A NOVEL AUTOMATED READING ASSISTANCE SYSTEM

J.J. Kang* and M. Eizenman**

University of Toronto/*,**Department of Electrical and Computer Engineering, **Department of Ophthalmology, *,**Institute of Biomaterials and Biomedical Engineering, Toronto, Canada

*jeff.kang@utoronto.ca, **eizenm@ecf.utoronto.ca

Abstract: A system to provide automated reading assistance during reading within a natural setting is described. The system detects when a reader encounters an unknown word, and immediately responds by vocalizing the word. The system allows for unrestricted head movement and motion of the reading material. In a pilot study with two subjects, the system provided assistance for 9 out of 10 unknown words and 1 out of 10 known words.

Introduction

During reading, we visually examine text and convert letters to sounds via cognitive processing that activates word recognition. When we encounter an unknown word, we apply learned conversion rules to map graphemes (letter units) to phonemes (sound units) [1]. However, unskilled readers may experience difficulty performing this conversion. Furthermore, many English words do not conform to standard conversion rules. Hence, it is desirable to hear unknown words pronounced to learn the proper letter to sound mappings. We describe a system that automatically detects when readers encounter an unknown word, and renders immediate assistance by vocalizing the word. We validate the system through experimentation in which subjects read from reading cards similar to cards used in language instruction.

Principle of Operation

According to the dual-route model of reading, visual word recognition is facilitated by two separate processing routes [2]. In the lexical route, the word's letter units are processed visually in parallel and a match is found within the reader's orthographic lexicon, containing all words the reader knows. In effect, the word is recognized as a single unit, rather than through a visual examination of its individual letters. A mapping from the orthographic lexicon to the phonological lexicon, i.e. from a word image to a pronunciation, is then performed. In the non-lexical route, a letter string is converted into a phoneme string via a grapheme-phoneme conversion governed by a set of mapping rules.

The lexical route is generally faster since each word is recognized as a whole. However, when the word is unknown to the reader, and thus not within the reader's orthographic lexicon, the lexical route fails and the reader must resort to the non-lexical route. The difficulty to pronounce an unknown word is evidenced by a longer processing time. To measure processing time we use a point-of-gaze estimation system. When processing time exceeds a predetermined threshold, the reading assistance system assumes that an unknown word has been encountered, and provides assistance in the form of automated vocalization of the viewed word.

System Description

An automated reading assistance system requires the following components: a means to measure per-word processing time during reading, a means to set the threshold that defines when an unknown word has been encountered, and a means to provide vocalization of words. In this section we describe a method to determine the processing time associated with each word.

To determine the viewed word during reading while allowing unrestricted head movement, we use a head-mounted eye-tracker along with a head-mounted camera [3, 4]. This camera captures images of the subject's field of view; in the discussion to follow, we refer to this camera as the scene camera. When using head-mounted eye-trackers, eye position is generally measured with respect to a coordinate system defined relative to the head (head coordinate system). The position of the scene camera's imaging plane within the head coordinate system is constant, yielding a constant relationship between the head coordinate system and the scene camera's image coordinate system. This allows eye position to be easily transformed to a point-of-gaze within the scene camera's image coordinate system.

Using this method, point-of-gaze is estimated with respect to a moving reference frame as the head moves. However, points of interest in the scene, such as the position of words on a reading card, are defined with respect to an object coordinate system that does not move, or moves independently of the head. Therefore, to identify what the subject is looking at, the relationship between the scene camera's image coordinate system and the object coordinate system must be determined to allow the point-of-gaze to be mapped to the reference frame of the points of interest. We determine this relationship by establishing correspondences between points in images obtained by the scene camera (containing projections of the reading card) and points on the reading card defined in the object coordinate system.

ISSN: 1727-1983 © 2005 IFMBE

The head-mounted eye-tracker was implemented based on the principle of tracking the pupil centre and corneal reflections (glints) to estimate eye position [5, 6]. The eye-tracker consists of a PC-based processing unit and an imaging unit mounted on a head-band carrying an eye-tracker assembly and a scene camera. The eye-tracker assembly consists of three primary components: a small infra-red (IR) CCD video camera (eye camera), two IR light-emitting diodes (LEDs), and a hot mirror that allows visible light to pass through while reflecting IR light. The two IR LEDs are mounted above the right eye; light is reflected by the angled hot mirror onto the eye, providing illumination and generating the two corneal reflections. Images of the eye are reflected by the hot mirror up to the eye camera which is also positioned above the eye. The scene camera has a field of view of 92.1°H × 69.1°V. A photograph of the eye-tracker is shown in Figure 1.

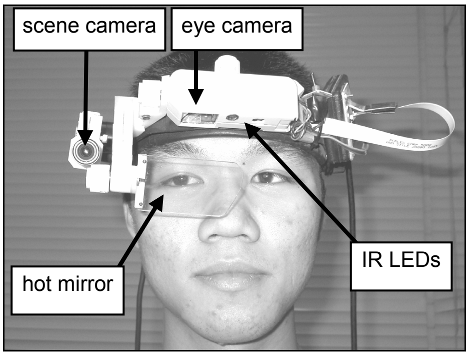

Figure 1: Head-mounted video-based eye-tracker

Images of the eye are processed at a rate of 50 Hz to locate the corneal reflections and the pupil centre. The relative positions of these eye features are used to estimate the point-of-gaze on the scene image.

The mapping of point-of-gaze from the scene camera's image coordinate system and the object coordinate system can be described in terms of projective geometry. Using homogenous coordinates, an arbitrary 3D object point $\mathbf{M} = (X, Y, Z, 1)^T$ is related to its a 2D image point $\mathbf{m} = (x, y, 1)^T$ via the linear mapping described by a 3×4 projection matrix $\mathbf{P}$:

$$S\mathbf{m} = \mathbf{PM} \tag{1}$$

where S is an arbitrary scalar [7]. Using a pinhole model of the scene camera, and assuming that all points $\mathbf{M}$ are located on the 2D surface of a reading card, defined by the Z=0 plane of the object coordinate system, the mapping may be simplified to:

$$S\mathbf{m} = \mathbf{HM'} \tag{2}$$

where $\mathbf{H}$ is a 3×3 homography matrix, and $\mathbf{M'} = (X, Y, 1)^T$ describes the point $\mathbf{M}$ on the reading card for which the Z-coordinate is zero and is omitted [8]. Given $N \geq 4$ point correspondences, $\mathbf{M}_i' \leftrightarrow \mathbf{m}_i$, $i = 1 \ldots N$, and the constraint that no N-1 of the N points are collinear, it is possible to calculate $\mathbf{H}$. We employ the Direct Linear Transformation (DLT) algorithm to solve for $\mathbf{H}$ [7]. A point-of-gaze estimate $\mathbf{POG}_{image} = (x_{image}, y_{image}, 1)^T$, described in the scene camera's image coordinate system is then mapped to $\mathbf{POG}_{object} = (X_{object}, Y_{object}, 1)^T$, the corresponding position on the reading card described in the object coordinate system using the relation:

$$\mathbf{POG}_{object} = S\mathbf{H}^{-1}\,\mathbf{POG}_{image}. \tag{3}$$

The task of automatically determining corresponding points between objects within a scene and its image is a classic problem in machine vision and photogrammetry commonly referred to as the correspondence problem [9]. A common approach to automating the measurement and identification of image points is to place coded targets within the scene [4, 10]. We adopt an approach using N circular coded targets, which are placed on the 2D surface of the reading card at known positions $\mathbf{M}_i'$, where $i = 1 \ldots N$, and $N \geq 4$. During reading, the scene camera captures images of the reading card and the coded targets from which the pixel positions of the targets, $\mathbf{m}_i$, are estimated. This establishes the N point correspondences from which $\mathbf{H}$ is calculated. Figure 2 shows a sample reading card displaying four targets.

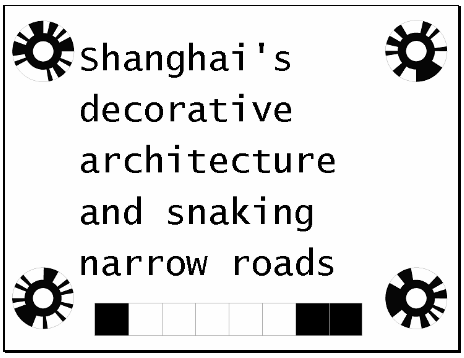
Shanghai's
decorative
architecture
and snaking
narrow roads

Figure 2: Sample reading card

In the context of a reading assistance system, reading cards must also be uniquely identified to allow more than one card to be used. Figure 2 shows the 8-bit barcode affixed to each reading card. This barcode is automatically extracted from the scene image during reading, and is decoded to identify the reading card being viewed.

After the reading card is identified, a lookup table containing the reading card's words is consulted. This table describes the position of each word with respect to the object coordinate system of the reading card. The viewed word is determined as the word that matches the position of the point-of-gaze, $\mathbf{POG}_{object}$.

Design of the Reading Card

For proper identification of the viewed words, words on the reading card must be separated by a distance greater than the resolution of the reading assistance system. The resolution is affected by two components:

(a) the accuracy of the point-of-gaze estimate, and (b) the accuracy of the mapping from the scene camera's image coordinate system to the object coordinate system.

The mapping method presented above assumes that the scene camera is an ideal pinhole camera. In practise, mapping errors are caused by non-linear lens distortions and measurement noise. Hence, the measured image point $\tilde{\mathbf{m}}_i = (\tilde{x}_i, \tilde{y}_i, 1)^T$ will deviate from the theoretical image point $\mathbf{m}_i = (x_i, y_i, 1)^T$, and this error will propagate to the homography matrix **H**.

To evaluate mapping performance in a typical experimental setup using reading cards (Figure 2), we constructed a validation card containing 9 targets (Figure 3) located at known object coordinates $\mathbf{M}_i'$ (i = 1…9). The card was held in a typical reading pose and imaged by the scene camera. Head movements and reading-card movements were made to simulate the motions expected during a reading task. A sequence of 1000 scene images was captured and the image coordinates of the targets $\tilde{\mathbf{m}}_i$ (i = 1…9) were recovered from each image.

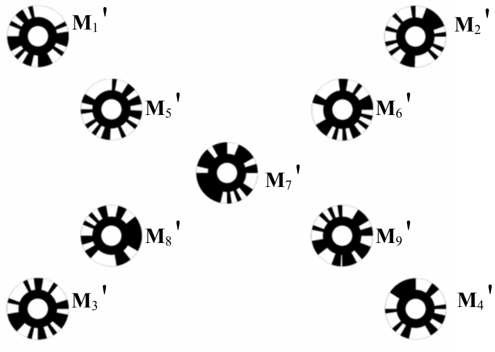

Figure 3: Validation card containing nine circular coded targets

For each image, the four point correspondences at the corners of the validation card, $\mathbf{M}_i' \leftrightarrow \tilde{\mathbf{m}}_i$ (i = 1…4), were used to estimate the homography as $\tilde{\mathbf{H}}$. These four targets serve the same function as the targets shown in the sample reading card. The computed homography was used to map the remaining five image points $\tilde{\mathbf{m}}_i$ (i = 5…9) to the corresponding reading material points $\tilde{\mathbf{H}}^{-1}\tilde{\mathbf{m}}_i$. Using the sequence of 1000 scene images, a total of 5000 points were mapped in this manner. For each mapped point, a mapping error, defined as the Euclidean distance between $\mathbf{M}_i'$ and $\tilde{\mathbf{H}}^{-1}\tilde{\mathbf{m}}_i$ (i = 5…9), was calculated. Figure 4 shows a histogram of the mapping errors.

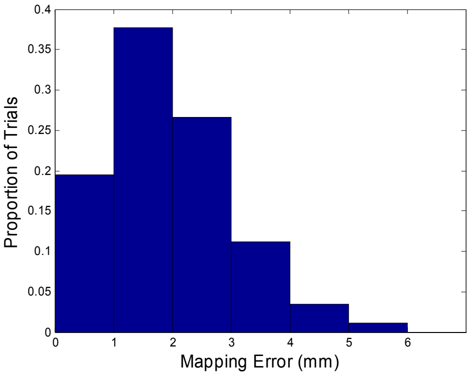

Figure 4: Histogram of mapping errors

The RMS mapping error for the 5000 points was found to be 2.29 mm. More than 98% of the points yielded a mapping error less than 5 mm. The head-mounted eye-tracker provides gaze estimates ($\mathbf{POG}_{\text{image}}$) with an accuracy of approximately one degree of visual angle. For a reading card placed 50 cm from the eye (a typical reading distance), the point-of-gaze is expected to have a maximum error of approximately 10 mm. The reading cards were designed such that all words are separated by at least 15 mm. For such a separation we expect that the viewed word will be correctly identified 98% of the time.

Detection of Unknown Words

The ability to identify the viewed word allows the reading assistance system to monitor the per-word processing time in real-time. However, to automatically trigger assistance for unknown words, the processing time must exceed a specific detection threshold. In this section, we describe a method to set this threshold.

To develop the detector, four skilled English readers read passages of text containing a mixture of known and unknown words while point-of-gaze was monitored. To minimize errors associated with head and reading-card motion, head position was stabilized using a chinrest, and simulated reading cards were presented on a computer screen.

Each subject read twenty passages aloud and twenty passages silently (approximately 1200 total words). After each passage, the subject was asked to indicate the unknown words in the passage. The processing time for each word was measured. Processing time was found to be slower for aloud reading for all subjects; therefore in the design of the detector, the processing times for silent and aloud reading were analyzed separately.

Since processing time is affected by word length [11], measured processing times were binned by word length. Words in each bin were further divided into two groups: known words and unknown words. This was done for measurements obtained during both silent and aloud reading. The mean processing time for each group was calculated. As expected, the results showed that processing time was longer for unknown words than known words for each word length. Figure 5 shows the mean processing times for Subject P.L. (aloud reading).

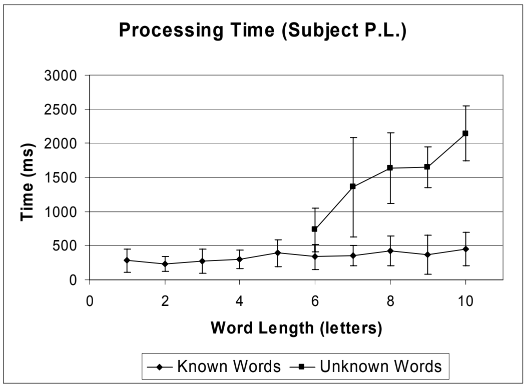

Figure 5: Processing times for Subject P.L. during aloud reading (one standard deviation is shown). This subject recognized all words shorter than six letters in length.

Using processing time as the detection criterion, we set a threshold to define the maximum processing time expected for a known word. Detection of an unknown word occurs when processing time exceeds this threshold. The Neyman-Pearson Criterion is used to set a threshold that maximizes the probability of detection while constraining the probability of a false alarm. The processing time for known words is modeled by a Gaussian random variable r ~ $N(\mu_k, \sigma_k^2)$, where μ_k and σ_k are estimated for each word length for individual subjects under either aloud or silent reading. The probability of false alarm, P_F, is given by

$$P_F = P\{r \geq T\} \tag{4}$$

where T is the processing time threshold above which a word is detected as an unknown word. We impose the constraint $P_F = \alpha$, where α is the acceptable probability of a false alarm. Hence, we can express T as

$$T = \sigma_k \sqrt{2}\mathrm{erf}^{-1}(1\text{-}2\alpha) + \mu_k \ . \tag{5}$$

For each value of $P_{F,}$ there is a set of threshold values (T), where each threshold value is determined for a particular word length. The process to determine the threshold values for each word length is time consuming. If the reading assistance system is used by many subjects, it is impractical to determine the exact threshold values for each subject. Therefore, we decided to approximate the threshold values by a simple function of word length, which would allow threshold values to be specified by a small number of parameters. Figure 6 shows the set of threshold values calculated for Subject P.L. (aloud reading). The shape of the threshold curve suggested that a linear approximation may be appropriate.

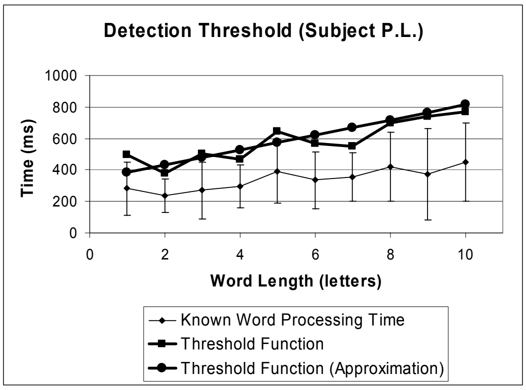

Figure 6: Detection threshold for Subject P.L. during aloud reading

The validity of this approximation was evaluated by applying the threshold values and their linear approximations to a set of processing times from which μ_k and σ_k were estimated (the training set). Detection performance, in terms of measured false alarm rate and measured detection rate, was not significantly affected by the use of the linear approximation. For values of P_F between 0.01 and 0.20, the change in false alarm rate and detection rate did not exceed 0.05.

Evaluation of Detector

In this section, we report the performance of the detector (using the linear approximation) on both the training set and a new test set. Figure 7 shows the measured false alarm rate for the training set as a function of the theoretical false alarm rate (P_F).

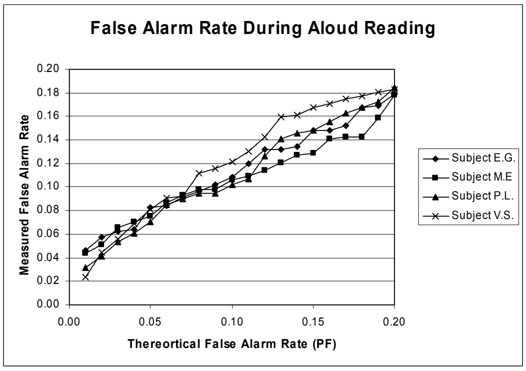

Figure 7: Measured false alarm rate for the training set (aloud reading)

The results showed that the measured false alarm rates closely followed P_F. This suggests that approximations introduced by: (a) modeling r as a Gaussian random variable and (b) the use of a linear approximation for the threshold function, did not significantly affect the ability to predict the false alarm rate.

The detector performance can also be evaluated in terms of the measured detection rate for the training set as a function of P_F, shown in Figure 8.

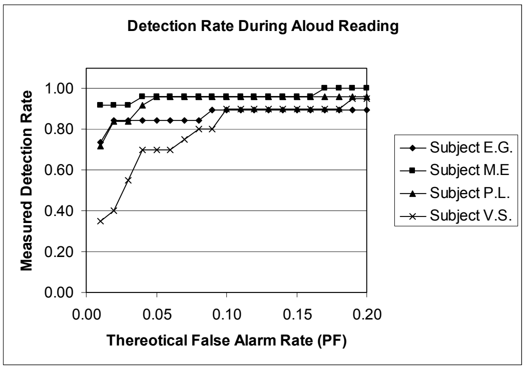

Figure 8: Measured detection rate for the training set (aloud reading)

As expected, the detection rate increases as P_F increases. Based on the above results, the value of P_F can be set to satisfy application-specific requirements for detection rate and false alarm rate. We consider a P_F value of 0.10 to be suitable since 9 out of 10 unknown words will be detected and vocalized. In the experiments to follow, we used this P_F value to evaluate system performance.

Detection performance was evaluated on a test set using the detection thresholds obtained from the training set. To obtain the test set, the same four subjects read an additional twenty passages silently and twenty passages aloud. Detection rates and false alarm rates for silent and aloud reading are summarized in Tables 1 and 2, respectively.

Table 1: Detection Performance for Silent Reading

	Detection Rate		False Alarm Rate	
Subject	Training Set	Test Set	Training Set	Test Set
E.G.	0.90	0.78	0.11	0.10
M.E.	0.96	1.00	0.14	0.16
P.L.	0.96	0.93	0.10	0.08
V.S.	0.64	0.74	0.10	0.12
Mean	0.87	0.86	0.11	0.12

Table 2: Detection Performance for Aloud Reading

	Detection Rate		False Alarm Rate	
Subject	Training Set	Test Set	Training Set	Test Set
E.G.	0.89	0.86	0.11	0.09
M.E.	0.96	1.00	0.11	0.12
P.L.	0.96	0.94	0.10	0.10
V.S.	0.90	0.86	0.12	0.11
Mean	0.93	0.92	0.11	0.10

The detection rates and the false alarm rates for the test sets and the training sets are similar. In general, the false alarm rate closely adhered to the specified P_F value of 0.10. The results suggest that the detection thresholds obtained from a small training set can be successfully applied when reading new text.

System Performance

An experiment using the head-mounted eye-tracker with the proposed mapping technique to determine the viewed word and the detection thresholds obtained above was performed to validate the principle of operation.

Two subjects read aloud from reading cards while sitting in a comfortable reading position (Figure 9). Head position was not restrained, allowing for natural head motions during reading. The subjects read aloud from reading cards, held at a comfortable reading distance (approximately 50 cm). The cards moved as the subject moved. P_F was set to 0.10. Detection of an unknown word triggered computer vocalization of the word.

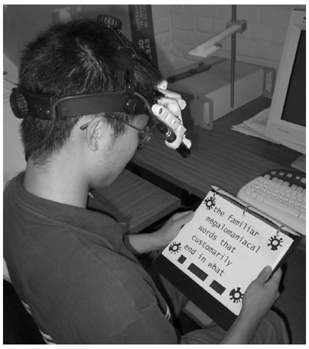

Figure 9: Subject wearing the eye-tracker in a typical reading pose

Detection performance for the two subjects is summarized in Table 3. The results are very similar to that obtained for aloud reading when head movement and reading card movement were constrained (Table 2). The results show that the detector can be used within a natural reading setting.

Table 3: Detection Performance within a Natural Reading Setting

Subject	Detection Rate	False Alarm Rate
M.E.	0.94	0.10
P.L.	0.95	0.09

Conclusions

We described a system that provides immediate vocalization of words that are unknown to the reader. We described a detector for which the false alarm rate can be controlled. The system provided vocalization assistance for 9 out of 10 unknown words (detection rate) and 1 out of 10 known words (false alarm rate). Vocalization of known words interrupts reading, therefore it is necessary to minimize the false alarm rate. This detection performance may be acceptable when using the reading assistance system to instruct unskilled English readers. In such a scenario the number of unknown words will be high compared to the number of known words, and the frequency of undesired vocalizations will be low.

To use the reading assistance system within a natural reading setting we have used a head-mounted eye-tracker. However, the system is not suitable for long reading tasks due to subject fatigue. It is possible to address this drawback using techniques that combine remote gaze estimation and reading-card motion estimation.

References

[1] EHRI, L. C. (1998): ‘Grapheme-phoneme knowledge is essential for learning to read words in English’, In Metsala and Ehri (Eds.). *Word recognition in beginning reading*, pp. 3-40, (Lawrence Erlbaum Associates, Hillsdale).

[2] COLTHEART, M., RASTLE, K., PERRY, C., LANGDON, R., and ZIEGLER, J. (2001): ‘The DRC model: A model of visual word recognition and reading aloud’, *Psychological Review*, **108**, pp. 204 - 258.

[3] WETZEL, P.A. KRUEGER-ANDERSON, G., POPRIK C., and BASCOM P. (1997): ‘An Eye Tracking System for Analysis of Pilots’ Scan Paths’, United States Air Force Armstrong Laboratory, Tech. Rep. AL/HR-TR-1996-0145.

[4] YU, L. H. and EIZENMAN, M. (2004): ‘A new methodology for determining point-of-gaze in head-mounted eye tracking systems.’ *IEEE Transactions on Biomedical Engineering*, **51**, pp. 1765-1773.

[5] YOUNG, L. R. and SHEENA, D. (1975): ‘Methods and designs: survey of eye movement recording methods,’ *Behavior Research Methods and Instrumentation*, **7**, pp. 397-429.

[6] DISCENNA A. O., DAS V., ZIVOTOFSKY A. Z., SEIDMAN S. H., and LEIGH R. J. (1995): ‘Evaluation of a video tracking device for measurement of horizontal and vertical eye rotations during locomotion’, *Journal of Neuroscience Methods*, **58**, pp. 98-94.

[7] HARTLEY, R. and ZISSERMAN, A. (2003): ‘Multiple View Geometry in Computer Vision, 2nd Ed.’ (Cambridge University Press, Cambridge).

[8] ZHANG, Z. (1999): ‘Flexible camera calibration by viewing a plane from unknown orientations’, Proc. of ICCV 1999, pp. 666-673.

[9] MARR D. and POGGIO T. (1976): ‘Cooperative computation of stereo disparity’, *Science*, **194**, pp. 283-287.

[10]GANCI G. and HANDLEY H. B. (1998): ‘Automation in videogrammetry’, *International Archives of Photogrammetry and Remote Sensing*, **32**, pp. 53-58.

[11]TRUESWELL, J. C., TANENHAUS, M. K., and GARNSEY, S. M. (1994): ‘Semantic influences on parsing: use of thematic role information in syntactic ambiguity resolution’, *Journal of Memory and Language*, **33**, pp. 284-318.