a continuous and constant manner.

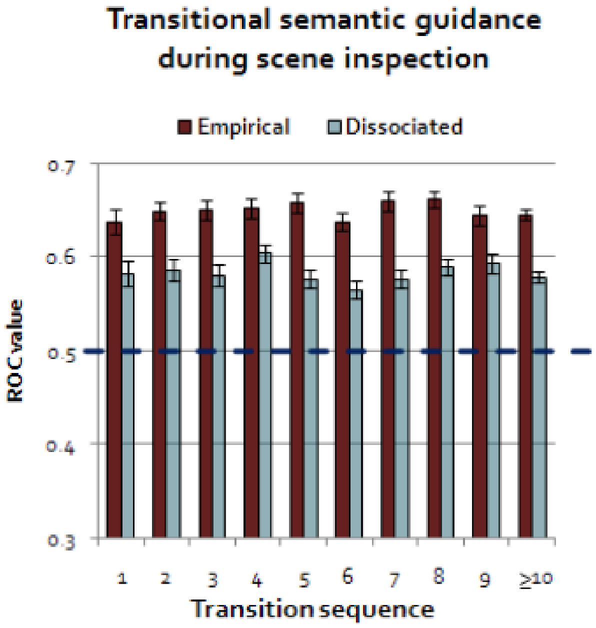


Figure 24: Temporal variation of semantic guidance during scene inspection indicated by the difference between the ROC values for the empirical data and the dissociated control case. The dashed line represents chance level, and error bars indicate standard error of the mean. Note that the rightmost column, labeled "≥10", includes the data for not only the tenth transition, but also for all subsequent ones.

To analyze in more detail regarding the timing of the guidance, the gaze transitions for each scene and each subject were separated into those visiting an object for the first time and those re-visiting an object. We found no significant difference in ROC values, t(9)=0.49, p>0.6, between the first-time visit (0.647±0.020) and re-visit groups (0.645±0.012). This finding suggests that transitional semantic guidance was not limited to revisiting of objects but also occurred, to the same extent, before an object was fixated for the first time.

While the results of Experiment 1 provide evidence for transitional semantic guidance during scene inspection, it also raises some questions. As discussed above, we can assume some level of object recognition to be necessary for accessing an object's semantic information (e.g., Torralba et al., 2006). Consequently, in order for transitional semantic guidance to take effect, it seems that such level of recognition must have been achieved for a set of potential target objects prior to the programming of a saccade. The guidance mechanism could then semantically relate these objects with the currently attended one and bias the selection of the next saccade target toward the most similar object.

However, as the Greedy Model and fixation re-visit analyses show, transitional guidance does not substantially decrease with greater eccentricity of saccade targets, even for large angles and for targets that have not previously been fixated. This finding seems to imply that recognition performance does not differ between objects at, for example, eccentricities of 1degree and 18 degree, which is clearly implausible. To explain this pattern of results, it should be noted that long saccades are rather infrequent; for example, only 19.4% of all saccades are longer than 8 degrees, and only 1.8% of them are longer than 18 degree. For most of these long saccades, it is conceivable to assume that a particularly salient peripheral object attracted the observer's attention prior to the saccade. This allocation of attention likely enabled at least some rudimentary processing of the object's visual information, possibly including some semantic analysis, before the saccade was programmed. In such situations, the semantic guidance mechanism may bias saccade-target selection toward either the peripheral object or one of the more central objects whose semantic information has already been accessed. Such a bias could prevent a large proportion of long saccades to peripheral objects that are likely unrelated to the currently fixated object based on the semantic information available. Thus, transitional semantic guidance could still exert a significant influence even on long saccades.

Experiment 2

Method

Participants. Ten subjects, who did not participate in Experiment 1, participated in Experiment 2, all of them were students at the University of Massachusetts Boston, aged between 19 to 40 years old, with normal or corrected-to-normal vision. Each of them received a $10 honorarium.

Apparatus and Materials. The apparatus and materials used in Experiment 2 were identical to those in Experiment 1.

Procedure. Subjects were instructed to search for objects whose name or description was shown prior to the scene. After a two-second presentation of the object name, the search scene was shown for five seconds. During each scene presentation, whenever subjects thought they had found a target object, they were to press a button while fixating on that object. Since there could be multiple target objects in the same scene, subjects were asked to continue searching for target objects until the trial ended (see Figure 18b). This task design allowed fixed five-second duration of scene presentation as in the scene inspection experiment (Experiment 1) and thereby enabled a useful between-experiments comparison of results. After scene presentation, the correct location of targets would be indicated or the text "Object does not exist" would be shown, in the target-present or the target-absent case, respectively. Search targets were randomly selected, then inspected, and possibly newly selected to avoid target objects that can be detected very easily. In each experiment, subjects performed 200 randomly ordered trials preceded by five practice trials. Target-present and target-absent cases were evenly distributed among the 200 trials.

Data Analysis. Besides the analysis of transitional guidance that was introduced in Experiment 1, the search task used in Experiment 2 motivated the additional study of a hypothesized second kind of semantic guidance, termed *target-induced semantic guidance*, influencing gaze distribution during scene search. This guidance reflects the extent to which semantic similarity between the target object and the objects in the search image determines the choice of fixated objects. Its computation for a given search scene-target pair only requires a single semantic saliency map, which represents the spatial configuration of semantic similarity between the target object and all non-target objects in the scene. As shown in Figure 25, the ROC value was measured for this saliency map as a predictor of all eye fixations made during a trial.

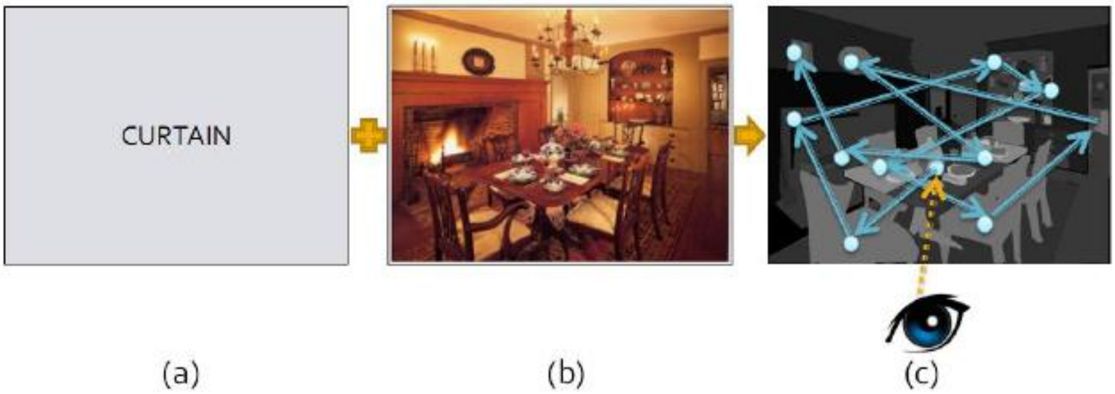


Figure 25: Example of *target-induced semantic guidance* computation in the scene search experiment. For each trial, a single semantic saliency map is generated based on the search target (a) and all objects in the scene (b). The ROC value for this map as a predictor of all fixated objects during that trial is taken as the guidance measure (c).

Results and Discussion

Basic Performance Measures. Subjects made an average of 16.2±1.6 fixations per trial in Experiment 2, with no statistical difference to Experiment 1 (15.5±4.0 fixations), t(18)=0.57, p>0.5. Among those fixations were 9.3±3.6 gaze transitions per trial between distinct objects. The average fixation duration was 301±87 ms, which was significantly greater than that measured in Experiment 1 (248±29ms), t(18)=2.67, p<0.05. Even if we exclude all fixations on targets in Experiment 2, which may have been prolonged by verification processes and by executing button presses, the resulting fixation duration (286±69ms) was still significantly greater than that in

Experiment 1, t(18)=3.27, p<0.05. This difference in fixation duration between scene inspection and scene search clearly differs from previous studies (e.g., Castelhano & Henderson, 2009; Vo & Henderson, 2009). A possible reason for the current pattern of results is that in the current scene inspection task, subjects were asked to memorize the objects in the scene. Given the large average amount of objects in the stimulus images that needed to be memorized in a short amount of time, subjects may have produced more saccades than they would have without any explicit task instruction.

The average saccade amplitude in Experiment 2 was 6.43±5.27˚, which was significantly larger than the one measured in Experiment 1 (5.81±4.30˚), t(18)=2.33, p<0.05. The subjects' response accuracy in Experiment 2, measured as the percentage of correctly identified target-present and target-absent cases, was 70.1%, which was very similar to Experiment 1 (72.0%). In target-present trials, subjects manually reported the detection of the first target after an average of 6.2±3.5 fixations.

Transitional Semantic Guidance and Control Cases. Analogous to Experiment 1, in Experiment 2 we excluded all saccades that did not transition from one object to a different one, which amounted to an elimination of 36.8% of the saccades, from all further analysis (25.9% within-object saccades and 10.9% saccades starting or landing outside of any labeled objects). As in Experiment 1, subsequent fixation analyses were not affected by this exclusion. Once again, we examined transitional semantic guidance through four ROC analyses based on (1) empirical data, (2) the random control case, (3) the dissociated control case and (4) the Greedy Model. As shown in Figure 26a, the ROC value for simulated random fixations, 0.504±0.104, was close to 0.5, indicating unbiased saliency maps. The ROC value for the dissociated gaze-scene pairs was significantly elevated (0.566±0.127) above the random-fixation ROC value, t(9)=17.10, p<0.001, revealing a proximity effect similar to the one observed in Experiment 1. Moreover, the ROC value for the empirical eye movements was slightly greater (0.583±0.134) than that for the dissociated case, t(9)=4.71, p<0.001. Even though this difference (0.017±0.012) was statistically significant, it was substantially smaller than the corresponding difference for scene inspection (0.063±0.016), t(18)=7.27, p<0.001.

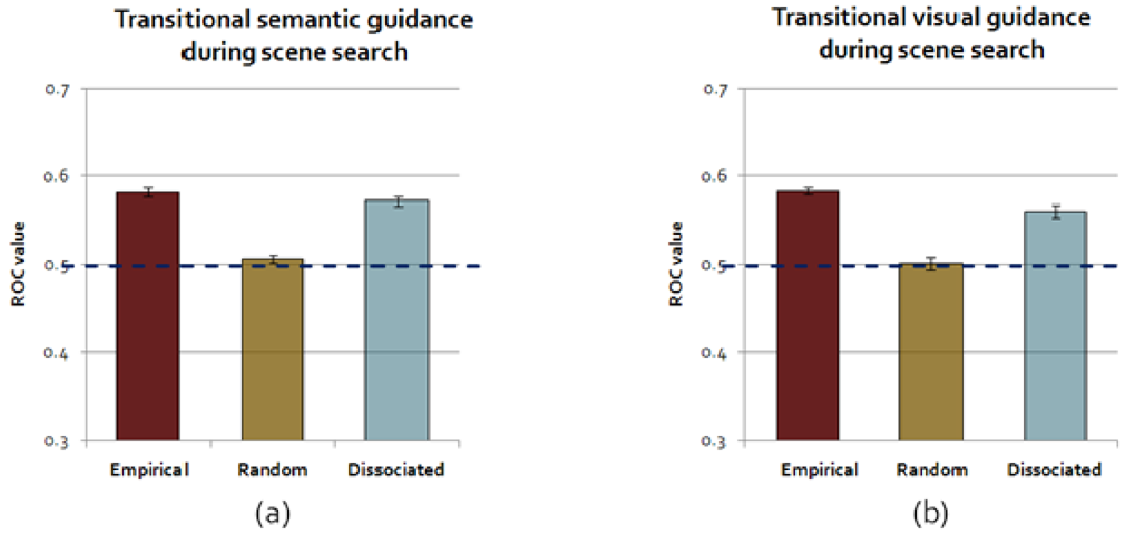


Figure 26: (a) Transitional semantic guidance and (b) transitional visual guidance in Experiment 2 (scene search) as measured by the ROC method, with dashed lines indicating chance level and error bars representing standard error of the mean.

In order to verify that the dissociated fixations did not bias the results by breaking the fixation-to-object mapping, we also applied the Greedy Model to the data of Experiment 2. Figure 27 illustrates the ROC values for the empirical and the modeled gaze transitions for different saccade amplitude intervals. Empirical transitions longer than 18 degrees (4.7% of the data) were pooled into a single data point, and for the model this was done with all transitions above 10 degrees, which also affected 4.7% of the respective data. The model's average saccade amplitude was 3.58±3.01 degrees. In contrast to Experiment 1, the ROC comparison between empirical and model data in Experiment 2 did not show a clear distinction. Comparing the mean values for saccade amplitudes below 10 degrees did not reveal a significant difference between the empirical (0.582) and model data (0.578), t(9)=1.18, p>0.25. Given this result and the small difference between empirical and dissociated ROC, the present data do not provide any conclusive evidence of transitional guidance during search. However, Figure 27 suggests that with greater amplitude, the model ROC decreases faster than the empirical ROC, allowing the speculation that there may be weak, long-range transitional semantic guidance effects during search.

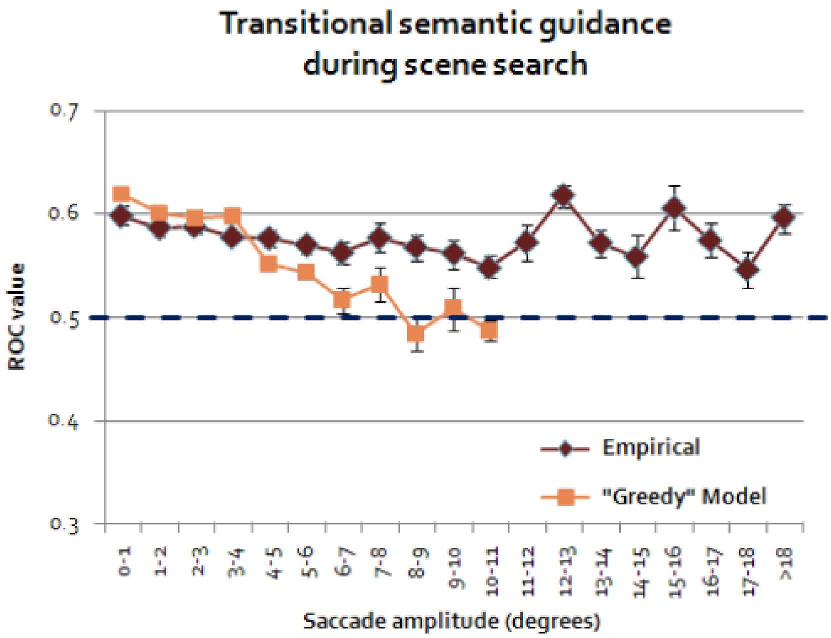


Figure 27: Comparison of transitional semantic guidance during scene search (Experiment 2) between empirical gaze transitions and transitions generated by the Greedy Model. Results are shown separately for different saccade amplitude (distance between transition starting point and endpoint) intervals. Note that all ROC values for saccades longer than 18 degrees and 10 degrees for the empirical and model data, respectively, were collapsed into one data point for each series. The dashed line indicates ROC chance level, and error bars show the standard error of the mean.

To investigate the contribution of low-level visual similarity to transitional guidance of eye movements during search, we computed ROC values based on visual similarity in the same manner as for Experiment 1. As illustrated in Figure 26b, an elevated ROC measure of visual guidance in the dissociated case (0.560±0.095) as compared to the random case (0.501±0.069), t(9)=7.00, p<0.001, demonstrated a proximity effect. There was also a slight effect of visual similarity guidance, as indicated by significant differences between the empirical case (0.583±0.058) and the dissociated case, t(9)=3.12, p<0.005.

Comparing Figures 26a and 26b, we find that, in contrast to Experiment 1, the ROC difference between the empirical and dissociated cases is greater for visual similarity than for semantic similarity. However, since both effects are very modest, we can only speculate about their behavioral relevance and underlying mechanisms. It is possible that both types of guidance slightly influence transitional eye movements during scene search, or that the transitional semantic guidance may, at least in part, be due to both transitional visual guidance and the correlation between the two similarity measures.

In summary, while there were statistically significant effects of both semantic and visual guidance on transitional eye movements in both scene inspection and scene search, the pattern of results differed noticeably between the two tasks. During scene inspection, subjects were guided much more strongly by semantic similarity as compared to low-level visual similarity. This finding suggests that during scene inspection tasks, subjects may inspect semantically similar objects consecutively to enhance scene memorization for later recall. In the scene search task, on the other hand, visual guidance is stronger than semantic guidance, but both influences are clearly weaker than that of semantic similarity in the scene inspection task. It seems that when subjects are assigned to a specific task, 'search for the target object', this task takes precedence over scene inspection. As a result, the strategy of gaze control may be shifted to a target-focused mode that is not aimed at object memorization. The slight transitional visual guidance found during search could be a result of subjects forming a visual template whose low-level features guide their search (cf. Schmidt & Zelinsky, 2009; Yang & Zelinsky, 2009).

Target-Induced Semantic Guidance. The analysis of target-induced semantic guidance during scene search was similar to the analysis of transitional semantic guidance. Target-induced semantic guidance represents the influence of semantic similarity between the search target and all non-target scene objects on fixation distribution. As shown in Figure 28, target-induced semantic guidance for the random control case (0.497±0.076) was close to 0.5, confirming that the target-based semantic saliency maps were unbiased. For the dissociated control case, target-induced semantic guidance was 0.506±0.145, which was very slightly, but significantly, greater than the value for the random case, t(9)=4.35, p<0.005, indicating a very small proximity effect in the scene search task. The empirical ROC value was significantly higher (0.637±0.159) than both the random, t(9)=18.79, p<0.001, and dissociated ones, t(9)=17.04, p<0.001. We can thus conclude that target-induced semantic guidance plays a significant role in scene search, independently of proximity effects.

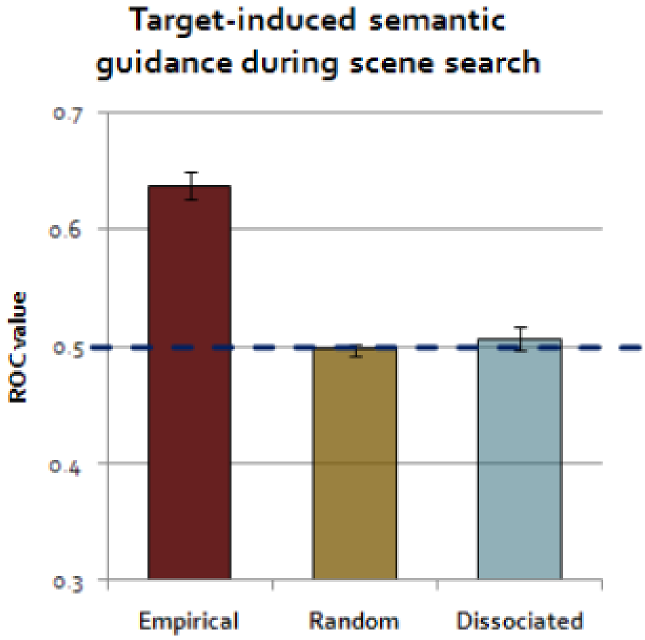


Figure 28: Target-induced semantic guidance during scene search, with a dashed line indicating chance level and error bars representing standard error of the mean.

Note that, in the search task, the target object was specified only by its verbal description, not by its visual features. Due to the large visual variation among those objects that match a given description, determining dependable and representative visual features of the target that could define visual similarity between the target and other objects in the scene is computationally infeasible. As a consequence, we did not attempt to compute target-induced visual similarity guidance in Experiment 2. Nevertheless, it is still possible that the ROC values for target-induced semantic guidance were partially due to visual guidance. For example, an observer searching for a fork may look at a knife not because the two are semantically similar, but because they look alike. While such an effect cannot be ruled out, the weak correlation (r=0.15) between visual and semantic similarity makes it seem unlikely that visual similarity, rather than semantic similarity, plays a major role in producing the current results.

Comparing the results for transitional and target-induced semantic guidance during scene search (Figures 26a and 28, respectively), it is noticeable that while transitional semantic guidance is hardly detectable (0.017±0.012), target-induced guidance is very pronounced (0.131±0.023). This finding further supports our interpretation that the specific priorities in the scene search task are responsible for reduced transitional semantic guidance as compared to the scene inspection task. More insight into this issue may be obtained by analyzing the time course of target-induced semantic guidance, which is reported in the following section.

Time Course of Target-Induced Semantic Guidance. Following the same grouping used in Experiment 1, we examined the time course of target-induced semantic guidance for each of the first nine gaze transitions after stimulus onset, covering 59.8% of all fixations, and an average value over 10th or later fixations.

As shown in Figure 29a, target-induced semantic guidance increased gradually during search in a given scene, followed by a decrease after approximately the sixth fixation. This pattern may be due to interference between visual saliency and semantic saliency. At the beginning of the search, although the target is specified by words, a visual representation of the target object, a ‘search-template’, may be generated and maintained in working memory in order to continuously match it with objects in the scene (e.g., Desimone & Duncan, 1995; Houtkamp & Roelfsema, 2009; Schmidt & Zelinsky, 2009; Wolfe, 1994; Yang & Zelinsky, 2009). Therefore, initially eye movements may be strongly governed by bottom-up and top-down processing of low-level visual saliency, with only little influence by semantic information. As search progresses, a better semantic understanding of the scene may develop, and if visual feature saliency by itself fails to detect the target, target-induced semantic guidance may start to dominate the control of eye movements. Conceivably, after the first target object in the scene has been found (as reported above, it occurs after an average number of 6.2±3.5 fixations), subjects may maintain strong semantic guidance to detect further targets. When no more targets can be found, guidance may slowly decrease.

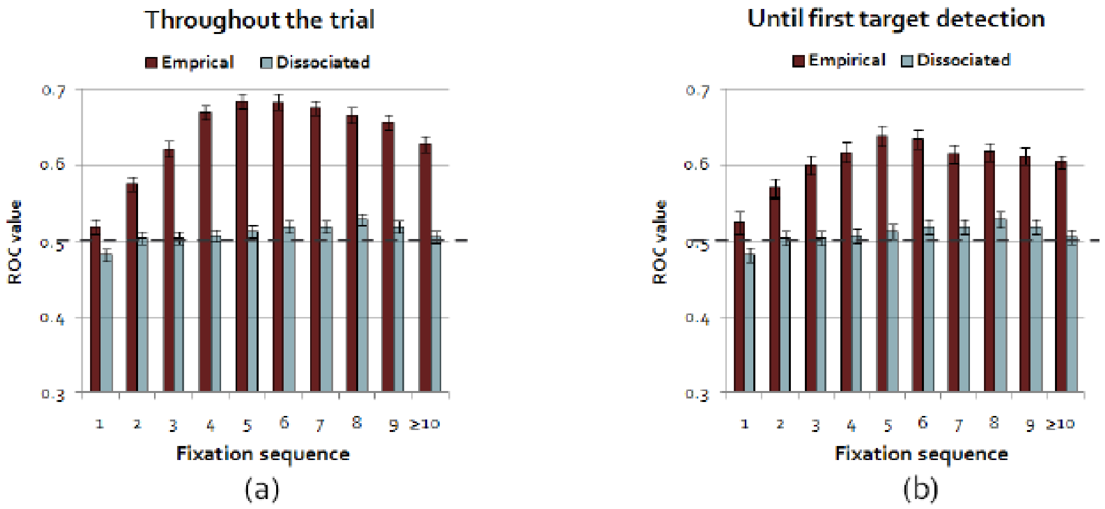


Figure 29: Temporal change of target-induced semantic guidance during scene search indicated by the difference between the ROC values for the empirical data and the dissociated control case, including fixations (a) throughout the trial and (b) from stimulus onset until first target detection. Dashed lines represent chance level, and error bars indicate standard error of the mean. The rightmost column, labeled "≥10", includes not only the data for the tenth transition or fixation, but also all subsequent ones.

We should consider the possibility, however, that the specific search task - requiring subjects to continue search for further targets after detecting the first one – may have artificially induced the steady increase in semantic guidance over time. The reason for such a data artifact could be that after the first target detection, subjects might transition back and forth between the first target and further target candidates. Due to the possibly high semantic similarity between these objects with the target label, such gaze behavior would likely increase target-induced guidance in the later stages of the search.

Our data revealed that subjects did in fact frequently revisit the first target object they detected; on average, this occurred 1.02±0.44 times per trial. However, this number does not appear large enough to suggest a significant impact of back-and-forth scanning behavior on guidance measurements. In order to rule out such potential bias from the data, we recomputed our analysis, but this time excluded all eye-movement data that were recorded in any given trial after the first target detection was reported. The resulting time-course of target-induced guidance (Figure 29b) shows slightly reduced ROC values but does not differ qualitatively from the initial one and thus supports the notion of a steady increase of guidance over the course of the search.

To analyze in more detail when transitional semantic guidance occurred, the gaze transitions for each scene and each subject were separated into those that visited an object for the first time and those that re-visited an object (regardless of whether this object was a target or not). We found no significant ROC difference, t(9)=1.83, p>0.1, between the first-time visit (0.576±0.020) and re-visit groups (0.592±0.019). This finding indicates that transitional semantic guidance was not limited to revisiting of objects but also occurred, to the same extent, before an object was fixated for the first time.

General Discussion

Previous studies on semantic effects on visual processing have focused on global contextual effects based on scene gist and eye fixation distribution, semantic effects in simple, artificial visual search tasks, or context effects based on co-occurrence or contextual cueing of objects. In contrast, the present work investigated semantic guidance of eye movements in real-world scenes, induced by the semantic similarity of scene objects to each other or to a search target.

We conducted two experiments to demonstrate semantic guidance of gaze transitions during scene inspection and semantic guidance of gaze distribution during scene search. To accomplish this, we introduced a novel interdisciplinary approach combining visual context research and linguistic research. Using eye-movement recording and linguistics-based LSA on object labels, we demonstrated that our visual scan paths in the inspection of everyday scenes are significantly controlled by the semantic similarity of objects. Our gaze tends to transition to objects that are semantically similar to the currently fixated one, basically unaffected by the time course of the inspection or whether an object is fixated for the first time or is revisited.

When interpreting the current data, we have to consider the possibility that, besides semantic guidance, contextual guidance may also have influenced the subjects' gaze transitions. While the dissociated control case allowed us to account for proximity effects, it did not fully control for the fact that semantically similar objects are often also located in contextually restrained parts of a scene in similar ways. For example, a spoon and a plate are often placed on a horizontal surface within the scene, such as a table. Eye movements during search in real-world scenes can be guided by such contextual factors (Castelhano & Henderson, 2007), relying on global scene statistics rather than the identification of individual objects and their semantics (Torralba et al., 2006). Since 'scene context' is ultimately built on the spatial layout of semantically related objects, it is difficult to rule out the possibility of contextual guidance in the current study. However, our data show that during scene inspection, transitional guidance by semantic similarity does not decrease with greater distance between the currently fixated object and the saccade target object. This finding is important because longer saccades should be more likely to move the observer's gaze beyond a contextually constrained part of the scene. Although such scene parts are sometimes large, e.g., the sky, we would expect at least a small reduction in average empirical ROC values for longer saccades if contextual guidance, and not semantic guidance, were the main factor driving the observed effects. This is clearly not supported by our data. Nevertheless, the contribution of contextual guidance to the effects observed in this study needs to be examined in future experiments.

While the current study demonstrates the existence of semantic guidance, it only allows a rough characterization of its underlying mechanisms. Clearly, it is impossible for observers to semantically analyze each individual scene object prior to their first eye movement in a scene. Instead, there has to be an iterative semantic exploration of the scene. The present data suggests that it involves parafoveal and even peripheral semantic analysis, since even long saccades tend to land on objects that are semantically similar to the previously fixated one. This finding is in line with several of the semantic inconsistence studies such as Underwood et al (2007), Becker et al (2007), and Bonitz and Gordon (2008), but it conflicts with others such as Vo et al (2009). The last study used well-controlled, computer-generated displays to control for some confounds in earlier work, and it did not find an effect of peripheral analysis. There are two possible reasons for the discrepancy between Vo et al's (2009) and the present data: First, Vo et al (2009) had subjects inspect a large number of scenes without presenting intermittent questions about the scene content. In contrast, our study required subjects to memorize a potentially large number of objects within five seconds of scene presentation, which may have induced a strategy of peripheral semantic analysis. Second, it is likely that the detection of semantic inconsistency differs from semantic analysis of individual visual objects. Guidance toward semantic information that is related to the currently attended information is a plausibly useful mechanism which allowing us in everyday life to efficiently explore the semantic content of a visual scene. Detection of semantic inconsistencies, however, is a rather unusual task as such inconsistencies rarely occur in the real world. It is thus possible that the human visual system has developed ability for at least a rough semantic analysis of peripheral objects, whereas the detection of semantic inconsistencies requires focal attention.

In a larger context, the transitional semantic guidance data may reveal a general mechanism of high-level attentional guidance by semantic association. As put, most prominently, by William James (1890), there are different "varieties of attention", among them visual attention and internal attention to our thoughts, with the latter variety producing trains of thought by association, i.e., transitions between semantically related concepts. The current study demonstrates that the former variety, visual attention, also proceeds by semantic association when exploring a visual scene and this is the first time that any such general attentional mechanism has been studied quantitatively.

Once a visual search task is involved, search seems to take precedence over scene inspection. In a search task, the fixation order is prioritized by similarity to the target, and as a result, guidance of individual gaze transitions by semantic factors almost disappears. However, the overall distribution of fixations during the search task shows strong target-induced semantic guidance - observers tend to inspect objects that are semantically similar to the search target. This result demonstrates that semantic bias of attention, as previously shown for artificial search displays (Huettig & Altmann, 2006; Yee & Sedivy, 2006), also exists during search in real-world scenes. Unlike transitional semantic guidance during scene inspection, target-induced guidance increases gradually during the time course of the search task. This increase in guidance is similar to, but slower than, the one observed in guidance of visual search by low-level visual features in real-world scenes (Hwang, Higgins & Pomplun, 2007). For such low-level guidance, it is assumed that observers first examine the overall composition of the scene in terms of its low-level visual features before using those features to guide their search (Pomplun, 2006). With regard to semantic guidance, a similar assumption seems plausible: It is possible that the progressively developing semantic understanding of the scene during the course of the search task is accompanied by an increased influence of target-induced semantic guidance on eye movements. Furthermore, it is likely that observers start their search under the guidance of a generic visual template that they create based on the verbal description of the target. This visual feature guidance may initially dominate the search process, at the cost of semantic guidance, until it either fails to detect the target, or more semantic context information becomes cognitively available, or both.

The current findings can be considered a first glimpse at the high-level, semantic mechanisms of attentional control in real-world situations. Further experiments are necessary to corroborate the current findings of semantic guidance by using explicit manipulations of semantic scene content and other semantic similarity measures than LSA. Future research should also address the dynamics of semantic guidance in more detail. It would be desirable to develop a dynamic model of semantic guidance that accounts for the iterative semantic exploration of real-world scenes and might be able to predict scanning behavior more accurately. Moreover, for a deeper understanding of these mechanisms, further research needs to address, in particular, the function of the observed semantic guidance. For instance, a crucial question to investigate is whether semantically ordered sequences of object inspections lead to better scene understanding or memorization as compared to random sequences. Furthermore, the processes underlying the gradual increase of target-induced guidance during search have to be examined. Will guidance increase even more slowly or stay at a marginal level in scenes showing unnatural arrangements of objects with no attainable global semantic understanding? Answering such questions promises to reveal the cognitive processes underlying semantic guidance and build a comprehensive, multi-level model of the control of visual attention. As argued above, such a model may generalize, at least in part, toward other "varieties" of attention.

CHAPTER 8

THE ATTRACTION OF VISUAL ATTENTION TO TEXTS IN REAL-WORLD SCENES

The results of Chapter 7 indicated that semantic factors affect where we look, which raises questions regarding how people process texts in real-word scenes, for instance, how do people locate and read signs or billboards that are embedded in a complex environment? Do semantic factors affect how fast we access texts, e.g., words vs. scrambled words, or English vs. Chinese texts for English vs. Chinese speakers?

Texts in real-world scenes were found to attract more attention than regions with similar size and position in free viewing task (Cerf, Frady, & Koch, 2009), but it is still an open question what factors would control such an attentional bias toward texts. It is possible that low-level visual saliency attracts attention (e.g., Itti, Koch & Niebur, 1998; Bruce & Tsotsos, 2006; Itti & Koch, 2001; Parkhurst, Law, & Niebur, 2002), or top-down control of visual attention (e.g., Hwang, Higgins, & Pomplun, 2009; Peters & Itti, 2007; Pomplun, 2006; Zelinsky, 2008). Another possibility is that texts typically carry higher saliency, luminance contrast, or edge information. Baddeley and Tatler (2006) suggested that different attention-attracting features are likely correlated in natural scenes (e.g., high spatial frequency edge-content information is typically associated with high contrast), and they found that edge-content information predicts the positions of fixations more accurately than do other features, such as contrast. The edge measures may thus be important factors that make texts more attractive than other objects.

Moreover, it is also possible that the typical *locations* of texts in the scene context are more predictable to contain important information and thus attract a disproportionate amount of attention. Torralba, Oliva, Castelhano, and Henderson (2006) suggested that scene context, i.e., the combination of objects that have been associated over time and are capable of priming each other to facilitate object and scene categorization, predicts the image regions likely to be fixated. Furthermore, Võ and Henderson (2009) claimed that scene syntax, i.e., the position of objects within the specific structure of scene elements, influences eye-movement behavior during real-world scene viewing. Such an effect would be in line with the studies of dependency among objects (e.g., the relative position of a plate and silverware; Oliva

& Torralba, 2007) and the contextual guidance model (Torralba et al., 2006), which predicts the expected location of the target in a natural search task based on global statistics from the entire image. Furthermore, Eckstein, Drescher, and Shimozaki (2006) recorded viewers' first saccades during search for objects that appeared in expected and unexpected locations in real world scenes, and they found the endpoints of first saccades in target-absent images to be significantly closer to the expected than the unexpected locations. Adding to the above results, an experiment by Mack and Eckstein (2011) investigated the effects of object co-occurrence on visual search, and it was found that viewers searched for targets at expected locations more efficiently than for targets at unexpected locations.

Finally, the *familiarity* of texts to viewers might also influence the attractiveness of texts; for example, observers' attention may or may not be attracted by the contents of an information board in a language that they do not understand. Cerf, et al. (2009) implied that attention to text may be developed through learning. If this assumption holds, a writing system familiar to viewers would be expected to catch more attention than an unfamiliar one. Here we have to distinguish between the meaning of texts that is inaccessible to observers who do not speak the given language, and their potential unfamiliarity with the writing system, i.e., the visual features of the texts. Both factors need to be investigated separately.

The goal of the present study was to investigate the contributions of low-level visual saliency, expected locations, specific visual features, and familiarity of texts to their ability to attract attention in real-world scene viewing. In order to test if texts are more attractive than other scene objects, in Experiment 1 an eye-tracking database of scene viewing by Judd et al. (2009) was first reanalyzed. In Experiments 2 to 5, new eye-movement data were collected and analyzed to study the factors that underlie the attraction of attention by texts.

Experiment 1: Reanalysis of Previous Data

Method

Participants. Judd and colleagues (2009) collected eye tracking data of 15 viewers. These viewers were males and females between the ages of 18 and 35. Two of the viewers were researchers on their project and the others were naive viewers.

Apparatus. All viewers sat at a distance of approximately two feet from a 19-inch computer screen of resolution 1280□1024 in a dark room and used a chin rest to stabilize their head. A table-mounted, video-based ETL 400 ISCAN eye tracker with a sampling rate of 240 Hz recorded their eye movements using a separate computer (see Judd et al., 2009). The images were presented at approximately 30 pixels per degree.

Stimuli. There were 1003 images in the database by Judd et al. (2009), and these images included both outdoor and indoor scenes. Some of these images were included in the freely available LabelMe image dataset (Russell et al., 2008) which contains a large number of scene images that were manually segmented into annotated objects. The locations of objects are provided as coordinates of polygon corners and are labeled by English words or phrases.

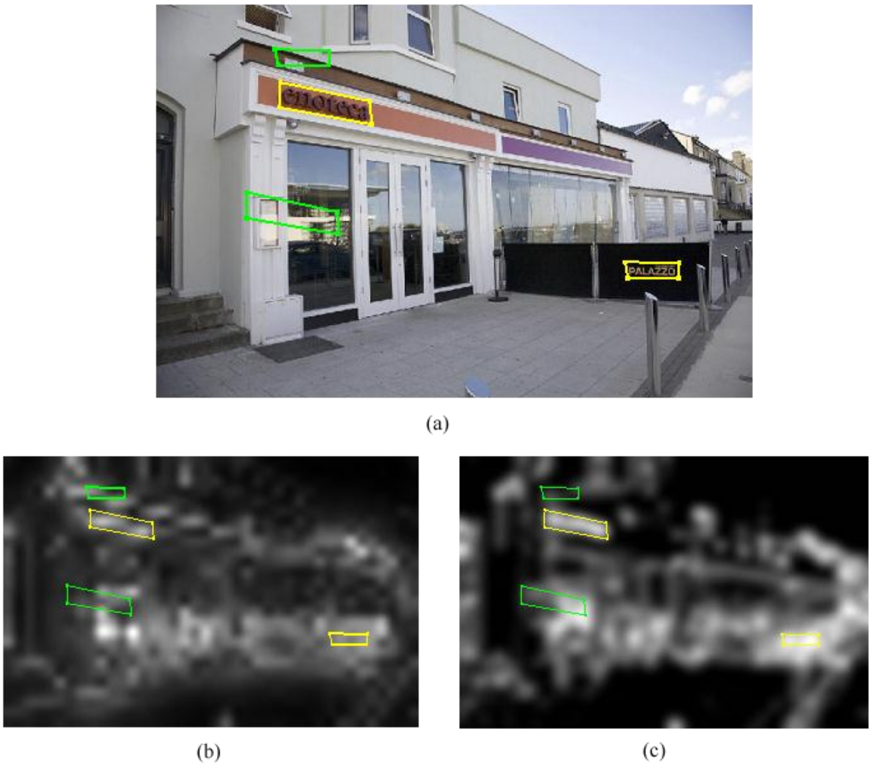
Procedure. All participants freely viewed each image for 3 seconds, separated by 1 second of viewing a gray, blank screen. To ensure high-quality tracking results, camera calibration was checked every 50 images. All images were divided into two sessions of 500 randomly ordered images. The two sessions were done on average at one week apart. After the presentation of every 100 images, participants were asked to indicate which images they had seen before to motivate them to pay attention to the images.

Analysis. The LabelMe dataset was used to identify and localize text in real-world scene stimuli. Out of the 1003 images we selected 57 images containing 240 text-related labels and another 93 images containing only non-text objects. Figure 30a shows one of the scene stimuli containing texts. The text-related labels included terms such as ‘text’, ‘banner’, or ‘license plate’. For the non-text objects, we excluded objects with text-related labels or background labels, e.g., ‘floor’, ‘ceiling’, ‘wall’, ‘sky’, ‘crosswalk’, ‘ground’, ‘road’, ‘sea’, ‘sidewalk’, ‘building’, or ‘tree’ since previous research has shown that viewers prefer looking at objects over background (Buswell, 1935; Henderson, 2003; Yarbus, 1967; Nuthmann & Henderson, 2010). It must be noted that the definition of background is not entirely clear (see Henderson & Ferreira, 2004). For example, objects labeled as ‘building’ or ‘tree’ may or may not be considered as background. To reduce the ambiguity, this study excluded ‘building’ and ‘tree’ from the set of non-text objects. The label ‘face’ was also excluded since faces have been shown to be particularly attractive (see Judd et al., 2009, for a review). There were 1620 non-text objects in the final selection. The images were rescaled to have a resolution of 1024×768 pixels (roughly 34×26 degrees of visual angle), and the coordinates of all objects were updated accordingly.

The raw eye movement data were smoothed using a computer program developed by Judd et al. (2009) that calculates the running average over the last 8 data points (i.e., over a 33.3 ms window). A velocity threshold of 6 degrees per second was used for saccade detection. Fixations shorter than 50 ms were discarded (see Judd et al., 2009).

In the analysis, several variables needed to be controlled for, such as the *eccentricity* and *size* of objects. It is known that these variables influence eye-movement measures, because observers tend to fixate near the center of the screen when viewing scenes on computer monitors (Tatler, 2007) and larger objects tend to be fixated more frequently. The eccentricity of an object (the distance from its center to the center of the screen) and its size (number of pixels) were calculated according to the coordinates provided by LabelMe. In order to control for low-level visual features in our analyses, we computed saliency, luminance contrast, and edge-content information of LabelMe objects. Saliency was calculated by the freely available computer software "Saliency Map Algorithm" (http://www.klab.caltech.edu/~harel/share/gbvs.php, retrieved on December 25, 2011) by Harel, Koch, and Perona (2006) using the standard Itti, Koch, and Niebur (1998) saliency map based on color, intensity, orientation, and contrast as shown in Figure 30b. The average saliency value of pixels inside an object boundary was used to represent object saliency. Luminance contrast was defined as the gray-level standard deviation of pixels enclosed in an object. For computing edge-content information, images were convolved with four Gabor filters, orientated at 0, 45, 90, and 135 degrees. Tatler et al. (2005) suggested to set the spatial frequency of the Gabor carrier to values between 0.42 and 10.8 cycles per degree, and we chose a value of 6.75 cycles per degree. All computations followed Tatler et al. (2005) and Baddeley and Tatler (2006) except that a popular boundary padding method, the built-in Matlab function "symmetric" was used and that the results were smoothed by a Gaussian filter (σ = 0.5 degrees). The average value of pixels inside an object boundary of the edge-content information map (shown in Figure 30c) was used to represent that object's edge-content information.

To derive matching control objects for all text objects, non-text objects were binned by eccentricity (smaller than 200, between 200 and 300, and greater than 300 pixels) and size (smaller than 1650, between 1650 and 5600, and greater than 5600 pixels). These ranges of eccentricity and size were selected to roughly include the same number of objects in each interval. Each text object was paired with one non-text object within the same size and eccentricity interval and matched in terms of saliency and luminance contrast as closely as possible. A text object and its non-text match were typically selected from different images.



(a)

(b)

(c)

Figure 30: (a) Texts (yellow polygons) and their paired control regions (green polygons) in one of the scene stimuli. The corresponding saliency and edge-content information are illustrated in (b) and (c).

Table 25: Average characteristics of text objects, non-text objects, and control regions. Size and eccentricity (Ecc.) are shown in pixels, and degrees of visual angle are shown in parentheses. Furthermore, saliency (Sal.), luminance contrast (LumC.), and edge-content information (EdgeC.) are presented.

	Size	Ecc.	Sal.	LumC.	EdgeC.
Experiment 1					
Text	2631 (2.92)	283 (9.43)	0.39	40	0.65
Non-Text	2828 (3.14)	292 (9.73)	0.40	40	0.64
Con. Region	2631 (2.92)	283 (9.43)	0.35	46	0.53.
Experiment 2					
Erased Text	2631 (2.92)	283 (9.43)	0.41	21	0.48
Non-Text	2676 (2.97)	293 (9.77)	0.41	24	0.57
Con. Region	2631 (2.92)	283 (9.43)	0.35	36	0.45
Experiment 3					
UncText H B	2351 (2.61)	288 (9.60)	0.20	10	0.22
UncText INH B	2723 (3.03)	281 (9.37)	0.36	55	0.59
UncText H	2351 (2.61)	288 (9.60)	0.25	34	0.43
UncText INH	2723 (3.03)	281 (9.37)	0.36	57	0.69
Non-Text H	2670 (2.97)	301 (10.03)	0.28	34	0.53
Non-Text INH	2746 (3.05)	284 (9.47)	0.38	57	0.69
Con. Region H	2351 (2.61)	287 (9.57)	0.26	40	0.50
Con. Region INH	2723 (3.03)	281 (9.37)	0.37	56	0.61

Additionally, for each text object, a control region in the same scene was set up that matched its counterpart exactly in its shape and size, and had identical eccentricity (Ecc.) and similar saliency (Sal.), luminance contrast (LumC.), and edge-content information (EdgeC.). The control regions could enclose non-text objects or backgrounds but did not intersect with any text objects. The characteristics of text objects, non-text objects, and control regions (Con. Region) are summarized in Table 25.

In order to measure the attraction of visual attention, object-based eye movement measures were used. We used one major measure, *fixation probability* (the probability of a fixation to land inside a text or non-text object or a control region during a trial), and two minor measures, *minimum fixation distance* (the shortest Euclidean distance from the center of the object or region to any fixation during a trial) and *first acquisition time* (the time from stimulus presentation to first target fixation). In every analysis, the major measure was used first in order to examine fixation preference, and subsequently the minor measures were used to support the major measure or to detect any effects when the major measure did not reveal any differences. One drawback of the fixation probability measure is that when there is no fixation landing inside an object boundary, the fixation probability for that object is 0 regardless of how closely a fixation approached it. The same drawback exists for first acquisition time; it may not be representative when fixation probability is low and only few data points become available. Minimum fixation distance was computed to overcome this drawback and provide convergent evidence for any attractiveness results. According to Nuthmann and Henderson (2010), viewers have a tendency to saccade to the center of objects in order to examine them. Their result may support the psychological validity of the measure of minimum fixation distance proposed in this study. Higher fixation probability, shorter first acquisition time, and shorter minimum fixation distance were considered to indicate stronger attraction of attention by a given object. A within-subject one-way analysis of variance (ANOVA) was used to examine the main effect of object category (texts vs. non-texts vs. control regions), and then Bonferroni corrected post-hoc tests were used for the comparison of conditions.

Results and Discussion

Fixation probability and minimum fixation distance of texts, non-texts and control regions are shown in Figure 31. The main effect of object category (texts vs. non-texts vs. control regions) on fixation probability was significant, F(2; 28) = 98.26, p < 0.001. Post-hoc tests revealed that the fixation probability of texts (M = 0.18, SD = 0.05) was significantly higher than the one of non-text objects (M = 0.08, SD = 0.02) and control regions (M = 0.03, SD = 0.01), both ps < 0.001. Furthermore, non-text objects had higher fixation probability than control regions, p < 0.001, which may be due to control regions not having an obvious boundary like text and non-text objects. This result is in line with the finding of Nuthmann & Henderson (2010) that viewers tend to fixate close to the center of objects (and therefore receive higher fixation probability), but not necessarily close to the centers of salient regions that do not overlap with real objects. In terms of the number of text objects in an image, we found that fixation probability decreases as their number increases, F(2; 42) = 25.52, p < .001, when all cases were categorized into bins of 1 to 4 (M = 0.25, SD = 0.07), 5 to 8 (M = 0.17, SD = 0.07), and more than 8 text objects (M = 0.09, SD = 0.03) with roughly the same number of cases in each bin. Post-hoc analysis indicated that all groups differed significantly, ps < .01. The results may be due to multiple text objects competing with each other, and the 3 second viewing may be insufficient for viewers to explore all text objects. Since we set up the same number of control regions for text objects in the same images, the number of text objects in an image should not influence the overall results.

We used minimum fixation distance instead of first acquisition time for additional analysis because average fixation probability was low (less than 0.2). The main effect of object category on minimum fixation distance was significant F(2; 28) = 106.06, $p < 0.001$. Minimum fixation distance was shorter for texts (M = 89.93, SD = 21.36) than for non-text objects (M = 115.79, SD = 28.05) and control regions (M = 137.31, SD = 26.03), ps < 0.001. Furthermore, non-text objects had shorter minimum fixation distance than control regions, $p < 0.001$. In summary, the consistency of these results suggests that texts were more attractive than both non-text objects and control regions.

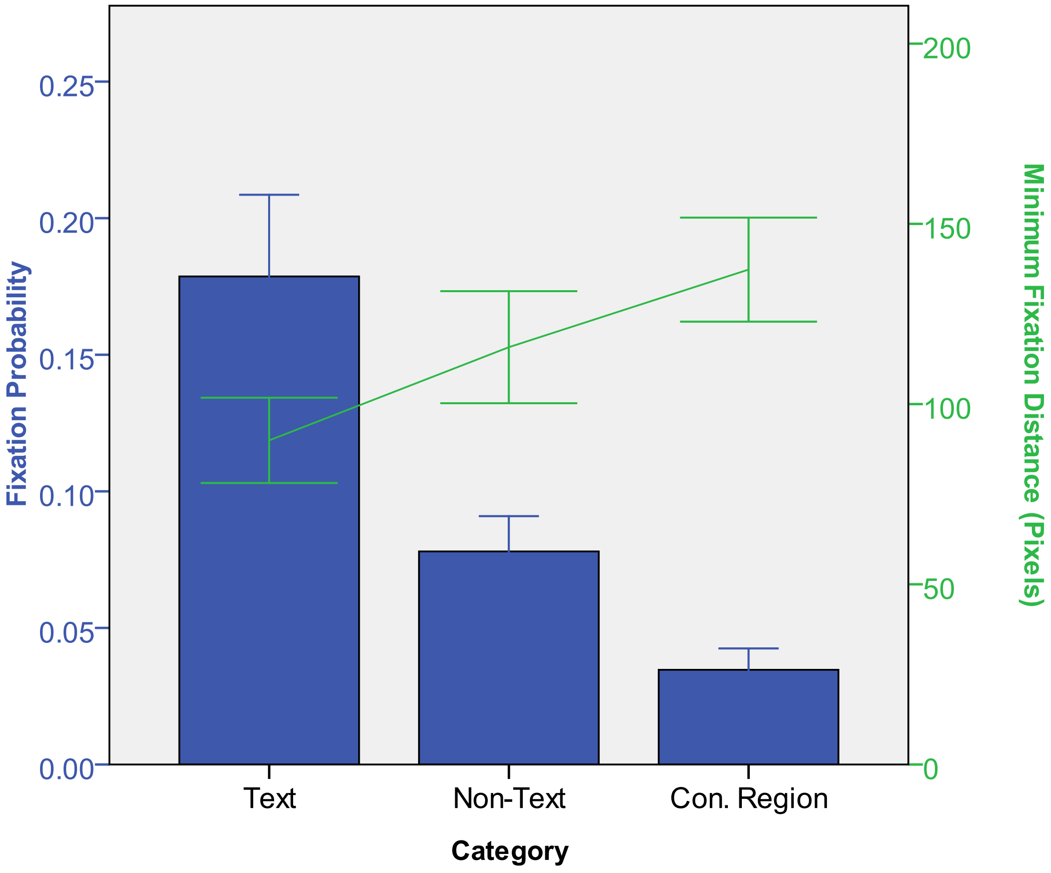


Figure 31: Fixation probability and minimum fixation distance of texts, non-texts, and control regions in Experiment 1. In this chart and all following ones, error bars are based on 95% confidence intervals.

The selected controls attempted to separate the contribution of low-level salience from high-level features such as expected locations, dependencies among objects or global statistics from the entire image, or unique visual features of texts to the allocation of visual attention. Texts, like faces, might have unique visual features that are unrelated to typical low-level visual saliency. Human observers may have developed "text detectors" during everyday scene viewing that are sensitive to these features and guide attention toward them. We will test how expected locations of texts affect eye movements in Experiment 2, and the potential influence of unique visual features of texts on attention will be examined in Experiment 3.

Experiment 2: Erased Text

To test whether the typical locations of text placement contribute to the attractiveness of texts, in Experiment 2 we “erased” the text parts from text objects and examined whether the observers’ attention was still biased toward these objects.

Method

Participants. Fifteen participants performed this experiment. All were students at the University of Massachusetts Boston, aged between 19 to 40 years old, and had normal or corrected-to-normal vision. Each participant received 10 dollars for participation in a half-hour session.

Apparatus. Eye movements were recorded using an SR Research EyeLink-II system with a sampling frequency of 500 Hz. After calibration, the average error of visual angle in this system is 0.5°. Stimuli were presented on a 19-inch Dell P992 monitor with a refresh rate of 85 Hz and a screen resolution of 1024×768 pixels.

Stimuli. The same 57 images and 240 text regions used in Experiment 1 were employed in Experiment 2. However, in Experiment 2, the "text parts" in text objects were removed manually, using the Adobe Photoshop 9.0 software, by replacing them with the background color of the texts as shown in Figure 32. This removal led to a reduction in average luminance contrast from 40 to 21 (see Table 25). Nonetheless, the average saliency was not affected by this text removal, due to the computation of saliency being based on center-surround differences in color, intensity, and orientation (see Itti, Koch & Niebur, 1998). Note that luminance contrast was computed exclusively within an object, but saliency was calculated according to the whole image, and the neighboring pixels of an object were taken into account. Therefore, a stop sign might still be salient without the text "stop" because of the color difference between the sign and its surroundings while its luminance contrast is reduced since there is minimal contrast inside the sign.

Procedure. After participants read the instructions, a standard 9-point grid calibration (and validation) was completed. Following two practice trials, participants viewed 130 stimuli in random order. They were instructed to freely inspect the scenes. At the start of each trial, a drift calibration screen appeared, and participants were instructed to look at the calibration dot that appeared in the center of the screen. After subjects had passed the drift correction, the stimuli were presented. Following a ten-second presentation of each scene, the stimulus disappeared and the calibration dot appeared again. In some cases, calibration and validation were performed once again to increase eye-tracking accuracy.

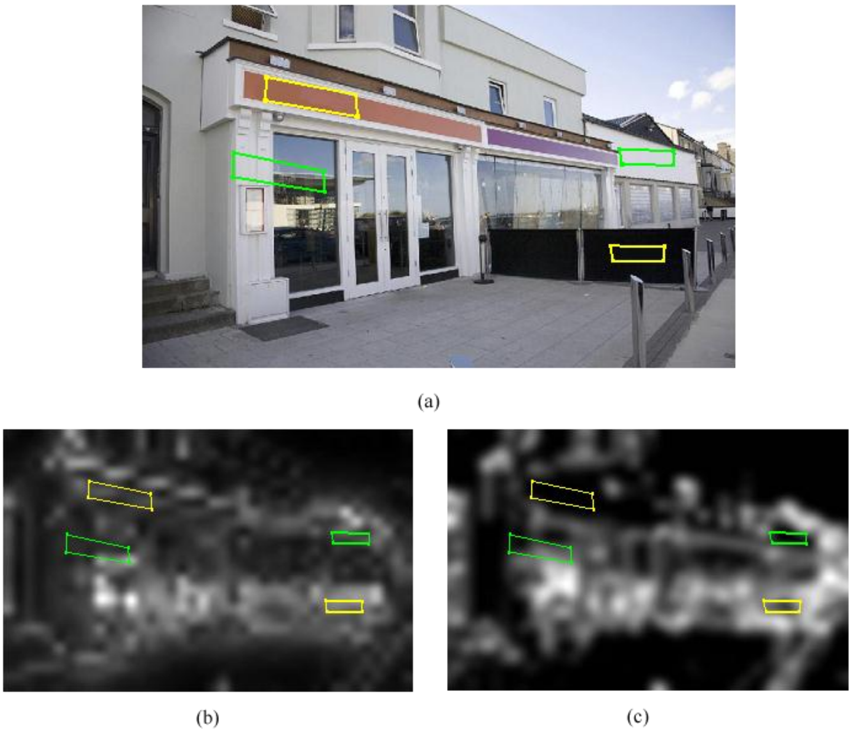


Figure 32: (a) Erased texts (yellow polygons) and their paired control regions (green polygons) in a sample stimulus for Experiment 2. The corresponding saliency and edge-content information are illustrated in (b) and (c). Note that the saliency and edge-content information of erased texts regions were reduced compared to Figure 30, and therefore the control regions were chosen differently.

Analysis. The raw eye-movement data were processed using the standard EyeLink parser (see EyeLink User Manual version 1.4.0 by SR Research). To investigate the attractiveness of texts during the initial visual scanning of the scenes, eye fixation data were only analyzed for the first 3 seconds of the viewing duration. In the same manner as performed in Experiment 1, non-text objects and control regions were chosen based on similar size, eccentricity, saliency, and luminance contrast (see Table 25). As mentioned above, the luminance contrast within the regions of removed texts was low due to these regions being "plain" after the text removal, but the saliency was affected less. For control regions, we were not able to match both saliency and luminance contrast, since these two variables were positively correlated, r = 0.34, for a randomly selected region from the given eccentricity. The luminance contrast of control regions (36) was higher than that of removed-text regions (21). We will further discuss this in the following section.

Results and Discussion

The main effect of object category (erased text vs. non-text vs. control region) on fixation probability was significant, F(2; 28) = 17.02, p < 0.001, as shown by a within-subject one-way ANOVA (see Figure 33). Post-hoc tests revealed that while erased texts (M = 0.07, SD = 0.02) had slightly higher fixation probability than non-text objects (M = 0.06, SD = 0.02), this difference was not statistically significant, p = 1.00. Both erased text and non-text objects received higher fixation probability than control regions (M = 0.03, SD = 0.01), both ps < 0.01.

For additional analysis, minimum fixation distance was used because average fixation probability was low (less than 0.1). The main effect of object category on minimum fixation distance was significant, F(2; 42) = 8.27, p < 0.01. A post-hoc test indicated that minimum fixation distance for erased texts was shorter than for non-text objects, t(14) = 5.06, p < 0.001 and for control regions, t(14) = 8.40, p < 0.001. Furthermore, minimum fixation distance for non-text objects was shorter than for control regions, t(14) = 2.35, p < 0.05. These results show that viewers did not fixate inside the boundaries of typical locations of text, which may be due to the plainness caused by text removal. However, the results of minimum fixation distance indicated that viewers paid a disproportionate amount of attention to the text removal regions within the scene.

The findings of Experiment 2 indicate that part of the attractiveness of texts derives from their prominent, expected locations in typical real-world images. This effect might be caused by dependencies among objects or global statistics within the entire scene. For example, viewers might recognize a store banner from its positions within the building layout, and they might be attracted by this banner region even without texts. However, Einhäuser and König (2003) pointed out that strong local reductions of luminance-contrast attract fixations. We consider this factor part of saliency because we found that the text removal regions still carried high saliency although their luminance contrasts were strongly reduced. We tried to match saliency between text removal regions and controls as much as possible in order to separate the contribution of low-level saliency from high-level features (i. e., expected location and special features of texts) to fixation positions.

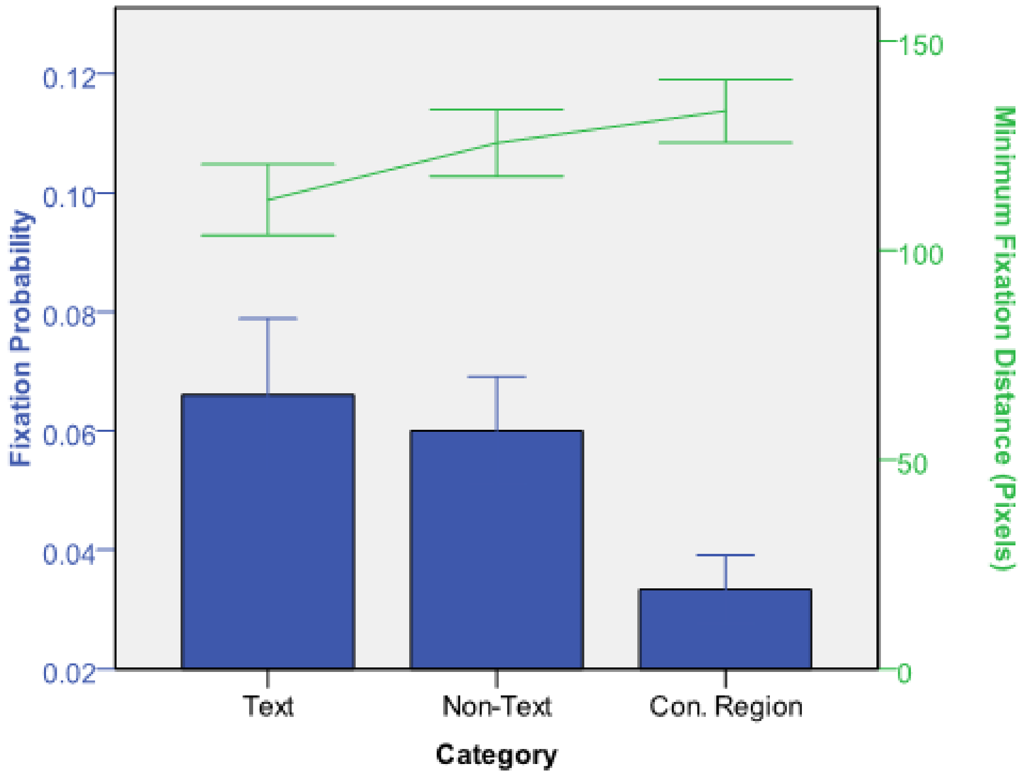


Figure 33: Fixation probability and minimum fixation distance of texts, non-texts, and control regions in Experiment 2.

Experiment 3: Unconstrained Text

To eliminate the influence of expected locations and test whether the unique visual features of text by themselves attract attention, Experiment 3 dissociated texts from their typical locations and placed them in front of homogeneous or inhomogeneous backgrounds. The purpose of using inhomogeneous backgrounds was to add visual noise (non-text patterns) to the unique visual features of text (text pattern), and we expected to find less attraction of attention by texts in front of such inhomogeneous backgrounds.

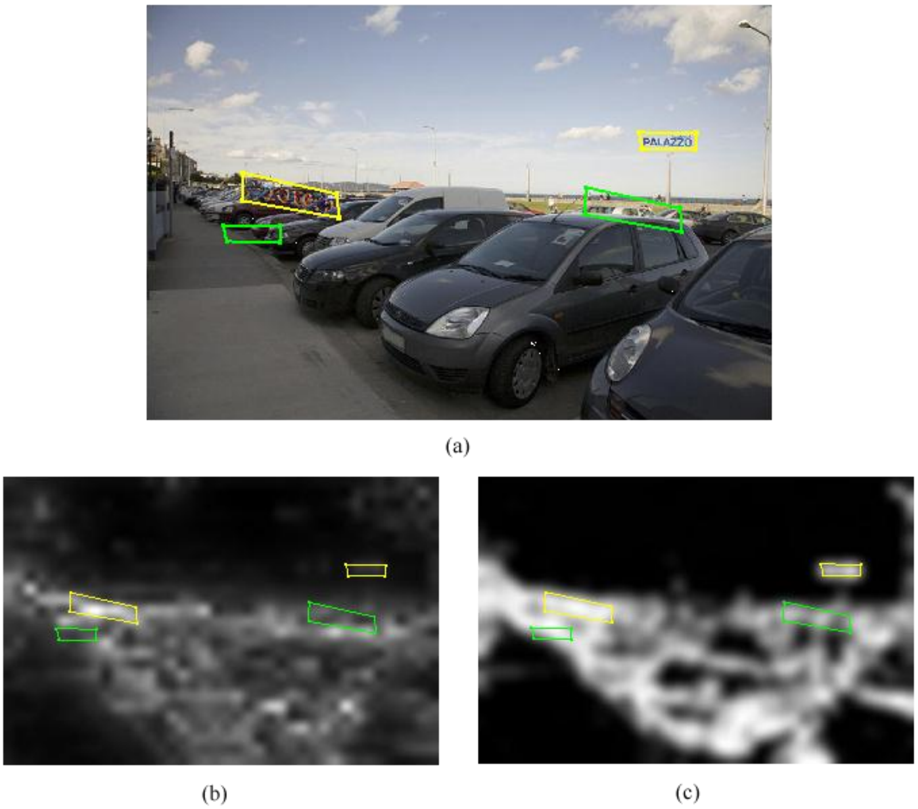
Method

Participants. An additional 15 students from the University of Massachusetts Boston participated in this experiment. None of them had participated in Experiment 2. All were students aged between 19 to 40 years old and had normal or corrected-to-normal vision. Ten dollars were received by each participant for a half-hour session.

Apparatus. Eye movements were recorded using an SR Research EyeLink Remote system with a sampling frequency of 1000 Hz. Subjects sat 65 cm from an LCD monitor. A chin rest was provided to minimize head movements. The spatial accuracy of the system is about 0.5 degrees of visual angle. Although viewing was binocular, eye movements were recorded from the right eye only. Other settings were the same as in Experiment 2.

Stimuli. To extract the “text part” of a text object, the difference in each of the RGB color components of every pixel in each text object between Experiments 1 and 2 was calculated. These patterns of color differences were recreated in other, randomly chosen scenes and placed in positions where the original size and eccentricity were maintained (see Figure 34). These unconstrained texts were prevented from overlapping with regions currently or previously occupied by texts. There were a total of 240 unconstrained text objects. Half of them were placed on homogeneous background, i.e., in regions with the lowest luminance contrast of all possible locations before placing the text parts, while the others were placed on inhomogeneous background, i.e., those areas with the highest luminance contrast. To prevent an unconstrained text from being placed on a computationally inhomogeneous but visually homogeneous background, e.g., half black and half white, the luminance contrast of a candidate region was calculated using 10×10 pixel windows covering the candidate region.

As discussed above, inhomogeneous backgrounds might cause visual noise that interferes with the unique visual features of texts and thereby reduces the attraction of the viewers' attention by such features. Table 25 shows the characteristics of the unconstrained text in front of homogeneous background before (UncText H B) and after (UncText H) the text parts were placed as well as those of the unconstrained texts in front of inhomogeneous background before (UncText INH B) and after (UncText INH) the text parts were placed.



(a)

(b)

(c)

Figure 34: (a) Unconstrained texts (yellow polygons) placed in front of homogeneous (right) and inhomogeneous backgrounds (left) and their paired control regions (green polygons) in one of the scene stimuli. The corresponding saliency and edge-content information are illustrated in (b) and (c).

Procedure. The procedure was identical to Experiment 2.

Analysis. The analyses were identical to Experiment 2. Three-second viewing durations were analyzed for unconstrained texts in front of homogeneous and inhomogeneous backgrounds. Each unconstrained text was paired with a non-text object and a control region using the same methods applied in Experiments 1 and 2. Table 25 lists the characteristics of paired non-text objects and control regions.

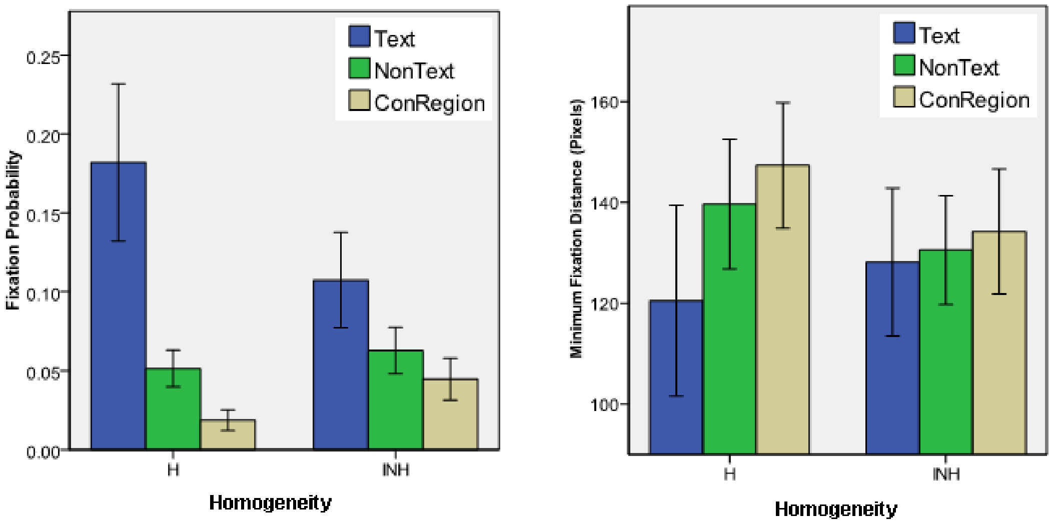
Results and Discussion

For fixation probability, a within-subject two-way (ANOVA) showed that the main effect of object category (texts vs. non-texts vs. control regions) was significant, F(2; 28) = 37.53, $p < 0.001$, the main effect of background (homogeneous vs. inhomogeneous) was also significant, F(1; 14) = 4.70, $p < 0.05$, and the interaction of object category and background was significant as well, F(2; 28) = 24.87, $p < 0.001$. As illustrated in Figure 35a, this interaction can be explained by the object category effect being more pronounced for homogeneous than for inhomogeneous background. A within-subject one-way ANOVA revealed that the main effect of object category for homogeneous background was significant, F(2; 28) = 38.68, $p < 0.001$. The fixation probability of unconstrained texts in front of homogeneous background (M = 0.18, SD = 0.09) was higher than for non-texts (M = 0.05, SD = 0.02) and control regions (M = 0.02, SD = 0.01), both ps < 0.001. The main effect of object category for inhomogeneous background was significant as well, F(2; 28) = 19.37, p < 0.001. The fixation probability for texts (M = 0.11: 0.11, SD = 0.05) was still significantly higher than for non-texts (M = 0.06, SD = 0.03) and control regions (M = 0.04, SD = 0.02), ps < 0.01, but the difference was not as large as for texts in front of homogeneous background.

For minimum fixation distance, a corresponding within-subject two-way (ANOVA) also revealed significant main effects of object category, F(2; 28) = 10.79, p < .001, and background, F(1; 14) = 18.07, p < 0.01, and their interaction was also significant, F(2; 28) = 11.77, p < .001. Within-subject one-way ANOVAs showed a significant main effect for homogeneous background, F(2:28) = 12.36, p < 0.001, and for inhomogeneous background, F(2; 28) = 3.56, p < 0.05. The post-hoc tests revealed that for homogeneous backgrounds, minimum fixation distance was significantly higher for unconstrained texts (M = 120.48, SD = 34.16) than for non-text objects (M = 139.64, SD = 23.21) and control regions (M = 147.29, SD = 22.51), ps < 0.05. For inhomogeneous background, minimum fixation distance of unconstrained texts (M = 128.12, SD = 26.49) was significantly higher than the one of control regions (M =

134.22, SD = 22.38), p < .05. As shown in Figure 35b, the trends were similar to fixation probability; unconstrained texts in front of homogeneous and inhomogeneous background received shorter distances than did control objects and regions and can therefore be considered more attractive.

To summarize, we found texts in front of homogeneous background (Text H) to be more attractive than texts in front of inhomogeneous background (Text INH; see Figures 35a and 35b). Regions with higher low-level saliency measures tend to receive more attention, but the opposite result was observed, i.e., Text INH was associated with higher saliency, luminance contrast, and edge-content information than Text H (see Table 25), but received less attention. Therefore, our data imply that the distinctive visual features of texts might be superior to low-level saliency measures in attracting attention.



(a) (b)

Figure 35: Fixation probability (a) and minimum fixation distance (b) of unconstrained texts in front of homogeneous (H) and inhomogeneous (INH) background, and the corresponding values for non-text objects and control regions.

It should be noted that participants being attracted by texts and actually "reading" texts are two different matters, and this study focused on how participants' attention was caught by texts. Text INH containing both text and non-text patterns may or may not be "recognized" as text due to the noise level and position, but they did draw more attention than controls (see Figure 35b).

Furthermore, it must be pointed out that the unconstrained texts could be considered as object-scene inconsistencies (specifically, syntactic violations and maybe semantic violations) since they were placed in unexpected locations in other scenes. Scene inconsistencies have been a highly debated issue, and previous studies either found them to influence initial eye movements (e.g., Loftus & Mackworth, 1978; Becker, Pashler, & Lubin, 2007; Bonitz & Gordon, 2008; Underwood & Foulsham, 2006; Underwood, Humphreys, & Cross, 2007; Underwood, Templeman, Lamming, & Foulsham, 2008) or failed to obtain evidence for such early detection (e.g., Gareze & Findlay, 2007; Rayner, Castelhano, & Yang, 2009; Võ & Henderson, 2009; 2011).

Regardless of this debate, it is clear that at least in some instances, a text placed in an unexpected location, e.g., floating in mid-air, captures attention, which may be due to its specific visual features or its unusual placement. The latter case would also apply to any non-text object placed in the same way. To resolve the potential issue of unusual placement of texts that arose in this experiment, in Experiment 4 we placed both texts and line drawings of the objects described by the texts in unexpected locations.

Experiment 4: Unconstrained Texts and Line Drawings

We placed an item-pair - a text and a drawing - in unexpected locations in a scene. If the text were found to attract more attention than the drawings, it would confirm the contribution of specific visual features of texts to their attractiveness. Texts and drawings were placed either in front of homogeneous or inhomogeneous *backgrounds*. We expected to observe similar results to the ones found in Experiment 3, that is, the attraction of visual features of texts being degraded by noise. In addition to comparing texts and drawings, we compared two *text-types,* namely texts (regular words) and their scrambled versions (i.e., all letters of the word being randomly rearranged in such a way that they did not form another English word), in order to test if higher-level processing, such as semantics, influences the attraction of attention.

Method

Participants. Twelve students from the University of Massachusetts at Boston participated. All were students with normal or corrected-to-normal vision and between 19 to 40 years old. Each participant received 10 dollars for a half-hour session.

Apparatus. The apparatus was the same as in Experiment 3.

Stimuli. Two hundred new natural-scene images, which were not used in

Experiments 1 to 3, were selected from the LabelMe dataset. Eighty out of these images were randomly selected to be superimposed with one item-pair each. The other 120 images were presented without any modification. There were 4 versions of the 80 superimposed images, resulting in 320 images for a counterbalanced design (i.e., one viewer only saw one of the 4 versions of the stimuli). Each observer viewed 80 item-pairs (cases). Figure 36 shows an example of all four versions of the same stimulus with items drawn on homogeneous background. For the placement of texts and line drawings, two different items (items A and B) were chosen for each scene, and their addition to the scene was performed in four different versions: either (1) a word describing item A (e.g., "sled" as shown in Table 26) and a drawing of item B, (2) a word describing item B (e.g., "yoyo") and a drawing of item A, (3) a scrambled version of a word describing item A (e.g., "dsle") and a drawing of item B, and (4) a scrambled version of a word describing item B (e.g., "yyoo") and a drawing of item A. The length of regular and scrambled words ranged between 3 and 11 letters (average: 6 letters). The eccentricity of the text or the drawing was randomly assigned and varied between 200 and 320 pixels (average: 253 pixels). The minimum polar angle, measured from the screen center, between the text and the drawing in each image was set to 60 degrees to avoid crowding of the artificial items. All texts and drawings were resized to cover approximately 2600 pixels. Table 27 shows the characteristics of texts and drawings in front of homogeneous (H) and inhomogeneous backgrounds (INH).

Table 26: Examples of texts (words and scrambled words) and object drawings used in Experiment 4.

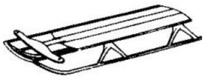
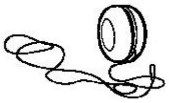
	Item A	Item B
Word (Scrambled Word)	sled (dsle)	yoyo (yyoo)
Object Drawing		

Table 27: Average characteristics of texts and drawings in Experiment 4.

	Size	Ecc.	Sal.	LumC.	EdgeC.
H					
Texts	2699 (3.00)	262 (8.73)	0.21	36.75	0.66
Drawings	2652 (2.95)	262 (8.73)	0.23	38.26	0.64
INH					
Texts	2700 (3.00)	258 (8.60)	0.32	51.64	0.78
Drawings	2652 (2.95)	258 (8.60)	0.33	52.15	0.79

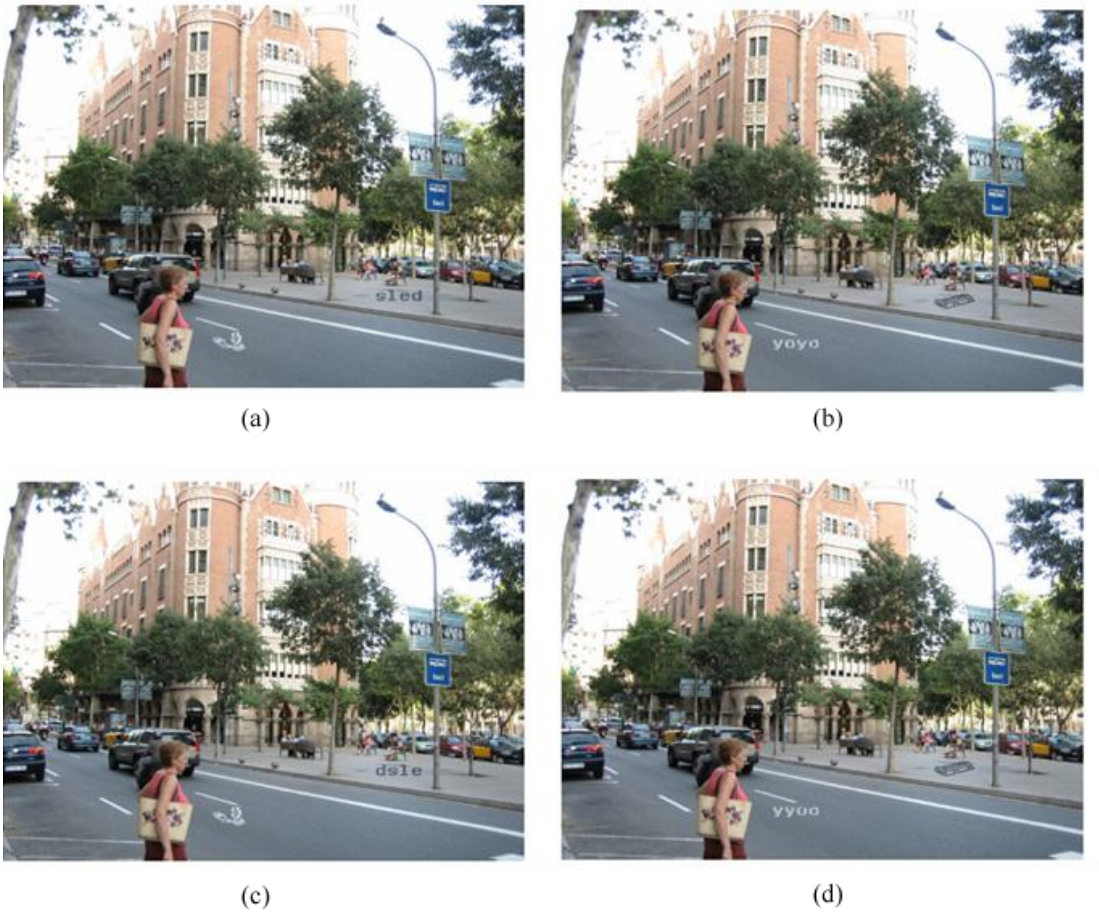


Figure 36: An example of the 4 stimulus versions of stimuli used in Experimnt 4, with words and drawings on homogeneous background. (a) Word of Item A (sled) vs. drawing of Item B, (b) word of Item B (yoyo) vs. drawing of Item A, (c) scrambled word of Item A (dsle) vs. drawing of Item B, and (d) scrambled word of Item B (yyoo) vs. drawing of Item A.

Procedure. Equal numbers of subjects viewed stimuli from conditions 1, 2, 3, and 4 in a counter-balanced design (described above), and each stimulus was presented for 5 seconds. The software “Eyetrack” developed by Jeffrey D. Kinsey, David J. Stracuzzi, and Chuck Clifton, University of Massachusetts Amherst, was used for recording eye movements. This software provides an easy-to-use interface for between-subject designs and has been widely used in the community of eye-movement researchers. Other settings were identical to Experiments 2 and 3.

Analysis. Fixation probability, minimum fixation distance, and first acquisition time were examined using a within-subject three-way ANOVA including item-type (texts vs. drawings), text-type (regular vs. scrambled), and background (homogeneous vs. inhomogeneous). There were 20 cases per condition. The fixation probability ANOVA served as the main analysis, while the ANOVAs for minimum fixation distance and first acquisition time were considered additional analyses. One participant was excluded from the analysis of first acquisition time since his fixation probability of drawings was 0 in one condition.

Results and Discussion

For fixation probability, the main effects of item-type and text-type did not reach significance, all Fs(1; 11) < 2.48, ps > 0.1, but the main effect of background did, F(1; 11) = 83.85, p < 0.001. Fixation probability was higher in front of homogeneous background than inhomogeneous background. All interactions among item-type, text-type, and background failed to reach significance, all Fs(1; 11) < 0.59, ps > 0.46. These results suggest that both texts and drawings drew more attention when they were presented on a clear background than when they were degraded by an inhomogeneous background.

For minimum fixation distance, a three-way ANOVA yielded main effects for item-type and background, both Fs(1; 11) > 33.17, ps < 0.001, but not for text-type, F(1; 11) = 0.35, p = 0.57. All interactions among item-type, text-type, and background were non-significant, Fs(1; 11) < 3.08, ps > 0.11. Minimum fixation distance was shorter for texts than drawings, and it was also shorter for homogeneous background than inhomogeneous background.

The results of the first acquisition time again demonstrated significant main effects of item-type and background, both Fs(1; 10) > 13.96, ps < 0.01, but not for text-type F(1; 10) = 1.42, p = 0.26. The interactions among item-type, text-type, and background were not significant, Fs(1; 10) < 2.56, p > 0.14, except for a marginal interaction between item-type and text-type, F(1; 10) = 3.60, p = 0.09. Surprisingly, items in front of inhomogeneous background seemed to receive fixations earlier than those in front of homogeneous background. It should be noted, however, that first acquisition time only accounted for items being fixated. When the background was homogeneous, the average fixation probability was over 0.55. In contrast, the average fixation probability was only around 0.35 when items were in front of inhomogeneous background. Here we analyze first acquisition time separately for homogeneous and inhomogeneous background because the fixation probabilities in these conditions were incompatible. For homogeneous background, a two-way ANOVA yielded a significant main effect of item-type, F(1; 10) = 7.61, $p < 0.05$, but not for text-type nor the interaction, both Fs(1; 10) < 2.50, ps > 0.15. For inhomogeneous background, there were no significant main effects of item-type and text-type, nor a significant interaction, all Fs(1; 10) < 0.24, $p > 0.62$. The results indicated that first acquisition time was shorter for texts than for drawings when the background was homogeneous, but no effect was found for inhomogeneous background. The averages and standard deviations of fixation probability, minimum fixation distance, and first acquisition time are shown in Figure 37.

The results of minimum fixation distance and first acquisition time were consistent with regard to texts receiving more attention than drawings, suggesting that the specific visual features of texts cause their attractiveness advantage. By definition, the scrambled words in Experiment 4 were no dictionary words, but it is important to note that their word length was controlled compared to their paired (regular) words. We did not find statistical differences between words and scrambled words in any of the measures, Fs(1; 11) < 2.48, ps > 0.1. These data suggest that the attention-capturing features of texts are operating at a low level so that the attraction of attention does not seem to depend on whether a word carries meaning or not.

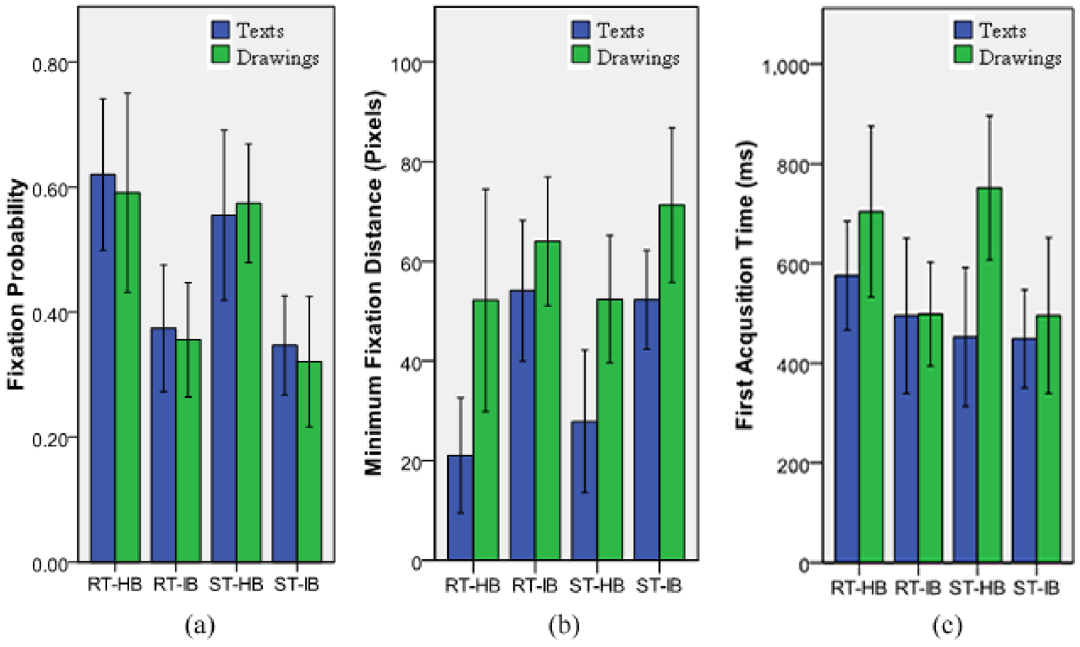


Figure 37: Results of Experiment 4 for texts and drawings. (a) Fixation probability, (b) minimum fixation distance, and (c) first acquisition time (RT: regular text, ST: scrambled text, HB: homogeneous background, and IB: inhomogeneous background).

The results of Experiment 4 confirmed that texts are more attractive than non-texts. Both words and scrambled words were found more attractive than line drawings depicting the corresponding objects. Because words and scrambled words yielded similar attractiveness results, the attraction of attention by texts seems to be caused by low-level visual features, not high-level semantics. This result raises important questions: Are these low-level features, such as the regular spacing and similarity of characters, specific to the observer's native writing system? Does a simple image transformation such as rotation by 180 degrees preserve their attractiveness? These questions were addressed in Experiment 5.

Experiment 5: Upside-Down English and Chinese Texts

To study the influence of the observers' familiarity with their native writing system, we carried out a further experiment by placing texts in Experiment 1 upside-down or replacing them with Chinese texts. These stimuli were presented to English speakers. The rationale for using upside-down English texts was to keep the low-level features such as regular spacing and similarity of letters but reduce possible influences of higher-level processing such as meaning. Chinese texts were chosen because they are visually dissimilar to texts in the English language and other alphabetic writing systems. Our hypothesis is that subjects may have developed specific “text detectors” for their native writing system during everyday life so that their attention would be biased toward words in that writing system.

After the conclusion of this experiment, we also received an opportunity to test native Chinese speakers. Since we found that turning texts upside-down did not affect attentional capture for English speakers, we decided to use exactly the same materials for the Chinese subjects without turning the Chinese texts upside-down for better comparability of results between the subject groups.

Method

Participants. In the group of non-Chinese English speakers, an additional 14 students from the University of Massachusetts at Boston participated in this experiment. All of them were native speakers of English, and none of them had learnt any Chinese or had participated in Experiments 1 to 4. For the group of Chinese speakers, 16 native speakers of Chinese were recruited at China Medical University, Taiwan. Each participant received 10 US dollars or 100 Taiwan dollars, respectively, for participation in a half-hour session. All had normal or corrected-to-normal vision.

Apparatus. Eye movements were recorded using SR Research EyeLink 1000

Remote systems both at the University of Massachusetts at Boston and at China Medical University, Taiwan. Other settings were the same as in Experiments 2 and 3.

Stimuli. As shown in Figure 38, the original texts from Experiment 1 were either rotated by 180 degrees or replaced by Chinese texts. Figure 38a illustrates C1, in which half of the original texts were rotated and the other half was replaced with Chinese texts. In C2, as demonstrated in Figure 38b, the upside-down texts in C1 were replaced with Chinese texts, and the Chinese texts in C1 were replaced with the original, but upside-down, English texts. Table 28 shows the characteristics of the upside-down and Chinese texts in C1 and C2. The characteristics of all upside-down and Chinese texts in C1 and C2 were very similar to those of the original texts in Experiment 1.

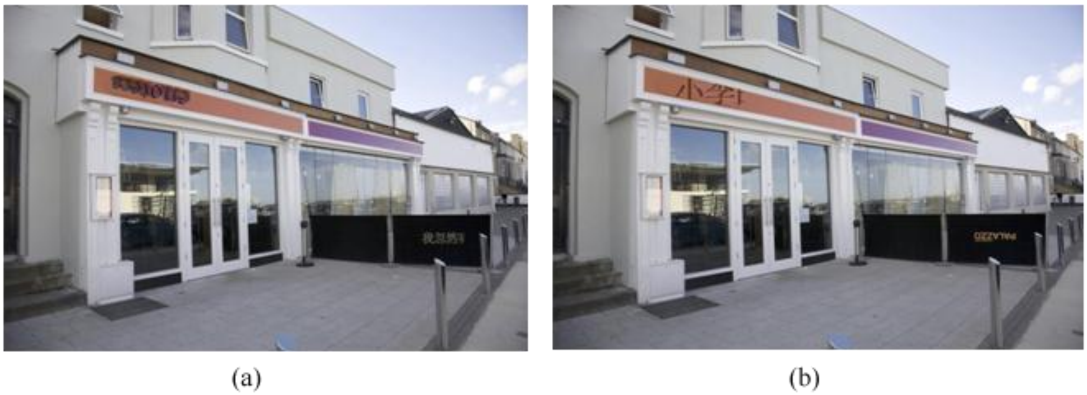


Figure 38: Example of upside-down and Chinese texts used in Experiment 5. (a) Version C1, in which half of the original texts were rotated and the other half was replaced with Chinese texts. (b) Version C2, in which the upside-down texts in C1 were replaced with Chinese texts, and the Chinese texts in C1 were replaced with upside-down texts.

Table 28: Average characteristics of upside-down and Chinese texts in each condition.

Experiment 5	Size	Ecc.	Sal.	LumC.
Upside-Down Text C1	2227 (2.47)	273 (9.10)	0.43	38
Chinese Text C2	2255 (2.50)	273 (9.10)	0.42	37
Upside-Down Text C2	3003 (3.34)	292 (9.73)	0.40	38
Chinese Text C1	2996 (3.33)	292 (9.73)	0.39	37

Procedure. The procedure was identical to Experiments 2 and 3 except that half of the subjects viewed condition 1 (C1) stimuli and the others viewed condition 2 (C2) stimuli in a between-subject counter-balanced design (described below). The same Eyetrack software as in Experiment 4 was used for recording eye movements.

Analysis. The analyses were identical to Experiments 2 and 3. Similar to Experiments 1 to 4, three-second viewing durations were analyzed for each trial. For English speakers, 7 subjects viewed C1 and 7 subjects viewed C2, and those data were combined so that upside-down English text and Chinese text for each item were viewed in a between-subject counter-balanced design. The same analysis was performed for Chinese speakers.

Results and Discussion

For English speakers, as shown in Figure 39, fixation probability was higher for upside-down texts than for Chinese texts, $t(13) = 5.62$, $p < 0.001$. This result suggests that upside-down English texts attract English speakers' attention more strongly than Chinese texts do. This trend is consistent with the results of minimum fixation distance, which was slightly shorter for upside-down texts (83.69) than for Chinese texts (88.16), but the difference failed to reach significance level, $t(13) =$

1.63, $p > 0.1$. A between-experiment comparison revealed that turning texts upside-down did not lead to any changes in their attraction of attention (see General Discussion for between-experiment analyses).

For Chinese speakers, the results were reversed as compared to English speakers; fixation probability was lower for upside-down English texts than for Chinese texts, $t(15) = 3.67$, $p < 0.01$. Minimum fixation distance was shorter for Chinese texts than for upside-down English texts, $t(15) = 4.46$, $p < 0.01$.

In the comparison between English and Chinese speakers, we found that Chinese texts were fixated equally often by both groups, but the upside-down texts were fixated more often by English speakers than by Chinese speakers. In other words, only the English speakers were biased toward their own native language. One possibility is that other factors played a role, such as expected locations, e.g., Chinese speakers might expect texts on vertical rather than horizontal signs given that most stimulus images were taken in North America and Europe. Nevertheless, based on the results of English speakers, Experiment 5 suggests that attraction of attention depends to some extent on the observer's familiarity with the writing system and language. The reason might be that English viewers have developed stronger "text detectors" for English texts during everyday life. The results may support the implication suggested in Cerf et al. (2009) that the allocation of attention to text is developed through learning.

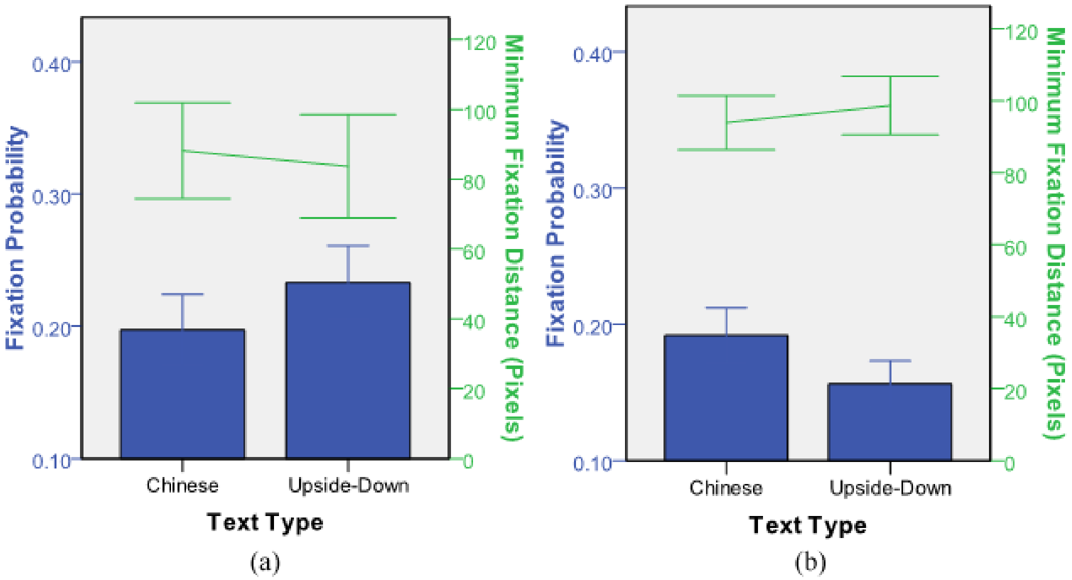


Figure 39: Fixation probability and minimum fixation distance of Chinese and upside-down English texts for (a) English readers and (b) Chinese readers.

General Discussion

In Experiment 1, we found that text objects were more attractive than non-text objects and control regions of similar size, eccentricity, saliency, and luminance contrast. Since we controlled for the typical saliency computed by color, intensity, orientation, and contrast, the results might be caused by high-level features (expected locations), special visual features of text, or both. Experiment 2 further investigated the attraction of attention by high-level features, and the results suggested that eye fixations were influenced by expected locations that might be assumed to be more informative. This finding has important implications for our understanding of attention in real-world scenes. First, it supports the concept of “contextual guidance” found by Torralba et al. (2006) and the influence of expected locations on visual attention as pointed out by Eckstein, et al. (2006). Second, and most importantly, it demonstrates that this factor does not only apply to search tasks but that expected locations play a role even in a free viewing task. By presenting the unique visual features of text in unexpected locations and in both fully visible and degraded variants, the results of Experiment 3 indicated that the specific visual features of texts were superior to features typically associated with saliency in their ability to attract attention, and their influence on attention was reduced by the noise caused by inhomogeneous background. However, the results obtained in Experiment 3 might also have been caused by the replacement of texts inducing oddness through semantic or syntactic violation. Experiment 4 provided convergent evidence for the contribution of the specific visual features to text attractiveness by placing texts and object drawings in unexpected locations and still finding stronger attentional capture by texts. In addition, Experiment 4 indicated that this capture might be caused by low-level visual features rather than high-level semantics since words and scrambled words yielded similar results. Experiment 5 further investigated how familiarity influences the attraction of attention by texts by presenting upside-down English and upright Chinese texts to native English and Chinese speakers. The results showed that viewers were biased toward their native language, which indicates that familiarity affects the allocation of attention. We conclude that both low-level specific visual features of texts and, to a lesser extent, high-level features (expected locations) contribute to the ability of texts to attract a disproportionate amount of visual attention in real-world scenes.

The results obtained from Experiment 1 might serve as a baseline for other experiments. In Experiment 2, fixation probability for erased texts (mean: 0.07) dropped in comparison to text objects in Experiment 1 (mean: 0.18), F(1; 28) = 35.82, $p < 0.001$, for a between-subject ANOVA. Minimum fixation distance was significantly longer for erased texts in Experiment 2 (mean: 111.98) than for texts in Experiment 1 (mean: 89.93), F(1; 28) = 10.53, $p < 0.01$. This result might be caused by the reduction of saliency and luminance contrast that accompanied the erasure of text. In Experiment 3, Fixation probability of the unconstrained texts in front of homogeneous background was not statistically different from that of texts in

Experiment 1 located in expected positions (both means: 0.18), F(1; 14) = 0.01, p > 0.9. This finding suggests that the specific text features might cause stronger attraction than expected locations. For Experiment 5, it is interesting to point out that the fixation probability of viewers' non-native language stimuli was considerably high (0.20 for upside-down texts viewed by Chinese speaker and 0.16 for upside-down texts viewed by Chinese speaker) compared to the text objects in Experiment 1 (0.18). This finding might imply that there are cross-language features of texts that capture attention, regardless whether the texts carry meaning or not. Moreover, turning English texts upside-down does not seem to significantly reduce their capture of English speakers' attention, which provides further evidence for the dominance of low-level factors in attracting attention to texts. However, those implications from between-experiment comparisons need to be verified in further well-controlled experiments, for example, an experiment containing regular, erased, upside-down texts in a between-subject counter-balanced design. Furthermore, to follow up on Experiment 2, another experiment could be conducted by erasing non-text regions by filling them with a background color, and then comparing them in terms of their attentional capture to text-removal regions. Both cases in such design cause strong reduction of luminance contrasts, but only text-removal regions occupy expected locations for texts. However, such investigations are beyond the scope of the current study and should be part of a follow-up study.

The free viewing task seems to be less constrained as compared to visual search or memorization tasks. Search and memorization tasks require specific top-down control of attention that might dominate task performance and therefore lead to different results from those obtained in the present study. However, during free viewing tasks, observers might attempt to retrieve as much information as possible, including deliberately looking for texts in order to make the scene more interpretable and contribute to its understanding and memorization. Therefore, although the task was free viewing and we included text-absent images in all experiments, we cannot rule out the possibility that observers may actually perform text searching and memorizing.

It would be interesting to see how texts are "read" in real-world scenes. In our previous study (Wang, Hwang, and Pomplun, 2010), fixation durations were found to be influenced by object size, frequency, and predictability, and we suggested that the recognition of objects in scene viewing shares some characteristics with the recognition of words in reading. It is important to analyze the underlying factors affecting processing time of texts in real-world scenes and compare the results to existing text reading studies (see Rayner, 2009).

There are other factors, i.e., scene context and scene syntax, which might affect expected locations. For instance, Torralba et al. (2006) developed a computational model of "contextual guidance" according to global scene statistics. Furthermore, Hwang, Wang, and Pomplun (2011) proposed "semantic guidance" during scene viewing which leads to a tendency towards gaze transitions between semantically similar objects in the scene. It was also found that "object dependency" (i.e., the statistical contingencies between objects, such as between a plate and silverware) can help viewers to predict the location of other objects from a given object or scene (Oliva & Torralba, 2007). For a better understanding of the attentional bias toward texts, it may thus be important to further extract the object dependency between texts and other objects from an image dataset such as LabelMe. Using the concepts of contextual guidance, semantic guidance, and object dependency, a computational model for human text detection could be developed.

There are many text-like patterns such as windows, fences, or brick walls that are easily misidentified as texts by artificial text detectors (see Ye, Jiao, Huang, & Yu, 2007). Furthermore, in Experiment 5 we found that English and Chinese-speaking viewers possess different preferences for the attraction of their attention to texts.

Future research could study the influences of specific visual features of texts to human viewers, using the analysis of eye movements. For example, such experiments could test the contribution of individual features of texts, e.g., orientations or arrangements of letters and strokes, to low-level attraction of human viewers' attention. Furthermore, it might be useful to further investigate the difference of special features between English and Chinese texts, as the results are potentially important for developing more efficient and general text detection algorithms.

CHAPTER 9

VISUAL ATTENTION IS ATTRACTED BY TEXT FEATURES EVEN IN SCENES WITHOUT TEXT

As we have seen in the preceding chapter, viewers' attention is disproportionately attracted by texts, especially by the ones they are familiar with. One possible reason is that viewers have *developed* a "text detector" in their visual system to bias their attention toward some specific text features. One way to verify this hypothesis is to add a text detector module to a visual attention model and test if the inclusion increases the model's ability to predict eye fixation positions. In a previous study, adding a module of manually-defined regions of texts was shown to improve the prediction of eye fixations in text-present images (Cerf et al., 2009). However, it is still unclear if viewers' attention is biased toward any non-text objects which share some features of texts, particularly in text-absent images. Therefore, an *automatic* text detector based on the recognition of specific text features is required to address this question.

Automatic text detection has been a hot topic in the fields of computer vision and pattern recognition for its practical applications. The special features of texts, e.g., the small variation of the stroke width (see Epshtein, Ofek, & Wexler, 2010; Jung, Liu, & Kim, 2009) or edge density have been used to develop text detectors. Although many text detection techniques, i.e., texture-based, region-based, and stroke-based methods, have been reported, many non-text objects, such as windows, fences, or brick walls, easily cause false alarms (see Ye, Jiao, Huang, & Yu, 2007, for a review).

For the presented study, we used an automatic text detector trained with Support Vector Machine (SVM) classifiers by specific text features including: contrast of strokes over background, width of strokes, joints of horizontal and vertical strokes, and stroke structure. The text detector was used to test whether it can improve the prediction of viewers' fixations.

In the present study, the stimuli and eye-movement data from Experiments 4 and 5 in Chapter 8 are re-analyzed. The goals of the present study are (1) to investigate the contribution of the automatic text detector to the prediction of eye fixations in real-world scenes, and (2) to verify the hypothesis that viewers' text detection skills are “trained” through exposure to language and affect attentional control even in text-absent scenes.

Experiment 1: Unconstrained Texts

We superimposed unconstrained texts onto real-world scenes, i.e., placed them in unexpected locations, in front of either homogeneous background, i.e., in regions with the lowest luminance contrast in the image before placing the text parts, or inhomogeneous background, i.e., those areas with the highest luminance contrast, and found that texts attracted more attention than non-text objects. This dataset is chosen for re-analysis in the present study since the stimuli contain both text-present and text-absent images. Two models, both including saliency and center-bias maps (channels), but one with and one without text-detector map are compared in order to determine whether the inclusion of the text detector improves the prediction of fixations, particularly in text-absent images.

Method

Participants, apparatus, stimuli, procedure are identical to Experiment 4 in Chapter 8.

Analysis. Two eye movement measures were taken: correlation (R) and Receiver Operating Characteristic (ROC). The Pearson correlation coefficient R between two maps is computed according to sampling points taken every 10 pixels along the x and y axes, and then the correlation coefficient between saliency/center-bias/text-detector and attentional maps (described below) are obtained. An example of a stimulus image and its attention, saliency, center-bias, and text-detector maps is shown in Figure 40. The computation of the ROC measure is described in Hwang, Higgins & Pomplun (2009). If a map had higher correlation or ROC values with regard to the subjects' fixations, the map was considered a better predictor of visual attention. The chance level is 0.5 for ROC and 0 for R.

Saliency was calculated by the freely available computer software "Saliency Map Algorithm" using the standard Itti, Koch, and Niebur (1998) saliency map based on color, intensity, orientation, and contrast. A center-bias map was obtained using a two-dimensional Gaussian distribution at the center of the screen with 3 degrees of visual angle (90 pixels in our experiment setting). The text-detector maps were computed using the automatic text detector which analyzes features such as variation of edge width and edge density.

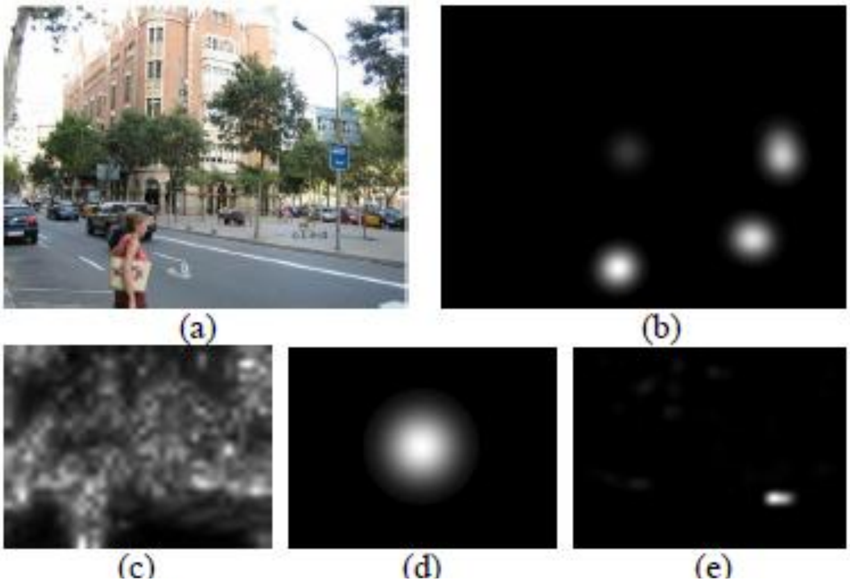


Figure 40. An example of (a) stimulus image, (b) attention (3-second viewing) (c) saliency, (d) center-bias, and (e) text-detector maps.

For the attentional map, we excluded the initial center fixation and included all other fixations within a given viewing duration. The attentional map was built according to each fixation in an image by a two-dimensional Gaussian distribution centered at the fixation point, where the standard deviation was one degree of visual angle to approximate the size of the human fovea. Then we simply summed up these Gaussian distributions for fixations weighted by their durations (see Pomplun, Ritter, & Velichkovsky, 1996).

We computed the attentional maps for each image inspected by each viewer for the initial 1.5, 2, …, 5 seconds. The averages of correlations and ROC values for each viewer were calculated for all, text-present, text-absent, text in front of homogeneous (H-BG), and text in front of inhomogeneous backgrounds (INH-BG) images, and an ANOVA and paired t-tests were performed to analyze the differences between these values.

Results and Discussion

Models with and without Text-Detector Maps. The average R and ROC values of all 12 viewers are shown in Table 30. Text-detector maps overlap attentional maps the best when the images contain text in front of homogeneous background, and the worst in text-absent images. These results are consistent with the finding by Judd et al. (2009) that object detectors by themselves do not predict attention well when the objects are absent and therefore should be used in conjunction with other features.

One-way ANOVAs with the factor “predictor” showed that the performances of Sali, Cen, TextDet, SC, and SCT maps differed significantly in all, text-present, H-BG, INH-BG, and text-absent images for R, all Fs(4; 55) > 3.64, ps < .05, and ROC, all Fs(4; 55) > 11.17, ps < .01. SC (without text-detector) obtained significantly lower measures than SCT (with text-detector maps) for all, text-present, H-BG, INH-BG, and text-absent images for R, all ts(11) > 3.93, ps < .01, and ROC, all ts(11) > 7.68, ps < .001. The results indicate that the text detector improved the prediction of viewers' visual attention. It is interesting to see that the SCT obtained higher R and ROC than the SC even in text-absent images. One plausible explanation is that some non-objects containing text-like features catch a disproportionate amount of attention.

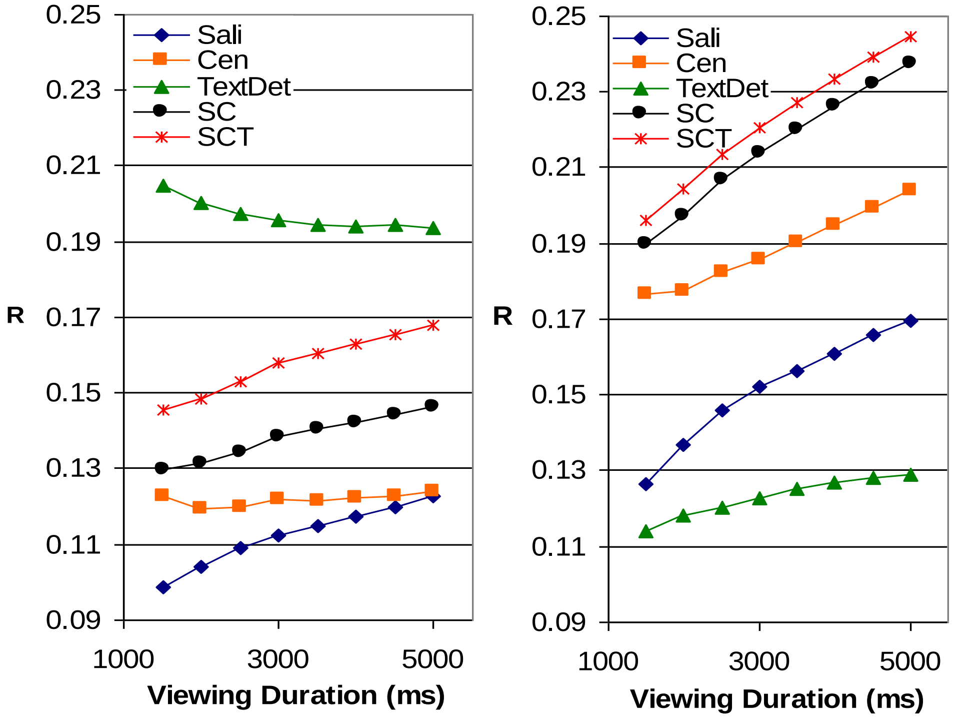
Text-Present vs. Text-Absent and H-BG vs. INH-BG Images. The five predictors were analyzed in one-way ANOVAs with the factor "image type," and the results demonstrate that both R and ROC values significantly differed in all, text-present, text-absent, H-BG, and INH-BG images, all Fs(4; 55) > 4.91, ps < .01, and all Fs(4; 55) > 4.72, ps < .01, respectively, except ROC for Cen, F(4; 55) = 0.92, p > .4. The text detector (TextDet) performed better for text-present images than text-absent ones with regard to R, t(11) = 10.67, p < .001 as well as ROC, t(11) = 5.66, p < .001. Homogeneous background images obtained higher values than inhomogeneous background images for both R, t(11) = 7.31, p < .001, and ROC, t(11) = 3.94, p < .01.

Table 29: The average R and ROC of saliency (Sali), center-bias (Center), text-detector (TextDet), saliency combined with center-bias (SC), and all combined (SCT) maps as predictors of the attentional maps for 3-second viewing. H-BG represents images in front of homogeneous background, and INH-BG represents images on inhomogeneous background.

	Sali	Cen	TextDet	SC	SCT
R -All	0.14	0.16	0.15	0.18	0.20
Text-Present	0.11	0.12	0.20	0.14	0.16
H-BG	0.09	0.10	0.24	0.10	0.12
INH-BG	0.14	0.15	0.15	0.17	0.19
Text-Absent	0.15	0.19	0.12	0.21	0.22
ROC - All	0.65	0.63	0.63	0.69	0.72
Text-Present	0.61	0.61	0.66	0.64	0.70
H-BG	0.55	0.60	0.67	0.58	0.67
INH-BG	0.67	0.62	0.64	0.70	0.72
Text-Absent	0.67	0.64	0.62	0.72	0.73

Visual Attention over Time. SCT outperformed SC (without text detector) for all viewing durations for R and ROC in both text-present images, both ts(11) > 9.68, ps < .001, and text-absent ones, both ts(11) > 3.93, ps < .01. The difference between

SCT and SC was larger in text-present images than in text-absent ones. In text-present images, the R of TextDet initially dominated but decreased over time, while the R of Sali increased (see Figure 41a).. These data suggest that texts are typically detected early during the inspection process and receive sustained attention while the viewers are reading them, thereby elevating the occurrence of text features near fixation. Later in the process, viewers tended to be guided more strongly by saliency as defined by the Itti and Koch algorithm. In text-absent images, the R of Sali, Cen, and TextDet increased over time, indicating that the corresponding mechanisms became more important during the later – likely more focused and fine-grained (Unema, Pannasch, Joos, & Velichkovsky, 2005) – stages of inspection. Clearly, Sali and Cen played more important roles when texts are absent.



(a)

(b)

Figure 41: Correlations for 1.5-, 2-, …, and 5-second viewing of (a) text-present and (b) text-absent images.

Experiment 2: English vs. Chinese Texts and Native Speakers

In Experiment 1, we showed that the addition of a text-detector map to saliency and center-bias maps makes the model a better predictor of viewers' visual attention. Our hypothesis is that viewers have developed a "text detector" because they are exposed to texts everyday and become sensitive to text-patterns. Wang and Pomplun (2012) found that native speakers of English and Chinese-speakers were both attracted by English and Chinese texts in real-world scenes but were attracted more strongly by the texts of their native languages. The reason might be that English and Chinese texts share some common features, such as the histogram of edge width, but also contain their unique features, e.g., Chinese texts usually contain vertical, horizontal, and diagonal strokes but fewer "curves" (such as in "O" or "G" in English). In Experiment 2, the dataset in Wang and Pomplun (submitted) was reanalyzed and our expectation was that the text detector (Lu, submitted) designed for English texts will perform better prediction of gaze fixations for English-speaking viewers than for Chinese-speaking ones.

Method

Participants, apparatus, stimuli, procedure are identical to Experiment 5 in Chapter 8.

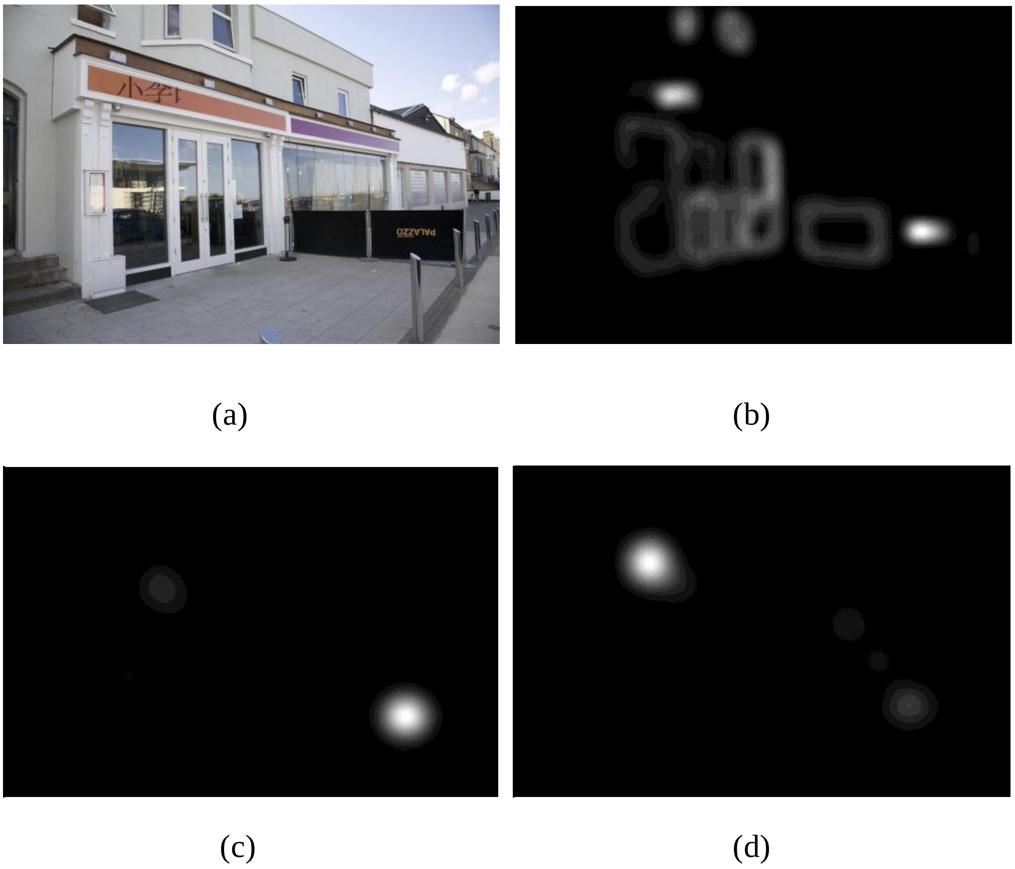
Analysis. The analyses were identical to Experiment 1.

Results and Discussion

Models with and without Text-Detector Maps. The average R and ROC of all 14 English-speaking and 16 Chinese-speaking viewers are shown in Table 30. For

English-speaking viewers, one-way ANOVAs showed that the Sali, Cen, TextDet, SC, and SCT maps performed differently in all, text-present, and text-absent images for R, all $Fs(4; 65) > 8.47$, $ps < .01$, and for ROC, all $Fs(4; 65) > 53.78$, $ps < .001$. SCT predicted attentional maps better than SC in all, text-present, and text-absent images for R, all $ts(13) > 3.49$, $ps < .01$, and ROC, all $ts(13) > 6.61$, $ps < .001$. For Chinese-speaking viewers, similar results were obtained - the performances of Sali, Cen, TextDet, SC, and SCT maps significantly differed for both R, all $Fs(4; 75) > 33.91$, $ps < .001$, and ROC, all $Fs(4; 75) > 22.86$, $ps < .001$. SCT yielded better prediction of attentional maps than SC for both R, all $ts(15) > 4.85$, $ps < .001$, and ROC, all $ts(15) > 5.29$, $ps < .001$. The results of SCT are consistent with Experiment 1 in that the text detector improved the prediction of viewers' visual attention, even in text-absent images.

Text-Present vs. Text-Absent Images. For English-speaking viewers, TextDet performed better in text-present images than in text-absent ones for both R, $t(13) = 6.41$, $p < .001$, and ROC, $t(13) = 5.58$, $p < .001$. For Chinese-speaking viewers, similar results were found: text-present images obtained higher R and ROC than text-absent ones, $t(15) = 4.97$, $p < .001$, and $t(15) = 7.35$, $p < .001$, respectively.



(a) (b)

(c) (d)

Figure 42: An example of (a) stimulus image, (b) text-detector map, (c), attentional map of an English-speaking viewer (5-second viewing), and (d) attentional map of a Chinese-speaking viewer (5-second viewing).

English vs. Chinese-Speaking Viewers. As shown in Figure 43, TextDet predicted English-speaking viewers' attention better than Chinese-speaking viewers' attention for all viewing durations in both text-present images, t(7) = 23.12, p < .001, and text-absent images, t(7) = 5.38, p < .01. These results indicate that the text detector that was designed for English texts performed better at predicting the allocation of attention for English-speaking viewers than for Chinese-speaking ones.

General Discussion

In Experiment 1, we found that adding a text detector to an attention model improved its prediction of viewers' visual attention, even in text-absent images. Our results suggest that non-text objects whose features resemble those of texts (such as high spatial frequency edges) catch a disproportionate share of attention. Based on the current data, it seems that the viewers' "biological text detectors" are somewhat similar to the artificial system and influence the viewers' distribution of attention when viewing real-world images. From a time-course analysis, it appears that the biological text detector influences the allocation of attention particularly strongly during later stages of image inspection when viewers are increasingly likely to attend to detailed local structures (see Unema et al., 2005) for semantic interpretation of perceived text.

Table 30: The average R and ROC of saliency (Sali), center-bias (Cen), text-detector (TextDet), saliency combined with center-bias (SC), and all combined (SCT) maps as predictors of attentional maps for 5-second viewing. En represents English-speaking viewers, and Ch means Chinese-speaking viewers.

	Sali	Cen	TextDet	SC	SCT
R (En)	0.17	0.17	0.14	0.20	0.21
Text-Present	0.15	0.16	0.16	0.19	0.21
Text-Absent	0.18	0.17	0.12	0.21	0.22
R (Ch)	0.17	0.16	0.12	0.19	0.20
Text-Present	0.15	0.15	0.14	0.18	0.19
Text-Absent	0.18	0.17	0.11	0.20	0.21
ROC (En)	0.69	0.61	0.60	0.72	0.73
Text-Present	0.68	0.62	0.63	0.71	0.73
Text-Absent	0.69	0.61	0.59	0.72	0.73
ROC (Ch)	0.68	0.60	0.60	0.70	0.71
Text-Present	0.67	0.61	0.62	0.69	0.71
Text-Absent	0.68	0.60	0.58	0.70	0.70

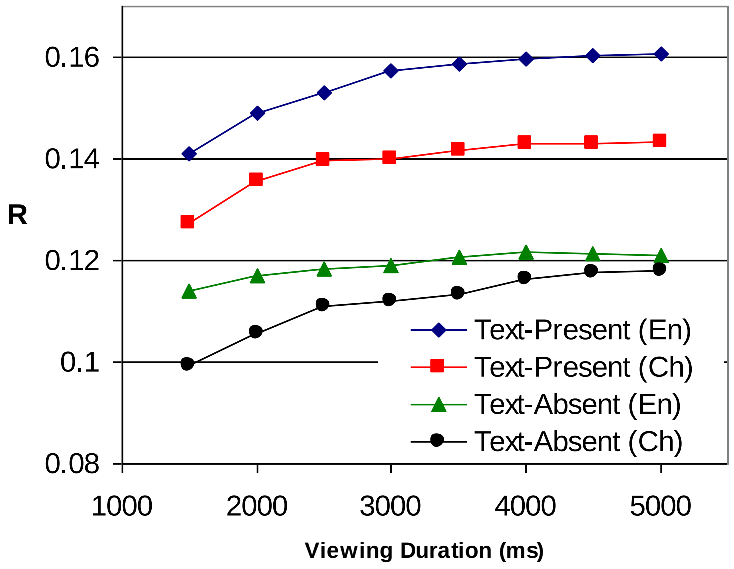


Figure 43. The R values of TextDet for 1.5-, 2-, …, and 5-second viewing of text-present and text-absent images by English-speaking (En) and Chinese-speaking (Ch) viewers.

Whereas the results of Experiment 1 could have been caused by the text detection algorithm being sensitive to visual features that generally attract attention, such as edge density, this interpretation becomes implausible given the results of Experiment 2. We found that the text detector designed for English texts predicted English-speaking viewers' attention better than Chinese-speaking viewers', supporting the hypothesis that viewers have developed a "text detector" that is sensitive to text patterns they are familiar with. It is interesting to see that the way we learn to read influences our allocation of visual attention in everyday life, even when there are no texts presented and we are not specifically looking for any texts.

While the present study has demonstrated the influence of language on visual attention in real-world scenes, further research needs to identify the visual features that underlie this effect. This could be achieved by using text detection algorithms for different writing systems and test their individual components as predictors of native and non-native speakers' attention in natural scenes. Besides a more comprehensive understanding of attentional control in humans, such studies may also result in technological advances. Human viewers can easily locate texts in natural scenes, performing clearly better than current text-detection techniques even when the texts are degraded by noise, rotated, distorted, or shown from unusual perspectives. Consequently, the results of this line of research, such as analyzing what features or local structures are actually learned by the biological text detector, might contribute to the development of more effective automatic text detectors, which could, for example, make a great difference to visually challenged people's lives.

CHAPTER 10

CONCLUSIONS

The results of this dissertation will broaden our understanding of cognitive processing during reading and real-world scene viewing. The most important outcomes of the current work are (1) the proposed computational models of using SVD and LSA, from visual encoding and word identification to contextual predictability in sentence and scene image processing, (2) the interface of linguistic and visual processing of words, objects, and scene texts, and (3) cross-linguistic investigations. Below I will discuss these findings and their potential for leading to important practical applications.

The Computational Models Using SVD and LSA

First of all, we found that that SVD is a powerful tool in determining what the most informative segments of Chinese characters are. Chapter 2 investigated how readers recognized Chinese characters that were degraded using SVD to identify the most important and least important segments. Reading was most impaired when subjects read sentences with the most important segments removed. On the other hand, reading was not impaired when the least important segments were removed, and it was only moderately impaired when randomly selected segments were removed. The outcomes of degraded character recognition may be adopted in the field of scene text detection since texts in real-world scenes are often partially occluded, shown in low resolution, or degraded by motion blur.

Subsequently, the two models based on LSA tackled the problem of subjective differences and ambiguity of transparency judgments. 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 based on a clustering analysis of morphological family members (e.g., “butterfingers” or “buttermilk” for “butter”). The proposed models account for polysemy of constituents and successfully predicted participants’ transparency ratings. They may thus explain differences between the processing of English and Chinese compounds.

For sentence processing, TP and LSA were employed to predict eye fixation times and to investigate the word predictability effect on early and late stage lexical processing. 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.

Furthermore, based on LSA as initial connection matrix weights, our LS model successfully represented the cognitive processes that are sensitive to semantic constraint. The connection matrix in the LS model can operationalize a variety of linguistic characteristics stored explicitly, or otherwise represented, in long-term memory. The proposed LS model is sensitive to both strong and subtle changes in contextual semantic constraint, and provides a means of investigating how language comprehension is affected by the activation of concepts in working memory.

The Interface of Linguistic and Visual Processing – Words, Objects, and Scene Texts

Taken together, during text reading, the duration of eye fixations was found to decrease with greater frequency and predictability of the currently fixated word. In Chapter 6, we extended these results to scene viewing by computing object frequency and predictability from both linguistic and visual scene analysis (LabelMe, Russell et al., 2008). LSA was applied to estimate predictability. In a scene-viewing experiment, we found that, for small objects, linguistics-based frequency, but not scene-based frequency, had effects on first fixation duration, gaze duration, and total time. Both linguistic and scene-based predictability affected total time. Similar to reading, fixation duration decreased with higher frequency and predictability. For large objects, we found the direction of effects to be the inverse of those found in reading studies. These results suggest that the recognition of small objects in scene viewing shares some characteristics with the recognition of words in reading.

Additionally, we used a novel interdisciplinary approach in Chapter 8, combining visual context research and linguistic research, and conducted two experiments of scene viewing and search to demonstrate semantic guidance of gaze transitions. Applying LSA to the object labels in LabelMe, we generated semantic saliency maps to predict the location of the next eye fixation. An ROC analysis revealed a tendency of observers to sequentially look at semantically similar objects during scene viewing and their attention being guided toward objects that are semantically similar to the target in a visual search task.

For scene texts, we conducted a series of experiments to investigate how texts attract visual attention. In Experiment 1, we found that scene texts were more attractive than non-text objects, and the results might be caused by high-level features (expected locations), special visual features of text, or both. Experiment 2 suggested that eye fixations were influenced by expected locations that might be assumed to be more informative. By presenting the unique visual features of text in unexpected locations and in both fully visible and degraded variants, the results of Experiment 3 indicated that the specific visual features of texts were superior to features typically associated with saliency in their ability to attract attention, and their influence on attention was reduced by the noise caused by inhomogeneous background. Experiment 4 provided convergent evidence for the contribution of the specific visual features to text attractiveness by placing texts and object drawings in unexpected locations and still finding stronger attentional capture by texts. In addition, Experiment 4 indicated that this capture might be caused by low-level visual features rather than high-level semantics since words and scrambled words yielded similar results.

Cross-Linguistic Investigations

This dissertation also demonstrated important findings regarding cross-linguistic investigations. We found that, similar to the letters in English words, not all strokes in Chinese characters are of equal importance to the word recognition process. Furthermore, word predictability tends to influence both English and Chinese sentence processing. During reading, our results from a computational perspective replicated prior research using experimental approaches to study Chinese reading, demonstrating that word predictability and word frequency influence how long readers fixate on words during reading.

Further cross-linguistic investigations in this dissertation also revealed interesting differences between the two languages. For the models of predicting semantic transparency, corroborating evidence from two different languages was presented by testing the stimuli used in prior compound word studies in English and Chinese. Model 1 compared the semantic similarity between a compound and each of its constituents, and Model 2 computed the semantic transparency of a constituent by its morphological families. The results suggested that Model 2 is in general a better approach than Model 1 for constituents of Chinese compounds, which may be due to the concept of a word not being as clearly defined in Chinese as in English, and Chinese readers possibly learning the polysemy of characters implicitly from polymorphemic words.

For the study of scene texts, English and Chinese speakers participating in the study viewed both upside-down English and normally oriented Chinese texts, which are visually dissimilar to texts in the English language and other alphabetic writing systems. Our results showed that viewers were biased toward the writing system of their native language, which indicates that familiarity affects the allocation of attention. We suggest that subjects may have developed specific "text detectors" for their native writing system during everyday life so that their attention would be biased toward words in that writing system.

Practical Applications

From our text detection studies, we conclude that both low-level specific visual features of texts and, to a lesser extent, high-level features (expected locations) contribute to the ability of texts to attract a disproportionate amount of visual attention in real-world scenes. When an automatic text detector was added to an attention model, the model showed improved prediction of viewers' visual attention, even in text-absent images. Our results suggest that non-text objects whose features resemble those of texts (such as high spatial frequency edges) catch a disproportionate share of attention. Based on the current data, it seems that the viewers' "biological text detectors" are somewhat similar to the artificial system and influence the viewers' distribution of attention when viewing real-world images. We also found that the text detector designed for English texts predicted English-speaking viewers' attention better than Chinese-speaking viewers', supporting the hypothesis that viewers have developed a "text detector" that is sensitive to text patterns they are familiar with. It is interesting to see that the way we learn to read influences our allocation of visual attention in everyday life, even when there are no texts presented and we are not specifically looking for any texts.

The results can be adopted to improve existing text detection algorithms. The state-of-the-art text detection accuracy in real-world photographs is only approximately 70% and the recognition accuracy even falls below 50% for a commonly used public dataset (see Jung, Liu, Kim, 2009; Lucas, Panaretos, Sosa, Tang, Wong, & Young, 2003; Epshtein, Ofek, & Wexler, 2010). The low text detection accuracy is due to (1) texts appearing in various sizes, colors, orientations, image contrasts and scene contexts and (2) frequently occurring image degradation caused by imaging artifacts such as blur, uneven illumination, and occlusion, and (3) some objects sharing similar features with texts, such as high contrast or consistent edge density, e.g., brick walls and window frames. However, human viewers can effortlessly overcome these obstacles. Therefore, studying how human viewers detect texts using experimental approaches, for example, by examining correlations between features and eye fixation data, may lead to more powerful automatic text detection algorithms..

There have been studies focusing on incorporating scene texts in robotics (Posner, Corke, & Newman, 2010) and assisted technologies for visually impaired or blind persons (Case, Suresh, Coates, & Ng, 2011; Yi & Tian, 2011). Text detection (spotting) is the first step to obtain the possible locations of text for further optical character recognition. Due to constraints with regard to computational resources and application features, a more effective text detection algorithm is required. Such an algorithm can help visually impaired persons, who can receive only limited hints of an object from its shape and material by touch and smell. Therefore, a practical application of a better guidance for visually impaired persons could be accomplished, for example, for reading a room number or a street sign, or identifying objects by spoken text labels, which will make a great difference to visually challenged people's lives in terms of reading and scene viewing.

REFERENCE LIST

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

Athanasiadis, T., Mylonas, P, Avrithis, Y. & Kollias, S. (2007). Semantic Image Segmentation and Object labeling, *IEEE Transactions on Circuits and Systems for Video Technology,* 17, 298-312.

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.

Baddeley, R. J. & Tatler, B. W. (2006). High frequency edges (but not contrast) predict where we fixate: A Bayesian system identification analysis. *Vision Research,* 46(18), 2824-2833.

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.

Balota, D. A., Pollatsek, A., & Rayner, K. (1985). The interaction of contextual constraints and parafoveal visual information in reading. *Cognitive Psychology,* 17, 364-390.

Becker, M. W., Pashler, H., & Lubin, J. (2007). Object-intrinsic oddities draw early saccades. *Journal of Experimental Psychology: Human Perception and Performance*, 33, 20–30.

Belke, E., Humphreys, G.W., Watson, D.G., Meyer, A.S. & Telling, A. (2008). Top-down effects of semantic knowledge in visual search are modulated by cognitive but not perceptual load. *Perception and Psychophysics*, 70, 1444-1458.

Berry, M.W., Dumais S.T., & Obrien G.W. (1995). Using linear algebra for intelligent information-retrieval. *SIAM Review*, 37, 573-595.

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

Biederman, I., Mezzanote, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14, 143–177.

Biederman, I. (1987). Recognition-by-Components: A Theory of Human Image Understanding. *Psychological Review*, 94(2), 115-147.

Bonitz, V. S., & Gordon, R. D. (2008). Attention to smoking-related and incongruous objects during scene viewing. *Acta Psychologica*, 129, 255–263.

Bosch, A., Munoz, X. & Marti, R. (2007) Review: Which is the best way to organize/classify images by content? *Image and Vision Computing*, 25, 778-791.

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.

Bruce, N. D. B., & Tsotsos, J. K. (2006). Saliency based on information maximization. *Advances in Neural Information Processing Systems*, 18, 155–162.

Buswell, G. T. (1935). *How people look at pictures*. Chicago: University of Chicago Press.

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.* London: Academic Press. 257-294.

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

Case, C., Suresh, B., Coates, A., & Ng, A. Y. (2011). Autonomous sign reading for semantic mapping. In *Proceedings of IEEE International Conference on Robotics and Automation (ICRA),* Shanghai, China, 3297-3303.

Castelhano, M. S., & Henderson, J. M. (2007). Initial scene representations facilitate eye movement guidance in visual search. Journal *of Experimental Psychology: Human Perception and Performance,* 33(4), 753-763.

Castelhano, M., Mack, M. L. & Henderson, J. M. (2009). Viewing task influences eye movement control during active scene perception. *Journal of Vision,* 9(3):6, 1-15.

Cerf, M., Cleary, D., Peters, R., Einhäuser, W., & Koch, C. (2007). Observers are consistent when rating image conspicuity. *Vision Research,* 47, 3052–3060.

Cerf, M., Frady, E. P., & Koch, C. (2009). Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision,* 9(12):10, 1–15.

Chan, H. C. (2008) Empirical comparison of image retrieval color similarity methods with human judgment. *Displays,* 29, 260-267.

Chen, A. Y. C., Corso, J. J., & Wang, L. (2008) HOPS: Efficient region labeling using Higher Order Proxy Neighborhoods. *The 18th International Conference on Pattern Recognition (ICPR),* 1-4.

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, 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, M. J., & Weekes, B. S. (2004). Effects of semantic radicals on Chinese character categorization and character decision. *Chinese Journal of Psychology, 46(2),* 179-195.

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, 4, 415-435.

Choi, F. 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.

Chun, M., M. & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology*, 36, 28-71.

Chun, M. M. & Phelps, E. A. (1999). Memory deficits for implicit contextual information in amnesic subjects with hippocampal damage. *Nature Neuroscience*. 2 (9):775-776.

Costello, F. J., & Keane, M. T. (2000). Efficient creativity: constraint-guided conceptual combination. *Cognitive Science*, 24(2), 299-349.

Craw, I. & Cameron, P. (1991). Parameterising images for recognition and reconstruction. In P. Mowforth (Ed.). *Proceedings of the British Machine Vision Conference*, Berlin: Springer Verlag.

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.

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

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Sciences,* 41, 391–407.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience,* 18, 193-222.

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.

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

Dumais, S. T., Furnas, G.W., Landauer, T. K., Deerwester, S.,&Harshman, R. (1988). Using latent semantic analysis to improve information retrieval. *Proceedings of SIGCHI Conference on Human Factors in Computing Systems*, 281–285.

Duchowski, A. T. (2002). A breadth-first survey of eye tracking applications. *Behavior Research Methods, Instruments, and Computers*, *34*, 455-470.

Eckstein, M. P., Dreseher, B. A., Shimozaki, S. S. (2006). Attentional cues in real world scenes, saccadic targeting, and Bayesian priors. *Psychological Science*. 17(11), 973-980.

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.

Einhäuser, W., & König, P. (2003). Does luminance-contrast contribute to a saliency map for overt visual attention? European Journal of Neuroscience, 17, 1089-1097.

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision*, 8(14):18, 1–26.

Elazary, L., & Itti, L. (2008). Interesting objects are visually salient. *Journal of Vision*, 8(3):3, 1–15.

Elden, L. (2007). *Matrix Methods in data Mining and Pattern Recognition.* Society of Industrial and Applied Mathematics, Cambridge.

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.

Engbert, R., Nuthmann, A., Richter, E., & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review*, 112, 777-813.

Epshtein, B., Ofek, E., Wexler, Y.: Detecting text in natural scenes with stroke width transform. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, San Francisco, USA, 2963-2970.

Evett, L. J., & Humphreys, G. W. (1981). The use of abstract graphemic information in lexical access. *Quarterly Journal of Experimental Psychology, 33A*, 325-350.

Fellbaum, C. (1998). *Wordnet: An Electronic Lexical Database*. Bradford Books.

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.

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

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

Felzenszwalb, P., McAllester, D., & Ramanan, D. (2008). A Discriminatively Trained, Multiscale, Deformable Part Model. *Computer Vision and Pattern Recognition (CVPR),* Anchorage, Alaska, USA, 1-8.

Findlay, J. M. (2004). Eye scanning and visual search. In J. M. Henderson & F. Ferreira (Eds.), *The Interface of Language, Vision, and Action: Eye Movements and the Visual World.* Psychology Press, 135–159.

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.

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.

Gareze, L., & Findlay, J. M. (2007). Absence of scene context effects in object detection and eye gaze capture. In R. van Gompel, M. Fischer, W. Murray, & R. W. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 537–562). Amsterdam: Elsevier.

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

Gough, P. B. (1972) One second of reading. In J. F. Kavanagh and I. G. Mattingly (Eds.), *Language by eye and by ear*. Cambridge, Mass: MIT Press.

Gollan T. H., Slattery T. J., Goldenberg D., Van Assche E., Duyck W., Rayner K. (2011). Frequency Drives Lexical Access in Reading but Not in Speaking: The Frequency-Lag Hypothesis. *Journal of Experimental Psychology: General*,140, 2, 186–209.

Gonzalez, R. C. & Woods, R. E. (2002). *Digital Image Processing*, 2nd edition, Upper Saddle River, NJ: Prentice Hall.

Grossberg, S., & Huang, T.-R. (2009). ARTSCENE: A neural system for natural scene classification. *Journal of Vision*, *9*(4):6, 1-19.

Harel, J., Koch, C., & Perona P. (2006). Graph-Based Visual Saliency, *Proceedings of Neural Information Processing Systems (NIPS)*.

Henderson, J. M., Weeks, P. A., & Hollingworth, A.(1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology: Human Perception and Performance*, 25, 210–228.

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, 7, 498–504.

Henderson, J. M., & Ferreira, F. (2004). Scene perception for psycholinguists. *In J. M. Henderson and F. Fer-reira (Eds.), The interface of language, vision, and action: Eye movements and the visual world* (pp. 1-58). New York: Psychology Press.

Houtkamp, R., & Roelfsema, P. R. (2009). Matching of visual input to only one item at any one time. *Psychological Research*, 73, 317-326.

Hubel, D. H. & Wiesel, T. N. (1962). Receptive fields, binocular interaction and ftunctional architecture in the cat's visual cortex. *Journal of Physiology*. 160, 106-154.

Hubel, D. H. & Wiesel, T. N. (1963). Receptive fields of cells in striate cortex of very young, visually inexperienced kittens. *Journal of Neurophysiology*. 26, 994-1002.

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.

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.

Huettig, F. & Altmann, G. T. M. (2006). Word meaning and the control of eye fixation: semantic competitor effects and visual world paradigm. *Cognition, 96*, B23-B32.

Hong, J.-F. & Huang C.-R.. 2006. Using Chinese Gigaword Corpus and Chinese Word Sketch in linguistic research. The 20th Pacific Asia Conference on Language, Information and Computation (PACLIC-20). Wu-Han: China Huazhong Normal University.

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.

Hwang, A. D., Higgins, E. C. & Pomplun, M. (2007). How Chromaticity Guides Visual Search in Real-World Scenes, *Proceedings of the 29th Annual Cognitive Science Society* (pp.371-378), Austin, TX: Cognitive Science Society.

Hwang, A. D., Higgins, E. C., & Pomplun, M. (2009). A model of top-down attentional control during visual search in complex scenes. *Journal of Vision*, 9(5), 1–18 (25).

Hwang, A. D., Wang, H. C., & Pomplun, M. (2011). Semantic guidance of eye movements in real-world scenes. *Vision Research*, 51, 1192-1205.

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.

Itti, L, Koch, C., & Niebur, E. (1998). A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20 (11): 1254-1259.

Itti, L., & Koch, C. (2001). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10-12), 1489-1506.

Jacobs, A.M. (1986). Eye movement control in visual search: How direct is visual span control? Perception & Psychaphysics, 39, 47-58.

James, W. (1890). *The principles of psychology*. New York: Henry Holt.

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 and Cognitive Processes*, 23, 1191-1223.

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.

Jordan, T. R., Thomas, S. M., Patching, G. R., & Scott-Brown, K. C. (2003). Assessing the importance of letter pairs in initial, exterior, and interior positions in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 29(5), 883-393.

Joubert, O. R., Fize, D., Rousselet, G. A., & Fabre-Thorpe, M. (2008). Early interference of context congruence on object processing in rapid visual categorization of natural scenes. *Journal of Vision, 8*(13):11, 1-18.

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look, *IEEE International Conference on Computer Vision (ICCV)*, Kyoto, Japan, 2106 - 2113.

Juhasz, B.J., & Rayner, K. (2003). Investigating the effects of a set of intercorrelated variables on eye-fixation durations in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 1312-1318.

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.* Oxford, UK: Elsevier, 373-389.

Jung, C., Liu, Q., Kim, J. (2009). A stroke filter and its application for text localization. *Pattern Recognition Letters,* 30(2), 114–122.

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.

Krauskopf, J. Lennie, P., & Sclar. G. (1990). Chromatic mechanisms in striate cortex of macaque. *Journal of Neuroscience,* 10, 646-669.

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.

Land, M. F. (2006). Eye movements and the control of actions in everyday life. *Progress in Retinal and Eye Research, 25*, 296-324.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25,* 259–284.

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 [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.

Lennie, P., Derrington, A. M., & Krauskopf, J. (1984). Chromatic mechanisms in lateral geniculate nucleus of macaque. *Journal of Physiology,* 357, 241-265.

Le Saux, B. & Amato, G. (2004). Image classifiers for scene analysis, *Proceedings of the International Conference of Computer Vision & Graphics (ICCVG),* Warsaw, Poland.

Levy, R. (2008). Expectation based syntactic comprehension. *Cognition 106,* 1126-1177.

Li, L. Socher, R. & Li, F. (2009). Towards Total Scene Understanding:Classification, Annotation and Segmentation in an Automatic Framework. *Computer Vision and Pattern Recognition (CVPR). 2009.*

Linderholm, T., Virtue, S., Tzeng, Y., & van den Broek, P. W. (2004). Fluctuations in the Availability of Information during Reading: Capturing Cognitive Processes using the Landscape Model. *Discourse Processes*, 37(2), 165-186.

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.

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

Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performance*, 4, 565–572.

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

Lucas, S.M., Panaretos, A., Sosa, L., Tang, A., Wong, S., Young, R. (2003). ICDAR 2003 robust reading competitions. In *Proceedings of International Conference on Document Analysis and Recognition (ICDAR)*, Edinburgh, UK, 682–687.

Mack, S. C., & Eckstein, M. P. (2011). Object co-occurrence serves as a contextual cue to guide and facilitate visual search in a natural viewing environment. *Journal of Vision*, 11(9):9, 1-16.

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.

Manginelli, A. A. & Pollmann, S. (2009). Misleading contextual cues: How do they affect visual search? *Psychological Research*, 73, 212-221.

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*. Erlbaum, 35-55.

Martelli, M., Majaj, N. J., & Pelli, D. G. (2005). Are faces processed like words? A diagnostic test for recognition by parts. *Journal of Vision*, 5, 58-70.

McClelland, J. L. & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review*, 88, 375-407.

McConkie, G. W. & Zola, D. (1979). Is visual information integrated across successive fixations in reading? *Perception & Psychophysics*, 25(3), 221-224.

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]

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

Moores, E., Laiti, L., Chelazzi, L. (2003). Associative knowledge controls deployment of visual selective attention. *Nature Neuroscience*, 6, 182-189.

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.

Morris, R. K. (1994). Lexical and message-level sentence context effects on fixation times in reading. *Journal of experimental psychology: Learning, Memory, &Cognition*, *20*, 92–103.

Najemnik, J. & Geisler, W.S. (2005). Optimal eye movement strategies in visual search. *Nature,* 434, 387-391.

Neider, M. B. & Zelinski, G. J. (2006) Scene context guides eye movements during visual search, *Vision Research*, 46, 614-621.

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.

Nuthmann, A., & Henderson, J. M. (2010). Object-based attentional selection in scene viewing. *Journal of Vision,* 10(8):20, 1–19.

Oliva, A. & Torralba, A. (2001). Modeling the Shape of the Scene: A holistic Representation of the Spatial Envelop. *International Journal of Computer Vision,* 42 (3), 145-175.

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Science,* 11, 520–527.

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.

Parkhurst, D. J., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual selective attention. *Vision Research,* 42, 107–123.

Peters, R. J. & Itti, L. (2007). Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2007),* 1-8.

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.

Pomplun, M. (2006). Saccadic selectivity in complex visual search displays. Vision Research, 46, 1886-1900.

Pomplun, M., Ritter, H., & Velichkovsky B., (1996). Disambiguating Complex Visual Information: Toward Communication of Personal Views of a Scene, *Perception*, 25, 8, 931-948.

Posner, I., Corke, P., & Newman, P. (2010). Using text spotting to query the world. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, Taipei, Taiwan, pp. 3181-3186.

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.

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.

Rasiwasia, N. & Vasconcelos, N. (2006). Scene Classification with Low-dimensional Semantic Spaces and Weak Supervision. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Anchorage.

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. (1998). Eye movement in reading and information processing: 20 years of research. *Psychological Bulletin, 24,* 372-422.

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., 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., & Kaiser, J. S. (1975). Reading mutilated text. *Journal of Educational Psychology, 67*, 301-306.

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.

Rayner, K., White, S., Johnson, R. & Liversedge, S. (2006). Raeding Wrods With Jubmled Lettres: There Is a Cost. *Psychological Science, 17,* 192-193.

Rayner, K., Castelhano, M. S., & Yang, J. (2009). Viewing task influences eye movements during active scene perception. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 35, 254–259.

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

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.

Russell, B. C., Torralba, A., Murphy, K. P. & Freeman, W. T. (2008), LabelMe: a database and web-based tool for image annotation, *International journal of computer vision, volume 77, issue1-3,* 157-173.

Schmidt, J. & Zelinsky, G.J. (2009). Search guidance is proportional to the categorical specificity of a target cue. *Quarterly Journal of Experimental Psychology*, 62 (10), 1904-1914.

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

Selfridge, O. G. (1959). Pandemonium: A paradigm for learning. In D. V. Blake & A. M. Uttley (Eds.), *The mechanisation of thought processes* (pp. 511-529). London: H. M. Stationery Office.

Shaoul, C., & Westbury, C. (2010). Exploring lexical co-occurrence space using HiDEx. *Behavior Research Methods*, 42(2), 393-413.

Shu, H., & Anderson, R. C. (1997). Role of radical awareness in the character and word acquisition of Chinese children. *Reading Research Quarterly,* 32(1), 78-89.

Slattery, T.J., Angele, B., Rayner, K., (2011). Eye movements and display change detection during reading. *Journal of Experimental Psychology: Human Perception and Performance*, 37, 1924-1938.

Stirk, J. A., & Underwood, G. (2007). Low-level visual saliency does not predict change detection in natural scenes. *Journal of Vision*, *7*(10):3, 1-10.

Strang G. (1993). Introduction to Linear Algebra, 2nd Edition, Wellesley-Cambridge Press.

Swain, M. J. & Ballard, D. H. (1991). Color Indexing, *Journal of Computer Vision,* 7(1), 11-32.

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

Taft, M. (1985). The decoding of words in lexical access: A review of the morphographic approach. *In D. Besner, T.G.Waller, & G.E. MacKinnon (Eds.) Reading research: Advances in theory and practice* (pp. 83-126), New York: Academic Press.

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

Tatler, B., Baddeley, R. & Gilchrist, l. (2005). Visual correlates of fixation selection: effects of scale and time. *Vision Research,* 45, 643-659.

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, *7*(14), 1-17.

Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, 30, 415–433.

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network: Computation in Neural Systems*. 14, 391-412.

Torralba, A., Oliva, A., Castelhano, M., & Henderson, J.M. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological Review*, 113, 766-786.

Traxler, M. J., Foss, D. J., Seely, R. E., Kaup, B., & Morris, R. K. (2000). Priming in sentence processing: Intralexical spreading activation, schemas, and situation models. *Journal of Psycholinguistic Research, 29*, 581-594.

Tseng, S. C., Chang, L. H., & Wang C. C. (1965). An informational analysis of the Chinese language: Ⅰ. The reconstruction of the removed strokes of the ideograms in printed sentence-texts [In Chinese]. *Acta Psychologica Sinica*. 10, 299-306.

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.

Turk, M. & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3, 71-86.

Tzeng, Y. (2007). Memory of narrative texts: How parts of Landscape model work. *Chinese Journal of Psychology*, *49* (3), 1.-25.

Tzeng, Y., van den Broek, P., Kendeou, P., & Lee, C. (2005). The computational implementation of the landscape model: Modeling inferential processes and memory representations of text comprehension. *Behavior Research Methods*, 37(2), 277-286.

Underwood, G., & Foulsham, T. (2006). Visual saliency and semantic incongruency influence eye movements when inspecting pictures. *Quarterly Journal of Experimental Psychology*, 59, 1931–1949.

Unema, P. J. A., Pannasch, S., Joos, M., & Velichkovsky, B.M. (2005). Time course of information processing during scene perception. *Visual Cognition*, 12(3), 473-494.

Underwood, G., Humphreys, L., & Cross, E. (2007). Congruency, saliency, and gist in the inspection of objects in natural scenes. In. R. P. G. van Gompel, M. H. Fischer, W. S. Murray, & R. L. Hill (Eds.), *Eye movements: A Window on Mind and Brain*, 564–579.

Underwood, G., Templeman, E., Lamming, L., & Foulsham, T. (2008). Is attention necessary for object identification? Evidence from eye movements during the inspection of real-world scenes. *Consciousness and Cognition*, 17, 159–170.

van den Broek, P. (2010). Using texts in science education: cognitive processes and knowledge representation. *Science*, 328, 453.

Viola, P. & Jones, M. (2004) Robust real-time face detection. *International Journal of Computer Vision*, 57(2), 137-154.

Võ, M. L.-H., & Henderson, J. M. (2009). Does gravity matter? Effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception. *Journal of Vision*, 9(3):24, 1-15.

Võ, M. L.-H., & Henderson, J. M. (2011). Object–scene inconsistencies do not capture gaze: evidence from the flash-preview moving-window paradigm. *Attention, Perception, & Psychophysics*, 73(6), 1742-1753.

Wang H. C., Hwang, A. D. & Pomplun, M. (2010). Object frequency and predictability effects on eye fixation durations in real-world scene viewing. *Journal of Eye Movement Research*, 3(3):3, 1-10.

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.

Wang, H. C. & Pomplun M. (2011). The attraction of visual attention to texts in real-world scenes. *The Annual Meeting of the Cognitive Science Society (Cogsci2011)*, 2733–2738.

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

Wisniewski, E. J. (1996). Construal and Similarity in Conceptual Combination. *Journal of Memory and Language,* 35, 434-453.

Wolfe, J. M. (1994). Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review*, 1, 202-238.

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(4), 1069-1075.

Yan, G., Bai, X., Zang, C., Bian, Q., Cui, L., Qi, L., Rayner, K. & Liversedge, S. (2012). Using stroke removal to investigate Chinese character identification during reading: evidence from eye movements. *Reading and Writing,* 25, 951-979.

Yang, H. & Zelinsky, G.J. (2009). Visual search is guided to categorically-defined targets. *Vision Research*, 49, 2095-2103.

Yarbus, A. L. (1967). *Eye movements and vision*. New York: Plenum.

Ye, Q., Jiao, J., Huang, J., & Yu, H. (2007). Text detection and restoration in natural scene images. *Journal of Visual Communication and Image Representation*. 18, 504-513.

Yee, E. & Sedivy, J. C. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 32, 1-14.

Yi, C. & Tian, Y. (2011). Assistive text reading from complex background for blind persons. In Proceedings of Camera-based Document Analysis and Recognition (CBDAR), Beijing, China, 15-28.

Zelinsky, G. J. (2008). A theory of eye movements during target acquisition. *Psychological Review,* 115 (4), 787–835.

Zhang, J. J., Wang, H. P., Zhang, M., Zhang, H. C. (2002). The effect of the complexity and repetition of the strokes on the cognition of the strokes and the Chinese characters. *Acta Psychologica Sinica,* 34(5): 449-453. (in Chinese)

Zhou, X., & Marslen-Wilson, W. (1995). Morphological structure in the Chinese mental lexicon. *Language and Cognitive Processes, 10* (6), 545-600.

Zhou, X., Marslen-Wilson, W. , Taft, M., & Shu, H. (1999). Morphology, orthography, and phonology in reading Chinese compound words. *Language and Cognitive Processes, 14* (5/6), 525-565.

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