Chapter 18

FOVEAL LOAD AND PARAFOVEAL PROCESSING: THE CASE OF WORD SKIPPING

SARAH J. WHITE

University of Leicester, UK
Abstract

Three experiments showed that localised foveal load does not modulate the probability of skipping the following 4–6 letter parafoveal word (word $n+1$). In Experiments 1 and 2 the preview of the word $n+1$ was always correct. In Experiment 3 the preview of the word $n+1$ was either correct or incorrect. Localised foveal difficulty did not significantly modulate the effect of preview on the probability of skipping the word $n+1$. The results suggest that the processes that produce modulations of parafoveal preprocessing by foveal load on reading time measures may not apply to the control of word skipping.
As we read, we preprocess text that has not yet been fixated. Such preprocessing results in a greater probability of skipping words that are short, frequent or predictable compared to words that are long, infrequent or unpredictable (for reviews see Brysbaert & Vitu, 1998; Rayner, 1998). There are two different approaches to explaining the mechanisms that control which words are fixated or skipped during reading. The first is to suggest that the processes that determine word skipping are the same or similar to those that influence reading time. The second is to suggest that the processes that determine reading times and word skipping are qualitatively different. The present study investigates this issue by examining whether foveal load modulates parafoveal preprocessing in the same way for both reading times and word skipping.

Studies have suggested that the amount of parafoveal preprocessing, as shown by reading times, is limited by foveal processing difficulty1 (Henderson & Ferreira, 1990; Kennison & Clifton, 1995; Schroyens, Vitu, Brysbaert, & d’Ydewalle, 1999; White, Rayner, & Liversedge, 2005). These studies used the boundary saccade contingent change technique, which involves altering the parafoveal preview (which may be correct or incorrect) such that the word is correct when it is subsequently fixated (Rayner, 1975). For example, Henderson and Ferreira (1990) compared reading times on critical words (e.g. despite) when the preview of that word was correct (e.g. despite) or incorrect (e.g. zqdiowy) and when the word prior to the critical word was either frequent (e.g. chest) or infrequent (e.g. trunk). The difference in reading times when the preview is correct or incorrect gives a measure of the extent to which preprocessing of the correct preview facilitates processing once the word is fixated, known as preview benefit (Rayner & Pollatsek, 1989). Henderson and Ferreira showed that preview benefits for the critical word were larger when the previous word was frequent compared to when it was infrequent. That is, parafoveal preprocessing was reduced (preview benefits were smaller) when foveal processing was difficult compared to when it was easy.

Two different accounts have been proposed to explain the finding that parafoveal preprocessing is limited by foveal load. One is based on serial processing of words and a second is based on parallel processing of multiple words. Critically, both these accounts suggest that foveal processing difficulty influences parafoveal preprocessing as shown by both reading times and word skipping.

Serial attention shift models, such as the E-Z reader model (Reichle, Pollatsek, Fisher, & Rayner, 1998; Reichle, Rayner, & Pollatsek, 1999, 2003), have adopted an architecture in which shifts of attention and programming of eye movements are de-coupled. After saccade programming to the following word (word \(n+1\)) has begun, linguistic processing of the fixated word (word \(n\)) continues. The time to process word \(n\) is influenced by foveal load. When processing of word \(n\) has finally been completed, and usually before saccade programming is complete, attention shifts to word \(n+1\) so that it can be preprocessed. Due to the de-coupling of saccade programming and attention, the time to preprocess word

---

1 These studies used incorrect preview conditions in which multiple letters were incorrect. Other studies which have used incorrect previews containing a single internal incorrect letter have not shown any modulation of preprocessing by foveal load (Drieghe, Rayner, & Pollatsek, 2005; White & Liversedge, 2006b).
n + 1 (whilst fixating word n) is restricted by the time required to complete processing of word n. Critically, the time to attend to word n + 1 influences the extent to which word n + 1 is preprocessed and therefore the amount of preview benefit for word n + 1 (as shown by reading times on word n + 1). Similarly, once attention has moved to word n + 1, if there is sufficient time before the saccade is executed, and if word n + 1 is identified quickly enough, then the saccade programme may be re-programmed to skip word n + 1. Therefore both reading times and word skipping are determined by the same mechanism which is influenced by foveal load (time to process word n). Consequently Reichle et al. predict that foveal load should reduce the probability of skipping the following word.

In contrast, in their Glenmore model, Reilly and Radach (2003) suggest that multiple words can be processed in parallel and that there is competition between words for activation related to linguistic processing. Consequently greater processing of the fixated word reduces processing of other words. As a result, foveal load reduces preview benefit for the following word. Reilly and Radach also suggest that each word has a salience value, such that saccades are directed to the word with greatest salience. These salience values are influenced by the same linguistic activation system that influences reading times. Therefore, although not explicitly stated, Reilly and Radach’s account also appears to suggest that word skipping is modulated by foveal load.

To summarise, empirical evidence suggests that foveal load modulates parafoveal preprocessing as shown by reading time preview benefit. Accounts based on both serial processing and parallel processing have been proposed that explain this phenomenon. Both these models also predict that foveal load modulates word skipping in a similar way as for preview benefits. However, other studies have suggested that the processes that determine when and where the eyes move can be different (Radach & Heller, 2000; Rayner & McConkie, 1976; Rayner & Pollatsek, 1981). Indeed, a number of models of eye movement control in reading have been developed in which different mechanisms determine when and where the eyes move. These models either do not predict that foveal load modulates preprocessing as in the case of SWIFT (Engbert, Longtin, & Kliegl, 2002; Engbert, Nuthmann, Richter & Kliegl, 2005; Kliegl & Engbert, 2003) and the Competition/Interaction model (Yang & McConkie, 2001), or they predict that both the when and where systems are modulated by foveal load, as in Glenmore (Reilly & Radach, 2003). Nevertheless, accounts which differentiate between mechanisms that determine when and where the eyes move highlight the possibility that although foveal load modulates preprocessing as shown by reading times, foveal load may not necessarily modulate where the eyes move (as shown by the probability of word skipping).

The issue of whether foveal load modulates the probability of word skipping is therefore critical not only for evaluating the architecture of current models but also for assessing the fundamental question of whether the mechanisms that determine when and where the eyes move are the same or different. Drieghe et al. (2005) investigated whether foveal load modulated the probability of skipping parafoveal three-letter words (see also Kennison & Clifton, 1995). For cases in which there was a correct preview of the parafoveal word, Drieghe et al. showed no significant effect of foveal load on the probability of skipping the following word. Despite the non-significant result, skipping rates were numerically
higher when there was low, compared to high, foveal load which is suggestive of the possibility that foveal load may modulate word skipping. Therefore it is important to examine whether foveal load does reliably influence word skipping. Also, as Drieghe et al. only tested the probability of skipping three letter words, it is important to test whether foveal load modulates the probability of skipping slightly longer words.

The present study includes three experiments that test whether localised foveal load influences the probability of skipping four- to six-letter parafoveal words. The manipulations of foveal load include orthographic regularity, spelling and word frequency. These manipulations are intended to influence the ease with which a specific word can be processed. Importantly, this study does not test whether general processing load modulates word skipping. General processing load may be modified by text difficulty (e.g. contextual factors) or reading strategy. For example, general processing load could modulate global parameters for eye movement control such that increased load might increase fixation durations or shorten saccade lengths (see Yang & McConkie, 2001). Within each of the experiments presented here, such general factors are controlled by using the same sentence beginnings up until the critical words across each of the experimental conditions.

For all of the experiments presented here, the analyses include only cases in which a single fixation was made on the foveal word and no regressions were made out of the foveal word. Refixations on the foveal word could modify factors which might influence word skipping, such as launch site and the quality of the parafoveal preview. Therefore, restricting the analyses to cases in which single fixations were made on the foveal word ensures that any differences in skipping probabilities could not be accounted for by differences in refixation probabilities. Overall, if foveal load influences word skipping then the probability of skipping the parafoveal word (word $n+1$) should be greater when the foveal word (word $n$) is easy, compared to difficult, to process.

1. Experiment 1

In Experiment 1, foveal load was manipulated by orthographic regularity. The foveal word (word $n$) was either orthographically regular (low foveal load e.g. *miniature*) or orthographically irregular (high foveal load e.g. *ergonomic*) and was followed by the parafoveal word (word $n+1$) (e.g. *chairs*).

Note that the foveal processing load manipulation in Experiment 1 has been shown to have a small (less than 0.5 character) but reliable influence on initial fixation positions on these words (White & Liversedge, 2006a). Therefore in the present study, initial fixations land nearer to the beginning of the foveal words that are difficult to process (orthographically irregular) than the foveal words that are easy to process (orthographically regular). Consequently, the launch site prior to skipping or fixating word $n-1$ may have been slightly further away for the high foveal load words compared to the low foveal load words. Launch site may influence skipping probabilities such that saccades launched from further away may be less likely to skip word $n+1$. Importantly, note that the direction of these effects would have facilitated an effect of foveal load on word skipping such that
when foveal load was high, word \( n + 1 \) would be less likely to be skipped. This additional factor would therefore have to be taken into account if orthographic modulations of foveal load were to modulate word skipping.\(^2\)

1.1. Method

1.1.1. Participants

One hundred and four native English speakers at the University of Durham were paid to participate in the experiment. The participants all had normal or corrected to normal vision and were naïve in relation to the purpose of the experiment.

1.1.2. Materials and design

The foveal word \( n \) had orthographically regular (low foveal load) or irregular (high foveal load) word beginnings and these two conditions were manipulated within participants and items. The parafoveal word \( n + 1 \) was identical for each of the conditions within each item. The foveal words were nine or ten letters long and the parafoveal words were five or six letters long. Sixty of the participants read half of the sentences entirely in upper case, this variable was not included in the analysis.\(^3\) There were 24 critical words in each condition. Word \( n \) and \( n + 1 \) were embedded roughly in the middle of the same sentential frame up to and including word \( n + 1 \). Each of the sentences was no longer than one line of text (80 characters). See Table 1 for examples of experimental sentences and critical words. Full details regarding the nature of the orthographic regularity manipulation and the construction of the stimuli lists can be found in White and Liversedge (2006a).

<table>
<thead>
<tr>
<th>Table 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examples of experimental sentences and critical words for each condition in Experiment 1. Word ( n ) is shown in italics. For each sentence frame, version “a” is the low foveal load (regular beginning word) condition and version “b” is the high foveal load (irregular beginning word) condition.</td>
</tr>
<tr>
<td>1a. Last friday the modern miniature chairs were placed in the dolls house.</td>
</tr>
<tr>
<td>1b. Last friday the modern ergonomic chairs were transported to the shops.</td>
</tr>
<tr>
<td>2a. He hated the heavy primitive tools that the farmer gave him to use.</td>
</tr>
<tr>
<td>2b. He hated the heavy pneumatic tools that were used to dig up the road.</td>
</tr>
<tr>
<td>3a. He knew that the clever candidates would produce impressive answers.</td>
</tr>
<tr>
<td>3b. He knew that the clever auctioneer would ask him about the valuable lots.</td>
</tr>
<tr>
<td>4a. She knew that the modern extension would add value to the house.</td>
</tr>
<tr>
<td>4b. She knew that the modern ointments would work if she could get them in time.</td>
</tr>
</tbody>
</table>

\(^2\) Similar reasoning also applies in Experiment 2. The misspellings used in Experiment 2 were also found to modulate saccade targeting (White & Liversedge, 2006b).

\(^3\) In Experiment 1, type case did not significantly influence reading times on the foveal word \( n \).
1.1.3. Procedure

Eye movements were monitored using a Dual Purkinje Image eye tracker. Viewing was binocular but only the movements of the right eye were monitored. The letters were presented in light cyan on a black background. The viewing distance was 70 cm and 3.5 characters subtended 1° of the visual angle. The resolution of the eye tracker is less than 10 min of arc and the sampling rate was every millisecond.

Participants were instructed to understand the sentences to the best of their ability. A bite bar and a head restraint were used to minimise head movements. The participant completed a calibration procedure and the calibration accuracy was checked after every few trials during the experiment. After reading each sentence the participants pressed a button to continue and used a button box to respond “yes” or “no” to comprehension questions.

1.1.4. Analyses

Fixations shorter than 80 ms that were within one character of the next or previous fixation were incorporated into that fixation. Any remaining fixations shorter than 80 ms and longer than 1200 ms were discarded. Five percent of trials were excluded due to either no first pass fixations on the sentence prior to word \( n-1 \) or tracker loss or blinks on first pass reading of word \( n-1 \) or \( n \).

1.2. Results and discussion

Paired-samples \( t \) tests were undertaken with participants \( (t_1) \) and items \( (t_2) \) as random variables. Eleven percent of trials were excluded due to first pass regressions made out of the foveal word and 30 percent of trials were excluded due to skipping or multiple first pass fixations on the foveal word.

1.2.1. Single fixation duration word \( n \)

Single fixation durations on word \( n \) were significantly longer in the high foveal load orthographically irregular condition \( (M = 366, \ SD = 135) \) than in the low foveal load orthographically regular condition \( (M = 314, \ SD = 100) \), \( t_1(103) = 9.63, \ p < 0.001; \ t_2(23) = 6.4, \ p < 0.001 \). The manipulation of the orthographic regularity of word \( n \) was clearly effective.\(^4\)

\(^4\) For all three of the experiments the foveal load manipulation for word \( n \) also significantly influenced first fixation durations and gaze durations on word \( n \).
1.2.2. Skipping probability word \( n + 1 \)

There was no difference in the probability of skipping word \( n + 1 \) on first pass between the high foveal load orthographically irregular condition (0.14) and the low foveal load orthographically regular condition (0.15), \( t_1(103) = 1.62, p = 0.107; t_2(23) = 1.02, p = 0.319 \). Regardless of whether the foveal word had caused reduced (e.g. miniature) or increased (e.g. ergonomic) foveal processing difficulty, the probability of skipping the following parafoveal word (e.g. chairs) was the same. The findings from Experiment 1 show that although orthographic regularity clearly increased processing time on word \( n \), this had no effect on the probability of skipping the following word.

2. Experiment 2

Orthographic regularity significantly influenced single fixation durations on the foveal word in Experiment 1. Nevertheless, perhaps this manipulation of foveal difficulty was not sufficiently strong to influence subsequent word skipping. Experiment 2 used a stronger manipulation of foveal difficulty. Previous studies have shown that misspellings cause disruption to reading due to the difficulty associated with understanding non-words as words (e.g. Zola, 1984). Therefore in Experiment 2 foveal difficulty was manipulated by spelling. The foveal words were either spelled correctly (low foveal load, e.g. performer) or incorrectly (high foveal load, e.g. pwrformer) and these were followed by the parafoveal words (e.g. stood). The method was the same as for Experiment 1 except where noted below.

2.1. Method

2.1.1. Participants

Forty-four native English speakers at the University of Durham participated in the experiment.

2.1.2. Materials and design

The foveal word \( n \) was either spelled correctly (low foveal load) or misspelled (high foveal load) and these two conditions were manipulated within participants and items. The foveal words were all nine or ten letters long and the parafoveal words were five or six letters long. Half the foveal words were preceded by a frequent word and half by an infrequent word, this variable was not included in the analysis. Except for word \( n - 1 \) (which had high or low word frequency), word \( n \) and word \( n + 1 \) were embedded in sentence frames which were otherwise identical for each condition. See Table 2 for examples of experimental sentences and critical words. Full details regarding the nature of the misspelling manipulation, the word frequency manipulation, and construction of the stimuli lists can be found in White and Liversedge (2006b, Experiment 1).
Ch. 18: Foveal Load and Parafoveal Processing: The Case of Word Skipping

Table 2

Examples of experimental sentences and critical words for each condition in Experiment 2. Word \( n \) is shown in italics. For each sentence frame, version “a” is the low foveal load (correctly spelled) condition and version “b” is the high foveal load (misspelled) condition.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1a.</td>
<td>After the circus act the famous <strong>performer</strong> stood to receive the applause.</td>
</tr>
<tr>
<td>1b.</td>
<td>After the circus act the famous <strong>pwrformer</strong> stood to receive the applause.</td>
</tr>
<tr>
<td>2a.</td>
<td>At the meeting the whole <strong>committee</strong> voted against the planning application.</td>
</tr>
<tr>
<td>2b.</td>
<td>At the meeting the whole <strong>ctmmittee</strong> voted against the planning application.</td>
</tr>
<tr>
<td>3a.</td>
<td>The tourists enjoyed talking to the young <strong>traveller</strong> about his many experiences.</td>
</tr>
<tr>
<td>3b.</td>
<td>The tourists enjoyed talking to the young <strong>tlaveller</strong> about his many experiences.</td>
</tr>
<tr>
<td>4a.</td>
<td>The brave explorers knew that the great <strong>endeavour</strong> would need a lot of effort.</td>
</tr>
<tr>
<td>4b.</td>
<td>The brave explorers knew that the great <strong>endeavour</strong> would need a lot of effort.</td>
</tr>
</tbody>
</table>

2.1.3. Procedure

Participants were instructed that some sentences would contain misspellings but that they should read and understand the sentences to the best of their ability.

2.1.4. Analyses

Seven percent of trials were excluded.

2.2. Results and discussion

The results were analysed in the same manner as for Experiment 1. Sixteen percent of trials were excluded due to first pass regressions made out of the foveal word and 30 percent of trials were excluded due to skipping or multiple first pass fixations on the foveal word.

2.2.1. Single fixation duration word \( n \)

Single fixation durations on word \( n \) were significantly longer in the high foveal load misspelled condition (\( M = 376, SD = 162 \)) than in the low foveal load correctly spelled condition (\( M = 307, SD = 92 \)), \( t_1(43) = 7.59, p < 0.001; t_2(47) = 6.98, p < 0.001 \). The manipulation of spelling accuracy on word \( n \) was clearly effective.

2.2.2. Skipping probability word \( n + 1 \)

There was no difference in the probability of skipping word \( n + 1 \) on first pass between the high foveal load misspelled condition (0.3) and the low foveal load correctly spelled condition (0.3), \( ts < 1 \). Regardless of whether the foveal word had caused reduced (e.g. **performer**) or increased (e.g. **pwrformer**) foveal processing difficulty, the probability of skipping the following parafoveal word (e.g. **stood**) was the same. Therefore, similar to Experiment 1, Experiment 2 showed no effect of foveal load on the probability of skipping the following word, even when a very strong manipulation of foveal processing difficulty was used.
3. Experiment 3

Foveal difficulty was manipulated by orthographic regularity in Experiment 1 and by spelling in Experiment 2. The findings of both these experiments suggest that localised foveal load does not modulate the probability of skipping the following word. In order to ensure that this finding is robust across a range of different types of localised foveal load, Experiment 3 used word frequency to modulate foveal processing difficulty. In addition, both Experiments 1 and 2 used long foveal words. Experiment 3 therefore tested whether the findings held for short foveal words. Furthermore, Experiment 3 manipulated the nature of the preview of the parafoveal word.

In Experiment 3 the foveal words (word \(n\)) had high word frequency (low foveal load, e.g. happy) or low word frequency (high foveal load, e.g. agile) and these were followed by the parafoveal word (e.g. girl). The boundary saccade contingent change technique (Rayner, 1975) was used such that the preview of the parafoveal word was either correct (e.g. girl) or incorrect (e.g. bstc). The reading time data for the parafoveal word in Experiment 3 are reported in White et al. (2005) (Group 1 data). When foveal load was low (e.g. happy), there was 47 ms gaze duration preview benefit for the subsequent word (e.g. girl) whereas when foveal load was high (e.g. agile) there was only 1 ms gaze duration preview benefit. Similar to Henderson and Ferreira (1990) these results show that the difficulty of the foveal word modulates preprocessing (preview benefit) for the following word. Therefore, in Experiment 3, foveal load is clearly modulating parafoveal preprocessing, at least as shown by when the eyes move.

If foveal load modulates the probability of skipping the following word then the correct preview of word \(n+1\) will be more likely to be skipped when the foveal word is easy (high frequency) compared to difficult (low frequency) to process. If this is the case then the influence of foveal load on the probability of skipping the visually dissimilar incorrect preview should provide further insight into the nature of such an effect. First, foveal load may influence the probability of skipping word \(n+1\) regardless of the characteristics of word \(n+1\). That is, foveal load should have the same effect on the probability of skipping word \(n+1\) both when the preview is correct (e.g. girl) and incorrect (e.g. bstc). Second, it might be argued that words should usually only be skipped if they are familiar (Reichle et al., 1998, 1999, 2003). Consequently the incorrect preview of word \(n+1\) (e.g. bstc) should be skipped only very rarely because it is an unfamiliar non-word. That is, there should be an interaction such that foveal load modulates the probability of skipping the correct, but not the incorrect, previews of word \(n+1\). The method for Experiment 3 is the same as for Experiment 1 except where noted below.

3.1. Method

3.1.1. Participants

Thirty-two students at the University of Massachusetts were paid or received course credit to participate in the experiment.
3.1.2. Materials and design

Two variables, foveal processing difficulty (word \(n\)) and parafoveal preview (word \(n+1\)), were manipulated within participants and items. The foveal word \(n\) was easy to process (high frequency, *happy*) or difficult to process (low frequency, *agile*). The preview of word \(n+1\) before it was first fixated was correct (*girl*) or incorrect (*bstc*). The foveal word \(n\) was either five or six letters long and the parafoveal word was always four letters long. Full details regarding the nature of the materials and construction of the stimuli lists can be found in White et al. (2005). See Table 3 for examples of experimental sentences and critical words.

3.1.3. Procedure

The eye contingent boundary technique was used (Rayner, 1975); the display changes occurred within 5 ms of detection of the boundary having been crossed. Sentences were displayed at a viewing distance of 61 cm and 3.8 characters subtended 1° of visual angle.

3.1.4. Analyses

Trials were excluded due to: (a) display changes happening too early, (b) tracker loss or blinks on first pass reading of words \(n\) or \(n+1\) and (c) zero reading times on the first part of the sentence. Seventeen percent of trials were excluded.5

Table 3

<table>
<thead>
<tr>
<th>Examples of experimental sentences and critical words for each condition in Experiment 3. Word (n) is shown in italics. For each sentence frame, version “a” is the low foveal load (high frequency) condition and version “b” is the high foveal load (low frequency) condition. The incorrect preview of word (n+1) is shown in parentheses</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a. Outside the school the <em>happy</em> girl (<em>bstc</em>) skipped around the other children.</td>
</tr>
<tr>
<td>1b. Outside the school the <em>agile</em> girl (<em>bstc</em>) skipped around the other children.</td>
</tr>
<tr>
<td>2a. The supporters cheered when the <em>local</em> team (<em>wtdr</em>) finally won the match.</td>
</tr>
<tr>
<td>2b. The supporters cheered when the <em>inept</em> team (<em>wtdr</em>) finally won the match.</td>
</tr>
<tr>
<td>3a. The cook ordered the <em>daily</em> food (<em>gkhn</em>) from the local market.</td>
</tr>
<tr>
<td>3b. The cook ordered the <em>bland</em> food (<em>gkhn</em>) from the local market.</td>
</tr>
<tr>
<td>4a. The child pestered the <em>green</em> fish (<em>jbws</em>) that was hiding behind the pondweed.</td>
</tr>
<tr>
<td>4b. The child pestered the <em>timid</em> fish (<em>jbws</em>) that was hiding behind the pondweed.</td>
</tr>
</tbody>
</table>

5 Note that a larger proportion of data was excluded in Experiment 3 compared to Experiments 1 and 2 because the display contingent change technique requires that additional data must be excluded due to display changes happening too early.
3.2. Results and discussion

Fourteen percent of trials were excluded due to first pass regressions made out of the foveal word and 22 percent of trials were excluded due to skipping or multiple first pass fixations on the foveal word. A series of 2 (word n foveal load: frequent, infrequent) by 2 (word n + 1 preview: correct, incorrect) repeated measures Analyses of Variance (ANOVAs) were undertaken with participants \((F_1)\) and items \((F_2)\) as random variables. Table 4 shows the mean single fixation durations on word \(n\) and the probability of skipping word \(n + 1\) for Experiment 3.

3.2.1. Single fixation duration word \(n\)

Single fixation durations on word \(n\) were significantly longer in the high foveal load infrequent condition than in the low foveal load frequent condition, \(F_1(1, 31) = 11.99, p < 0.01; F_2(1, 40) = 18.37, p < 0.001\). There was no effect of the preview of word \(n + 1, F_1(1, 31) = 1.16, p = 0.29; F_2 < 1,\) and no interaction between the frequency of word \(n\) and the preview of word \(n + 1, F_1(1, 31) = 1.81, p = 0.188; F_2(1, 40) = 2.99, p = 0.092,\) for single fixation durations on word \(n\). Therefore the manipulation of foveal load on word \(n\) was clearly effective and the preview of word \(n + 1\) did not significantly influence reading times on word \(n\).

3.2.2. Skipping probability word \(n + 1\)

The parafoveal words were more likely to be skipped when the preview was correct (.185) compared to when it was incorrect (.135), this effect was significant across participants, \(F_1(1, 31) = 4.91, p = 0.03,\) but not items, \(F_2(1, 40) = 2.63, p = 0.113.\) Importantly, the effect of preview on the probability of word skipping indicates that the linguistic characteristics of parafoveal words influence whether they are subsequently fixated. The parafoveal word \(n + 1\) was skipped on 17 percent of trials when the foveal word \(n\) was high frequency and 15 percent of trials when the foveal word was low frequency. Foveal load had no significant effect on the probability of skipping the parafoveal word \((F's < 1)\)

<table>
<thead>
<tr>
<th>Probability of skipping word (n + 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word (n)</td>
</tr>
<tr>
<td>------------------------------------</td>
</tr>
<tr>
<td>Preview of word (n + 1)</td>
</tr>
<tr>
<td>Frequent</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Infrequent</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Table 4

Experiment 3. Single fixation durations on word \(n\). Standard deviations in parentheses.
and there was no interaction between foveal load and the parafoveal preview, $F_1 < 1; F_2(1, 40) = 1.51, p = 0.226$.

As in Experiments 1 and 2, there was no difference in the probability of skipping a correct preview of word $n + 1$ when there was low (0.18) compared to high (0.19) foveal load. Therefore, regardless of whether the foveal word had caused reduced (e.g. happy) or increased (e.g. agile) foveal processing difficulty, the probability of skipping the following parafoveal word (e.g. girl) was the same. However, note that the word $n+1$ was numerically less likely to be skipped when there was high foveal load and an incorrect parafoveal preview, compared to the other conditions. The nature of this interactive pattern is similar to that shown by Drieghe et al. (2005) (see Drieghe et al. for an extended discussion of possible explanations). Critically, the absence of any effect of foveal load on the probability of skipping the correctly spelled word $n + 1$ suggests that whatever might have caused, the numerical effect for incorrect previews does not hold during normal reading of correctly spelled text.

4. General discussion

All three of the experiments presented here show no effect of localised foveal load on the probability of skipping four to six letter words. The results are consistent with Drieghe et al.’s (2005) finding that for correctly spelled words, there was no significant difference between the probability of skipping three-letter words when there was high, compared to low, foveal load. Although Drieghe et al. showed a numerical difference in skipping probabilities, the fact that none of the experiments here showed more than a 0.01 difference in skipping probabilities for correctly spelled words suggests that there is no reliable effect of foveal load on the probability of word skipping when reading normal text. These findings contrast with studies which demonstrate that foveal load modulates parafoveal preprocessing as shown by preview benefit (Henderson & Ferreira, 1990; Kennison & Clifton, 1995; Schroyens et al., 1999; White et al., 2005). Furthermore, the fact that the same experiment demonstrated such effects for reading times (as reported in White et al.) but shows no such effects for word skipping (Experiment 3) is particularly poignant. Together, these findings provide strong evidence for the notion that localised foveal load modulates preview benefits but not the probability of skipping the following word.

Although the experiments presented here suggest that localised foveal load does not influence the probability of word skipping, as noted in the Introduction this does not preclude the possibility that general processing load may have a global influence on skipping rates. Indeed, note that the skipping probabilities are higher in Experiment 2 than in either Experiment 1 or 3. This could be because the sentence beginnings included context relevant to the remainder of the sentence in Experiment 2, and because word $n + 1$ tended to occur later in the sentence in Experiment 2 compared to the other experiments. Such differences could have influenced general processing load, which may have modulated skipping probabilities.
The current study suggests that localised foveal load modulates parafoveal preprocessing as shown by reading times, but not as shown by word skipping. It is possible that foveal load may influence word skipping, but this effect may be so very small that it is undetectable in standard reading experiments. Alternatively, reading times and word skipping may be influenced by qualitatively different processes. The latter suggestion would be inconsistent with accounts which suggest that foveal load influences reading times and word skipping by the same mechanism (Reichle et al., 1998, 1999, 2003) or due to a common input (Reilly & Radach, 2003). However, if necessary, such models might be adapted such that the word skipping mechanism operated independent of foveal load.

The possibility that reading times and word skipping are controlled by qualitatively different processes supports the notion that different processes might determine when and where the eyes move (Rayner & McConkie, 1976). Indeed, White and Liversedge (2006b) also showed that foveal difficulty does not modulate parafoveal orthographic influences on where words are first fixated. This finding suggests that saccade targeting to a word is also independent of foveal processing load. However, note that the processes that determine word skipping and saccade targeting may be different to other types of “where” decisions such as refixations and regressions.

The findings presented here indicate that words may be preprocessed qualitatively differently for the mechanisms that determine reading times and word skipping. Parafoveal preprocessing that is limited by foveal load influences the mechanisms that determine reading times. In contrast, word skipping mechanisms may be influenced by parafoveal preprocessing that occurs regardless of foveal load. For example, the processes that determine reading times may be sensitive to the progress of word recognition and sentence comprehension processes. Such language comprehension processes might be limited by processing load such that a parafoveal word is preprocessed to a lesser extent when there is high, compared to low, foveal load. Therefore, reading time preview benefits for parafoveal words reflect reduced preprocessing of words when there is high, compared to low, foveal load. In contrast, the processes that determine word skipping may acquire information from parafoveal text in an automatic manner, independent of comprehension difficulty, such that words can be skipped even if they have not been recognised.

The processes that determine word skipping may be different from those which determine reading times because word skipping may only be influenced by parafoveal information. Note that as a result, the eyes may sometimes become “out of sync” with the location of attention (the progress of sentence comprehension). For example, when reading the phrase “agile girl”, the high frequency word girl may be skipped before the low frequency word agile has been fully processed. Consequently processing of skipped words may continue on subsequent fixations. This suggestion is consistent with the finding that there are more regressions following skips, compared to first pass fixation, of words (Vitu, McConkie, & Zola, 1998). Such an automatic word targeting mechanism may sometimes move the eyes away from what needs to be processed. However, a system based on simple parafoveal linguistic processing may be most optimal for selecting which words to fixate given the very limited time periods available for saccade programming.
To summarise, the results suggest that qualitatively different mechanisms might determine parafoveal preprocessing as shown by reading times (preview benefit) and word skipping. Future accounts of eye movement control in reading may need to adopt an architecture in which there is separate processing of, and possibly inputs to, the mechanisms that determine preview benefits and word skipping.

Acknowledgements

The findings reported here were presented at the Sixth European Workshop on Language Comprehension, St. Pierre d’Oléron, France (2004). The author acknowledges the support of a University of Durham Studentship; Experiments 1 and 2 form part of the author’s PhD thesis. Experiment 3 was undertaken whilst the author was on a research visit at the University of Massachusetts, Amherst, supported by a Study Visit Grant from the Experimental Psychology Society. This research was also supported by a Biotechnology and Biological Sciences Research Council Grant 12/S19168. Thanks to Simon P. Liversedge, Hazel I. Blythe, Alan Kennedy and Keith Rayner for their comments on an earlier version of this chapter. Other data from the Experiments are reported elsewhere. See White and Liversedge (2006a, 2006b) and White et al. (2005) for further details of Experiments 1, 2 and 3 respectively.

References


Chapter 19

THE FLEXIBILITY OF LETTER CODING: NONADJACENT LETTER TRANSPOSITION EFFECTS IN THE PARAFOVEA

REBECCA L. JOHNSON

University of Massachusetts, Amherst, USA
Abstract

Previous experiments have shown that transposed-letter (TL) nonwords (e.g., jugde for judge) produce significant priming relative to orthographic controls (e.g., jupte). In fact, masked priming experiments indicate that TL effects exist even when the letter manipulations are nonadjacent, as long as the transposed letters are both consonants (Perea & Lupker, 2004a). This chapter presents data from a new study in which nonadjacent TL effects and the differential effects of vowels and consonants are explored during sentence reading using an eye-contingent display change paradigm. Results indicate that TL effects exist when nonadjacent letter positions are manipulated, suggesting that the coding of letter identities within a word is not specific to the absolute letter position, but is, instead, much more flexible. However, unlike the results of Perea and Lupker, those from the present study indicate that vowels and consonants pattern similarly.
Many current models of visual word recognition assume that letter positions are encoded very early in visual word recognition, even before the encoding of letter identities. Such models include the Multiple Read-Out Model (Grainger & Jacobs, 1996), the Dual Route Cascaded model (Coltheart, Rastle, Perry, Ziegler, & Langdon, 2001), the Interactive Activation Model (McClelland & Rumelhart, 1981), and the Activation Verification Model (Paap, Newsome, McDonald, & Schvaneveldt, 1982). These models all assume a “channel-specific” coding scheme for the processing of letter identities. That is, letter positions are encoded first, followed by the encoding of letter identities within each specific letter position.

One sharp criticism that has been made against these models is that they fail to account for the fact that transposed-letter (TL) nonwords (e.g., jugde) have been found to be more similar to their base words (e.g., judge) than nonwords in which two letters are substituted with other letters (e.g., jupte). This transposed-letter effect is well documented across a number of tasks including naming (Andrews, 1996; Christianson, Johnson, & Rayner, 2005), lexical decision (Andrews, 1996; Chambers, 1979; Forster, Davis, Schoknecht, & Carter, 1987; Holmes & Ng, 1993; O’Connor & Forster, 1981; Perea & Lupker, 2003a, 2003b, 2004a, 2004b; Perea, Rosa, & Gómez, 2005; Schoonbaert & Grainger, 2004), semantic categorization (Taft & van Graan, 1998), and normal silent reading (Johnson, Perea, & Rayner, 2007). Models employing a channel-specific coding scheme incorrectly predict that these two nonwords are equally similar to one another, because in both cases, three of the five letters are in their correct letter position. Findings from such experiments have helped to argue against models of word recognition that suggest a “channel-specific” encoding of letters. It appears that the encoding of letter identities within a word is not dependent upon absolute letter position, but is much more flexible.

While the majority of these studies have found TL effects at the foveal level (i.e., where all stimuli fell within 2° of visual angle around the point of fixation), these effects have recently been found to exist in normal silent reading (Johnson et al., 2007) in which transpositions occurred in the parafovea (i.e., the area extending 4° to the left and 4° to the right beyond the foveal area). Johnson et al. used an eye-contingent display change technique (the boundary paradigm, Rayner, 1975) to manipulate the parafoveal preview readers received prior to fixating on a given target word (Figure 1). The stimuli from Perea and Lupker (2003a) were embedded into sentences and the prime conditions served as parafoveal previews of the target word. Parafoveal previews fell into one of five conditions: (1) identical to the target word (clerk as the preview of clerk), (2) a transposition of two internal letters (celrk), (3) a substitution of two internal letters (cohrk), (4) a transposition of the two final letters (clekr), or (5) a substitution of the two final letters (clefn). Johnson et al. found that the TL effects obtained in masked priming also exist during normal silent reading, where the potential priming information is located to the right of fixation in the parafovea. That is, parafoveal previews involving a transposition of two adjacent letters led to shorter fixation durations than previews involving a substitution of two adjacent letters. For short (five-letter) words, this pattern was true for both internal and final-letter manipulations, but for longer (seven-letter) words, there was no difference
Greg put the wild *newor* in a vase at his grandmother’s house.

Note: The asterisk located below each sentence indicates the reader’s fixation location. At the onset of the sentence, the target word (here shown in bold) is replaced with one of the three parafoveal previews (in this example, the transposed-letter preview, *flewor*). When the reader’s eyes cross the invisible boundary (located just to the left of the space immediately preceding the target word), the parafoveal preview of the target word changes to the target word (here, *flower*) and remains as such until the participant indicates that they have finished reading the sentence.

Figure 1. Example sentence employing the boundary paradigm.

between the transposed and substituted letter (SL) conditions at the word-final position, likely due to acuity constraints.

Thus, it appears that letter identity information can be extracted from the fovea (and the parafovea from the first five letters of the word to the right of fixation) independent of absolute letter position. These experiments also suggest that the encoding of specific letter positions follows some time after the encoding of letter identities. What is unclear, however, is the extent to which letter position does not matter. In the experiments presented so far, all of the TL conditions involved a single transposition of two adjacent letters.1

In a series of experiments using masked-priming techniques, Perea and Lupker (2004a) explored the nature of TL effects with nonadjacent letter manipulations in Spanish. In addition, manipulations involved either vowels (e.g., *anamil* as the prime for *animal*) or consonants (e.g., *caniso* as the prime for *casino*). The results indicated that when the letter manipulations involved consonants, TL primes led to faster lexical

1 Interestingly, research indicates that transposing the first two letters in a word causes disruption during normal silent reading (Johnson, Perea, & Rayner, 2007) and in masked-priming (Chambers, 1979). Chambers found that word-initial TL nonwords (e.g., *omtor* for *motor*) were less similar to their base words than word-internal TL nonwords (*liimt* for *limit*).
decision times than SL primes. However, when the letter manipulations involved vowels, there was no significant TL effect. Thus, TL-nonwords involving nonadjacent transpositions can activate the lexical representation of their base words, but only under certain circumstances.

The differential patterning of TL effects among consonants and vowels was further explored by Perea and Lupker (2004b) using English stimuli. Form priming effects were obtained for adjacent consonant transpositions (e.g., hosre–horse vs honce–horse) and for adjacent consonant–vowel transpositions (brcik–brick vs brsok–brick), but there were no priming effects for adjacent vowel transpositions (draem–dream vs droim–dream). These results, then, also support the differences in TL effects across vowels and consonants.

However, in English, the spelling-to-sound correspondences for consonants are much more regular than those for vowels. It would follow that consonants should be coded and processed more rapidly than vowels. This has led many researchers to hypothesize that these two types of letters are processed differently in reading (Berent & Perfetti, 1995). In fact, there has been much data from response time tasks (Perea & Lupker, 2004a, 2004b), silent reading tasks (Lee, Rayner, & Pollatsek, 2001, 2002), and brain-damaged patients (Caramazza, Chialant, Capasso, & Miceli, 2000) that suggest that vowels and consonants do play different roles in visual word recognition. For example, Berent and Perfetti (1995) and Lee et al. (2001, 2002) have data suggesting that at the foveal level, consonants play a greater role than vowels in the early stage of visual word recognition. The contribution of vowel information is just as strong as that of consonants, but plays a role much later in lexical identification.

In light of this previous research, the goal of the current experiment was twofold. First, I sought to investigate whether TL effects exist during normal silent reading when letter manipulations involve nonadjacent letter positions. Although letter identity can be encoded independent of absolute letter position while reading sentences, it could be the case that letter identities can only be encoded outside of their correct letter position when they are displaced one letter position to the left or right (to positions N−1 and N+1). If, however, readers are able to extract useful identity information from the parafovea that falls outside of this region (i.e., in this case, two character positions from the correct location, to positions N−2 and N+2), we would expect to find shorter fixation durations on target words (e.g., flower) preceded by a TL parafoveal preview (e.g., flewor) rather than a SL parafoveal preview (e.g., flawur). Such findings would provide even more support against models suggesting channel-specific encoding strategies.

Secondly, I sought to explore the differential patterning of TL effects that has been found in masked priming lexical decision tasks among vowels and consonants using a parafoveal preview experiment. All previous research on the differential roles of vowels and consonants in visual word recognition has addressed the patterning of these two letter groups when the stimuli (and experimental manipulations) were presented in foveal vision. If vowels and consonants are also processed differently in the parafovea, we might expect to see different patterns of TL-effects for these two types of letters. In the present experiment, TL nonwords (e.g., flewor and fosert) and SL nonwords (e.g., flawur and
fonewt) were presented as parafoveal previews of their base words (e.g., flower and forest, respectively) to explore the role of vowels and consonants in parafoveal processing.

1. Method

1.1. Participants

Thirty-three members of the University of Massachusetts Amherst community who were native speakers of American English participated in the experiment. All participants had normal vision or wore soft contact lenses and were naïve to the purpose of the experiment. At the completion of the experiment, they received course credit or monetary compensation for their time.

1.2. Apparatus

Single-line sentences appeared one at a time on a 15-inch NEC MultiSync 4FGe monitor. Participants were seated 61 cm from the monitor, and at this distance, 3.8 letters equaled 1° of visual angle. The display was refreshed every 5 ms. Eye movements were recorded using a Generation V Fourward Technologies Dual Purkinje Eyetracker interfaced with a Pentium computer. Although reading took place binocularly, eye movements were sampled every millisecond from only the reader’s right eye.

1.3. Stimuli

Thirty-six six-letter target words were embedded into single-line sentences no longer than 76 characters. Target words never occupied the sentence-initial or sentence-final word position and represented a variety of word classes and word frequencies. Three parafoveal preview conditions were created for each target word. In the identity condition, the parafoveal preview was identical to the target word (e.g., flower as the preview of flower). In the TL condition, the preview involved the transposition of the third and fifth letters (e.g., flewor). Finally, in the SL condition, the preview involved the substitution of the third and fifth letters (e.g., flawur). The replacement letters in the SL condition were visually similar to the two transposed letters. That is, vowels were substituted with vowels, consonants were substituted with consonants, ascending letters were substituted with ascending letters, and descending letters were substituted with descending letters. The TL condition and SL condition always maintained the overall word shape as presented in Courier font. For example, fosert was used as a TL nonword for forest (both the letters s and r are neither ascending nor descending) but furute was not used as a TL nonword for future (the letter r is neither ascending nor descending but the letter t is ascending).

In addition to the three parafoveal preview conditions, two types of target words were used. Target words either included (1) vowels at letter positions 3 and 5 (e.g., flower),
or (2) consonants at letter positions 3 and 5 (e.g., \textit{forest}). The two target word groups were matched for word frequency using both the Francis and Kučera (1982) frequency count and the Celex Lexical Database (Baayen, Piepenbrock, & Gulikers, 1995). Francis and Kučera frequencies for the 18 vowel words ranged from 1 to 340 per million (mean = 76). For the 18 consonant words, frequencies ranged from 1 to 301 per million (mean = 76). The frequencies of these two groups did not differ significantly from each other using either of the two frequency counts ($t$'s < 1).2

Previous research has found that when words are highly predictable from their previous context, they are often skipped (Rayner, 1998). Thus, in order to maximize the likelihood that target words would be fixated, the context leading up to each target word was neutral. In a predictability norming procedure, ten participants were presented with the beginning part of each sentence (up to the target word) and asked to predict the next word in the sentence. All target words were found to be unpredictable from their previous context (mean predictability score = 4.7%). There was also no significant difference in the predictability scores across the two word types ($t$ < 1).

In order to ensure that all of the target words fit well within their sentence context, the sentences were also normed for understandability. Ten participants were asked to rate from one (not understandable) to seven (very understandable) how well each target word fit within its sentence frame. All target words were judged to be highly understandable (mean = 6.5). In addition, there were no significant differences in understandability across the two word types ($t$ < 1). The experimental sentences (including the three parafoveal preview conditions) for each of the two word types are presented in the Appendix.

1.4. Design and procedure

In order to reduce head movements during the experiment, a bite bar and a forehead rest were used. The initial calibration then took place (which lasted roughly 5 min), followed by a practice session involving eight sentences. The experimental session then followed. Each experimental sentence appeared one at a time (in random order) along the center row of the monitor. Readers were told to read each sentence silently at a comfortable pace and to press a response key when finished. In order to investigate the amount of parafoveal information the readers are gaining about a target word before fixating it, the boundary paradigm (Rayner, 1975) was used (see Figure 1). Prior to the presentation of the sentence, a fixation box appeared at the leftmost part of the screen. The experimenter then initiated the onset of the trial in which the sentence appeared on the screen with the first letter of the sentence at the location of the fixation box.

2 Another possible difference between the two word type conditions or the three parafoveal preview conditions includes the mean bigram textual frequency and the mean trigram textual frequency. Although the mean bigram and mean trigram type frequencies of the identity previews (56.93 and 10.24, respectively) were significantly greater than those of the TL (40.38 and 3.84) and SL previews (32.72 and 3.03), there were no significant differences in the mean bigram or trigram type frequencies across the two word type conditions ($p$'s > 0.24).
The target word appeared in one of the three preview conditions. When the readers moved their eyes to fixate on the target word (crossing the invisible boundary located just to the left of the space immediately preceding the target word), the display changed so that the preview changed to the target word. The display change occurred during the saccade, and the target word then remained throughout the remainder of the trial. Between each trial, the accuracy of the initial calibration was checked before the experimenter initiated the next trial.

Parafoveal preview was a within-subject and within-item variable; word type was a within-subject and between-item variable. Each participant read all 36 experimental sentences (18 of which included a consonant target word and 18 of which included a vowel target word). Items were counterbalanced so that there were 12 sentences in each of the three preview conditions. Thus, there were three counterbalancing conditions. Experimental sentences were presented in random order along with 78 filler sentences. Comprehension questions followed 16% of the trials to ensure that participants were carefully reading the sentences. All of the readers scored above 89% accuracy on the questions (mean = 97%). The entire experimental procedure took less than 30 min.

2. Results

The amount of time spent fixating a word is thought to reflect the time it takes to process that word (Rayner, 1998; Rayner & Pollatsek, 1989). Given that readers can extract useful information from the parafovea prior to fixating a word, parafoveal previews that provide more useful information will lessen the subsequent time the reader spends directly fixating the target word. If the different parafoveal preview conditions provide more or less useful information, we would expect to see differences in fixation times on the target words themselves. Three common measures of the amount of time spent on the target word are first fixation duration, single fixation duration, and gaze duration. First fixation duration is the amount of time spent on the initial fixation of the target word, regardless of whether there is more than one fixation on it. In contrast, single fixation duration is the amount of time spent on the initial fixation of the target word given that there was only one fixation on the first pass reading of the word. Gaze duration is the sum of all fixation durations on the target word before the reader leaves the word.

Trials were eliminated from data analysis if (1) the display change was triggered too early, (2) tracker loss occurred during a trial, or (3) the participant blinked while fixating the pre-target word, target word, or post-target word. In cases in which adjacent fixations fell within one character of one another, and one of the fixations was short (less than 80 ms), the two fixations were pooled (see Rayner, 1998). In addition, extremely short (less than 80 ms) isolated fixations and extremely long (greater than 800 ms) fixations were eliminated from the data. Altogether, 15.8% of the data were eliminated. The mean first fixation durations, single fixation durations, and gaze durations for each of the three parafoveal preview conditions in each of the two word type conditions are shown in Table 1.
For each of the three dependent fixation duration measures, a 2 (word type: vowels or consonants) by 3 (parafoveal preview: identity control, transposed letters, or substituted letters) Analysis of Variance (ANOVA) was conducted on the data. Error variance was calculated over participants ($F1$) and over items ($F2$). In addition, planned comparisons were run to compare fixation duration in the TL condition to the respective identity condition and SL condition across the two word types.

The main effect of parafoveal preview was highly significant both by participants and by items across all three viewing duration measures (first fixation: $F1(2, 64) = 6.62, p < 0.01; F2(2, 68) = 7.46, p < 0.01$; single fixation: $F1(2, 64) = 13.87, p < 0.001; F2(2, 68) = 12.90, p < 0.001$; gaze duration: $F1(2, 64) = 10.33, p < 0.001; F2(2, 68) = 7.50, p < 0.01$). For first fixation duration and single fixation duration, this main effect was due to significantly longer viewing durations on target words preceded by SL previews when compared to both identity previews (first fixation: $t1(32) = 3.20, p < 0.01; t2(35) = 3.43, p < 0.01$; single fixation: $t1(32) = 4.75, p < 0.001; t2(35) = 4.58, p < 0.001$) and TL previews (first fixation: $t1(32) = 3.32, p < 0.01; t2(35) = 2.87, p < 0.01$; single fixation: $t1(32) = 4.45, p < 0.001; t2(35) = 3.14, p < 0.01$). For first fixation duration, there was no significant difference between the identity condition and the TL condition (both $t's < 1$), and for single fixation duration, the difference between these two conditions was significant only by items ($t1(32) = 1.64, p = 0.11; t2(35) = 2.15, p < 0.05$).

In contrast, for gaze duration, this main effect was the result of significantly shorter viewing durations for identity previews when compared to both TL previews ($t1(32) = 3.29, p < 0.01; t2(35) = 2.73, p < 0.01$) and SL previews ($t1(32) = 4.00, p < 0.001; t2(35) = 3.77, p < 0.001$). The difference between the TL condition and the SL condition was not significant either by participants or by items (both $p's > 0.12$).

The main effect of word type was not significant across any of the viewing duration measures (all $F's < 1$, all $p's > 0.5$). Critically, the interaction between parafoveal preview and word type was also not significant across any of the dependent measures (all $F's < 1$, all $p's > 0.55$). That is, the same pattern of parafoveal preview facilitation was seen in words in which vowels were transposed as in words in which consonants were transposed.

### Table 1

Means as a function of word type and parafoveal preview (standard errors in parentheses)

<table>
<thead>
<tr>
<th></th>
<th>First fixation</th>
<th>Single fixation</th>
<th>Gaze duration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>V</td>
<td>C</td>
<td>V</td>
</tr>
<tr>
<td>Identity</td>
<td>281 (8.9)</td>
<td>278 (8.1)</td>
<td>283 (8.7)</td>
</tr>
<tr>
<td>TL</td>
<td>283 (7.8)</td>
<td>288 (8.8)</td>
<td>290 (7.9)</td>
</tr>
<tr>
<td>SL</td>
<td>301 (8.2)</td>
<td>310 (8.6)</td>
<td>312 (7.5)</td>
</tr>
</tbody>
</table>

*Note:* All durations for first fixation duration, single fixation duration, and gaze duration are in ms. Word type involved the manipulation of either vowels (V) or consonants (C).
For vowels, the 18 ms TL effect for first fixation duration was significant by participants ($t_{1(32)} = 2.22, p < 0.05$) and marginally significant by items ($t_{2(17)} = 1.6, p = 0.064$). The 22 ms TL effect for single fixation duration was significant both by participants and by items ($t_{1(32)} = 2.90, p < 0.01$; $t_{2(17)} = 1.76, p < 0.05$). However, the 9 ms TL effect for gaze duration was not significant in either analysis (both $p$'s > 0.19). For consonants, a similar pattern arose. The 22 ms TL effect for first fixation was significant by participants and by items ($t_{1(32)} = 2.46, p < 0.01$; $t_{2(17)} = 2.42, p < 0.05$). The 25 ms TL effect for single fixation was significant by participants and by items ($t_{1(32)} = 2.79, p < 0.01$; $t_{2(17)} = 2.62, p < 0.01$). The 17 ms TL effect for gaze duration was marginally significant by participants ($t_{1(32)} = 1.33, p = 0.096$) and not significant by items ($p = 0.1$).

3. Discussion

The results from the current experiment are straightforward. When looking at early measures of visual word recognition (i.e., first fixation duration and single fixation duration), viewing duration measures were shortest for the identity condition and the TL condition and significantly longer for the SL condition. These findings indicate that letter identity is encoded early in visual word recognition. Previews that contained accurate letter identity information (i.e., the identity condition and the TL condition) led to significantly shorter early viewing duration times than previews containing inaccurate letter identity information (i.e., the SL condition). Furthermore, the fact that TL previews led to shorter viewing durations than SL previews indicates that letter identities can be encoded outside of their specific letter position. Encoding of letter identities is thus much more flexible than channel-specific models assume. These results are consistent with previous research showing TL effects during visual word recognition tasks and extend the work of Johnson et al. (2007) to show that TL effects also exist with nonadjacent transpositions.

In the analysis of a later measure of visual word recognition (namely, gaze duration), it becomes apparent that the role of accurate letter position is important in the encoding process. While the TL effect seen in earlier measures was no longer significant, the identity condition led to significantly shorter viewing durations than either of the other two conditions. These findings, thus, also challenge the aforementioned models of visual word recognition, which suggest that letter positions are encoded prior to letter identities. It appears the importance of letter identity occurs earlier in visual word recognition than the role of letter position. Others, too, have postulated that letter positions take longer to encode than letter identities (Adams, 1979). As these results indicate, it is quite likely that when a reader encounters a TL nonword such as caniso, he or she begins to encode and activate the lexical representations of each of the component letters (including the n and s, which are in their incorrect letter location). At some later point, however, the processing system must reorder the letters (i.e., those that are in incorrect letter positions)
Models assuming a channel-specific coding scheme for letter identities cannot account for the current findings. Furthermore, as Perea and Lupker (2004a, 2004b) pointed out, the presence of TL effects in nonadjacent manipulations causes especially great problems for interactive-activation models, which rest on the assumption that the bottom-up encoding of letter identities and letter positions is non-noisy. Perea and Lupker argued that another form of coding scheme, rather than a channel-specific coding scheme, is needed to be able to fully account for TL effects. Some more recent models of visual word recognition can account for the TL effects found in previous studies as well as those in the current experiment because they allow for “noise” in the system, although in differing ways. Three such models are the SOLAR model (Davis, 1999), the SERIOL model (Whitney, 2001), and the Overlap model (Gómez, Perea, & Ratcliff, 2003).

The SOLAR model (Davis, 1999) employs a spatial coding scheme to assign letter positions different activation levels according to their location within the letter string. According to the model, the first letter position receives the highest level of activation, followed by the second letter position and so forth. Letter strings with different letter identities in different letter positions, then, receive different patterns of activation. This leads to the successful distinction between anagrams like stop, pots, opts, post, tops, and spot. The model can account for TL effects because it also includes a separate parameter to measure the amount of similarity in the set of letter nodes. Since the nonword caniso shares the same set of six letters as casino, it is more similar to its base word than the nonword caviro.

The SERIOL model (Whitney, 2001) also assigns varying activation levels to successive letter positions. In addition, it relies on the activation of bigram nodes to encode words. The word casino, for example, can be broken down into 15 bigrams (ca, cs, ci, cn, co, as, ai, etc.). The nonword caniso shares 12 of these bigrams, while the nonword caviro shares only 6 of these. Thus, nonadjacent TL nonwords are successfully predicted to be more similar to their base words than SL nonwords.

As one reviewer noted, the fact that the TL condition differed from the identity condition in gaze duration (a pattern not seen in first fixation or single fixation) could indicate that in order for absolute (or even relative) letter position encoding to occur, more than one fixation on the word is necessary. While this is a possible explanation for the pattern of data in this experiment, the current results are inconclusive to address this hypothesis because gaze durations also include trials in which only a single fixation occurred. In the current experiment, only 6% of the target words received more than one fixation on the first pass reading of the word, making subanalyses of these items impossible. Furthermore, although the first fixation durations and single fixation durations between the TL condition and the identity condition were not always significant, the pattern of means suggests a trend in the same direction as that seen in gaze duration. Therefore, the differences seen in early versus late measures are likely due to the increased amount of time spent fixating on the target word rather than on the number of fixations the target word receives.

As reported by Perea and Lupker (2004a), the SOLAR model would predict that the identity condition (e.g., casino) would yield the highest similarity score to its base word (1.00), followed by the TL condition (caniso, 0.83), and lastly the SL condition (caviro, 0.54). Likewise, the SERIOL model predicts the same pattern of results with the following similarity scores: identity, 1.00; TL, 0.88; SL, 0.49.
Finally, the Overlap Model (Gómez, Perea, & Ratcliff, 2003) can account for TL effects because it assumes that letter representations extend into neighboring letter positions. The encoding activation of a given letter is greatest at its correct letter position, but this activation also encompasses other letter positions as well. Specifically, the encoding activation is represented as a normal distribution with activation decreasing as a letter appears further from its correct letter position. The Overlap Model would also successfully predict the pattern of results from the current experiment since activation of a given letter not only is greatest at its correct letter position (thus predicting greatest facilitation from the identity condition), but also extends out to neighboring letter positions (thus predicting more facilitation from TL nonwords than SL nonwords).

As indicated by the lack of a significant interaction, this pattern of results was nearly identical when consonants were manipulated as when vowels were. This suggests that, at least at the parafoveal level, the processing of consonants and vowels is similar in terms of the encoding of letter identities and letter positions. This is in contrast to the previous literature that has found differences between the role of consonants and vowels when stimuli are presented in foveal vision (Berent & Perfetti, 1995; Lee et al., 2001, 2002; Perea & Lupker, 2003b, 2004a, 2004b). Based on several foveal TL experiments manipulating vowels and consonants in both English and Spanish, Perea and Lupker (2004a) concluded that the absolute letter position of vowels may be more important than the absolute letter position of consonants. While this may be the case for foveally presented stimuli, the current results indicate that the flexibility of letter coding in the parafovea extends to include both vowels and consonants. It is thus likely that the parafoveal transposed letter effects found in the current study and also by Johnson et al. (2007) are likely to occur at a very low level of visual word recognition, before the encoding of a vowel/consonant label and the phonological attachment of letters to sounds. Clearly, more research should be conducted to assess the role of consonants and vowels at the parafoveal level to evaluate whether these two letter types have differing roles prior to fixation.

While the current study indicates that TL effects exist when nonadjacent letter positions are manipulated, there are still many questions that need to be addressed to explore the limitations of letter identity and letter position coding. Whether these parafoveal priming effects would exist (and how strong they would be) if manipulations were made further than two letter positions away (e.g., cnsiao for casino) or if multiple manipulations were made (e.g., csanio for casino) is currently unknown. Knowing exactly how far these TL effects extend and which factors mediate these effects can be helpful for testing models and informing certain parameters. For example, in the Overlap Model, knowing exactly how far TL effects extend would help to set the parameters of the encoding activation at neighboring letter positions to better define the shape and spread of the bell-shaped curves.

In summary, although nonadjacent TL effects have previously been found only to exist at the foveal level when consonants are manipulated, the current findings indicate that these effects hold for both consonants and vowels presented in the parafovea. Although consonants and vowels may play different roles in foveal visual word recognition,
the evidence presented here indicates that these two letter groups pattern similarly at the parafoveal level. The extraction of letter identity from the parafovea is not strictly dependent upon specific letter position, but rather, letter identity is encoded prior to letter position. These findings challenge models of visual word recognition that assume a channel-specific letter coding scheme and support models that allow more flexibility in the coding of letter identities. Specifically, the current findings support the SOLAR model, the SERIOL model, and the Overlap model. Future studies, then, should be designed to directly test opposing predictions that these three models make.

Acknowledgements

This research was supported by Grant HD26765 from the National Institute of Health to Keith Rayner. Thanks to Keith Rayner, Barbara Juhasz, Simon Liversedge, and one anonymous reviewer for their helpful comments on an earlier draft.

Appendix

A.1. Vowels

Greg put the wild (flower, flewor, flawur) flower in a vase at his grandmother’s house. Some athletes shy away from a tough (league, leugae, leogie) league with harsh competition. Consider the age and intelligence level of your typical (reader, reedar, reodur) reader when writing. Octavian attained the supreme (status, stutas, stotis) status as the first Roman Emperor in 27 BC. The magazine sought a careful (editor, edotir, edater) editor and great headline writer for the job. Samantha put the sharp (weapon, weopan, weepun) weapon outside of the young child’s reach. Our high school soccer team finished their amazing (season, seosan, seisun) season last weekend. In Roman circuses, the most popular (animal, anamil, anomul) animal for arena use was the lion. Jesus appointed Simon Peter as the first notable (leader, leedar, leodur) leader of the church. Lowering the risk of disease is a worthy (reason, reosan, reusin) reason to breastfeed your baby. Last night, a brown (weasel, weesal, weosil) weasel ran across the highway in front of our car. When the rich (deacon, deocan, deucen) deacon began abusing his authority, he was asked to step down.
The candles and lovely (violet, violot, vialit) **violet** blossoms added a nice touch to the reception.
The Johnsons have a tall (spiral, sparil, sporel) **spiral** staircase in the foyer of their home.
It is possible to have a droopy upper (eyelid, eyiled, eyalud) **eyelid** following cataract surgery.
The most useful (shovel, shevol, shuval) **shovel** for digging up fossils has a semi-pointed front edge.
After driving on the rough (gravel, greval, grivul) **gravel** road for an hour, our car was filthy.
Kimberly cut out the latest (coupon, coopun, coapen) **coupon** from the paper to use at the grocery.

**A.2. Consonants**

The animal hid in the deep (forest, fosert, fonewt) **forest** until it was safe for it to come out.
The inappropriate racial (remark, reramk, recaxk) **remark** led to much anger and violent reactions.
The child’s sudden (desire, derise, dewice) **desire** for seclusion and inactivity concerned his mom.
I bought a fresh (bottle, boltte, boktbe) **bottle** of cough medicine when the expiration date arrived.
Among many things, the Census collects data on family (income, inmoce, inroxe) **income** and housing.
While the boy was in surgery, the anxious (parent, panert, pamect) **parent** paced in the waiting room.
Especially in older adults, extreme (stress, stsers, stcens) **stress** can cause many health problems.
We brought some pure (spring, spnirg, spmicg) **spring** water on our hiking trip to the Rockies.
After much searching, Jen bought the rare (record, rerocd, rewosd) **record** from an online auction.
Ross plotted the path of the rocket from the initial (moment, monemt, morect) **moment** it took off.
Cassie pulled the warm (waffle, walffe, watfke) **waffle** out of the toaster and added syrup and jam.
The popular hangout for young adults is the busy (tavern, tarevn, tanemn) **tavern** on the corner.
Is it better to be a silent (coward, corawd, cosam) **coward** or an outspoken radical?
Lyla said she had a quick (errand, ernard, ersacd) **errand** to run before going to dinner with us.
Doctors recommend that you maintain a healthy (intake, inkte, inlabe) *intake* of dietary fiber.

For their honeymoon, Matt and Barb stayed at a nearby (resort, rerost, recomt) *resort* on the beach.

Theodore poured the thick (cement, cenemt cesert) *cement* to secure the basketball goal in place.

After I saw the empty (pantry, partny, pawtvy) *pantry* and bare refrigerator, I went to the grocery.

**References**


Chapter 20

EYE MOVEMENTS AND SPOKEN LANGUAGE PROCESSING

MICHAEL K. TANENHAUS

University of Rochester, USA
Abstract

This chapter provides an overview of recent research that uses eye movements to investigate both spoken language comprehension and language production. Issues of data analysis and linking hypotheses are addressed in what is now commonly referred to as the “visual world” paradigm, including issues that arise in comparing results to those from eye-tracking reading studies. It is argued that eye-tracking reading studies primarily use fixation duration as a processing load measure, whereas visual world studies use the location and timing of fixations as a representational measure. Considerations are raised about how the presence of a visual scene and use of actions might influence results from visual world studies.
Eye movements have been one of the most widely used response measures in psycholinguistic studies of reading for more than a century (see Ferreira & Henderson, 2004; Rayner, 1998). In contrast, it is only within the last decade that eye movements have become a commonly used response measure in studies of spoken language processing. In these studies, participants’ eye movements to real world objects or to pictures in a display or scene are monitored, typically using a head-mounted eye-tracker, as the participants follow instructions, listen to sentences, or generate utterances about the “visual world.”

Psycholinguists now use the visual-world eye-movement method to study both language production and language comprehension, in studies that run the gamut of current topics in language processing. Eye movements are a response measure of choice for studies addressing many classical questions about spoken language processing in psycholinguistics: e.g., is the processing of stop consonants categorical? (McMurray, Tanenhaus, & Aslin, 2002); does context influence the earliest moments of temporary lexical and syntactic ambiguity resolution? (Dahan & Tanenhaus, 2004; Spivey, Tanenhaus, Eberhard, & Sedivy, 2002); what is the locus of frequency effects in spoken word recognition? (Dahan, Magnuson, & Tanenhaus, 2001); what factors influence the time course with which anaphoric expressions, such as pronouns, are resolved? (Arnold, Eisenband, Brown-Schmidt, & Trueswell, 2000) and, for bilingual speakers, does a word spoken in one language activate the lexical representations of similar sounding words in the other language (Spivey & Marian, 1999; Ju & Luce, 2004).

The use of eye movements has also opened up relatively uncharted territory in language comprehension and language production. In comprehension these include real-time sentence processing in children (Trueswell, Sekerina, Hill, & Logrip, 1999); the role of common ground in online processing (Hanna, Tanenhaus, & Trueswell, 2003; Keysar, Barr, Balin, & Brauner, 2000); how listeners make use of disfluencies in real-time language processing (Arnold, Tanenhaus, Altmann, & Fagnano, 2004; Baily & Ferreira, 2003, this volume); and how participants in a conversation coordinate their referential domains (Brown-Schmidt, Campana, & Tanenhaus, 2005; Tanenhaus, & Brown-Schmidt, to appear). In production these include, the locus of disfluency effects (Griffin, 2004b), and the interface between message formulation and utterance planning (Bock, Irwin, Davidson, & Levelt, 2003; Brown-Schmidt & Tanenhaus, in press; Griffin & Bock, 2000). Finally, the visual-world approach has spawned a new family of studies investigating the interface between action and language and between vision and language (Altmann & Kamide 2004; Chambers, Tanenhaus & Magnuson, 2004; Knoeferle, Crocker, Scheepers, & Pickering, 2005; Spivey & Geng, 2001; Spivey et al., 2002).

Why is the visual world paradigm becoming so widely used? First, in contrast to reading, time-locked, relatively natural measures of spoken language processing have been hard to come by. Many of the most widely used tasks for studying spoken language comprehension present only a snapshot of processing at a single point in time, require meta-linguistic judgments, and interrupt the flow of the speech input. In contrast, eye movements provide a sensitive, implicit measure of spoken language processing in which the response is closely time-locked to the input without interrupting the speech stream. Second, the eye movement paradigm can be used with simple natural tasks such as
picking up and moving objects, making it well suited for studies with young children (Trueswell et al., 1999) and with special populations (Novick, Trueswell, & Thompson-Schill, 2005; Yee, Blumstein, & Sedivy, 2000). It also makes the paradigm well suited to investigations of language within an embodiment framework (e.g., Spivey, Richardson, & Fitneva, 2004). Third, the coupling of a visual world with language makes it possible to ask questions about real-time interpretation, especially questions about reference that would be difficult to address, and perhaps intractable, if one were limited to measures of processing complexity for written sentences or spoken utterances (cf. Sedivy, Tanenhaus, Chambers, & Carlson, 1999). It also makes it possible to examine questions at the interface between language, perception, and action (see chapters in Henderson & Ferreira, 2004 and Trueswell & Tanenhaus, 2005). Fourth, eye movements can be used to study issues about the relationship between real-time message planning and utterance planning (Bock, Irwin, & Davidson, 2004; Brown-Schmidt & Tanenhaus, in press; Griffin, 2004a).

Finally, the paradigm allows one to study real-time language production and comprehension in natural tasks involving conversational interaction. This makes it possible to bridge the two dominant traditions in language processing research: the “language-as-action” tradition, which has focused on natural interactive conversation while generally ignoring questions about the time course of real-time language processing, and the “language-as-product” tradition, which has focused on the time course of processing while being primarily limited to de-contextualized language (Clark, 1992; Tanenhaus & Trueswell, 2005).

As with any new paradigm, excitement about novel findings and new arenas of investigation must be tempered with concerns about the nature of the paradigm itself, including task-specific strategies, and the assumptions that link the behavioral measure to the hypothesized underlying mechanisms. This chapter is divided into four sections. The first section provides an introduction to how eye movements are used to study spoken language processing, beginning with a brief review of some of the foundational studies and concluding with an overview of how eye movement data are analyzed. The second section makes some observations about similarities and differences between how eye movements are used to study similar psycholinguistic issues in reading and spoken language comprehension. The third section raises issues about when using a circumscribed visual world might introduce distortions that limit the degree to which results from visual world studies can be generalized to language processing in less constrained settings. The chapter then concludes with a brief overview of the other contributions to this part of the volume.

1. Some foundational studies

1.1. Comprehension

The use of eye movements as a tool for studying spoken language comprehension was pioneered by Roger Cooper (1974) in a remarkable article, presciently titled The control
of eye fixation by the meaning of spoken language: a new methodology for the real-time investigation of speech perception, memory and language processing. Cooper tracked participants’ eye movements as they listened to stories while looking at a display of pictures. He found that participants initiated saccades to pictures that were named in the stories, as well as pictures associated to words in the story. Moreover, fixations were often generated before the end of the word.

Tanenhaus, Spivey-Knowlton, Eberhard, and Sedivy (1995) initiated the recent surge of interest in what they dubbed the “visual world paradigm”, now sometimes referred to as the action-based version of the visual world paradigm. Taking advantage of the advent of accurate lightweight head-mounted eye-trackers, they examined eye movements as participants followed instructions to perform simple tasks with objects in a workspace. They found that varying the number of potential referents for a temporarily ambiguous prepositional phrase (e.g., Put the apple on the towel in the box in a display containing four objects, a towel, a box, an apple on another towel, and either a second apple, or another object such as a pencil) determined whether the ambiguous prepositional phrase (on the towel) was initially parsed as a goal argument (where to put the apple) or as a modifier (the location of the apple to be moved), as predicted by Altmann and Steedman (1988). (A more complete report of the results from Tanenhaus et al. (1995) is presented in Spivey, Tanenhaus, Eberhard, & Sedivy, 2002.)

Trueswell et al. (1999) replicated the Tanenhaus et al. (1995) study with adults and with five- and eight-year-old children, finding important developmental differences. The developmental differences, and the fact that the paradigm could be adapted for use with young children, have laid the foundation for studies of online sentence processing in preliterate children.

Eberhard, Spivey-Knowlton, Sedivy, & Tanenhaus (1995) demonstrated that fixations to entities referred to in an instruction are remarkably time-locked to the unfolding utterance, with fixations to a target referent in a display of competitors occurring as soon as continuous integration of constraints provided by the unfolding speech and the visual display could, in principle, distinguish the referent from its competitors. These results were obtained both for simple instructions (Touch the starred red square) and for complex instructions (Put the five of hearts that’s below the eight of clubs above the three of diamonds). This “point-of disambiguation” logic is widely used in studies of reference resolution (e.g., Sedivy et al., 1999; Chambers, Tanenhaus, Eberhard, Filip, & Carlson, 2002). As in studies of eye movements in natural tasks in vision (Hayhoe & Ballard, 2005), for the complex instructions, a high proportion of the saccades are generated to entities at the point in the instruction where they become task relevant.

Building on initial results by Spivey-Knowlton (1996), Allopenna, Magnuson, & Tanenhaus (1998) demonstrated that the timing of fixations to a pictured referent, and competitors with different types of phonological overlap, was sufficiently time-locked to the input to trace the time course of lexical access. Allopenna et al. also showed that a simple linking hypothesis could be used to map fixations onto computational models of lexical activation, thus laying the foundation for the growing body of work that uses the visual world paradigm to study spoken word recognition.
Altmann and Kamide (1999) made an important addition to the visual world paradigm by demonstrating linguistically mediated anticipatory eye movements using a task like Cooper’s in which the participants’ primary task was to listen to a description of an upcoming event involving entities depicted in a display. As participants heard sentences such as the boy will eat the cake, they made anticipatory eye movements to a picture of cake before the offset of eat, when the other depicted objects were not eatable. Anticipatory eye movements are now widely used as a dependent measure, typically with this so-called, passive listening, variant of the visual-world paradigm.

1.2. Production

The foundation for eye movement studies of language production comes from two studies: one by Meyer, Sleiderink, and Levelt (1998), the other by Griffin and Bock (2000). Meyer et al. had participants name sequences of objects. Eye gaze was tightly coordinated with the speech. Participants fixated a to-be-named object about 1 s prior to the onset of naming. This eye–voice lag is similar to the time it takes to initiate naming an object in isolation (Rossion & Pourtois, 2004; Snodgrass & Yuditsky, 1996), suggesting that the eye–voice delay reflects word preparation. About 150 ms before the onset of speech, participants launched a saccade to the next object.

Griffin and Bock (2000) presented participants with a simple event rendered as a line drawing that could be described with either an active or passive sentence, such as a woman shooting a man, or lightening striking a church. The sequence of eye movements reflected the order of constituents in the utterance. Speakers looked at pictured objects about 800 ms to 1 s before naming them. Once speaking began, the sequence and timing of fixations was controlled by the utterance, rather than perceptual properties of the input, suggesting that the speaker had completed message planning prior to beginning to speak (also see Bock, Irwin, Davidson, & Levelt, 2003).

The contribution by Wheeldon, Meyer, and van der Meulen (this volume) builds on the line of research initiated by Meyer et al. to ask whether fixation patterns differ for correct pronunciations and pronunciations with anticipation errors. They concluded that the timing of fixations does not differ for correctly and incorrectly produced names.

2. Data analysis and linking assumptions

In order to briefly describe how eye movement data are analyzed in comprehension, I will use Experiment 1 from Allopenna et al. (1998). This experiment will also prove useful for subsequent discussion of some of the methodological concerns that arise in visual world studies. Allopenna et al. (1998) evaluated the time course of activation for lexical competitors that shared initial phonemes with the target word (e.g., beaker and beetle) or that rhymed with the target word (e.g., beaker and speaker). Participants were instructed to fixate a central cross and then followed a spoken instruction to move one of
four objects displayed on a computer screen with the computer mouse (e.g., Look at the cross. Pick up the beaker. Now put it above the square).

2.1. Data analysis

A schematic of a sample display of pictures is presented in Figure 1, upper left. The pictures include the target (the beaker), the cohort (the beetle), a picture with a name that rhymes with the target (speaker), and the unrelated picture (the carriage). For purposes of illustrating how eye movement data are analyzed, we will restrict our attention to the target, cohort, and unrelated pictures. The particular pictures displayed are used to exemplify types of conditions and are not repeated across trials.

Five hypothetical trials are shown in the upper right portion of the figure. The 0 ms point indicates the onset of the spoken word *beaker*. The grey line begins at about 200 ms – the earliest point where one would expect to see signal-driven fixations. On trial one, the hypothetical participant initiated a fixation to the target about 200 ms after the onset of the word, and continued to fixate on it (typically until the hand brings the mouse onto the target). On trial two, the fixation to the target begins a bit later. On trial three, the first fixation is to the cohort, followed by a fixation to the target. On trial four, the first fixation is to the unrelated picture. Trial five shows another trial where the initial fixation is to the cohort. The graph in the lower right hand portion of the figure illustrates the proportion of fixations over time for the target, cohort, and unrelated pictures, averaged across trials and participants. These fixation proportions are obtained by determining the

![Figure 1. Schematic illustrating proportion of fixation curves. (See Color Plate 2.)](image-url)
proportion of looks to the alternative pictures at a given time slice and they show how the pattern of fixations change as the utterance unfolds. The fixations do not sum to 1.0 as the word is initially unfolding because participants are often still looking at the fixation cross.

Proportion of fixation curves would seem to imply that eye movements provide a continuous measure – a misconception that my colleagues and I may have sometimes contributed to in some of our papers. It is more accurate to say that eye movements provide an approximation to a continuous measure. For example, the assumption behind linking fixations to continuous word recognition processes is that as the instruction unfolds, the probability that the listener’s attention will shift to a potential referent of a referring expression increases with the activation of (evidence for) its lexical representation, with a saccadic eye movement typically following a shift in visual attention to the region in space where attention has moved. Because saccades are rapid, low cost, low-threshold responses, a small proportion of saccades will be generated by even small increases in activation, with the likelihood of a saccade increasing as activation increases. Thus, while each saccade is a discrete event, the probabilistic nature of saccades ensures that with sufficient numbers of observations, the results will begin to approximate a continuous measure. For an insightful discussion, including the strengths and weaknesses of eye movements compared to a truly continuous measure, tracking the trajectories of hand movements, see Spivey, Grosjean, and Knoblich (2005) and Magnuson (2005).

Researchers often define a window of interest, as illustrated by the rectangle in the proportion graph. For example, one might want to focus on the fixations to the target and cohort in the region from 200 ms after the onset of the spoken word to the point in the speech stream where disambiguating phonetic information arrives. The proportion of fixations to pictures or objects and the time spent fixating on the alternative pictures (essentially the area under the curve, which is a simple transformation of proportion of fixations), can then be analyzed. Because the duration of each fixation is likely to be 150–250 ms, the proportion of fixations in different time windows is not independent. One way of increasing the independence is to restrict the analysis to the proportion of new saccades generated to pictures within a region of interest. In future research it will also be important to explore using additional statistical techniques that are designed to deal with dependent measures that contain temporal dependencies.

Finally, there are other potential measures that can provide additional information. For example, the duration of fixations to a picture or entity in a scene has proved particularly useful in production studies. Comprehension researchers have occasionally restricted analyses to the initial saccade, especially in tasks where the participant is likely to be looking at a fixation point when the critical input arrives. In addition, as discussed later, one can examine a variety of contingent fixations.

To date, there has been relatively little debate about the relative merits of different measures, especially in comparison to eye movement reading studies (but, cf. Altmann & Kamide, 2004). One likely reason is that the visual world paradigm is new enough so that there have been few, if any, studies that I am aware of where investigators have examined similar issues, but come to different conclusions, using different measures.
A second reason is that examining timing may be more straightforward in visual world studies than in reading studies for reasons that we will discuss later.

In Figure 2, the graph in the upper left quadrant shows the data from the Allopenna et al. (1998) experiment. The figure plots the proportion of fixations to the target, cohort, rhyme, and unrelated picture. Until 200 ms, nearly all of the fixations are on the fixation cross. These fixations are not shown. The first fixations to pictures begin at about 200 ms after the onset of the target word. These fixations are equally distributed between the target and the cohort. These fixations are remarkably time-locked to the utterance: input-driven fixations occurring 200–250 ms after the onset of the word are most likely programmed in response to information from the first 50–75 ms of the speech signal. At about 400 ms after the onset of the spoken word, the proportion of fixations to the target began to diverge from the proportion of fixations to the cohort. Subsequent research has established that cohorts and targets diverge approximately 200 ms after the first phonetic input that provides probabilistic evidence favoring the target, including coarticulatory information in vowels (Dahan, Magnuson, Tanenhaus, & Hogan, 2001; Dahan & Tanenhaus, 2004).

Figure 2. Top left: data from Allopenna, et al., (1998); Bottom left: Trace simulations; Top right: linking hypothesis; Bottom right predicted proportion of fixations.
Shortly after fixations to the target and cohort begin to rise, fixations to rhymes start
increase relative to the proportion of fixations to the unrelated picture. This result
discriminates between predictions made by the cohort model of spoken word recognition
and its descendents (e.g., Marslen-Wilson, 1987, 1990, 1993), which assume that any
feature mismatch at the onset of a word is sufficient to strongly inhibit a lexical candidate,
and continuous mapping models, such as TRACE (McClelland & Elman, 1986), which
predict competition from similar words that mismatch at onset (e.g., rhymes). The results
strongly confirmed the predictions of continuous mapping models.

2.2. Formalizing a linking hypothesis

We can now illustrate a simple linking hypothesis between an underlying theoretical
model and fixations. The assumption providing the link between word recognition and
eye movements is that the activation of the name of a picture determines the probability
that a subject will shift attention to that picture and thus make a saccadic eye movement
to fixate it. Allopenna et al. formalized this linking hypothesis by converting activations
into response strength, following the procedures outlined in Luce (1959). The Luce choice
rule is then used to convert the response strengths to response probabilities. The graph
in the lower left quadrant of Figure 2 shows the activation values for beaker, beetle,
carriage, and speaker, generated by a TRACE simulation. The equations used in the
linking hypothesis are shown in the upper right hand quadrant of the figure.

The Luce choice rule assumes that each response is equally probable when there is no
information. Thus when the initial instruction is look at the cross or look at picture X,
we scale the response probabilities to be proportional to the amount of activation at each
time step using the equations presented in the top right hand corner of the figure, where
max_act is the maximum activation at a particular time step, max_act_overall is a constant
equal to the maximum expected activation (e.g., 1.0), i is a particular item and d_t is the
scaling factor for each time step. Thus the predicted fixation probability is determined
both by the amount of evidence for an alternative and the amount of evidence for that
alternative compared to the other possible alternatives.

Finally, we introduce a 200 ms delay because programming an eye movement takes
approximately 200 ms (Matin, Shao, & Boff, 1993). In experiments without explicit
instructions to fixate on a particular picture, initial fixations are randomly distributed
among the pictures. Under these conditions, the simple form of the choice rule can be used
(see Dahan, Magnuson, & Tanenhaus, 2001; Dahan, Magnuson, Tanenhaus, & Hogan
2001). Note that the Allopenna et al. formalization is only an approximation to a more
accurate formalization of the linking hypothesis which would predict the probability that a
saccade would be generated at a particular point in time, contingent upon (a) the location
of the previous fixation and perhaps the several preceding fixations; (b) time from the
onset of the last fixation and (c) the current goal state of the listener’s task – which can
be ignored in a simple “click” task like the Allopenna et al. paradigm. In a more complex
task, such as assembling a piece of furniture or preparing a recipe, goal states might
have a more complex structure, with several sub-goals competing for the capture of local attention at any point in time.

When the linking hypothesis is applied to TRACE simulations of activations for the stimuli used by Allopenna et al., it generates the predicted fixations over time shown in Figure 2, bottom right hand corner. Note that the linking hypothesis transforms the shape of the functions because it introduces a non-linear transformation. This highlights the importance of developing and using explicit linking hypotheses (see Tanenhaus, 2004), when evaluating the goodness of fit with behavioral data, as opposed to merely assuming a monotonic relationship. The fixations over time to the target, the cohort competitor, and a rhyme competitor closely match the predictions generated by the hypothesis linking activation levels in TRACE to fixation proportions over time.

2.3. Action-contingent analyses

One useful feature of the action-based approach is that the behavioral responses reveal the participants’ interpretation. This allows for interpretation-contingent analyses in which fixations are analyzed separately for trials on which participants choose a particular interpretation. Two recent applications illustrate how interpretation-contingent analyses can be used to distinguish between competing hypotheses.

McMurray et al. (2002; McMurray, Aslin, Tanenhaus, Spivey, & Subik, 2005) used a variation on the Allopenna et al. task to investigate the hypothesis that lexical processing is sensitive to small-within category differences in Voice-Onset Time (VOT). The stimuli were synthesized minimal pairs that differed only in voicing, such as bomb/palm and peach/beach. VOT varied in steps sizing from 0 to 40 ms (voiced sounds, such as /b/ have shorter VOTs than unvoiced sounds such as /p/). McMurray et al. found gradient increases in looks to the cross-category competitor as the VOT moved closer to the category boundary, with clear linear trends even when the trials with VOTs abutting the category boundary were excluded. While these results are consistent with the hypothesis that lexical processing is sensitive to within-category variation, the results could also be accounted for without abandoning the traditional assumption that within-category variation is quickly discarded.

The categorical interpretation would go as follows. If we make the plausible assumption that there is noise in the system, then as VOT approaches the category boundary, listeners are more likely to incorrectly categorize the input. Assume a category boundary of approximately 18 ms, which is what McMurray et al. found with their synthesized stimuli. For trials with a VOT of 20 ms, perhaps 20% of the stimuli might be perceived as having a VOT of less than 18 ms. With a VOT of 25 ms, the percentage might drop to 12%, compared to 8% for trials with a VOT of 30 ms and 4% for a VOT of 35 ms etc. Thus, the proportion of looks to the cross-category competitor might increase as VOT approaches the category boundary because the data will include increasingly more trials where the target word was misheard as the cross-category competitor and not because the underlying system responds in a gradient manner.
McMurray et al. were able to rule out this alternative explanation by filtering any trials where the participant clicked on the cross-category picture. For example, if the VOT was 25 ms, and the participant clicked on the picture of the bomb, rather than the palm, then the eye movement data from that trial would be excluded from the analyses. McMurray et al. found that looks to the cross-category competitor increased as VOT approached the category boundary, even when all “incorrect” responses were excluded from the analyses, thus providing strong evidence that the system is indeed gradient.

A second illustration comes from recent studies by Runner and his colleagues (e.g., Runner, Sussman, & Tanenhaus, 2003; in press) investigating the interpretation of reflexives and pronouns in so-called picture noun phrases with possessors, e.g., \textit{Harry admired Ken’s picture of him/himself}. Participants were seated in front of a display containing three male dolls, Ken, Joe, and Harry, each with distinct facial features. Digitized pictures of the doll’s faces were mounted in a column on a board directly above each individual doll. The participant was told that each doll “owned” the set of pictures directly above him; that is, the three pictures in the column above Joe were Joe’s pictures, the pictures in the column above Ken were Ken’s pictures, etc.

Binding theory predicts that the reflexive, \textit{himself}, will be interpreted as referring to Ken’s picture of Ken in instructions such as \textit{Pick up Harry. Now have Harry touch Ken’s picture of himself}. Runner et al. found that looks to both the binding-appropriate and inappropriate referents began to increase compared to an unrelated picture in the same row, beginning about 200 ms after the onset of the reflexive. This result suggests that both binding-appropriate and inappropriate referents are initially considered as potential referents for a reflexive. Moreover, participant’s choices showed frequent violations of classic binding for reflexives: on approximately 20% of trials with reflexives, participants had Harry touch Ken’s picture of Harry. However, one might argue that the early looks to binding-inappropriate referents came from just those trials on which the participant mistakenly arrived at the “incorrect” interpretation. Runner et al. were able to rule out this interpretation by analyzing just those trials where the participant made the binding-appropriate response, finding that there was still an increase in looks to the inappropriate referent, compared to controls. Thus, both the binding-appropriate and inappropriate referents compete, with the binding-appropriate referent (probabilistically) preferred.

One clear limitation of the action-based paradigm is that the linguistic stimuli must be embedded in instructions, which can limit the experimenter’s degrees of freedom. Beginning with Altmann and Kamide (1999), a number of researchers using eye movements to study spoken language comprehension have begun to use variations on the original Cooper procedure in which participants simply listen to the input (e.g., Arnold et al., 2000; Boland, 2005; Knoeferle et al., 2005). This procedure places fewer constraints on both the experimenter and the participant. Without an explicit action, the participant’s attention is less constrained, thus increasing the likelihood of anticipatory eye movements, which are extremely useful for inferring expectations generated by the listener. Some important applications of listening without an explicit action include Kaiser and Trueswell (2004), who examined expectations driven by different word orders, Boland
(2005), who compared verb-based expectations for adjuncts and arguments, the line of research initiated by Altmann and colleagues (e.g., Altmann & Kamide, 1999; Kamide, Altmann, & Haywood, 2003) and the effects of visually based information on expectation about thematic role assignment initiated by Knoeferle and her colleagues (Knoeferle et al., 2005; Knoeferle, this volume).

Few studies have directly compared the action-based and non-action-based versions of the paradigm. However, to a first approximation, it appears that when anticipatory eye movements are excluded, the timing of fixations to potential referents may be slightly delayed in passive listening tasks compared to action-based tasks. The data from action-based tasks is somewhat cleaner than the data from listening tasks, perhaps because a high proportion of the fixations are task-relevant.

2.4. Other contingent analyses

Investigators have also used a variety of other contingent analyses. For example, one can compare the time to launch a saccade to a referent contingent upon the participant having fixated that referent or a potential competitor during preview (Dahan & Tanenhaus, 2005; Dahan et al., this volume). Likewise, one can compare the time to launch a fixation from a competitor, once disambiguating information arrives.

In production, one can also examine the timing of the onset of an utterance, or the form of an utterance, contingent upon whether or not the participant has looked at a critical part of the display. For example, Brown-Schmidt and Tanenhaus (in press) examined the timing and form of referential descriptions with contrast (e.g., the little peach) contingent upon whether or not, and, if so, when the participant (first) fixated the contrast referent (e.g., another bigger peach). Griffin (2004b) has compared the duration of fixations to pictures as a function of whether or not the picture was named correctly, a strategy also adopted by Wheeldon et al., this volume to compare fluent and disfluent productions.

3. Comparing visual world and eye movement reading studies

Many of the issues that have been investigated for decades using eye movements in reading, in particular issues in lexical processing and sentence processing are now being investigated using eye movements with spoken language. Although some aspects of these processes will differ in reading and spoken language because of intrinsic differences between the two modalities, scientists investigating issues such as syntactic ambiguity resolution and reference resolution using eye movements in reading and eye movements in spoken language believe they are testing theoretical claims about these processes that transcend the modality of the input. Thus, the psycholinguistic community will increasingly be faced with questions about how to integrate results from visual world studies with results from studies of eye movements in reading and sometimes how to reconcile conflicting results.
This section addresses some of these issues of interpretation by working through two examples of how similar questions would be asked in eye-tracking reading and eye-tracking listening studies. We then turn to some cases where the results support different conclusions. This will lead us into a discussion about ways of addressing closed set issues within the visual world paradigm.

3.1. Processing load vs representational measures

In considering these examples, it will be useful to make a distinction between behavioral measures of language processing which measure processing difficulty and measures that probe representations. The distinction is more of a heuristic than a categorical distinction because many response measures combine aspects of both. Processing load measures assess transient changes in processing complexity, and then use these changes to make inferences about the underlying processes and representations. Representational measures examine when during processing a particular type of representation emerges and then use that information to draw inferences about the underlying processes and representations. Neither class of measure or its accompanying experimental logic is intrinsically preferable to the other; the nature of the question under investigation determines which type of response measure is more appropriate.

The majority of studies that use eye movements to examine reading make use of eye movements as a processing load measure. The primary dependent measure is fixation duration. The linking hypothesis between fixation duration and underlying processes is that reading times increase when processing becomes more difficult. In contrast, the majority of studies that use eye movements with spoken language processing use eye movements as a representational measure. The primary dependent measure is when and where people fixate as the utterance unfolds.

3.2. Lexical ambiguity

In a well-known series of studies, Rayner and colleagues (e.g., Rayner & Duffy, 1986) have examined whether multiple senses of homographs, such as bank, ball and port are accessed during reading, and if so, what are the effects of prior context and the frequency with each sense is used. Processing difficulty compared to an appropriate control is used to infer how ambiguous words are accessed and processed. For “balanced” homographs with two more or less equally frequent senses, fixation duration is longer compared to frequency-matched controls, resulting in the inference that the multiple senses are competing with one another. This ambiguity “penalty” is reduced or eliminated for biased homographs when a “dominant” sense is far more frequent than a “subordinate” sense and when the context strongly favors either one of two equally frequent senses or the more frequent sense. Note that while these results do not provide clear evidence about time course per se, the overall data pattern allows one to infer that multiple senses are accessed, with the dominant sense accessed more rapidly. One can get crude time course information by separately analyzing the duration of the initial fixation and using that as a
measure of relatively early processes. More detailed information about time course can be obtained by using fixation duration as a measure, but using variations on the fast priming methods, introduced by Sereno and Rayner (1992).

Studies using the visual world paradigm adopt a similar approach to that used by Allopenna et al. Potential referents associated with the alternative senses are displayed and the time course of looks to these referents is used to infer degree of activation and how it changes over time. For balanced homophones, looks are found to the referents of both senses. For biased homophones, looks to the more frequent sense begin earlier than looks to the less frequent sense (Huettig & Altmann, 2004). This pattern is similar to those obtained in classic studies using cross-modal priming from the 1970s and early 1980s (Swinney, 1979; Tanenhaus, Leiman, & Seidenberg, 1979; for review see Simpson, 1984 and Lucas, 1999). These results do not provide direct information about processing difficulty, though one might infer that competing senses would result in an increase in complexity.

Thus, while the eye movements reading studies do not provide direct information about time course and visual world studies do not provide direct information about processing difficulty, the results from reading studies that use a processing load strategy and visual world studies that probe emerging representations converge on the same conclusions.

3.3. Syntactic ambiguity

Beginning with the classic article by Frazier and Rayner (1982), eye tracking in reading has been the response measure of choice for psycholinguists interested in syntactic processing. Frazier and Rayner’s approach was to examine the processing of temporarily ambiguous sentences, using reading times within pre-defined regions to infer if and when the reader had initially pursued the incorrect interpretation. For syntactic ambiguities, which involved disambiguating the phrase that could be “attached” to a verb phrase in a subordinate clause (the initially preferred interpretation) as the subject of the main clause, Frazier and Rayner found an increase in fixation duration and an increase in regressive eye movements from the disambiguating region. For current purposes we will focus on fixation duration because it is most clearly a processing load measure. The question of how to interpret regressions is more complex and beyond the scope of this chapter. The increase in fixation duration was interpreted as evidence that processing had been disrupted, thereby leading to the inference that readers had initially chosen the argument interpretation. Frazier and Rayner also introduced several different measures that divided fixations within a region in different ways. For example, “first pass” reading times include all fixations beginning with the first fixation within a region until a fixation that leaves a region, and are often used as a measure of early processing.

Studies examining syntactic ambiguity resolution with the visual world paradigm use the timing of looks to potential referents to infer, if and, if so, when, a particular analysis is under consideration. For example, in one-referent contexts (an apple on a towel, a towel, a box and a pencil) and instructions such as Put the apple on the towel in the box, Spivey et al. (2002) found that looks to the false goal (the towel without the apple)
began to increase several hundred milliseconds after the onset of *towel*. In contrast, in two-referent contexts (two apples, one on a towel and one on a napkin), fixations to the apple on the towel begin to increase several hundred milliseconds after the onset of *towel*. This pattern of results suggests that the prepositional phrase *on the towel* is initially considered a goal argument in the one-referent context and a noun-phrase modifier in the two-referent context. Information about time course is relatively straightforward with the visual world logic because fixations can be aligned with the input, allowing strong inferences about what information in the input was likely to have triggered the fixation. The reason one can align fixations and the input is, of course, that the input unfolds sequentially. Nonetheless, as will be discussed shortly, timing is complicated by issues of when and how the participant has encoded information from the display or scene.

Timing is less straightforward in eye-tracking reading when the measure is fixation duration and the fixations are divided into multiple word regions. Most of the complexities in inferring time course in reading studies arise because the sequence of fixations need not correspond to the linear order of the words, including when they have first been encountered. This is especially the case when one considers that arguments about timing often depend on defining regions of text and then partitioning fixations into categories in ways that separate the measure from when the input was first encountered.

One way to appreciate this is to compare results from a recent study by Clifton et al. (2003) with Trueswell, Tanenhaus, and Garnsey (1994). Both studies examined the effects of thematic bias on reading times to temporarily ambiguous, reduced-relatives clauses compared to unambiguous full-relative baselines. Trueswell et al. concluded that thematic constraints had immediate effects, whereas Clifton et al. argued that initial syntactic processing was unaffected by thematic context. When the same measures are compared, e.g., first pass reading times, Clifton et al. replicate the results found by Trueswell et al. However, when fixations are examined using a different measure (regression path), the evidence is somewhat more consistent with delayed effects of context. Which conclusion one endorses depends on the measure one emphasizes. Note, however, that it is difficult to make strong time course claims unless one includes only those fixations that preserve the order in which the information is first encountered (i.e., fixations on a word after a subsequent word has been fixated would be excluded and the regions would correspond to single words). Nonetheless, if we exclude the effects of context, which we will return to shortly, the conclusions that emerge for reading and visual world studies are similar for the prepositional phrase attachment structures which have been examined using both methods. The preferred analysis that is inferred from looks to potential referents is similar to the preferred analysis that is inferred by fixation durations.

4. Effects of display

On the basis of the two ambiguity cases we have compared thus far, it might seem that I am arguing that visual world studies have intrinsic advantages over reading studies
when it comes to asking questions about the time course with which interpretations arise because of two factors: (1) fixations and the input can be easily aligned and (2) the logic of representational measures is more direct. However, we have not yet considered the single factor that most complicates the interpretation of visual world studies of language processing – the need to use a display. The use of a visual display is what allows one to use fixations to probe emerging representations. However, it introduces two kinds of complexities. First, the encoding of the display can introduce contingencies. For example, the timing of looks to a potential referent at point $t$ could be affected by whether or not that referent has been fixated on during time $t-x$, either during preview or as the sentence unfolds. Thus the likelihood of a fixation may be contingent on both the input and the pattern of prior fixations. This, of course, has the potential to complicate inferences about time course, in much the same way that re-reading after a regression can complicate the interpretation of fixation duration data in eye movement reading studies. The contribution to this volume by Dahan, Tanenhaus, and Salverda begins to address this issue by examining how having fixated a potential referent during preview affects the likelihood that it will be fixated when it is temporarily consistent with the input, viz. its name is a cohort of the intended referent.

The second factor is that use of a display with a small number of pictured referents or objects and a limited set of potential actions creates a more restricted environment than language processing in most natural contexts, while at the same time imposing more demands on the participant than most psycholinguistic tasks. In order to address these closed set issues, we will consider two cases: the first from spoken word recognition and the second from reference resolution.

### 4.1. Spoken word recognition

In the Allopenna et al. paradigm, the potential response set on each trial is limited to four pictured items. If participants adopted a task-specific verification strategy, such as implicitly naming the pictures, then the unfolding input might be evaluated against these activated names, effectively bypassing the usual activation process, and leading to distorted results. Even if participants do not adopt such a strategy, the visual world methodology might be limited if the effects of the response alternatives mask effects of non-displayed alternatives (e.g., neighborhood effects in the entire lexicon). This would restrict its usefulness for investigating many issues in spoken word recognition, in particular issues about the effects of lexical neighborhoods, i.e., the set of words in the lexicon that are similar to the target word. Here, an analogy might be helpful. Researchers often use lexical priming paradigms to probe for whether an exemplar of a particular class of lexical competitor is active, e.g., cohorts or rhymes. However, these paradigms are not well suited for asking questions about the aggregate effects of the number and frequency of potential competitors. In order to investigate this class of question, researchers have found it more useful to measure response time to a target word, e.g., auditory lexical decision, which more closely approximates a processing load measure.
4.2. Implicit naming

The issue of implicit naming has been addressed most directly by Dahan and Tanenhaus (2005) in a study that varied the amount of preview time, 300 or 1000 ms, for four-picture displays with minimal phonological overlap between the names of the distractors and the target. On a subset of the trials, two of the pictures were visually similar (e.g., a picture of a snake and a coiled rope) and the instruction referred to one of the pictures (e.g., *click on the snake*). The particular pictures chosen as the two referents shared some features associated with a prototypical visual representation of one or both words. For example, the pair *snake–rope* was selected because the picture of a coiled rope shares some features with the visual representation most often associated with the concept of a snake. When selecting pictures, we sought to minimize their visual similarity so that the objects could be easily differentiated. For example, we chose a snake in a non-coiled position and a rope that was coiled. Thus, visual similarity was maximized between the prototypical visual representation of one of the concepts, the referent, and the picture associated with the other concept, the competitor, and minimized between the competitor picture and the picture of the referent concept.

Several aspects of the results provide strong evidence against implicit naming. With longer preview, one would expect increased likelihood of implicit naming. However, preview duration did not affect the magnitude of visual similarity effects (looks to visually similar competitors). Moreover, even in the 1000 ms condition, the magnitude of visual similarity effects was not affected by whether or not the competitor was fixated during preview; the naming hypothesis predicts that effects would be eliminated or weakened with preview because the encoded name of the picture would not match the unfolding target. Finally, similarity effects were larger when the target had a competitor that was chosen to share visual features of its prototype representation compared to when that competitor was the referent. Thus visual similarity effects were due to the fit between the picture and the conceptual representation of the picture, not simply surface visual confusability. This suggests that mapping of the word onto its referent picture is mediated by a visual/conceptual match between the activated lexical form of the target and the picture. This hypothesis is further supported by additional analyses of the effects of fixation to a competitor during preview on the likelihood that it will be re-fixated during the speech input reported by Dahan et al. in this book and evidence that a spoken word triggers looks to potential referents when the participant is engaged in a visual search task to identify the location of a dot when it appears on a random location within a schematic scene (Salverda & Altmann, 2005).

4.3. Sensitivity to hidden competitors

Perhaps the strongest test of the sensitivity of visual world studies comes from studies that look for effects of non-displayed or “hidden competitors”. A recent study by Magnuson, Dixon Tanenhaus, and Aslin (in press) illustrates this approach. Magnuson et al. examined the temporal dynamics of neighborhood effects using two different metrics: neighborhood...
density, a frequency-weighted measure defined by the neighborhood activation model (Luce & Pisoni, 1998), and a frequency-weighted measure of cohort density. The referent was displayed along with three semantically unrelated pictures, with names that had little phonological overlap with the referent (all names were monosyllabic). Crucially, none of the referent’s neighbors were either displayed or named throughout the course of the experiment. The results showed clear effects of both cohort and neighborhood density, with cohort density effects dominating early in the recognition process and neighborhood effects emerging relatively late. Proportion-of-fixation curves showing the effects of frequency, cohort density and neighborhood density are presented in Figure 3.

These results demonstrate that the processing neighborhood for a word changes dynamically as the word unfolds. It also establishes the sensitivity of the paradigm to the entire lexicon. To a first approximation then, when competitors are displayed, the paradigm can be used to probe specific representations, however, the aggregate effects of competitors can be observed in the timing of fixations to the target referent.

Magnuson et al.’s results complement Dahan, Magnuson, Tanenhaus, and Hogan’s (2001) finding that misleading coarticulatory information delays recognition more when it renders the input temporarily consistent with a (non-displayed) word, compared to when it does not. In addition, simulations using the Allopenna et al. linking hypothesis successfully captured differences between the effects of misleading coarticulatory information with displayed and non-displayed competitors. Whether the non-displayed competitor logic can be extended to higher-level sentence processing remains to be seen.

4.4. Sentence processing

Much trickier issues about the effects of the display come into play in higher-level processing. For example, one could argue that in the Tanenhaus et al. (1995) study displaying an apple on a towel and an apple on a napkin increases the salience of a normally less accessible sense compared to circumstances where the alternative referents are introduced linguistically. One could make a similar argument about the effects of action on the rapidity with which object-based affordances influence ambiguity resolution in studies by Chambers and colleagues (Chambers et al. 2002, 2004). In these studies, the issue of implicit naming seems prima facie to be less plausible. However, one might be concerned about task-specific strategies. For example, in Chambers et al. (2002), participants were confused, as indexed by fixations when they were told to *Pick up the cube. Now put the cube in the can*, and there were two cans. The confusion was reduced or eliminated, however, when the cube would only fit in one of the cans.

Note that one might attribute this to problem solving, and not as Chambers et al. argued to the effects of action and affordance on referential domains. The problem-solving argument would be that participants were able to ignore the competitor as soon as they picked up the cube (and learned that they were supposed to put it in something) because they were simply looking for a possible goal. However, if this were the case, then the same pattern of results would be predicted to occur with the instruction used an indefinite article, e.g., *Pick up the cube. Now put it in a can*. However, participants were now
Figure 3. Top to bottom, Time course of frequency effects, cohort density and neighborhood density. From Magnuson, Dixon, Tanenhaus, and Aslin (in press).
confused when the cube would only fit in one of the cans, despite the fact that there was still only one possible action. Thus the referential domain in conjunction with the referential conditions of the article influenced processing even though the article had no effect on the possible action. This strategy of pitting linguistic effects against potential problem-solving strategies that would bypass linguistic processing is one way to evaluate whether or not the task and the display introduce problematic task-specific processing strategies.

Perhaps the most general caution for researchers using the visual world paradigm in both production and comprehension is to be aware that while the visual world displays entities that can be used to infer the representations that the listener is developing, it also serves as a context for the utterance itself. Note that the fact that information in a display affects processing is not itself any more problematic than the observation that reference resolution, e.g., is affected by whether or not potential referents are introduced linguistically in a prior discourse. Both types of contexts can make potential referents salient.

One sometimes encounters the argument that the visual world paradigm can be informative about language processing only if gaze patterns to a potential referent in a display are not affected by the other characteristics of the display. This argument is no more or less valid than the comparable argument that fixations in reading can only inform us about word recognition or reference resolution if fixations to a word are unaffected by the context in which the fixated word occurs. What is crucial, however, is whether the nature of the interactions with the display sheds light on language processing or whether it introduces strategies that mislead or obscure the underlying processes.

Two examples might help illustrate this point. The first is taken from Tanenhaus, Chambers, and Hanna (2004, see also Tanenhaus & Brown-Schmidt, to appear) and illustrates how the visual world paradigm has clarified our understanding of pre-nominal scalar adjectives and definite reference. The second uses another hypothetical example to contrast examples of results that would and results that would not suggest that the visual display was leading to strategies that distort “normal” language comprehension.

Prior to visual world studies, a standard psycholinguistic account of the processing of the sentence, After putting the pencil below the big apple, James moved the apple onto the towel, would have gone something like this. When the listener encounters the scalar adjective big, interpretation is delayed because a scalar dimension can only be interpreted with respect to the noun it modifies (e.g., compare a big building and a big pencil). As apple is heard, lexical access activates the apple concept, a prototypical apple. The apple concept is then modified resulting in a representation of a big apple. When apple is encountered in the second clause, lexical access again results in activation of a prototypical apple concept. Because apple was introduced by a definite article, this representation would need to be compared with the memory representation of the big apple to decide whether the two co-refer.

Now, consider how real time interpretation proceeds in a context which includes a pencil, two apples, one small prototypical red apple, the other a large misshapen green apple, and a towel, taking into account recent results from visual world studies. At big,
the listener’s attention would be drawn to the larger of the two apples, because a scalar adjective signals a contrast among two or more entities of the same semantic type (Sedivy et al., 1999). Thus apple will be immediately interpreted as the misshapen green apple, even though a more prototypical red apple is present in the display. And, when the apple is encountered in the second clause, the red apple would be ignored in favor of the large green apple, because that apple has been introduced into the discourse, but not as the most salient entity (Dahan, Tanenhaus, & Chambers, 2002).

The standard account does not make sense when we try to generalize it to processing in the context of concrete referents. In contrast, the account that emerges from visual world research does generalize to processing in the absence of a more specific context. In particular, the scalar big would still be interpreted as the member of a contrast set picked out by size; it is just that the contents of the contrast set are not instantiated until the noun, apple has been heard. And, any increase in processing difficulty when the apple is processed would not be reflecting an inference to establish that the referent is the big apple, but rather the shift in discourse focus from the previously focused entity (the pencil) to a previously mentioned entity (the apple). In this example, then, the display clearly changes processing, but in ways that clarify (but do not distort) the underlying processes.

Now consider the discourse: The man returned home and greeted his pet dog. It/ The beast/A beast then began to lick/attack him. Well-understood principles of reference assignment mandate that it should refer to the dog, and a beast to an animal other than the dog. Now imagine the same discourse in the context of a display containing a man standing in front of an open door to a hut in a jungle village, a dog with a collar, a tiger, and a rabbit. Compared to appropriate control conditions, we would expect a pattern of looks indicating that it was interpreted as the dog, and a beast as the tiger, regardless of whether the verb was lick or attack. If, however, it were interpreted as the tiger when the verb was attack, or if a beast were to be interpreted as the dog for lick, then we would have a clear case of the display distorting the comprehension process. This conclusion would be merited because these interpretations would violate well-understood principles of reference resolution. Now consider the definite noun phrase, the beast. This referential expression could either refer to the mentioned entity, the pet dog, or it could introduce another entity. In the discourse-alone condition, a listener or reader would most likely infer that the beast refers to the dog, because no other entity has been mentioned. However, in the discourse with display condition, a listener might be more likely to infer that the beast refers to the tiger. Here the display changes the interpretation, but it does not change the underlying process; the display simply makes accessible a potential unmentioned referent, which is consistent with the felicity conditions for the type of definite reference used in the discourse. Indeed, we would expect the same pattern of interpretation if the tiger had been mentioned in the discourse.¹

¹ This observation is due to Gerry Altmann. The example presented here is adapted from one presented by Simon Garrod in a presentation at the 2003 Meeting of the CUNY Sentence Processing Conference.
Thus far investigations of the effects of using a display and using a task have not uncovered any evidence that the display or the task is distorting the underlying processes. To the contrary, the results have been encouraging for the logic of the visual world approach. However, it will be crucial in further work to explore the nature of the interactions between the display, the task, and linguistic processing in much greater detail. Moreover, the ability to control and understand the context in which the language is being produced and understood, which is one of the most powerful aspects of the visual world paradigm, depends in large part on developing a better understanding of these interactions.

5. Conclusion

This chapter has provided a general introduction to how psycholinguists are beginning to use eye movements to study spoken language processing. We have reviewed some of the foundational studies, discussed issues of data analysis and interpretation, and discussed issues that arise in comparing eye movement reading studies to visual world studies. We have also discussed some of the issues that arise because the visual world introduces a context for the utterance.

The following chapters each contribute to issues that we have discussed and illustrate the wide range of questions to which the visual world paradigm is now being applied. Dahan, Tanenhaus, and Salverda examine how preview affects the likelihood that a cohort competitor will be looked at as a target word unfolds, contributing to our understanding of the effects of the display on the inferences we can draw about spoken word recognition in visual world studies. Bailey and Ferreira, 2005 show how the visual world paradigm can be used to investigate expectations introduced by disfluency, extending investigations of spoken language processing to the kinds of utterances one frequently encounters in real life, but infrequently encounters in psycholinguistic experiments. Wheeldon, Meyer, and van der Meulen (this book) extend work on eye movements in production by asking whether fixations can shed light on the locus of speech errors in naming. The surprising results have important theoretical and methodological implications for the link between fixations and utterance planning. Finally Knoeferle explores how thematic role assignment is affected by when information in a scene becomes relevant, contributing to our understanding of the interplay between the scene and the unfolding language. She also provides a general framework for understanding these interactions.

Acknowledgments

This work was supported by NIH grants DC005071 and HD 27206. Thanks to Delphine Dahan, Anne Pier Salverda, and Roger Van Gompel for helpful comments.
References


Ch. 20: Eye Movements and Spoken Language Processing


Ferreira, F. & Bailey, K. G. B., This volume.


Wheeldon, L. R., Meyer, A. S., & van der Muelen, F. F. this volume.

Fixation proportions over time

<table>
<thead>
<tr>
<th>Target</th>
<th>Cohort</th>
<th>Unrelated</th>
</tr>
</thead>
<tbody>
<tr>
<td>beaker</td>
<td>beetle</td>
<td>carriage</td>
</tr>
</tbody>
</table>

Look at the cross. Click on the beaker.

Color Plate 2. Schematic illustrating proportion of fixation curves. (See Figure 1, Chapter 20, p. 449.)

Color Plate 3. Top left: Original scene. Top middle: Model-determined salient regions in the scene. Top right: Fixation locations from all participants. Bottom: Scene with salient regions and participant fixations overlaid. Red dots show participant fixations within a salient region. Red tails mark saccade paths that originated in a non-salient region. Green dots denote participant fixations outside of the salient regions. (See Figure 1, Chapter 25, p. 543.)
Chapter 21

THE INFLUENCE OF VISUAL PROCESSING ON PHONETICALLY DRIVEN SACCADIES IN THE “VISUAL WORLD” PARADIGM

DELPHINE DAHAN
University of Pennsylvania, USA

MICHAEL K. TANENHAUS and ANNE PIER SALVERDA
University of Rochester, USA
Abstract

We present analyses of a large set of eye–movement data that examine how factors associated with the processing of visual information affect eye movements to displayed pictures during the processing of the referent’s name. We found that phonetically driven fixations are affected by display preview, by the ongoing uptake of visual information, and by the position of pictures in the visual display. Importantly, lexical frequency associated with a picture’s name affects the likelihood of refixating this picture and the timing of initiating a saccade away from this picture, thus supporting the use of eye movements as a measure of lexical processing.
Eye movements have increasingly become a measure of choice in the study of spoken-language comprehension, providing fine-grained information about how the acoustic signal is mapped onto linguistic representations as speech unfolds. Typically, participants see a small array of pictured objects displayed on a computer screen, hear the name of one of the pictures, usually embedded in an utterance, and then click on the named picture using a computer mouse. Participants’ gaze location is monitored. Of interest are the saccadic eye movements observed as the name of the picture unfolds until the appropriate object is selected. Early research revealed that, as the initial sounds of the target picture’s name are heard and processed, people are more likely to fixate on an object with a name that matches the initial portion of the spoken word than on an object with a non-matching name. Moreover, fixations to matching pictures are closely time-locked to the input, with signal driven–fixations occurring as quickly as 200 ms after the onset of the word (Allopenna, Magnuson, & Tanenhaus, 1998; also see Cooper, 1974; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995).

Subsequent research has established that eye movements are a powerful tool for investigating the processes by which speech is perceived and interpreted, especially the time course of these processes. Allopenna et al. (1998) showed that the proportion of looks to each picture in the display over time can be closely predicted by the strength of the evidence that the name of the object is being heard. Strength of evidence for each object’s name was computed by transforming word activations predicted by a connectionist model of spoken-word recognition, TRACE (McClelland & Elman, 1986), into fixation proportions over time using the Luce choice rule (Luce, 1959) over the set of four word alternatives. Subsequent work has shown that eye movements are extremely sensitive to fine-grained phonetic and acoustic details in the spoken input (Dahan, Magnuson, Tanenhaus, & Hogan, 2001; McMurray, Tanenhaus, & Aslin, 2002; Salverda, Dahan, & McQueen, 2003).

The use of eye movements to visual referents as a measure of lexical processing requires the use of a circumscribed “visual world”, which is most often perceptually available to listeners before the target picture’s name is heard. This world provides the context within which the input is interpreted because, in most studies, the referent object is present on the display. Furthermore, for eye movements directed to visual referents to reflect processing of the spoken signal, information extracted from each type of stimulus must interact at some level. These aspects raise two interrelated questions: At what level does information extracted from the visual display interact with processing of the spoken input, and does the influence of the display limit the degree to which the results will generalize to less constrained situations?

Here, we lay out three possible ways by which visual and spoken stimuli might interact to constrain gaze locations. One possibility is that previewing the display before the spoken input begins provides a closed set of phonological alternatives against which a phonological representation of the speech input is later evaluated. This view assumes that the phonological forms associated with the pictured objects have been accessed before the spoken input begins, either because listeners implicitly prename the pictures to themselves or because the pictures automatically activate their names. When the spoken signal becomes available, no lexical processing per se is initiated. Instead, participants
match the phonological representation of the spoken input with the phonological form associated with each location on the display. The proportion of looks to each picture, then, reflects the goodness of match between the phonological representation of the input and the phonological form associated with the picture, bypassing normal lexical processing (see Henderson & Ferreira, 2004). For instance, participants’ fixations to the picture of a candle would reflect the close match between the phonological form /kændl/ associated with the picture’s location on the display and the first sounds of the spoken input /kæn/.

Findings from the “visual world” eye-tracking paradigm could not then be generalized to spoken-word recognition in less constrained situations.

Another possibility also assumes an implicit naming of the pictures, but differs from the first one by hypothesizing that the speech signal activates lexical-form representations. Thus, eye movements would reflect the goodness of match between activated lexical representations and phonological forms associated with each picture location. This view differs from the previous one by predicting that the degree of activation of lexical representations should modulate the probability of fixating displayed pictures. The impact of lexical factors, such as frequency and neighborhood density (i.e., the number and frequency of words that are phonologically similar to the spoken word) should be observable.

A third possibility assumes no implicit naming of the pictures. Displayed pictures would be associated with visual, and probably also conceptual, representations. The spoken input would activate lexical-form representations, which in turn activate semantic representations. Eye movements would reflect the goodness of match between activated semantic representations and the visual/conceptual representations associated with each picture location. On this view, then, the effects observed in the paradigm are not mediated by names of the pictures that have been accessed during preview.

Two findings argue against the hypothesis that speech processing in the “visual world” paradigm bypasses lexical processing. First, the probability with which a picture with a name that matches the input is fixated over time varies as a function of the lexical frequency of its name (Dahan, Magnuson, & Tanenhaus, 2001). Second, the time course of looks to the target picture is affected by the degree of match of its name to non-displayed words (Dahan, Magnuson, Tanenhaus, & Hogan, 2001), including the phonological neighborhood density of its name, i.e., the number and frequency of similar-sounding words in the lexicon (Magnuson, 2001; Magnuson, Tanenhaus, Aslin, & Dahan, 2003; Magnuson, Dixon, Tanenhaus, & Aslin, in press). These findings indicate that the set of lexical alternatives considered during the processing of the spoken word extends beyond those associated with the pictures present on the display. Taken together, these two findings demonstrate that eye movements in the “visual world” paradigm reflect the involvement of the lexicon during speech processing.

A recent result from our laboratories cannot be easily accounted for if one assumes that eye movements solely reflect the match between the activated lexical-form representations and preactivated names of the displayed objects, as assumed in the second view just described. Upon hearing a spoken word (e.g., snake), listeners are more likely to temporarily fixate on the picture of an object that shares visual features but no phonological similarity with the referent’s name (e.g., a rope) than on the picture of a visually and
phonologically unrelated object (e.g., a couch; Dahan & Tanenhaus, 2005; see Huettig & Altmann, 2004, for a similar finding). Moreover, these looks are not delayed compared to looks to pictures with matching names. Thus, the probability of fixating the displayed pictures reflects, at least to some degree, the mapping of lexical–semantic representations onto conceptual and visual representations associated with these pictures.

The findings just reviewed are important for answering questions about potential limitations of the “visual world” paradigm. However, as use of the paradigm grows, it becomes increasingly important to understand how preview and other characteristics of the display influence the nature of the picture/speech interaction. Previous research has always included some preview with the display, although its duration has varied across studies. Moreover, most previous research has not compared trials with fixations to critical pictures during preview and trials with no such fixations (one exception is Dahan and Tanenhaus [2005], who reported a similar visual-similarity effect on trials with and without a fixation to the visual-competitor picture during preview). Lexical factors that have been shown to account for the probability of fixating a critical picture overall may have differential effects when the picture was fixated before the onset of the relevant spoken input and when it was not. Another aspect that has not been examined is the position of pictures in the display. Picture positions have typically been randomized. However, the position of a picture may affect the probability that it will be fixated during preview, or merely attended without being overtly fixated. It is therefore of interest to evaluate the effect of the position of a critical picture on fixation probability. Finally, the speech/picture matching process can be based on representations associated with each picture location established prior to the speech input, whereby later influencing the probability of launching a fixation to a given location, or while the picture is being fixated, whereby affecting the duration of the current fixation. Past research has rarely, if ever, distinguished these dependent variables, let alone evaluated how lexical factors might differentially influence them. Here, we report analyses on a data set (from a study conducted in Dutch for different purposes, see Dahan & Gaskell, in press), where two factors were systematically varied: whether or not listeners were able to preview the display before the target picture’s name began, and the lexical frequency of the name of the target and the name of a displayed onset-overlapping competitor (high frequency/low frequency or low frequency/high frequency). By varying lexical frequency and preview time, we are able to examine the nature of the interaction between information extracted from the visual display and the output of phonetic processing.

1. Method

1.1. Participants

Participants were 69 college students from the University of Nijmegen, the Netherlands. Thirty-nine participants took part in the no-preview version of the experiment, and thirty in the preview version.
1.2. Materials

Twenty-eight pairs of picturable Dutch nouns overlapping at onset were selected. One of the nouns had a high frequency of occurrence, and the other, a low frequency (e.g., *koffie* [coffee] and *koffer* [suitcase]). Based on the CELEX database (Baayen, Piepenbrock, & van Rijn, 1993), the high-frequency items had an average log frequency per million of 1.7 ($\sigma = 0.6$), compared to 0.8 for the low-frequency items ($\sigma = 0.5$). In order to form a four-item display, two additional phonologically unrelated picturable nouns were associated with each onset-overlapping pair (e.g., *hond* [dog] and *spiegel* [mirror]). Each noun was matched for frequency with one item of the pair. The high frequency–matched distractors had an average log frequency of 1.7 ($\sigma = 0.5$), and the low frequency–matched distractors, of 0.7 ($\sigma = 0.5$). Finally, 70 filler trials were constructed; 35 were composed of four phonologically and semantically unrelated words; the other 35 trials included two onset-overlapping words, neither of them playing the role of target during the experiment.

Black-and-white line-drawing pictures were assigned to each of the 392 words. A picture-naming task was administered to an independent group of 15 Dutch speakers. Naming responses to each item of an experimental pair were coded to evaluate the identification of the pictured object and the use of its intended name. On average, high- and low-frequency pictures were recognized and labeled as intended and equally so (respectively 95 and 94% correct picture identification, and 88 and 87% correct picture labeling). Spoken stimuli were recorded by a female native speaker of Dutch. The average duration of the experimental target words was 528 ms (538 ms for high-frequency words, 518 ms for low-frequency words).

1.3. Design and procedure

The frequency of the target word and the competitor word was varied for each experimental item pair. On a given trial, the target picture was high-frequency, and its competitor was low-frequency (i.e., the high-frequency condition), or vice versa (i.e., the low-frequency condition). For a given participant, half of the 28 experimental trials were tested in the high-frequency condition, the other half in the low-frequency condition.

Participants were seated at a comfortable distance from a computer screen. Eye movements were monitored with an SMI Eyelink system, sampling at 250 Hz. The head-mounted eye tracker was first fitted onto the participant’s head, and a brief calibration procedure was performed. On each trial, a central fixation point appeared on the screen for 500 ms, followed by a blank screen for 600 ms. Then, a $5 \times 5$ grid with four pictures, four geometric shapes, and a central cross appeared on the screen (see Figure 1), either concurrently with or 500 ms before the presentation of the referent’s name (i.e., the no-preview or preview conditions, respectively). Prior to the experiment, participants were instructed that they would hear a word referring to one of the pictured objects on the screen. Their task was to click on the picture and move it above or below the geometric shape adjacent to it, using the computer mouse. Positions of the pictures were randomized across four fixed positions of the grid. The positions of the geometric shapes were...
Figure 1. Example of an experimental display (koffie [coffee], koffer [suitcase], hond [dog], spiegel [mirror]).

fixed. The edges of the pictures were approximately 4 cm apart; the distance between
the central cross and the closest edge was roughly 3 cm. (One centimeter corresponded
to approximately one degree of visual arc.) Participants were under no time pressure to
perform the action. After the participant moved the picture, the experimenter pressed a
button to initiate the next trial. Every five trials, a central fixation point appeared on the
screen, allowing for automatic drift correction.

2. Results and discussion

The data were first parsed into fixations and saccades. Saccade onsets and offsets were
automatically detected using the thresholds for motion (0.2), velocity (30/s), and acceler-
ation (8000/s²). Fixation duration corresponded to the time interval between successive
saccades. Fixation location was assessed by averaging the \( x \) and \( y \) coordinates of the
fixation’s samples, and by superimposing the fixation location onto the displayed grid
and pictures. Fixations that fell within the grid cell containing a picture were hand-coded
as fixations to that picture. All other fixations were coded as fixations to the grid, with-
out further distinction. Fixations were coded from the beginning of the trial (i.e., the
appearance of the display) until the target picture was fixated and clicked on.

Twenty experimental trials (accounting for 1% of the data) were excluded from the
analyses because of poor calibration or track loss (5 trials), failure to fixate on the target
object while or before clicking on it (10 trials) or selecting the wrong object (5 trials).
Does preview affect eye-movement behavior? When the spoken word was presented concurrently with the appearance of the display (i.e., the no-preview condition), the mean number of fixations per trial, including the initial fixation, was 4.5 (4.2 in the high-frequency condition and 4.8 in the low-frequency condition). When the pictures were displayed 500 ms before the onset of the spoken word (i.e., the preview condition), 4.8 fixations occurred from spoken-word onset until the end of the trial (4.5 in the high-frequency condition, and 5.2 in the low-frequency condition).

A close look at gaze locations reveals a clear effect of preview. Figure 2 illustrates the distribution of fixations to each of five possible locations, starting from the fixation concurrent with the onset of the spoken word (fixation 0) until the end of the trial. These locations were: (1) the target picture; (2) the competitor picture (i.e., the picture

![Figure 2. Number of fixations to each location (target, competitor, matched distractor, other distractor, or elsewhere) for each fixation performed during a trial, in the no-preview and preview conditions.](image-url)
with a name that overlapped with the spoken word at onset); (3) the matched distractor picture (i.e., the distractor matched for frequency with the competitor); (4) the other distractor picture; and (5) elsewhere on the grid. In the no-preview condition, the grid was fixated on the vast majority of all initial fixations (fixation 0). On the next fixation, all four pictures were fixated equally frequently ($\chi^2(3) = 6.15, p > 0.10$). A preference for fixating the target over the other picture locations emerged only on the next fixation (fixation 2, $\chi^2(3) = 72.1, p < 0.0001$). Despite the phonetic overlap between the initial portion of the spoken word and the name of the competitor picture, participants showed no greater tendency to launch a fixation to the competitor than to the matched distractor ($\chi^2(1) = 1.8, p > 0.10$).

The fixation distribution in the preview condition showed a different pattern. On fixation 0 (the fixation concurrent with or immediately following the onset of the spoken word), all four pictures received a roughly equal number of fixations ($\chi^2(3) = 1.3, p > 0.50$). On fixation 1, however, the target picture received more fixations than the other pictures ($\chi^2(3) = 22.6, p < 0.001$). As more fixations were realized, the proportion of fixations to the target increased; importantly, the proportion of fixations to the competitor became larger than that to the matched distractor at fixations 2 and 3 ($\chi^2(1) = 11.3, p < 0.001$ and $\chi^2(3) = 4.2, p < 0.05$, respectively). Thus, only when participants were able to briefly preview the display were more fixations launched to the competitor than to its matched distractor.

In order to test the influence of preview on gaze location throughout the trial, we established, for each trial, whether the competitor picture was fixated or not after the onset of the speech input. In the no-preview condition, the competitor picture was fixated at least once on 526 of the 1072 trials (49%) compared to 516 of the trials (48%) for the matched distractor. In order to compare the probability of fixating the competitor and distractor pictures while preserving the assumption of independence between observations, we restricted our analysis to the trials with a fixation to only one of these two pictures. The number of trials with a fixation to the competitor but not to the distractor was not significantly greater than the number of trials with a fixation to the distractor but not to the competitor (265 vs 255, $\chi^2(1) < 1$). Thus, when participants had not been pre-exposed to the display, they were equally likely to fixate the competitor or the distractor upon hearing the target picture’s name. When participants had 500 ms of preview, however, the competitor received more fixations than its matched distractor. The number of trials with at least one fixation to the competitor after spoken-word onset, after excluding the trials where the competitor happened to be fixated at spoken-word onset (113 trials), was 343 out of 727 trials (47%). By comparison, the number of trials with at least one fixation to the distractor after spoken-word onset, after excluding the 128 trials with a straddling fixation to the distractor, was 242 out of 712 trials (34%). The number of trials with a fixation to the competitor but not to the distractor was 174; the number of trials with a fixation to the distractor but not to the competitor was 94 ($\chi^2(1) = 23.9, p < 0.0001$).

Thus, while preview had little effect on the total number of fixations per trial, it did influence where these fixations were directed. Without time to preview the display, participants may not have apprehended the display rapidly enough to direct their attention...
(and thus their gaze) toward pictures that matched the input during the brief portion where
the speech signal was ambiguous between target and competitor interpretations. One may
argue that the lack of preview did not offer participants the opportunity to prename the
pictures. However, as will soon become clear, an examination of how preview affected
gaze location as the spoken input became available speaks against this interpretation.

*Does having fixated a picture during preview affect the likelihood of refixating it
later in the trial?* To clarify the nature of the information extracted when fixating on a
picture during preview, we examined how previewing a picture affected the likelihood
of refixating that picture after spoken-word onset. For each of the 727 trials in the
preview condition with no fixation to the competitor concurrent with the onset of the
spoken word, we determined whether the competitor had been fixated during preview
(i.e., before spoken-word onset) and after spoken-word onset (see Table 1). This trial
categorization was done separately for high-frequency trials (i.e., where the competitor
was of low frequency) and low-frequency trials (i.e., where the competitor was of high
frequency). For comparison, trials were similarly categorized according to fixations to
the matched distractor before or after spoken-word onset. Overall, having fixated the
competitor before the onset of the spoken word did *not* affect the likelihood of fixating it
afterwards ($\chi^2(1) < 1$). However, there was a significant interaction when we considered
only the low-frequency trials (i.e., when the competitor had a high-frequency name)
($\chi^2(1) = 5.2, p < 0.05$), but not when we considered only the high-frequency trials (i.e.,
when the competitor had a low-frequency name) ($\chi^2(1) = 1.3, p > 0.20$). Thus, when the
competitor picture had a high-frequency name (i.e., in low-frequency trials), participants
were more likely to fixate that picture as the spoken input unfolded if they had previously

<table>
<thead>
<tr>
<th>Fixation after word onset?</th>
<th>High-frequency trials</th>
<th>Low-frequency trials</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Fixation before</td>
<td>Competitor</td>
<td></td>
</tr>
<tr>
<td>word onset?</td>
<td>no</td>
<td>187</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>15</td>
</tr>
<tr>
<td>Distractor</td>
<td>no</td>
<td>223</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>15</td>
</tr>
</tbody>
</table>

1 A log linear model where the three-way interaction term was omitted yielded a significantly poorer fit to the
data than the model that included it ($L^2 = 5.9, p < .05$).
fixated it. In contrast, if the picture had a low-frequency name, having fixated the competitor picture did not affect the likelihood of refixating it. There was no influence of previewing the matched-distractor picture on the probability of fixating it after the onset of the spoken word on either type of trial (High- vs Low-Frequency trials, see Table 1). Thus, regardless of frequency, fixating the distractor picture during preview did not affect the likelihood of refixating it.

The interaction between the frequency of a picture’s name and the likelihood of refixating it when the spoken input is consistent with this name suggests that the information extracted during picture preview interacts with the outcome of lexical processing, that is, at a level where lexical-frequency biases operate. This finding argues against a view in which the phonetic input is mapped onto pre-activated picture names, thereby bypassing normal lexical processing of the spoken word. However, a possible objection to this conclusion could be raised if one were to assume that during preview, people may leave a fixated picture before having activated a name for it, and that this is more likely to occur for pictures with low-frequency names than for pictures with high-frequency names because low-frequency words are accessed more slowly than high-frequency words. On this view, as the spoken input becomes available, people would orient their attention to pictures with a pre-activated name that matches the phonetic input, and thus would be more likely to refixate high-frequency pictures, for which the name would more often be available, than low-frequency pictures, for which the name would not be available. Note that this view predicts some (albeit weak) tendency to refixate the picture of a low-frequency competitor, and the data did not support this prediction (in fact, numerically, the effect was in the opposite direction).

Nonetheless, we addressed this objection by examining the duration of fixations to competitor pictures with high- or low-frequency names that occurred after spoken-word onset. In order to neutralize the impact of preview, we restricted this analysis to the no-preview condition. On 96% of the trials, as the pictures appeared, the participant’s fixation was on a location of the grid without a picture. Thus, by the time the spoken word began, participants had not fixated any of the pictures. We examined the durations of the fixations to competitor pictures that immediately followed the onset of the spoken word. Of the 1072 fixations, 99 were launched to competitor pictures with high-frequency names and 97 to competitor pictures with low-frequency names. (For inferential statistics purposes, analyses were limited to participants who made such fixations on both high- and low-frequency trials.) Fixations to competitors with high-frequency names were on average 208 ms long, as opposed to 186 ms for fixations to competitors with low-frequency names ($t_{(30)} = 2.2, p < 0.05$). If people were to name the picture they are currently fixating in order to evaluate the match between the speech input and the picture’s name, we would have expected to observe longer fixations to competitor pictures with low-frequency names than to pictures with high-frequency name, or no difference between fixations to the two types of pictures if lexical frequency has a negligible effect on fixation duration. Instead, we observed longer fixations to high-frequency competitors than to low-frequency competitors. This finding, together with the influence of a picture name’s frequency on the likelihood of refixation reported above, argues against the hypothesis.
that people prename the pictures during preview or name the currently fixated picture in order to evaluate the match between picture name(s) and the phonetic input.

Longer fixations to competitor pictures with high-frequency names than to competitor pictures with low-frequency names and a greater tendency to refixate high-frequency pictures than low-frequency pictures once spoken input becomes available suggest that the match between lexical representations activated by the spoken input and information extracted from the display (either during preview or during an on-going fixation) is high for high-frequency pictures, but relatively weak for low-frequency pictures. While the results reported so far argue against the first view discussed in the Introduction, they cannot alone distinguish between the other two. Recall that the second view posits that pre-activated picture names are matched to word forms activated by the speech input. The third view, on the other hand, posits no picture prenaming; it assumes that semantic representations associated with word forms activated by the speech input are matched to visual and conceptual representations associated with the pictures’ locations. Both views are consistent with the frequency-modulated gaze behavior just reported. However, only the third one can also account for the Dahan and Tanenhaus (2005) results, where listeners tended to fixate a picture that shares visual but little phonological similarity with the concept associated with the target picture’s name more than they fixated visually and phonologically unrelated pictures. Thus, the nature of the representation that mediates the mapping between lexical processing and information extracted from picture preview appears to be semantic/conceptual, rather than phonological.

How does the position of the pictures in the display affect fixations? Results reviewed so far have revealed the importance of previewing the display early in the trial. Signal-driven fixations to pictures with names that are temporarily consistent with the spoken input require some exposure with the display. While some information can be extracted parafoveally, fixating a given picture during preview may increase the likelihood of fixating this picture later in the trial. Given the importance of having extracted information on a given location, it would be useful to know whether fixations are directed to particular spatial locations more often than others, both during preview and as the spoken word unfolds. Although picture locations are typically randomized in eye-movement studies to minimize the risk of confounds with location, choosing the spatial arrangement of competitors and distractors might maximize the sensitivity of the paradigm.

Figure 3 summarizes the distribution of fixation locations over the course of a trial, distinguishing fixations to the upper left picture, the upper right picture, the lower left picture, the lower right picture, or elsewhere on the display. A similar pattern emerges for both preview and no-preview conditions. There is a strong tendency to initially fixate on the picture located in the upper left cell, and then to move to another picture by performing either a horizontal saccade (and landing on the upper right picture) or a vertical saccade (and landing on the lower left picture).

A potential implication of the unbalanced distribution of fixations across picture locations is that whether a competitor picture is fixated or not during a trial may strongly depend on its position in the display. To address this question, we determined the position of the competitor picture on trials where it was fixated after spoken-word onset and on
trials where it was not fixated. We conducted this analysis on the preview condition trials. To neutralize the impact of preview on the likelihood of revisiting the picture, we excluded trials where the competitor had been fixated before or concurrently with the onset of the spoken word. We established the distribution of the remaining 677 trials...
Table 2

Distribution of trials in the preview condition, distinguishing whether the competitor or the matched-distractor picture was fixated or not after spoken-word onset, as a function of the position of the picture (upper left, upper right, lower left, lower right). (Note: Trials on which either picture was fixated before or concurrently with the onset of the spoken word were excluded.)

<table>
<thead>
<tr>
<th>competitor or distractor position</th>
<th>upper left</th>
<th>upper right</th>
<th>lower left</th>
<th>lower right</th>
</tr>
</thead>
<tbody>
<tr>
<td>competitor fixation?</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no</td>
<td>71</td>
<td>83</td>
<td>88</td>
<td>119</td>
</tr>
<tr>
<td>yes</td>
<td>90</td>
<td>85</td>
<td>85</td>
<td>56</td>
</tr>
<tr>
<td>distractor fixation?</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no</td>
<td>105</td>
<td>110</td>
<td>126</td>
<td>108</td>
</tr>
<tr>
<td>yes</td>
<td>43</td>
<td>41</td>
<td>63</td>
<td>79</td>
</tr>
</tbody>
</table>

Across all the four picture positions, distinguishing whether the competitor was fixated after spoken-word onset or not. A similar trial classification was established for fixations to the matched distractor (Table 2). Analyses revealed that the distribution of trials across all four possible positions of the competitor picture affected whether the competitor was fixated or not ($\chi^2(3) = 22.1, p < 0.0005$). This did not interact with trial frequency (the log linear model omitting the three-way interaction did not significantly differ from the saturated model, $L^2 = 2.5, p > 0.40$). A relationship between distractor-picture position and whether the distractor was fixated or not after spoken-word onset was also found ($\chi^2(3) = 10.5, p < 0.05$). Thus, the position of each picture on the display affected whether or not that picture was fixated as the speech input unfolded. However, the influence of picture position was different between the two types of pictures. In particular, there were fewer trials with a fixation to the competitor when the competitor was in the lower right position than when it was in any of the other positions. By contrast, there were more trials with a fixation to the distractor when this distractor was in the lower right position than when it was in any of the other positions. To formally evaluate this interaction, we considered the distribution of trials with a fixation to either the competitor or the distractor (but not to both) across all four possible picture positions. There was a significant interaction between the type of picture fixated and its position ($\chi^2(3) = 16.9, p < 0.0005$): When fixated, the competitor tended to be in the upper left or right position, and much less so in the lower right position; by contrast, the fixated distractor tended to be in the lower left or right positions, and less so in the upper left and right positions.

These results suggest the following interpretation. Early in the trial, people’s attention is drawn toward the upper positions of the display and some information about the objects displayed there may be acquired even if no fixation is launched to these locations. Upon hearing the spoken word, people are more likely to fixate on these positions if the competitor is located in one of these positions because of the match between the picture and the representation temporally activated by the spoken word. By contrast, if one of these positions is occupied by the distractor, participants’ attention is not drawn
toward these already explored locations when hearing the spoken word and participants continue to explore the display, thus fixating on the lower positions of the display. This finding is important because it demonstrates that the specific position of a competitor picture will affect whether this picture will be fixated as spoken input becomes available. Experimenters might want to control more systematically (rather than randomize) where critical pictures are located in the display, or perhaps include position as a factor during data analysis.

3. Conclusions

Examining the role of picture preview and display characteristics in the set-up typically used in eye-movement studies of spoken-word recognition revealed a number of interesting findings that shed light on the nature of the representations that mediate phonetically driven fixations. First, some preview is required to observe signal-driven fixations to competitors that are temporarily consistent with the input. When the display is briefly available before the speech input begins, participants are able to extract relevant visual and/or conceptual information associated with the pictured objects, even when they have not fixated on any of the pictures. This information guides their subsequent saccades to relevant pictures. Without preview, initial fixations mainly serve the purpose of extracting visual or conceptual information about the displayed pictures, and thus fail to reflect the temporary activation of lexical candidates triggered by the early portion of the spoken input.

Second, whether or not a picture is fixated during preview influences the likelihood that it will be refixated during speech processing, but only if this picture is consistent with a high-frequency interpretation of the speech input. Furthermore, the first fixation to a picture with a name that matches the unfolding speech input is longer when the picture name is high frequency compared to when it is low frequency. These two new findings provide strong evidence that even in the very circumscribed set of possible referents that the display offers, the speech signal is processed in terms of possible lexical interpretations, rather than directly matched to pre-activated picture names (as assumed by the first view introduced earlier). This extends Dahan et al.’s (2001) results by showing effects of competitor frequency even when the competitor picture was fixated during preview, casting doubts on the view that participants engage in covert picture naming.

Third, some spatial locations attract a disproportionate number of fixations, especially early in the trial, and the position of pictures in the display, in conjunction with the speech input, affects whether or not the picture is fixated.

Overall, the contingent analyses reported here shed light on how the display might influence lexical processing and help mitigate concerns about potential task-specific strategies that using displays of pictures might in principle have introduced. The results are encouraging for researchers who want to exploit the properties of the eye-tracking paradigm to explore issues of representation and process in spoken-word recognition.
D. Dahan et al.

Acknowledgments

This work was supported by grants from the National Institutes of Health (DC 005071 to Michael K. Tanenhaus) and the National Science Foundation (Human and Social Dynamics 0433567 to Delphine Dahan).

References


Chapter 22

THE PROCESSING OF FILLED PAUSE DISFLUENCIES IN THE VISUAL WORLD

KARL G. D. BAILEY
Andrews University, USA

FERNANDA FERREIRA
University of Edinburgh, UK

Eye Movements: A Window on Mind and Brain
Copyright © 2007 by Elsevier Ltd. All rights reserved.
Abstract

One type of spontaneous speech disfluency is the filled pause, in which a filler (e.g. *uh*) interrupts production of an utterance. We report a visual world experiment in which participants’ eye movements were monitored while they responded to ambiguous utterances containing filled pauses by manipulating objects placed in front of them. Participants’ eye movements and actions suggested that filled pauses informed resolution of the current referential ambiguity, but did not affect the final parse. We suggest that filled pauses may inform the resolution of whatever ambiguity is most salient in a given situation.
The most common type of overt interruption of fluent speech, or disfluency, is the filled pause (Bortfield, Leon, Bloom, Schober, & Brennan, 2001). Speakers produce filled pauses (e.g. *uh* or *um*) for a variety of reasons, such as to discourage interruptions or to gain additional time to plan utterances (Schacter, Christenfeld, Ravina, & Bilous, 1991). While speakers may benefit from producing filled pauses because they gain planning time, listeners may also use the presence of filled pauses to inform language comprehension (Bailey & Ferreira, 2003; Brennan & Schober, 2001; Brennan & Williams, 1995; Clark & Fox Tree, 2002). Thus, given the prevalence of filled pauses, and the use of such pauses by listeners, a complete model of language comprehension should account for how these disfluencies are handled.

In order to construct such a model of disfluency processing, it is necessary to describe and test possible hypotheses about how disfluencies might affect language comprehension. Evidence that supports one such hypothesis, cueing of upcoming structure, comes from a series of experiments involving grammaticality judgments (Bailey & Ferreira, 2003). This hypothesis is built on the observation that filled pauses occur in a particular distribution with respect to syntactic (Clark & Wasow, 1998), semantic (Schacter et al., 1991) or pragmatic (Smith & Clark, 1993) structure. In the case of syntactic structure, filled pauses (and other disfluencies, such as repetitions) are most likely to occur immediately prior to the onset of a complex syntactic constituent (Clark & Wasow, 1998; Ford, 1982; Hawkins, 1971; Shriberg, 1996). Filled pauses are also likely after the initial word in a complex constituent, especially after function words (Clark & Wasow, 1998). Thus, the cueing hypothesis assumes that listeners might be able to use the presence of a recent filled pause to predict that an ambiguous structure should be resolved in favor of a more complex analysis (Bailey & Ferreira, 2003). In a garden path utterance like [1] below, the filled pause might act as a “good” cue, because it correctly predicts the ultimate structure of the utterance; in [2], the filled pause might be a “bad” cue, because it leads the listener to predict the onset of a new constituent.

[1] While the man hunted the *uh uh* deer ran into the woods.
[2] While the man hunted the deer *uh uh* ran into the woods.

Grammaticality judgments supported this cueing hypothesis: [1] was judged grammatical more often than [2], which suggested that [1] is easier to process (Bailey & Ferreira, 2003). However, [1] and [2] confound “good” and “bad” cues with the presence of delay between the ambiguous head noun and the disambiguating verb. This type of delay has led to the same pattern of results in utterances with lexical modifiers (i.e., prenominal adjectives and relative clauses) in place of the disfluencies in [1] and [2] (Ferreira & Henderson, 1991). To avoid this confound, Bailey and Ferreira (2003) tested whether filled pauses that did not introduce delays between temporarily ambiguous head nouns and disambiguating verbs might also affect grammaticality judgments of spoken utterances depending on their location. Disfluencies were placed in two different locations in coordination ambiguity utterances prior to the onset of the temporarily ambiguous head noun. The “good” cue location in [3] below was consistent with the ultimate sentence coordination structure, while the “bad” cue in [4] was consistent with an noun phrase coordination structure (based on the...
assumption that listeners take a disfluency to be indicative of an upcoming complex constituent).

[3] Sandra bumped into the busboy and the uh uh waiter told her to be careful.
[4] Sandra bumped into the uh uh busboy and the waiter told her to be careful.

Participants were more likely to judge an utterance with a “good” cue disfluency (as in [3]) as grammatical than an utterance with a “bad” cue (as in [4]). This pattern of results was replicated with environmental noises replacing the disfluencies, but, importantly, not with adjectives, suggesting that it is the presence of a non-propositional interruption that is the cue, not the form of that interruption.

However, the results of Bailey and Ferreira (2003), while promising, are based on offline judgments following the end of the utterance. In other words, the grammaticality judgment task makes it possible to see that filled pauses have had an effect consistent with the cueing hypothesis by the time the utterance is finished, but it is not possible to chart the time course of that effect, nor to observe when processing of the disfluency takes place.

A recently rediscovered methodology that allows spoken language comprehension to be monitored on a moment by moment basis is the visual world paradigm (Cooper, 1974; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). In this paradigm (henceforth, the VWP), participants listen to utterances while viewing a concurrent array of clip art images on a computer screen (e.g. Altmann & Kamide, 1999) or while interacting with a set of objects within reach (e.g. Tanenhaus et al., 1995). The objects or images which make up the constrained visual world and the relationships between them serve as a context for a concurrent referring utterance (Tanenhaus et al., 1995). Inferences about language comprehension are drawn from listeners’ eye movement patterns: The eyes are naturally directed to objects that are related to concurrent language processing (Cooper, 1974). In the VWP, utterances can be presented without distortion and it is not necessary to instruct listeners to look at objects which are related to concurrent speech.

Two particular patterns of eye movements have been used to draw inferences about comprehension: anticipatory and confirmatory eye movements. Anticipatory eye movements (Altmann & Kamide, 1999, 2004; Kamide, Altmann, & Haywood, 2003) are saccades launched to objects before they are directly referenced by the utterance. Confirmatory eye movements (e.g. Spivey, Tanenhaus, Eberhard, & Sedivy, 2002) are made in response to a direct reference to an object and can include fixations on possible referents of a constituent (Tanenhaus et al., 1995) or on disconfirmed competitors (Sedivy, Tanenhaus, Chambers, & Carlson, 1999; Kamide et al., 2003). The presence of confirmatory eye movements is most easily seen in the probability of fixating a given object because participants may simply continue to fixate an object that they were already looking at due to an anticipatory eye movement launched prior to direct reference.

The cueing hypothesis would predict that eye movements during a filled pause should reflect a more complex parse of material currently being processed, and that saccades would be launched to objects consistent with that analysis. Confirmatory eye movements during a later ambiguous referring expression would then identify which of a set of possible parses had been selected and the time course of that selection (as the probability of fixating a given object rises and falls).
Ch. 22: The Processing of Filled Pause Disfluencies in the Visual World

In this chapter, we will present an experiment that directly tests whether a cueing mechanism can modulate the interpretation of a fully ambiguous utterance in the presence of a fully ambiguous visual world. As described earlier, the position of a disfluency can affect the probability that an utterance is judged grammatical (Bailey & Ferreira, 2003), suggesting that disfluencies may cue the parser to expect a certain structure. The strongest form of the cueing hypothesis, then, predicts that a fully ambiguous utterance will immediately be interpreted differently based solely on the location of a disfluent interruption. We do not find evidence for this strong hypothesis, but do find support for a weaker form, in which the disfluent interval introduced by the filled pause may allow the parser to further process any existing ambiguities. Depending on the demands of the task, final interpretations may or may not be affected by the disfluency cue. Nevertheless, we suggest that filled pauses provide a unique window on sentence processing in general, because they show what ambiguities are relevant at that point in the utterance.

1. Experiment

In order to test whether filled pauses can change the interpretation of an otherwise fully ambiguous utterance, the concurrent visual world must not constrain the interpretation of that utterance. Previous studies using otherwise fully ambiguous prepositional phrase ambiguities (Spivey et al., 2002; Tanenhaus et al., 1995) used visual worlds that constrained the interpretation of utterances such as [5] below. The objects in these displays required participants to arrive at the same semantic representation as in the disambiguated utterance [6], and disallowed the interpretation in [7].

[5] Put the apple on the towel in the box.
[6] Put the apple that’s on the towel in the box.
[7] Put the apple on the towel that’s in the box.

Two different constrained display types have been used (Spivey et al., 2002). The first, referred to as the one-referent display, contained a target object (e.g. an apple on a towel), a distractor object (e.g. a frog on a mitten), a goal location (e.g. a box), and a distractor location (e.g. a towel). The second, two-referent, display was identical to the first, except that the distractor object matched the target object in part (e.g. an apple on a mitten). Note that in both displays, the only possible action in response to [5] is for the apple that is on the towel to be placed in the empty box because there is no towel inside a box.

In order to modify these displays so that they did not constrain the interpretation of our utterances, we replaced the distractor location described above (a towel by itself) with a modified goal (e.g. a towel in a box as opposed to the unmodified goal, an empty box; see Figure 1). Thus, in the one-referent display, it would be possible to place an apple that is on a towel into an empty box (a modified theme interpretation) in response to [5] or an apple onto a towel that is in a box (a modified goal interpretation). In the case of the two-referent display, of course, the display still constrains the interpretation of the utterance because of the presence of a second apple. The modified goal interpretation in [7] is possible, but only if the listener violates syntactic or discourse constraints and...
uses the phrase “on the towel” twice: once to identify which apple to pick up (the apple that’s on the towel) and once to identify where to place the apple (the towel that’s in the box). These interpretations are unlicensed because a single constituent cannot play more than one role in a sentence; nevertheless, we have observed that participants occasionally behave as if that is the interpretation that they have obtained, perhaps because they have engaged in “good-enough processing” (Ferreira, Bailey, & Ferraro, 2002).

According to the strong form of the cueing hypothesis, disfluencies placed before one of the two possible modified noun phrases in [5] should bias the parser to prefer the corresponding structure. Specifically, a filled pause placed as shown in [8] is predicted to yield a modified theme interpretation similar to [6], whereas a filled pause placed as in [9] is predicted to yield modified goal interpretations similar to [7].

[8] Put the uh uh apple on the towel in the box.
[9] Put the apple on the uh uh towel in the box.

For the strong form of the cueing hypothesis to be supported, these interpretations should be seen both immediately (in saccades to appropriate goal objects during filled pauses), as the utterance unfolds (in fixations during the ambiguous noun towel), and in the overall interpretation of the utterance (the participants’ actions).

1.1. Materials and methods

1.1.1. Participants

Sixteen participants from the Michigan State University community participated in this experiment in exchange for credit in an introductory psychology course or payment ($7.00). All participants were native speakers of English, and had normal hearing and normal or corrected to normal vision.
1.1.2. Materials

Twenty-four prepositional phrase ambiguity utterances were constructed for this experiment using nouns from a set of 12 possible target objects and 12 possible goal objects. Utterances were recorded and digitized at 10 kHz using the Computerized Speech Laboratory (Kay Elemetrics), and then converted to wav format. Each utterance was recorded in two ways: once as an utterance with two disfluencies, as in [10] below, and once as a fluent utterance with two instances of that’s, as in [11].

[10] Put the uh uh apple on the uh uh towel in the box.
[11] Put the apple that’s on the towel that’s in the box.

Utterances like [8] and [9] were created from [10], and like [6] and [7] were created from [11] by excising the appropriate disfluency or word. Participants are relatively insensitive to the removal of disfluencies from utterances (Brennan & Schober, 2001; Fox Tree, 1995, 2001; Fox Tree & Schrock, 1999; ) and thus this procedure was used to control the prosody of the various utterances. The removal of a single disfluency or word from an utterance did not result in utterances that participants found odd or strange. In the experiment, each participant heard only one version of any given utterance.

Forty-eight filler utterances were also recorded and grouped with the 24 critical utterances into trials of 3 utterances each. A further 72 utterances were recorded to create 24 trials composed of only fillers. The types and proportions of syntactic structures used in the filler utterances and the interleaving of filler and critical trials were identical to those used in Spivey et al. (2002). Filled pauses occurred on half of filler trials and were placed at a variety of different locations within the sentences.

Displays consisted of a 2 × 2 grid (see Figure 1), and objects were set up according to the description provided by Spivey et al. (2002), so that depending on the height and posture of a given participant, the center of each object (or set of objects) was separated by 10–15° of visual angle from the center of each of its adjacent neighbors (note that previous studies did not report the angular distance between objects). In experimental trials, the possible theme referents (the target and distractor objects) were always on the left, and were each placed equally in both the proximal and distal positions across trials. The possible goal referents (modified and unmodified) were always on the right, and likewise were each placed equally in both the proximal and distal positions across trials. The possible theme and goal referents for filler utterances were equally likely to occur in any of the four positions, and any object on the table could be referenced as a target object in filler utterances. In all, 48 displays were created, one for each set of three utterances. Of the 24 critical displays seen by any participant, 12 were two-referent displays and 12 were one-referent displays. A new random ordering of trials adhering to the interleaving requirements was created for every fourth participant in this experiment.

1.1.3. Apparatus

The eyetracker used in this experiment was an ISCAN model ETL-500 head-mounted eyetracker (ISCAN Incorporated) with eye and scene cameras located on a visor.
Participants were able to view 103° of visual angle horizontally and 64° vertically. No part of the object display was occluded at any time by the visor. Eye position was sampled at 30 Hz and merged with scene video data. Eye position in this merged video was later hand-coded relative to Regions of Interest (henceforth, ROIs) frame by frame, starting with the onset of each critical utterance and ending with the movement of an object to a new location.

1.1.4. Procedure

After a participant was introduced to the objects and apparatus, and had provided informed consent, the eyetracker was placed on the participant’s head and adjusted. Depending on the height of each participant, participants either stood or were seated at a table. Participants’ eye positions were calibrated to the scene by recording pupil and corneal reflection locations while they looked at nine predetermined targets. The sentence comprehension task was introduced to the participant via three practice utterances involving the movement of a single object from one location to another. The practice utterances did not contain any lexical or syntactic ambiguities.

Immediately before beginning a trial, the experimenter set up the appropriate objects in front of the participant, which allowed 20–30 s of view time prior to the onset of the first utterance in the trial (as in Spivey et al., 2002 and Trueswell, Sekerina, Hill, & Logrip, 1999). Participants were instructed to respond as quickly as possible prior to practice trials, but were not reminded thereafter. In addition, no “Look at the center cross” instruction was given prior to the start of each trial (cf. Spivey et al., 2002), as pilot studies indicated that participants tended to perseverate in fixating the center cross when this instruction was given.

Design. The four utterance types (theme and goal disfluencies, and theme and goal modifiers) were combined with the two displays (one and two referent) to create eight unique conditions for this experiment. Three trials in each condition were presented to each participant, for a total of twenty-four critical trials. Each display occurred in each condition an equal number of times over the course of the experiment.

1.2. Results and discussion

The analysis of eye tracking data presented here differs somewhat from previous studies using this version of the VWP. These studies (Spivey et al., 2002; Tanenhaus et al., 1995; Trueswell et al., 1999) calculated probabilities of fixating particular objects at each sampling interval during arbitrary time segments that did not take into account variations in word length across individual utterances. Probabilities in this study, on the other hand, were calculated separately for each ROI and each word in each utterance and then averaged (see Altmann & Kamide, 2004, who described this procedure). These probabilities were then arcsine-transformed (Winer, 1971) and submitted to a 2 (cue location: theme or goal) by 2 (cue type: disfluency or modifier) by 2 (number of possible theme referents: one or two) ANOVA. In addition, behavioral responses to disfluent
instructions in the current experiment were classified as either modified goal (e.g. towel in box) directed or unmodified goal (e.g. empty box), and were submitted to a 2 (number of referents) by 2 (location of disfluency) ANOVA. Unambiguous controls were not included in the behavioral analysis as participants moved an object to the appropriate goal on over 90% of trials.

Participants were more likely to move a target object to the unmodified goal ($F_{1,15} = 23.6, p < 0.001$) in the two-referent display (70.8% of trials with a theme disfluency; 64.6% with a goal disfluency) than in the one-referent display (37.5% of trials with a theme disfluency; 35.4% with a goal disfluency). The location of the disfluency had no effect ($F < 1$) on participants’ actions and there was no significant interaction between the number of referents and disfluency location ($F < 1$). The effect of number of referents on the final interpretation of the utterance is not surprising, as the two-referent display should have constrained the interpretation of the utterance (due to the presence of two apples), while the one-referent display should not have. However, the lack of effect of disfluency location on the final interpretation of the utterances, even in the one-referent display, is evidence against the strong form of the cueing hypothesis, and suggests that disfluencies were not interpreted as strong predictors of the syntactic parse.

Eye movement patterns, on the other hand, did support a form of the cueing hypothesis. Figure 2 shows graphs representing the probability of fixating and launching a saccade to each ROI for each condition in the experiment. Gray polygons represent the probability of fixation on, and black lines the corresponding probability of launching a saccade to that ROI for each word. Each point on the polygons and line graphs corresponds to a single word in each utterance. Content words, disambiguating function words, and disfluencies are indicated above the fixation polygons. The eight conditions form rows, while the four ROIs form columns.

An effect of number of referents is presented in Figure 2; the different display types elicited different patterns of fixation, especially on the distractor and the modified goal. Consistent with previous studies (Spivey et al., 2002; Tanenhaus et al., 1995; Trueswell et al., 1999; the incorrect goal in previous studies corresponds to our modified goal), there is a significant increase in the probability of fixation on the modified goal in the one-referent display relative to the two-referent display during the word towel ($F_{1,15} = 36.2, p < 0.001$; $F_{2,23} = 30.6, p < 0.001$). This difference was found for all utterance types, including theme modifiers, which should rule out the modified goal as a possible referent of towel because of the preceding that’s. Main effects of cue location ($F_{1,15} = 6.29, p < 0.03$; $F_{2,23} = 5.12, p < 0.04$), and cue type ($F_{1,15} = 7.64, p < 0.02$; $F_{2,23} = 7.22, p < 0.02$) were also present. The effect of cue type was due to an increased proportion of looks to the modified goal in the disfluency conditions, which would be expected if the language comprehension system treated those utterances as more ambiguous than the unmodified goal. A significant interaction between number of referents and cue location ($F_{1,15} = 6.75, p < 0.02$; $F_{2,23} = 7.99, p < 0.02$) was present, but interactions between cue type and number of referents ($F < 1$), between cue type and location ($F_{1,15} = 3.54, p > 0.05$; $F_{2,23} = 3.61, p > 0.05$), and between all three variables ($F < 1$) were nonsignificant. This pattern (Figure 3) is consistent with the prediction that...
theme disfluencies and modifiers should elicit fewer looks to the modified goal (being consistent instead with a modified theme) than the corresponding modify goal utterances, but only in the one-referent display, where the identity of the theme has already been ascertained and the eye movement system is not engaged in deciding between the target and distractor objects. However, separate analyses of the disfluent conditions found only an effect of number of referents ($F_{1,15} = 29.8, p < 0.001; F_{2,23} = 24.9, p < 0.001$). The effect of cue location ($F < 1$) and the interaction between number of referents and cue location ($F_{1,15} = 2.69, p > 0.1; F_{2,23} = 2.29, p > 0.1$) were not significant, suggesting that the modifier conditions were carrying the overall interaction between the number of referents and cue location. This would suggest that disfluencies were not interpreted by the parser in the same ways as modifiers.

Similar patterns (Figure 4) are also present in the saccade data to the modified goal during towel, consistent with confirmatory saccades as the source of fixation patterns.
Main effects of number of theme referents \((F_{1,15} = 39.6, p < 0.001; F_{2,23} = 36.8, p < 0.001)\), cue location \((F_{1,15} = 4.68, p < 0.05; F_{2,23} = 2.63, p > 0.1)\), and cue type \((F_{1,15} = 5.03, p < 0.05; F_{2,23} = 3.76, p > 0.05)\) are again present (although the latter two effects are significant only by participants), as well as a marginal interaction between number of referents and cue location \((F_{1,15} = 3.18, p = 0.095; F_{2,23} = 4.17, p = 0.053)\). All other interactions were nonsignificant \((F < 1)\). Anticipatory saccades, however, may also have contributed to the probability of fixating the modified goal during towel, as saccades were also launched to this object during the word on, suggesting that these saccades may have been launched based on the expected arguments of the verb put (which requires both a theme and a goal when used imperatively; cf. Altmann & Kamide, 1999, 2004).

Separate analyses of disfluent conditions again revealed a main effect of number of referents \((F_{1,15} = 14.4, p < 0.01; F_{2,23} = 15.4, p < 0.01)\), but only a marginal effect of cue location by participants \((F_{1,15} = 3.23, p < 0.1; F_{2,23} = 1.56, p > 0.1)\) was present. The interaction between number of referents and cue location was nonsignificant \((F = 1.23; F = 1.87)\). The marginal effect of cue location tentatively suggests that disfluencies may have some immediate effect on the parser; however, it is clear that the display itself had a much greater impact on eye-movement patterns.
K. G. D. Bailey and F. Ferreira

Figure 3. Probability of fixation on the modified goal (e.g. towel in a box) during the word towel for each of the eight utterance and display conditions. 1REF and 2REF refer to the number of possible theme referents in the display; theme bias and goal bias refer to the locations of the disfluencies; disfluency and modifier refer to cue types.

Figure 4. Probability of saccade launch to the modified goal (e.g. towel in a box) during the word towel for each of the eight utterance and display conditions. 1REF and 2REF refer to the number of possible theme referents in the display; theme bias and goal bias refer to the locations of the disfluencies; disfluency and modifier refer to cue types.
Disfluencies, then, do not affect the final interpretation of utterances in this study, contrary to what was suggested by previous experiments employing grammaticality judgment tasks (Bailey & Ferreira, 2003), nor do they appear to be strongly biasing the parser during the utterance. The strong form of the cueing hypothesis must therefore be rejected in favor of a hypothesis that can account for both current and previous results. A possible modification might suggest that further processing of the most salient current ambiguities (whether lexical, syntactic, referential, or discourse related) may occur during the disfluency, or that less salient ambiguities may become more salient. As a result, participants’ eye movements may reflect the processing of possible resolutions of current ambiguities, and, in cases where the discourse context does not constrain the parse, the final parse may be affected (e.g. Bailey & Ferreira, 2003).

Evidence that ambiguities (not necessarily syntactic) are processed during disfluencies can be seen in the probability of saccade launch to each of the four objects during the filled pauses in the theme and goal disfluency conditions (Figure 5). The probability of launching a saccade (as opposed to probability of fixation) is sensitive to changes in visual attention (and by inference, cognitive operations) during a disfluency (Altmann & Kamide, 2004). As expected, participants are more likely to launch a saccade to the modified goal during the goal disfluency than the theme disfluency, regardless of condition, as indicated by a main effect of cue location (marginal by items; $F_{1,15} = 5.70, p < 0.05; F_{1,23} = 3.21, p = 0.086$), a nonsignificant main effect of number of referents, and a nonsignificant interaction between number of referents and cue location. This pattern is consistent with the fact that the goal disfluency occurs later in the utterance, often after the theme has been unambiguously identified.

![Figure 5](image-url)

Figure 5. Probability of saccade launch to each of the four regions of interest for each of the disfluent conditions. 1REF and 2REF refer to the number of possible theme referents in the display; theme disfluency and goal disfluency refer to the locations of the disfluencies.
The pattern of results for the distractor object is more complex: Only the interaction between number of referents and cue location is significant ($F_{1,15} = 5.31, p < 0.05$, $F_{2,23} = 5.89, p < 0.03$).

Looks to the distractor are more likely in the theme-disfluency one-referent display condition and the goal-disfluency two-referent display condition. These results are key because they indicate why the display exerts such a powerful influence on parsing in the structurally ambiguous disfluency conditions. When a disfluency occurs later in the utterance (in the goal disfluency conditions), the distractor is still a possible candidate theme in the two-referent condition (leading to many looks to the distractor and few to either goal), as opposed to the one-referent condition (leading to fewer looks to the distractor and more looks to either goal in anticipation of the speaker identifying the goal). Moreover, because the theme disfluency occurs after the word the, participants could expect immediate resolution of the referential ambiguity in the one-referent condition, while immediate resolution is not expected until the goal disfluency in the two-referent condition. The interaction between number of referents and cue location thus shows that the language comprehension system is sensitive to immediately upcoming ambiguity resolution.

This ambiguity resolution hypothesis, moreover, suggests a mechanism by which we can account not only for the results in this experiment but also for previous syntactic cueing results (Bailey & Ferreira, 2003). Recall that in the current experiment, the number of possible continuations for any phrase that began with the was limited by co-present objects (the preceded the disfluency in all of the critical items in this study, as well as in the Bailey & Ferreira, 2003, study). Thus, it was easy for the listener to predict the actual object that would be referenced. However, in Bailey and Ferreira’s (2003) study, no context was present. As a result, the number of possible lexical continuations at a disfluency was very large (limited only by the context introduced by the initial utterance fragment). On the other hand, the number of possible syntactic continuations was relatively small. Thus, the parser could have used the disfluent delay to consider less preferred structures, rather than possible lexical items. The grammaticality judgment task may have been sensitive to the occasions on which the parser identified the ultimately correct (less preferred) structure during the disfluency. “Bad” disfluencies may have occurred too early (i.e. [4]) to provide enough information to deduce possible structures, or so late (e.g. [2]) that the parser had committed to a single parse. “Good” disfluencies (i.e. [1] and [3]) may have occurred just late enough that less preferred structures could be identified, but not so late that multiple structures were no longer being considered. In essence, then, the nature of the grammaticality judgment task (which focuses participants on syntactic structure, with relatively little context) may have affected the way in which disfluencies were interpreted.

In the current experiment, however, it was possible for the listener to pick out the complete set of possible lexical continuations for any disfluency and interpret the disfluency as referential uncertainty. The final parse may therefore have been driven by the biases of the parser and the constraints of the display only (i.e. cue location did not lead to differences in commitments), leading to a null effect of disfluency location.
The processing done by the language comprehension system during a disfluency, then, may amount to identifying concurrent ambiguities, but the type of ambiguity that receives further processing depends on the partial parse and discourse context, and on the number of alternative continuations to be considered. Moreover, a particular type of ambiguity may be more salient than others in certain experimental settings (e.g., referential ambiguity in the VWP). This suggests that the language comprehension system uses the delay in propositional input and the distributional cues provided by disfluencies in very flexible ways that fit the comprehension goals of the listener.

Finally, these results suggest that filled pauses may provide an opportunity for studying the relative saliency of a variety of ambiguities during processing in different experimental paradigms, in that they provide a natural interruption of propositional input during which otherwise obscured ambiguity resolution processes may continue to run and thus be more easily observed. In fact, identifying the processes at work in the VWP may be especially important, as a model of cognition and language processing in this paradigm is necessary to ground and guide further study. Additional research is also needed to examine the processes that occur during filled pauses, to test the delay hypothesis described in this chapter, and to further understand the relationship between eye-movement patterns and language comprehension processes in the VWP.

References


Chapter 23

SPEECH-TO-GAZE ALIGNMENT IN ANTICIPATION ERRORS

LINDA R. WHEELDON, ANTJE S. MEYER and FEMKE VAN DER MEULEN

University of Birmingham, UK
Abstract

When speakers correctly name several objects, they typically fixate upon all objects in the order of mention and look at each object until about 150 ms before the onset of its name. The duration of gazes to objects that are to be named depends, among other things, on the ease of name retrieval (e.g., the gazes are longer when name agreement is low than when it is high, Griffin, 2001) and the ease of word form encoding (e.g., gazes are longer for objects with long names than with short names, Meyer et al., 2003). This suggests that speakers plan the object names sequentially and do not plan far ahead. We examined whether the same gaze pattern was found when speakers made anticipation errors, or whether such errors were associated with atypically short inspection times of the incorrectly named object or skipping of these objects. The gaze patterns preceding errors were indistinguishable from those preceding correct utterances. The methodological and theoretical implications of this finding are discussed.
It is generally agreed that we generate spoken utterances incrementally: We plan the first part of our utterance before we begin to speak and plan the rest while we are talking (Kempen & Hoenkamp, 1987; Levelt, 1989). Incremental language production allows speakers to use both their own time and the listener’s time in the most efficient way, but it is taxing for two reasons: First, speakers must co-ordinate speech planning and speech output in time, and second, they must distribute their processing resources over several concurrent activities. How speakers meet these challenges is the main question we have been addressing in an ongoing research programme. In many of the experiments in this project, we have asked speakers to name sets of pictured objects in utterances such as “kite doll tap sock whale globe” (Figure 1). We recorded their eye movements and carried out detailed analyses of the relationship between the speakers’ eye movements and their speech output. In these experiments we deliberately minimised grammatical encoding requirements in order to focus on one key component of speech production – lexical retrieval – and its co-ordination with visual information uptake and speech output.

These studies have generated a remarkably consistent picture of the relationship between eye gaze and speech. Speakers usually fixate upon each object they name in the order of mention. Their eyes run slightly ahead of the overt speech, with the saccade from

Figure 1. An example of a display used in Experiments 1 and 2. On the screen, each object covered 3° of visual angle in its longest dimension. The objects were arranged on a virtual oval with a width of 20° and a height of 16°.
one object to the next occurring about 150–200 ms before the onset of the first object’s name. The gaze duration for an object (defined as the time between the onset of the first fixation and the offset of the last fixation on the object) depends on the time speakers require to identify the object, select its name and retrieve the corresponding morphological and phonological form (Meyer & Lethaus, 2004; Meyer, Roelofs & Levelt, 2003; Meyer, Sleiderink & Levelt, 1998; Meyer & van der Meulen, 2000; see also Griffin, 2001, 2004a; Griffin & Bock, 2000). In other words, speakers only initiate the shift of gaze to a new object after they have planned the name of the current object to the level of phonological form.

It is perhaps not too surprising to find that speakers usually look at the objects they name; in many cases this is simply necessary to identify the objects. In addition, looking at the objects in the order of mention may support the conceptual ordering and linearisation processes that must be carried out to talk about a spatial arrangement in a sequence of words (e.g., Griffin, 2004a). The more surprising observation is that the speakers’ gaze durations usually exceed the time required to identify the objects. All current models of word planning assume that lexical access is based on conceptual, not visual, representations (e.g., Levelt, 1999; Levelt, Roelofs & Meyer, 1999; Rapp & Goldrick, 2000). Therefore one might expect that little would be gained from looking at the objects for longer than necessary to identify them. Different accounts for the speakers’ prolonged gazes have been put forward in the literature, which may not be mutually exclusive (for further discussion see Griffin, 2004a; Meyer & Lethaus, 2004). One hypothesis is that gazes to the referent objects play a supporting role in lexical activation (Griffin, 2004a; Meyer, van der Meulen & Brooks, 2004; but see Bock, Irwin, Davidson & Levelt, 2003).

A related hypothesis we are currently exploring is that speakers fixate upon an object they are about to name until they have completed all encoding processes for that object that are not automatic, i.e., all processes that require processing capacity. A commonly held view in the speech production literature is that speakers need processing resources to plan the content of their utterances (e.g., Goldman-Eisler, 1968) but that lexical access is an automatic process (e.g., Levelt, 1989). However, recent experimental studies suggest that lexical retrieval processes might also require processing capacity. Ferreira and Pashler (2002) reported a series of dual-task experiments in which participants had to name target pictures and simultaneously categorise tones as high, medium or low in pitch. Both the picture-naming latencies and the tone-discrimination latencies were affected by the ease of semantic and morphological encoding of the picture names. This suggests that the response to the tone was only prepared after semantic and morphological encoding had been completed, and that these processes are capacity demanding. Recent dual-task experiments carried out in our own lab (Cook & Meyer, submitted) suggest that the generation of the phonological code of words also requires processing capacity. In addition, processing resources are necessary for self-monitoring processes, which compare the phonological representation of an utterance to the conceptual input (Hartsuiker & Kolk, 2001; Postma, 2000; Wheeldon & Morgan, 2002). Thus the only processes that are likely to be automatic appear to be phonetic encoding and articulatory planning processes.
The late shift of eye gaze, occurring after the completion of the phonological encoding processes, is compatible with the hypothesis that speakers focus on each object they name until they have completed the capacity-demanding processes.

If speakers planned their utterances strictly sequentially, dealing with one word at a time, as the speakers’ eye movements in the multiple-object naming task suggest, they should never commit anticipatory errors, such as (1) and (2) below (from Fromkin, 1973), in which words or parts of words appear earlier than they should. However, such errors occur quite regularly, and in most error corpora stemming from healthy adult speakers by far outnumber perseverations such as (3) (from Fromkin, 1973).

(1) “a leading list” (instead of “a reading list”, p. 243)
(2) “mine gold” (instead of “gold mine”, p. 256)
(3) “a phonological fool” (instead of “a phonological rule”, p. 244).

The only plausible way of accounting for word and sound anticipations is to assume that speakers do not always plan utterances strictly sequentially, but that at least occasionally several words are activated simultaneously. Current models of sentence planning see anticipatory errors as indicative of a forward looking system in which several words of an utterance are being processed in parallel, though usually with differing priorities (e.g., Dell, Burger & Svec, 1997).

There are different ways of reconciling the evidence gained from analyses of speech errors and from eye tracking studies. First, speakers may plan their utterances further ahead when they produce connected speech than when they name individual objects. In the latter case, there is little to gain from advance planning because the names of the objects can be selected independently of each other (e.g., Levelt & Meyer, 2000; Smith & Wheeldon, 1999). Though it may be the case that speakers use different planning strategies for sentences and lists of object names, anticipatory errors are not confined to sentence production, but occur as well when speakers recite lists of words (Dell, Reed, Adams & Meyer, 2000; Shattuck-Hufnagel, 1992) or, as will be shown below, name sets of objects.

Second, in a multiple-object naming task, the object a person is fixating upon may not be the only one whose name is currently activated. Speakers may, for instance, process an extrafoveal object in parallel with the foveated object, as recent results by Morgan and Meyer (2005) suggest. They asked participants to name triplets of objects. During the eye movement from the first to the second object, the object initially shown in the second position (the interloper) was replaced by a new object (the target). The relationship between interloper and target was varied. The gaze duration for the target was shorter when interloper and target were identical or had homophonous names (e.g., animal bat/baseball bat) than when they were unrelated (fork/baseball bat). This demonstrates that the participants had processed the interloper, which they only viewed extrafoveally, and built upon this information when processing the target. One account of this preview effect is that speakers inspect the objects they name sequentially but process them in parallel: A speaker looking at one object and planning its name gives processing priority to that object but simultaneously allocates some processing resources to the next object he/she will have to name. Consequently, that object can be recognised and its name can
become activated prior to fixation. On this account, anticipatory errors can arise when the names of extrafoveal objects become highly activated too early and interfere with the selection of the name of the foveated object. More generally, errors can arise in a naming task because in addition to the name of the object that a speaker intends to name next, other object names can become highly activated and compete with the target name to be selected and produced (see also Morsella & Miozzo, 2002; Navarrete & Costa, 2005).

Finally, the tight co-ordination of eye gaze and speech may be a typical feature of correct utterances (which were the only utterances included in our earlier analyses), whereas a different type of co-ordination may be seen in anticipatory errors. Perhaps anticipatory errors arise because speakers sometimes fail to inspect the objects in the order in which they should mention them or because they get ahead of themselves and initiate the shift of visual attention and processing resources to a new object before the name of the present object has become sufficiently activated to be selected. In other words, errors may not always be associated with the premature automatic activation of object names, as discussed before, but may be caused by ill-timed voluntary shifts of eye gaze and attention. The goal of the present study was to discriminate between these two options by comparing the gaze patterns occurring when speakers named sets of objects correctly and when they made anticipation errors.

Our study is similar to a study by Griffin (2004b), who analysed 41 ordering and selection errors from several picture-naming experiments conducted with young and older adults. Griffin also tested the hypothesis that errors would be accompanied by atypical gaze patterns, in particular abnormally short gazes to incorrectly named objects. She found no evidence in support of this hypothesis. We used a larger error corpus than Griffin, which was more homogeneous, as all errors stemmed from two closely related experiments carried out with undergraduate students. In addition, we included only anticipation errors. In our experiments, the speakers named the same sets of objects on up to 20 trials, which allowed us to compare the gaze pattern observed on error trials to the pattern observed when the same speaker named the same set of objects correctly. This had not been possible in Griffin’s study.

1. Method

Materials. The errors were collected from two experiments that required participants to name sets of six objects each presented in a circle on a computer screen, as shown in Figure 1. Forty-eight line drawings of objects had been selected from a picture gallery available at the University of Birmingham. Twenty-four objects had monosyllabic names and the others had disyllabic names, which were mono-morphemic and stressed on the first syllable. From this pool of items, we constructed four displays of six objects each with monosyllabic names and four displays of six objects each with disyllabic names. The names of the objects within a set were semantically and phonologically unrelated (see Appendix).
**Ch. 23: Speech-to-Gaze Alignment in Anticipation Errors**

**Apparatus.** The experiment was controlled by the software package NESU provided by the Max Planck Institute for Psycholinguistics, Nijmegen. The pictures were presented on a Samtron 95 Plus 19-inch screen. Eye movements were monitored using an SMI Eyelink-1 eye tracking system. Throughout the experiment, the x- and y-coordinates of the participant’s point of gaze for the right eye were estimated every 4 ms. The positions and durations of fixations were computed using software provided by SMI. Speech was recorded using a Sony ECM-MS907 microphone and a Sony TCD-D8 DAT recorder.

**Participants.** Twenty-four undergraduate students of the University of Birmingham participated in each experiment. They were native speakers of English and had normal or corrected-to-normal vision and received payment or course credits for participation.

**Procedure.** The participants were tested individually, seated in a sound-attenuated booth. At the beginning of the experiment, they received a booklet showing the experimental pictures and the expected names. They studied the pictures and their names and then completed a practice block in which they named the objects shown individually. Any naming errors were corrected by the experimenter. Then the headband of the eye tracking system was placed on the participant’s head and the system was calibrated.

In Experiment 1, speakers named the objects in clockwise order, starting at the 12 o’clock position. A fixation point was presented in this position at the beginning of each trial for 700 ms. Speakers were told that on the first presentation of the display, they should aim to name the objects accurately, and on the following seven presentations they should try to name the same objects as quickly as possible. The experimenter terminated the presentation of the objects as soon as the participant had completed the name of the sixth object. The inter-trial interval was 1250 ms. Each display was presented on eight successive trials. Twelve participants named the objects with monosyllabic names, and twelve named those with disyllabic names. After all displays had been tested once, there was a short break, after which all displays were presented and named again.

In Experiment 2, each participant named each monosyllabic and each disyllabic sequence on twenty successive trials. Twelve participants did so at a fast speech rate (3.5 words/second for monosyllabic items and 2.8 words/second for disyllabic items) and twelve at a slower rate (2.3 words/second and 1.8 words/second for monosyllabic and disyllabic items respectively). This allowed us to separate the effects of speech rate and repetition of the materials. In a training block, participants were first extensively trained to produce practice sequences at these rates along with the beat of a metronome and then learned to maintain the rates after the metronome had been switched off. During the main experiment, speakers were instructed to maintain the speed they had been trained to use and they heard a tone at the moment when they should have completed the last word of a sequence. Each of the eight displays was presented on twenty successive trials. The first two trials using each display were considered practice trials and were not included in the analyses. The monosyllabic and disyllabic items were tested in separate blocks, with six participants within each speed condition beginning with the monosyllabic and six participants with the disyllabic items.
2. Results and discussion

One goal of the experiments was to study how the speakers’ speech-to-gaze alignment would change when they used different speech rates and after they had extensively practised producing the object names. The relevant findings are reported in full elsewhere (Meyer, van der Meulen & Wheeldon, in prep). Very briefly, we found that the participants usually fixated upon each of the six objects. The likelihood of speakers naming objects without fixating upon them first was less than 10% and did not change much with practice or when different speech rates were used. The participants’ eyes ran slightly ahead of their overt speech, i.e., they usually initiated the saccade from one object to the next shortly before they began to say the name of the first object. In Experiment 1, the participants’ speech rate increased slightly over the repetitions of the materials, as one would expect. Simultaneously, the gaze durations for the objects decreased, as did the time between the end of the gaze to an object and the onset of its name (the eye–speech lag). In other words, with practice, the co-ordination of eye gaze and speech became tighter. In Experiment 2, the participants maintained a fairly constant fast or slower speech rate across the repetitions of the materials. Their gaze durations for the objects remained stable as well, but the eye–speech lag decreased slightly over the repetitions. All in all, the temporal co-ordination between the speakers’ eye gaze and their speech was not greatly affected by practice or the requirement to use different speech rates.

The present study focuses on the participants’ speech errors. All utterances were transcribed and the errors were coded following Dell (1986). The error rates were similar for the monosyllabic and the disyllabic sequences (7.6 and 7.3% respectively in Experiment 1 and 7.7 and 11.8% respectively in Experiment 2); therefore the data for these stimuli are collapsed in the analyses that follow.

Table 1 shows the rates of the main error types in each experiment, as well as the percentage of each error type in the total corpus. Thirty-five per cent of the errors were hesitations or

<table>
<thead>
<tr>
<th>Table 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>The percentages of correct responses and of the different error types in Experiments 1 and 2 are shown along with the percentage of occurrences of each error type in the error corpus. Numbers are collapsed for monosyllabic and disyllabic sequences</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Response type</th>
<th>Experiment 1</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Frequency</td>
<td>%</td>
<td>Frequency</td>
<td>%</td>
<td>Frequency</td>
</tr>
<tr>
<td>Correct responses</td>
<td>1421</td>
<td>92.5</td>
<td>6236</td>
<td>90.2</td>
<td>277</td>
</tr>
<tr>
<td>Hesitations and pauses</td>
<td>28</td>
<td>1.8</td>
<td>249</td>
<td>3.6</td>
<td>182</td>
</tr>
<tr>
<td>Non-contextual errors</td>
<td>33</td>
<td>2.1</td>
<td>149</td>
<td>2.2</td>
<td>18</td>
</tr>
<tr>
<td>Perseverations</td>
<td>9</td>
<td>0.6</td>
<td>9</td>
<td>0.1</td>
<td>18</td>
</tr>
<tr>
<td>Exchanges</td>
<td>6</td>
<td>0.4</td>
<td>18</td>
<td>0.3</td>
<td>24</td>
</tr>
<tr>
<td>Anticipations</td>
<td>39</td>
<td>2.5</td>
<td>208</td>
<td>3.0</td>
<td>247</td>
</tr>
<tr>
<td>Others</td>
<td>0</td>
<td>0.0</td>
<td>43</td>
<td>0.6</td>
<td>43</td>
</tr>
<tr>
<td>Total errors</td>
<td>115</td>
<td>7.5</td>
<td>676</td>
<td>9.8</td>
<td>791</td>
</tr>
</tbody>
</table>
Ch. 23: Speech-to-Gaze Alignment in Anticipation Errors

Figure 2. An example of the speech-to-gaze-alignment during the production of an anticipatory error. The order and duration of the gazes to each object are shown as well as the order and duration of the spoken object names. The speech error section is shaded grey. As can be seen, the speech error is not reflected in the gaze pattern.

pauses, and 23% were non-contextual errors, i.e. speakers produced object names that were not part of the current sequence. Among the ordering errors, anticipations were far more frequent than perseverations. This is consistent with the patterns reported in other speech error studies (e.g., Dell, Burger & Svec, 1997). In 86% of the anticipation errors, the object name that was initiated too early was the name of the next object in the sequence. A typical gaze pattern on a trial on which an anticipation error occurred is illustrated in Figure 2.

Our main question was whether the speakers’ speech-to-gaze alignment in sequences including an anticipatory error would be different from the alignment in correct utterances. To answer this question, we carried out detailed analyses of a subset of anticipation errors, fulfilling the following criteria: (a) The trial including the error was preceded and followed by a trial in which the same utterance was produced correctly. This criterion was used because the speech-to-gaze alignment on error trials was to be compared to the alignment on trials in which the same objects were named correctly by the same speaker at a similar speech rate and after a similar number of repetitions of the materials. Using this criterion meant that we could not consider errors arising on the first or last repetition of a sequence. (b) The eye movements were recorded for all three trials (i.e. there were no technical errors) and the speaker fixated upon at least three of the six objects on each trial. This criterion was adopted because it makes little sense to say that an object was skipped when the other objects of the sequence were not looked at either. We call the object that was not named correctly on the error trial the target.

There were 104 triplets of trials that met both criteria. As Table 2 shows, on most of these error trials the participants uttered the onset consonant or the first few segments of a later object name and then interrupted and corrected themselves.1 These errors accounted for 71% of the errors analysed. In almost all of the remaining errors (28% of all errors) the participants uttered the entire name of an upcoming object. Participants stopped and corrected themselves on only 30% of these whole word anticipation trials. On the remaining trials, the target name was simply omitted, i.e. speakers fluently produced five instead of six object names.

1 The complete error corpus can be obtained from the authors.
Table 2
A break-down of the corpus of 104 anticipatory errors entered for analysis by error type and distance between the target and the anticipated word. In the examples, the anticipated speech is shown in bold and the source of the anticipation is underlined

<table>
<thead>
<tr>
<th>Error type</th>
<th>Examples</th>
<th>Distance from target</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Next word</td>
</tr>
<tr>
<td></td>
<td></td>
<td>word</td>
</tr>
<tr>
<td>Part of word anticipation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single phoneme</td>
<td><em>Owl</em> mask web <em>s-corn</em> sword brush</td>
<td>16</td>
</tr>
<tr>
<td>Two phonemes</td>
<td><em>Lamp</em> coin rope <em>st-bat</em> straw pie</td>
<td>33</td>
</tr>
<tr>
<td>Three or more phonemes</td>
<td><em>Penguin</em> car-ladder whistle carrot</td>
<td>14</td>
</tr>
<tr>
<td>Whole word anticipation</td>
<td><em>Whistle eh</em> ladder whistle carrot</td>
<td>25</td>
</tr>
<tr>
<td>Phoneme anticipation</td>
<td><em>Cord-corn</em> sword</td>
<td>1</td>
</tr>
</tbody>
</table>

One hypothesis to be examined was that the target object was less likely to be looked at on error trials than on the preceding and following trials. This hypothesis was not borne out: There were two error trials on which the target object was not fixated upon before the corresponding utterance was initiated, compared to three preceding and five following trials on which the corresponding object was not looked at before it was named. In other words, on the majority of the trials, the speakers looked at the target object before naming it – correctly or incorrectly. Interestingly, on all 20 error trials on which the target name was omitted, all six objects were inspected in the order in which they should have been mentioned. In other words, the participants looked at the target object, as they did on correct trials, but failed to mention it.

The second hypothesis was that the target object would be inspected for a shorter time before the onset of an error than before the onset of a correct name. The mean pre-speech gaze durations (the time intervals between the onset of the gaze to the target object and the onset of the corresponding expression) are shown in Table 3. We included only those cases where the target was fixated upon on the error trial, the preceding and the following trial (94 triplets of trials). On seven trials the target was only inspected after the onset of the target name. This happened most often on the trial following the error (four cases). For the remaining trials, we computed the pre-speech gaze duration. On most of these trials, the saccade to the following object occurred before the onset of the target name. Therefore, the pre-speech gaze duration was defined as the time interval between the onset of the first fixation to the target and the end of the last fixation. On 70 trials, the saccade to the next object followed the onset of the target name. In these cases the pre-speech gaze duration was terminated by the onset of the correct or incorrect target name. Contrary to the hypothesis, the mean pre-speech gaze duration did not differ significantly between error trials (312 ms) and the preceding and following trials with (means: 321 and 312 ms respectively, $F < 1$).
Table 3
Mean gaze durations, standard deviations (in ms) and number of observations (n) are shown for the error trials and the correct trials that preceded and followed them. The total gaze duration for the error picture is broken down into gaze duration preceding speech onset and gaze duration following speech onset. The lag between the end of the last fixation on the target and the onset of its name is also shown. Note that the total speech duration is not the sum of pre- and post-speech durations as they are calculated over different subsets of errors.

<table>
<thead>
<tr>
<th>Dependent measure</th>
<th>Preceding trial</th>
<th>Error trial</th>
<th>Following trial</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>(n)</td>
</tr>
<tr>
<td>Total gaze duration</td>
<td>334</td>
<td>139</td>
<td>(94)</td>
</tr>
<tr>
<td>Pre-speech gaze duration</td>
<td>321</td>
<td>132</td>
<td>(92)</td>
</tr>
<tr>
<td>Post-speech onset gaze duration</td>
<td>87</td>
<td>53</td>
<td>(14)</td>
</tr>
<tr>
<td>Eye–speech lag</td>
<td>104</td>
<td>122</td>
<td>(92)</td>
</tr>
</tbody>
</table>

The total gaze duration, however, differed significantly between the three trial types, with longer gaze durations being observed to objects named incorrectly than to objects named correctly, $F(2, 272) = 6.2, p < 0.01$. This result was due to differences in the participants’ speech-to-gaze alignment after the onset of the referring expressions. On 83% of all trials, the saccade to the next object occurred before the onset of the target name. Therefore, the participants were normally looking at the next object when they initiated the correct or incorrect target name. However, occasionally, the order of events was reversed, and the participants were still looking at the target when they initiated its name. This was more likely to be the case on error trials (34%) than on the trials preceding and following them (17%). Considering only the trials on which participants were still looking at the target object after the onset of its name, we found that this post-speech gaze duration was longer on error trials than on correct trials, $t(68) = 4.28, p < 0.01$. In addition, speakers were more likely to inspect the target object again, after having fixated another object, when they had made an error than when the utterance had been correct: There were 16 regressions to the target on error trials, compared to a total of three on preceding and following trials. These differences in the gaze patterns between correct and error trials most likely arose because participants detected and corrected their errors. Griffin (2004b) reported similar findings and more extensive analyses of eye movements during error detection and repair.

3. Conclusions

Our experiments demonstrated that speakers make anticipation errors when they name sequences of objects as they do in normal conversational speech. Indeed, most of the ordering errors we observed involved the early intrusion of part of an upcoming picture.
name in the sequence. Since most of the errors were corrected by the speakers, we assume that they selected the correct concept, but that the name of the next object was activated and selected prematurely. Our error data can be explained within standard frameworks, which propose that the early activation of lexical or segmental representations leads to competition with target representations and occasionally results in the selection of incorrect words or sounds (e.g., Dell, Burger & Svec, 1997). Our aim was to determine whether this early activation was in any part due to early voluntary shifts of visual attention and eye gaze to objects whose names were produced too early. If so, we should observe atypical gaze patterns on error trials compared to correct trials. However, similar to Griffin (2004b), our data show that the gaze pattern leading up to anticipation errors is indistinguishable from the pattern leading to correct utterances. Anticipation errors were not accompanied by skipping of objects or by early saccades away from the target object.

Where then did the early activation of object names come from? In our experiments, the objects were sized and arranged such that a speaker fixating upon one object could easily identify the neighbouring objects. As discussed in the Introduction, it is possible that speakers allocated their attention in a distributed way across the object they were about to name and the following object (e.g., Cave & Bichot, 1999; Morgan & Meyer, 2005). Most of their attentional resources would be allocated to the object to be named first with the remainder allocated to the following object. Occasionally the representations pertaining to the next object could become highly activated and would be selected instead of those pertaining to the target. The fact that the participants named the same objects on many successive trials may have facilitated such extrafoveal processing of upcoming objects. In addition, the participants may have built up a working memory representation of the name sequences, which may also have contributed to the early activation of object names.

A methodological implication of our research and the research carried out by Griffin (2004b) is that speech errors, and, more generally, the content of a speaker’s utterance cannot be predicted on the basis of the gaze pattern. Errors occurred on trials in which the gaze pattern was entirely regular. Gaze patterns therefore do not provide us with the whole picture regarding the availability of words within the speech production system. They are not a reliable predictor of utterance content.

A theoretical implication of our results is that at least in multiple-object naming tasks, anticipatory errors are not linked in any obvious way to the timing of voluntary shifts of eye gaze and the speakers’ focus of attention. These errors do not occur as a result of a particular planning strategy. However, one conclusion that should not be drawn from these findings is that the speakers’ eye movements and their speech planning are unrelated. We found here, as in earlier studies, that on most trials, the speakers looked at all objects they named, in the order in which they should name them. Moreover, in earlier research we showed that the gaze durations for objects that were to be named depended on the total time speakers required to identify them and to plan the utterances about them to the level of phonological form (e.g., Levelt & Meyer, 2000; Meyer & van der Meulen, 2000; Meyer, Roelofs & Levett, 2003). We have proposed that the reason why speakers look at the objects until they have completed all of these processes is that these processes require processing resources. In other words, speakers initiate a saccade to a new object as soon as
they have completed all capacity-demanding processes for the present object. If this view is correct, eye monitoring can be used to track the time course of the speaker’s allocation of processing resources: They demonstrate when and for how long one object is the focus of the speaker’s attention and when the focus of attention is moved to a new object.

Acknowledgements

The research was supported by Economic and Social Research Council Grant R000239659 to the first and second authors.

Appendix: Materials of Experiments 1 and 2

A.1. Monosyllabic object names

<table>
<thead>
<tr>
<th>lamp</th>
<th>coin</th>
<th>rope</th>
<th>bat</th>
<th>straw</th>
<th>pie</th>
</tr>
</thead>
<tbody>
<tr>
<td>pin</td>
<td>toe</td>
<td>spoon</td>
<td>leaf</td>
<td>bow</td>
<td>rat</td>
</tr>
<tr>
<td>owl</td>
<td>mask</td>
<td>web</td>
<td>corn</td>
<td>sword</td>
<td>brush</td>
</tr>
<tr>
<td>kite</td>
<td>doll</td>
<td>tap</td>
<td>sock</td>
<td>whale</td>
<td>globe</td>
</tr>
</tbody>
</table>

A.2. Disyllabic object names

<table>
<thead>
<tr>
<th>lemon</th>
<th>toilet</th>
<th>spider</th>
<th>pencil</th>
<th>coffin</th>
<th>basket</th>
</tr>
</thead>
<tbody>
<tr>
<td>saddle</td>
<td>bucket</td>
<td>penguin</td>
<td>ladder</td>
<td>whistle</td>
<td>carrot</td>
</tr>
<tr>
<td>barrel</td>
<td>wardrobe</td>
<td>monkey</td>
<td>statue</td>
<td>rabbit</td>
<td>garlic</td>
</tr>
<tr>
<td>sausage</td>
<td>dragon</td>
<td>robot</td>
<td>tortoise</td>
<td>candle</td>
<td>orange</td>
</tr>
</tbody>
</table>

References


Chapter 24

COMPARING THE TIME COURSE OF PROCESSING INITIALLY AMBIGUOUS AND UNAMBIGUOUS GERMAN SVO/OVS SENTENCES IN DEPICTED EVENTS

PIA KNOEFERLE

Saarland University, Germany
Abstract

The Coordinated Interplay Account by Knoeferle and Crocker (accepted) predicts that the time course with which a depicted event influences thematic role assignment depends on when that depicted event is identified as relevant by the utterance. We monitored eye movements in a scene during the comprehension of German utterances that differed with respect to when they identified a relevant depicted event for thematic role assignment and structuring of the utterance. Findings confirmed the coordinated interplay account: Gaze patterns revealed differences in the time course with which a depicted event triggered thematic role assignment depending on whether that event was identified as relevant for comprehension early or late by the utterance.
The monitoring of eye movements in scenes containing objects has revealed that the type of visual referential context rapidly influences the initial structuring and interpretation of a concurrently presented utterance (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). Gaze patterns have further shown that people established reference to an object more rapidly when the scene contained a contrasting object of the same kind than when it did not (Sedivy, Tanenhaus, Chambers, & Carlson, 1999). In more recent research, scenes depicted objects (e.g., a motorbike) and characters (e.g., a man). Semantic information provided by a verb (ride) combined with properties of that verb’s thematic agent (man) enabled the rapid anticipation of the verb’s theme (Kamide, Altmann, & Haywood, 2003).

Studies by Knoeferle, Crocker, Scheepers, and Pickering (2005) have extended this work by examining the influence of richer visual environments that contained explicitly depicted agent–action–patient events in addition to characters and objects. For initially structurally ambiguous German sentences, eye movements revealed the rapid, verb-mediated effects of the depicted events on the resolution of structural and thematic role ambiguity.

While there is a growing body of experimental research that examines the effects of increasingly rich, multi-modal settings on online utterance comprehension, findings from this research have not yet been fully and explicitly integrated into theories of online language comprehension (see Knoeferle, 2005). Existing theories of comprehension rather describe the workings of one cognitive system – the language system (e.g., Crocker, 1996; Frazier & Clifton, 1996; MacDonald, Pearlmutter, & Seidenberg, 1994; Townsend & Bever, 2001; Trueswell & Tanenhaus, 1994). Such ‘non-situated’ comprehension theories account for how various kinds of linguistic and world knowledge are integrated incrementally during comprehension. Knoeferle et al. (2005), however, have argued that to account for the influence of depicted events on thematic role assignment and structural disambiguation, it is necessary to adopt a framework that relates the language system to the perceptual systems, and that offers a suitably rich inventory of mental representations for objects and events in the environment.

General cognitive frameworks (e.g., Anderson et al., 2004; Cooper & Shallice, 1995; Newell, 1990), theories of the language system (e.g., Jackendoff, 2002), and approaches to embodied comprehension (e.g. Barsalou, 1999; Bergen, Chang, & Narayan, 2004; Chambers, Tanenhaus, & Magnuson, 2004; Zwaan, 2004) represent an important step in this direction. Many of these approaches provide frameworks and mental representations for the integration of information from the language comprehension and visual perception systems. Some embodied frameworks, moreover, offer an inventory of comprehension steps (e.g., the activation of a word) in situated settings (e.g., Zwaan, 2004).

These existing approaches do not yet, however, explicitly characterize the time course of situated sentence comprehension and its relation to the time course of visual processes. In brief, an integral part – a processing account – of a theory on situated sentence comprehension is still missing in current theory development.

The first step in characterizing the online interaction of language comprehension, and attention in the scene has been made by Tanenhaus et al. (1995) (see Cooper, 1974). Their research showed the close time-lock between understanding of a word and inspection of appropriate referents in a scene (see Roy & Mukherjee, 2005, for relevant
modelling research). Altmann and Kamide (1999) further investigated the time lock between comprehension and attention, and demonstrated that attention to an object can even precede its mention.

What the fixation patterns in these studies do not, however, permit us to determine is whether such a close temporal coordination also holds between the point in time when a relevant object/event is identified by the utterance, and the point in time when that object/event influences comprehension. Findings by Knoeferle et al. (2005) add insights regarding this question: Importantly, gaze patterns revealed that the effects of depicted events were tightly coordinated with when the verb identified these events as relevant for comprehension.

In German, a case-marked article can determine the grammatical function and thematic role of the noun phrase it modifies. Both subject(NOM)–verb–object(ACC) (SVO) and object(ACC)–verb–subject(NOM) (OVS) ordering are possible. The studies by Knoeferle et al. (2005) investigated comprehension of initially structurally ambiguous SVO/OVS sentences when neither case-marking nor other cues in the utterance determined the correct syntactic and thematic relations prior to the sentence-final accusative (SVO) or nominative (OVS) case-marked noun phrase.

For early disambiguation, listeners had to rely on depicted event scenes that showed a princess washing a pirate, while a fencer painted that princess. Listeners heard *Die Prinzessin wäscht/malt den Pirat/der Fechter.* (‘The princess (amb.) washes/paints the pirate (ACC)/the fencer (NOM)’). The verb in the utterance identified either the washing or the painting action as relevant for comprehension. Events either determined the princess as agent and the pirate as patient (washing, SVO), or the princess as patient and the fencer as agent (painting, OVS). Eye movements for the SVO compared with the OVS condition did not differ prior to the verb. After the verb had identified the relevant depicted event, and before the second noun phrase resolved the structural and thematic role ambiguity, more anticipatory inspection of the patient (the pirate) for SVO than OVS, and more eye movements to the agent (the fencer) for OVS than SVO revealed which thematic role people had assigned to the role-ambiguous character (the princess). The observed gaze patterns reflected rapid thematic role assignment and structural disambiguation through verb-mediated depicted events.

Based on these and prior findings, Knoeferle (2005) and Knoeferle and Crocker (accepted) explicitly characterized the temporal relationship between sentence comprehension and attention in a related scene as a ‘Coordinated Interplay Account’ (CIA) of situated utterance comprehension. The CIA identifies two fundamental steps in situated utterance comprehension. First, comprehension of the unfolding utterance guides attention, establishing reference to objects and events (Tanenhaus et al., 1995), and anticipating likely referents (see Altmann & Kamide, 1999). Once the utterance has identified the most likely object or event, and attention has shifted to it, the attended scene information then rapidly influences utterance comprehension (for further details, see Knoeferle and Crocker (accepted)).

If the coordinated interplay outline is correct, then the time course with which a depicted event affects structural disambiguation and thematic role assignment depends
on when that depicted event is identified as relevant for comprehension by the utterance. For earlier identification of a relevant depicted event, we would expect to find an earlier influence of that event on comprehension than for cases when identification of a relevant event takes place comparatively later.

The present experiment directly tests the temporal-coordination prediction of the CIA. We examine thematic role assignment and the structuring of an utterance for two types of German sentences that differ with respect to when the utterance identifies relevant role relations in the scene. For initially ambiguous German SVO/OVS sentences, the first noun phrase is ambiguous regarding its grammatical function and thematic role. Scenes were role-ambiguous (i.e., the first-named character was the agent of one, and at the same time the patient of another event). Only when the verb identified one of the two events as relevant for comprehension did it become clear whether the role-ambiguous character had an agent or patient role. The earliest point at which a thematic role can be assigned to the first noun phrase in this scenario is shortly after the verb has identified a relevant depicted event and its associated role relations. We would expect gaze patterns to reveal the influence of depicted events post-verbally for ambiguous sentences if predictions of the coordinated interplay outline are valid (see Knoeferle et al., 2005).

In contrast, for unambiguous sentences, nominative and accusative case-marking on the determiner of the first noun phrase for SVO and OVS sentences respectively should permit identification of the relevant depicted role relations earlier, shortly after the first noun phrase: Object case-marking on the determiner of the first noun phrase can be combined with the event scene, showing the referent of the first noun phrase as the patient of one depicted event. For unambiguous subject-initial sentences, the subject case on the determiner of the first noun phrase can be used to identify its referent as an agent. If listeners rely on this early identification of relevant thematic relations in the scene, gaze patterns should indicate disambiguation tightly temporally coordinated one region earlier than for the ambiguous sentences (on the verb).

1. Experiment

1.1. Method

1.1.1. Participants

There were 32 participants. All were native speakers of German, and had normal or corrected-to-normal vision and hearing.

1.1.2. Materials and design

Figure 1 shows a complete image set for one item. Figures 1a and 1b were presented with initially structurally ambiguous canonical (SVO) and non-canonical (OVS) sentences (see Table 1, sentences 1a and 1b). In addition we included unambiguous sentences that were
Table 1
Example item sentence set for the images in Figure 1

<table>
<thead>
<tr>
<th>Image</th>
<th>Condition</th>
<th>Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fig. 1a</td>
<td>SVO ambiguous 1a</td>
<td>Die Frau Orange tritt in diesem Moment den Sir Zwiebel.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The Ms Orange (ambiguous) kicks currently the Sir Zwiebel (object).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘Ms Orange kicks currently Sir Onion.’</td>
</tr>
<tr>
<td>Fig. 1a</td>
<td>OVS ambiguous 1b</td>
<td>Die Frau Orange schlägt in diesem Moment der Sir Apfel.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The Ms Orange (ambiguous) hits currently the Sir Apple (subject).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘Sir Apple hits currently Ms Orange’.</td>
</tr>
<tr>
<td>Fig. 1a</td>
<td>SVO unambiguous 2a</td>
<td>Der Herr Orange tritt in diesem Moment den Sir Zwiebel.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The Mr Orange (subject) kicks currently the Sir Zwiebel (object).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘Mr Orange kicks currently Sir Onion.’</td>
</tr>
<tr>
<td>Fig. 1a</td>
<td>OVS unambiguous 2b</td>
<td>Den Herrn Orange schlägt in diesem Moment der Sir Apfel.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The Mr Orange (object) hits currently the Sir Apple (subject).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘Sir Apple hits currently Mr Orange.’</td>
</tr>
<tr>
<td>Fig. 1b</td>
<td>SVO ambiguous 1a</td>
<td>Die Frau Orange schlägt in diesem Moment den Sir Apfel.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The Ms Orange (ambiguous) hits currently the Sir Apple (object).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘Ms Orange hits currently Sir Apple.’</td>
</tr>
<tr>
<td>Fig. 1b</td>
<td>OVS ambiguous 1b</td>
<td>Die Frau Orange tritt in diesem Moment der Sir Zwiebel.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The Ms Orange (ambiguous) kicks currently the Sir Onion (subject).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘Sir Onion kicks currently Ms Orange’.</td>
</tr>
<tr>
<td>Fig. 1b</td>
<td>SVO unambiguous 2a</td>
<td>Der Herr Orange schlägt in diesem Moment den Sir Apfel.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The Mr Orange (subject) kicks currently the Sir Apple (object).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘Mr Orange hits currently Sir Apple.’</td>
</tr>
<tr>
<td>Fig. 1b</td>
<td>OVS unambiguous 2b</td>
<td>Den Herrn Orange tritt in diesem Moment der Sir Zwiebel.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The Mr Orange (object) kicks currently the Sir Onion (subject).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>‘Sir Onion kicks currently Mr Orange’.</td>
</tr>
</tbody>
</table>
presented with the same images (2a and 2b in Table 1, Figure 1a). The unambiguous versions were created by replacing the ambiguous feminine noun phrase (Die Frau Orange (amb.), ‘the Ms Orange’) of sentences 1a and 1b with an unambiguous subject (NOM) or object (ACC) case-marked masculine noun phrase (Der Herr Orange (NOM), ‘the Mr Orange (subj)’; Den Herrn Orange (ACC), ‘The Mr Orange (obj)’).1

Each image was presented in two versions for counter-balancing reasons, resulting in two images and eight sentences for an item (Table 1, Figure 1). Stereotypicality and plausibility biases were absent in our materials and could not enable disambiguation. The corresponding words for the conditions of an item were matched for length and frequency of lemmas (Baayen, Piepenbrock, & Gulikers, 1995). To exclude the influence of inter-national cues on early scene-based disambiguation for ambiguous sentences (e.g., 1a and 1b), the ambiguous sentences were cross-spliced up to the second noun phrase for half of the items. To give an example: For the ambiguous OVS sentence 1b (Figure 1a), the beginning of the SVO sentence 1a (Fig. 1b) was spliced in before the second noun phrase. For ambiguous SVO sentences (e.g., 1a, Fig. 1a), the beginning of the OVS sentence 1b (Fig. 1b) was spliced in prior to the second noun phrase. Crossing ambiguity (ambiguous, unambiguous) with sentence type (SVO, OVS) created four conditions (ambiguous SVO/OVS, and unambiguous SVO/OVS).

For the initially structurally ambiguous sentences utterance-based ambiguity resolution could only occur through case-marking on the sentence-final noun phrase. Concurrently presented depicted events, however, showed who-does-what-to-whom, and offered role information for earlier thematic role assignment and structural disambiguation. The verb differed between the SVO condition (tritt, ‘kicks’, sentence 1a) and the OVS condition (schlägt, ‘hits’, 1b, Table 1, Fig. 1a). While the role-ambiguous character (the orange) was the agent of a kicking-event for SVO sentences (orange-kicking-onion), it was the patient of the apple-hitting event (apple-hitting-orange) in OVS sentences (Fig. 1a). If people can use the depicted events rapidly after the verb identified them as relevant, then we would expect more eye movements to the patient for SVO than OVS sentences, and more looks to the agent for OVS than SVO sentences during the post-verbal adverb (see Knoeferle et al., 2005).

For the unambiguous conditions, in contrast to the ambiguous ones, subject and object case-marking on the determiner of the first noun phrase contributed towards identifying the referent of the first noun phrase as either the agent of the kicking-event for SVO (Table 1, 2a, Fig. 1a), or as the patient of the hitting-event for OVS sentences (Table 1, 2b, Fig. 1a). While unambiguous subject case-marking on the determiner of the first noun phrase does not entirely disambiguate the thematic role of that noun phrase (sentences may be passive), this ambiguity is resolved as soon as the main verb appears in second position. As the verb unfolds, its information about relevant events can be combined with the information about thematic roles provided by case-marking. As a result, depicted

1 We chose fruit and vegetable characters for all stimuli in the present study with the future goal of conducting the experiment with young children. Keeping scenes simple would minimize changes in the materials between the adult and child studies.
P. Knoeferle

events were available earlier for thematic role assignment. If identification of depicted events and their effects upon thematic interpretation are tightly temporally coordinated, we would expect the same gaze patterns, but one region earlier than for ambiguous sentences (i.e., on the verb rather than post-verbally).

There were 24 experimental items. Each participant saw an equal number of SVO and OVS items, and an equal number of initially ambiguous and unambiguous sentences in an individually randomized list. There were 48 filler items. Eight started with an adverbial phrase and images showed two characters with only one performing an action; eight started with an unambiguously case-marked noun phrase, and images had four characters, with two performing an action; eight started with an unambiguously case-marked noun phrase and scenes showed no events; eight described an event that involved one character; sixteen started with a noun phrase coordination, and scenes showed five characters. The fillers ensured that sentences did not always begin with a noun phrase; that there was not always a verb in second position; that there was not always a depicted event. With respect to word order, 32 filler items had a subject–object constituent order and agent–patient relations; eight were dative-initial, and eight were passive sentences. One list contained 24 experimental and 48 filler items. Item trials were separated by at least one filler.

1.1.3. Procedure

An SMI EyeLink I head-mounted eye tracker monitored participants’ eye movements with a sampling rate of 250 Hz. Images were presented on a 21-inch multi-scan colour monitor at a resolution of \(1024 \times 768\) pixels concurrently with spoken sentences. Prior to the experiment, the experimenter instructed participants to try to understand both sentences and depicted scenes. There was no other task. Next, the camera was set up to track the dominant eye of participants. Calibration and validation of the eye tracker was performed manually using a nine-point fixation until both were successful. These procedures were always repeated after approximately half of the trials; if necessary they were performed more often. For each trial, the image appeared 1000 ms before utterance onset. The first experimental item for each participant was preceded by three filler items. Between the individual trials, participants fixated a centrally located fixation dot on the screen. This allowed the eye-tracking software to perform drift correction if necessary. The experiment lasted approximately 30 min.

1.1.4. Analysis

The eye-tracker software recorded the XY coordinates of participants’ fixations. The coordinates were converted into distinct codes for the characters and background of each image so that participants’ fixations were mapped onto the objects of an image (the background, the orange, the onion, the apple, and the distractor objects for Fig. 1a). Characters were coded depending on their event role for the analyses. For the sentences describing Fig. 1a, for instance, the orange would be coded as ‘ambiguous’ (acting and
being acted upon), the apple as ‘agent’, the onion as ‘patient’, and the two pot plants as ‘distractors’. Consecutive fixations within one object region (i.e., before a saccade to another region occurred) were added together and counted as one inspection. Blinks and out-of-range fixations were added to previous fixations; contiguous fixations of less than 80 ms were pooled and incorporated into larger fixations. We report proportion of inspections and inferential analyses of the number of inspections for individual time regions (Figs 2–5). We rely on this attentional measure since previous studies have shown that it reflects online comprehension processes (e.g., Altman & Kamide, 1999; Sedivy et al., 1999; Tanenhaus et al., 1995).

The data presented for the individual time regions (Figs 2–5) are based on exact computations of these regions for each individual trial. Word onsets and offsets were marked for the first noun phrase, the verb, the adverb, and the second noun phrase in each item speech file. We computed the proportion of cases per sentence condition for which inspections started within a time region (‘NP1’, ‘VERB’, ‘ADV’, ‘NP2’). For the inferential analysis of inspection counts during a time region we used hierarchical log-linear models. These combine characteristics of a standard cross-tabulation chi-square test with those of ANOVA. Log-linear models neither rely upon parametric assumptions concerning the dependent variable (e.g., homogeneity of variance) nor require linear independence of factor levels, and are thus adequate for count variables (Howell, 2002).

Inspection counts for a time region were subjected to analyses with the factors target character (patient, agent), sentence type (SVO, OVS), and either participants ($N = 32$) or items ($N = 24$). We report effects for the analysis with participants as $LR\chi^2(subj)$ and for the analysis including items as a factor as $LR\chi^2(item)$.

### 1.2. Results

We describe gaze patterns and report inferential analyses for the four analysis regions (NP1, VERB, ADV, NP2) in the ambiguous (graphs marked ‘a’) and unambiguous conditions (graphs marked ‘b’) (see Figs 2–5). For Figs 2–5, ‘ambiguous’ refers to the role-ambiguous character (the orange in Fig. 1a), ‘patient’ to the patient of the action performed by the role-ambiguous character (the onion); ‘agent’ refers to the other agent (the apple), and ‘distr’ to the distractor objects (the two pot plants) (Fig. 1a).

#### 1.2.1. NP1 region

Figure 2 presents gaze patterns for the NP1 region. People mostly inspected the ambiguous character that is referred to by the first noun phrase. There were no differences in gaze patterns between the initially ambiguous and unambiguous conditions (see Figs 2a and b). Furthermore, there were equally many looks to the patient and agent characters for both ambiguous and unambiguous conditions.

Log-linear analyses for the ambiguous (all $LR\chi^2 < 1$, Fig. 2a) and unambiguous conditions (all $LR\chi^2(subj) < 1$, $ps > 0.2$ by items, Fig. 2b) showed no significant main
effects or interactions for sentence type (SVO, OVS) and target character (agent, patient) during the first noun phrase.

1.2.2. Verb region

Gaze patterns for the verb region are presented in Fig. 3. During the verb region, inspections to the first-mentioned ambiguous character (orange) have decreased. Eye-movement patterns to the two target characters (patient, agent) in the ambiguous and unambiguous conditions started to diverge. In the ambiguous conditions, gaze patterns did not yet indicate thematic role assignment and structural disambiguation. Rather, people began to direct their attention more to the patient of the action performed by the ambiguous character than to the other agent (Fig. 3a). This replicates findings from Knoeferle et al. (2005) and Knoeferle and Crocker (accepted), and could be due to visual factors (the first-named ambiguous character is oriented towards the patient), or linguistic expectation of the patient triggered by a main clause preference (e.g., Bever, 1970). The agent was inspected equally often for ambiguous SVO and OVS sentences (Fig. 3a). Log-linear analyses for the ambiguous conditions during the verb region confirmed that the main effect of target character (patient, agent) was significant ($LR\chi^2(subj) = 46.51, df = 1, p < 0.0001; LR\chi^2(item) = 46.51, df = 1, p < 0.0001$). There was no significant effect of sentence type ($LR\chi^2 < 1$), and no interaction of sentence type and target character ($ps > 0.1$). The difference between inspections to the background/ambiguous character for OVS compared with SVO sentences in the ambiguous conditions was not significant ($ps > 0.2$).

For the unambiguous conditions people also inspected the patient more often than the agent during the verb. At the same time, however, gaze patterns reflected early thematic role assignment and structuring of the utterance during the verb (one region earlier than
Ch. 24: Processing sentences in depicted events

Background

(a) (b)

Ambiguous Patient

Agent

Distractor

Entities

0.0

0.1

0.2

0.3

0.4

0.5

0.6

Inspection proportions

Figure 3. Proportions of inspections to characters on the verb for the ambiguous (a) and unambiguous (b) conditions.

for the initially ambiguous conditions of this study; see also Knoeferle et al. (2005)). People clearly began to inspect the patient more often for unambiguous SVO than OVS sentences, and the agent more often for unambiguous OVS than for SVO sentences during the verb (Fig. 3b). We suggest this is due to the combined use of unambiguous case-marking on the first noun phrase identifying the ambiguous character as either agent (SVO) or patient (OVS) and the relevant depicted event. As soon as case-marking on the determiner of the first noun phrase assigns an agent (SVO) or a patient role (OVS) to the ambiguous character (the orange), and once the verb identifies the relevant depicted event, people anticipate the relevant patient (SVO)/agent (OVS) associated with that event. Log-linear analyses for the verb region corroborated the observed gaze patterns by revealing a main effect of target character (agent, patient) ($LR_{\chi^2}(subj) = 14.89, df = 1, p < 0.001$; $LR_{\chi^2}(item) = 14.89, df = 1, p < 0.001$), and a significant interaction of sentence type (SVO, OVS) with target character (agent, patient) ($LR_{\chi^2}(subj) = 31.30, df = 1, p < 0.0001$; $LR_{\chi^2}(item) = 21.47, df = 1, p < 0.0001$). The three-way interaction between target character (patient, agent), sentence type (SVO, OVS), and ambiguity (ambiguous, unambiguous) was only significant by items ($LR_{\chi^2}(item) = 8.74, df = 1, p < 0.01$) ($p > 0.4$ by participants).

1.2.3. Adverb region

Figure 4 shows the eye-movement patterns during the adverb region. In the ambiguous conditions gaze patterns now resemble those of the unambiguous conditions: People looked more often to the patient for SVO than for OVS sentences, and to the agent for OVS compared with SVO sentences, suggesting that at this point in time (post-verbally), the depicted events had enabled thematic role assignment for initially ambiguous sentences.
too (Fig. 4a). Log-linear analyses for the ambiguous conditions corroborated this view by revealing a significant interaction between sentence type (SVO, OVS) and target character (agent, patient) \( (LR/SL\chi^2/subj/) = 26.85, df = 1, p < 0.0001; LR/SL\chi^2/item/) = 20.18, df = 1, p < 0.0001)\). Contrasts confirmed more looks to the patient than OVS conditions \( (LR/SL\chi^2/subj/) = 9.59, df = 1, p < 0.01; LR/SL\chi^2/item/) = 6.11, df = 1, p = 0.02). People inspected the agent more often for OVS than for SVO sentences \( (LR/SL\chi^2/subj/) = 29.86, df = 1, p < 0.0001; LR/SL\chi^2/item/) = 23.35, df = 1, p < 0.0001). Differences in gaze proportions to the distractors for OVS compared with SVO conditions were not significant \( (p's > 0.5)\).

For the unambiguous conditions, gaze patterns continued to reflect disambiguation and thematic role assignment during the adverb region (Fig. 4b). Log-linear analyses of the adverb region for the unambiguous conditions (Fig. 4b) showed a significant interaction of sentence type (SVO, OVS) and target character (patient, agent) \( (LR/SL\chi^2/subj/) = 30.67, df = 1, p < 0.0001; LR/SL\chi^2/item/) = 44.92, df = 1, p < 0.0001)\). People inspected the patient more often for SVO compared with OVS sentences \( (LR/SL\chi^2/subj/) = 32.48, df = 1, p < 0.0001; LR/SL\chi^2/item/) = 33.76, df = 1, p < 0.0001). Contrasts for more looks to the agent in OVS compared with SVO sentences were not significant \( (p's > 0.06)\). Differences between OVS and SVO sentence types in inspections to the background were also not significant.

1.2.4. NP2 region

Figure 5 presents inspections to characters for the NP2 region. During the second noun phrase (Fig. 5a), the disambiguation gaze patterns continued for the ambiguous and unambiguous conditions. For the ambiguous conditions, analyses confirmed that the interaction between sentence type (SVO, OVS) and target character (patient, agent) was significant \( (LR/SL\chi^2/subj/) = 36.45, df = 1, p < 0.0001; LR/SL\chi^2/item/) = 35.23, df = 1, p < 0.0001)\).
Ch. 24: Processing sentences in depicted events

Background

(a) (b)

Ambiguous Patient Agent Distractor

Entities

0.0

0.1

0.2

0.3

0.4

0.5

0.6

Inspection proportions

0.0

0.1

0.2

0.3

0.4

0.5

0.6

Inspection proportions

Figure 5. Proportions of inspections to characters on NP2 for the ambiguous (a) and unambiguous (b) conditions.

Contrasts confirmed a significantly higher proportion of inspections to the patient for SVO than for OVS sentences \( LR\chi^2(subj) = 37.86, \text{df} = 1, p < 0.0001 \). Contrasts for looks to the agent were not significant \( (p's > 0.1) \). For the unambiguous conditions, analyses for the second noun phrase region (Fig. 5b) revealed a significant interaction of sentence type (SVO, OVS) and target character (patient, agent) \( LR\chi^2(subj) = 28.65, \text{df} = 1, p < 0.0001; LR\chi^2(item) = 28.82, \text{df} = 1, p < 0.0001 \). People inspected the agent more often for OVS than for SVO conditions \( (LR\chi^2(subj) = 17.92, \text{df} = 1, p < 0.0001; LR\chi^2(item) = 18.58, \text{df} = 1, p < 0.0001) \), and the patient more often for SVO compared with OVS sentences \( (LR\chi^2(subj) = 22.77, \text{df} = 1, p < 0.0001; LR\chi^2(item) = 24.51, \text{df} = 1, p < 0.0001) \). Differences of OVS and SVO conditions in inspections to the background/ambiguous character were not significant \( (ps > 0.2) \).

2. General discussion

The key finding was that the point in time when the utterance identified a relevant depicted event was closely temporally coordinated with the point in time when that depicted event triggered thematic role assignment and structuring of the utterance. As predicted by the Coordinated Interply Account (CIA), an early mediation of the relevant depicted event through unambiguous case-marking on the determiner of the first noun phrase triggered an earlier thematic role assignment than a later, verb-based mediation of the relevant depicted event in initially ambiguous sentences.

For the unambiguous conditions, case-marking on the sentence-initial noun phrase allowed listeners to identify the noun-phrase referent as either the agent of a depicted kicking action, or as the patient of a hitting action performed by another agent (Fig. 1a).
During the ensuing verb, people’s anticipatory eye movements to the patient and agent of the relevant depicted event revealed thematic role assignment for SVO and OVS sentences respectively. As predicted by the CIA, these gaze patterns were observed shortly after the first noun phrase, and thus tightly temporally coordinated with the case-marked mediation of a relevant agent–action–patient or a patient–action–agent event.

For initially structurally ambiguous sentences, the relevant depicted event was only identified after the sentence-initial ambiguous noun phrase, once the verb had mediated the relevant depicted action and its associated role relations. During the ensuing post-verbal region, more eye movements to the patient for SVO than OVS and more inspections to the agent for OVS than SVO sentences revealed thematic role assignment and structural disambiguation. As predicted by the CIA, disambiguation in this late-identification case occurred later than in the ambiguous conditions, and again tightly temporally coordinated, shortly after the verb.

Despite the close temporal coordination, eye movements revealed thematic role assignment with a certain delay (one region after the utterance identified relevant role relations in the scene). The recording of event-related potentials, however, has revealed the effects of depicted events on structural disambiguation of initially structurally ambiguous utterances during the verb itself (one region earlier than for the ambiguous conditions of this study) (Knoeferle, Habets, Crocker, & Münte, in preparation). These findings suggest an even tighter coordination of comprehension and the influence of depicted events at the neural than attentional level.

The coordinated interplay account builds on and is compatible with prior research. Gaze patterns in studies by Tanenhaus et al. (1995) showed that utterance interpretation rapidly directs attention in the scene (see also Cooper, 1974), and that the type of visual referential context (one possible referent, two possible referents) influences the incremental structuring of the utterance. The rapidity of scene effects was confirmed by the fact that eye movements differed between the two visual context conditions from the onset of the utterance (see also Spivey, Tanenhaus, Eberhard, & Sedivy, 2002).

What the fixation patterns in the studies by Tanenhaus et al. (1995) do not, however, permit us to determine is whether scene information influenced structuring and interpretation of the utterance in a manner closely time-locked to, or independently of, when the utterance identified that scene information as relevant. On one interpretation, comprehension of the utterance (i.e., the apple) directed attention in the scene, and this triggered the construction and use of the appropriate referential context (one referent, two referents). A second interpretation is that people acquired the referential context temporally independently of – perhaps even prior to – hearing the utterance, and then accessed that context much as they would access a prior discourse context (e.g., Altmann & Steedman, 1988). The fact that eye-movement patterns differed from the start of the utterance between the two contexts in Tanenhaus et al. (1995) renders it impossible to decide between these two interpretations.

While not directly aimed at investigating the utterance-mediated influence of the scene, findings by Sedivy et al. (1999) provide a clearer picture than gaze patterns in Tanenhaus et al. of when, in relation to its identification as relevant by the utterance, scene information
influences comprehension. Studies by Sedivy et al. demonstrated that the time course of establishing reference to objects in a scene depended on whether there was referential contrast between two same-category objects (two glasses) or not. In the referential contrast condition, the scene contained two tall objects (a glass and a pitcher), a small glass, and a key. In the no-referential-contrast condition, the scene contained two tall objects (a glass and a pitcher), a key, and a file folder, however, no contrasting object of the category ‘glass’. Gaze patterns between the two context types did not differ while participants heard *Pick up the*. Only after people had heard *Pick up the tall*, did they look more quickly at the target referent (the tall glass) than at the other tall object (a pitcher) when the visual context displayed a contrasting object of the same category as the target (a small glass) than when it did not. These results suggest that scene information influences comprehension tightly temporally coordinated with its identification as relevant by the utterance.

The CIA is further compatible with findings by Kamide, Scheepers, and Altmann (2003). Participants in their studies inspected images showing a hare, a cabbage, a fox and a distractor object while hearing *Der Hase frisst gleich den Kohl* (‘The hare (NOM/subj) eats soon the cabbage (ACC/object)’) or *Den Hasen frisst gleich der Fuchs* (‘The hare (ACC/object) eats soon the fox (NOM/subject)’). Anticipatory eye movements to the cabbage and fox during the post-verbal region indicated that the nominative (subject) and accusative (object) case-marking on the article of the first noun phrase together with world knowledge about ‘what is likely to eat what’ extracted at the verb allowed anticipation of the correct post-verbal referent.

Unlike the ambiguous conditions in our study, however, Kamide et al. (2003) observed no effects of unambiguous case-marking (subject, object) on the anticipation of target characters (hare, fox) for the verb region itself. We suggest this difference results from the kinds of information that scenes provided in these two studies (see Henderson & Ferreira, 2004, for discussion on different scene types). In the study by Kamide et al. (2003), scenes showed no explicitly depicted events. In such scenes, an unambiguous subject or object case-marked sentence-initial noun phrase leaves who-is-doing-what-to-whom underspecified: The hare could either be the patient of a future fox-eating action in a passive sentence or the agent of an eating-cabbage action; alternatively, no action between hare and fox might occur.

In contrast, for the present study, when people heard the unambiguously object case-marked noun phrase *Den Herrn Orange* (obj) (‘The Mr Orange (obj)’), scenes proffered an explicitly depicted action (hitting) of which Mr Orange was the patient, and of which another character was the agent (Mr Apple, Fig. 1a). There is probably some underspecification in such depicted event scenes too. A depicted hitting-event could be identified as relevant by various verbs, among them *hit, bash, or beat*. What is no longer underspecified in explicitly depicted agent–action–patient events, however, is who-is-doing-what-to-whom. Object case-marking combined with a depicted hitting-action enabled people to identify the orange as the patient shortly after the first noun phrase, as revealed by anticipatory eye movements to the agent of the hitting event. A subject case-marked sentence-initial noun phrase together with an explicit agent–action–patient event enabled thematic role assignment of an agent role to the orange during the verb.
While the above discussion stresses the role of the utterance in identifying those objects/events that are most informative for comprehension, we do not exclude a more independent use of information from a scene. In increasingly complex settings, however, it will be difficult to inspect all objects in a scene. As a result, the strong guiding role that the CIA accords to the utterance may become even more important for directing attention to informative objects and events. Moreover, once the utterance has identified scene information as relevant, other objects and events that are related to that scene information can also inform comprehension.

Findings from the present experiment clearly confirmed the Coordinated Interplay Account prediction that there is a tight temporal coordination between when the utterance identifies a relevant depicted event, and when that scene event influences thematic role assignment and structuring of the utterance (see also Mayberry, Crocker, & Knoeferle, 2005, for relevant modelling research). Discussion of prior research showed that the CIA is further compatible with, and even predicts, the temporal coordination between utterance comprehension, attention in the scene, and the influence of scene information on comprehension for a range of existing studies.

Taken together the present experimental findings and theoretical discussion represent the first step towards an online processing account of situated sentence comprehension.

Acknowledgements

This research was funded by a PhD scholarship to Pia Knoeferle awarded by the German research foundation (DFG). We would like to thank Matthew W. Crocker for his continued support. Further thanks go to Nicole Kühn for help with data acquisition and analysis.

References


Mahwah, NJ: Lawrence Erlbaum.


Chapter 25

VISUAL SALIENCY DOES NOT ACCOUNT FOR EYE MOVEMENTS DURING VISUAL SEARCH IN REAL-WORLD SCENES

JOHN M. HENDERSON and JAMES R. BROCKMOLE

University of Edinburgh

MONICA S. CASTELHANO

University of Massachusetts

MICHAEL MACK

Vanderbilt University
Abstract

We tested the hypothesis that fixation locations during scene viewing are primarily determined by visual salience. Eye movements were collected from participants who viewed photographs of real-world scenes during an active search task. Visual salience as determined by a popular computational model did not predict region-to-region saccades or saccade sequences any better than did a random model. Consistent with other reports in the literature, intensity, contrast, and edge density differed at fixated scene regions compared to regions that were not fixated, but these fixated regions also differ in rated semantic informativeness. Therefore, any observed correlations between fixation locations and image statistics cannot be unambiguously attributed to these image statistics. We conclude that visual saliency does not account for eye movements during active search. The existing evidence is consistent with the hypothesis that cognitive factors play the dominant role in active gaze control.
During real-world scene perception, we move our eyes about three times each second via very rapid eye movements (saccades) to reorient the high-resolving power of the fovea. Pattern information is acquired only during periods of relative gaze stability (fixations) due to a combination of central suppression and visual masking (Matin, 1974; Thiele, Henning, Buishik, & Hoffman, 2002; Volkman, 1986). Gaze control is the process of directing the eyes through a scene in real time in the service of ongoing perceptual, cognitive, and behavioral activity (Henderson, 2003; Henderson & Hollingworth, 1998, 1999).

There are at least three reasons that the study of gaze control is important in real-world scene perception (Henderson, 2003; Henderson & Ferreira, 2004a). First, human vision is active, in the sense that fixation is directed toward task-relevant information as it is needed for ongoing visual and cognitive computations. Although this point seems obvious to eye movement researchers, it is often overlooked in the visual perception and visual cognition literatures. For example, much of the research on real-world scene perception has used tachistoscopic display methods in which eye movements are not possible (though see Underwood, this part; Gareze & Findlay, this part). While understanding what is initially apprehended from a scene is an important theoretical topic, it is not the whole story; vision naturally unfolds over time and multiple fixations. Any complete theory of visual cognition, therefore, requires understanding how ongoing visual and cognitive processes control the direction of the eyes in real time, and how vision and cognition are affected by where the eyes are pointed at any given moment in time.

Second, eye movements provide a window into the operation of selective attention. Indeed, although internal (covert) attention and overt eye movements can be dissociated (Posner & Cohen, 1984), the strong natural relationship between covert and overt attention has recently led some investigators to suggest that studying covert visual attention independently of overt attention is misguided (Findlay, 2004; Findlay & Gilchrist, 2003). For example, as Findlay and Gilchrist (2003) have noted, much of the research in the visual search literature has proceeded as though viewers steadfastly maintain fixation during search, allocating attention only via an internal mechanism. However, visual search is virtually always accompanied by saccadic eye movements (e.g., see the chapters by Hooge, Vlaskamp, & Over, this part; Shen & Reingold, this part). In fact, studies of visual search that employ eye tracking often result in different conclusions than do studies that assume the eyes remain still. As a case in point, eye movement records reveal a much richer role for memory in the selection of information for viewing (e.g. McCarley, Wang, Kramer, Irwin, & Peterson, 2003; Peterson, Kramer, Wang, Irwin, & McCarley, 2001) than research that uses more traditional measures such as reaction time (e.g. Horowitz & Wolfe, 1998). To obtain a complete understanding of the role of memory and attention in visual cognition, it is necessary to understand eye movements.

Third, because gaze is typically directed at the current focus of analysis (see Irwin, 2004, for some caveats), eye movements provide an unobtrusive, sensitive, real-time behavioral index of ongoing visual and cognitive processing. This fact has led to enormous insights into perceptual and linguistic processing in reading (Liversedge & Findlay, 2000; Rayner, 1998; Sereno & Rayner, 2003), but eye movements are only now becoming
a similarly important tool in the study of visual cognition generally and scene perception in particular.

1. Fixation placement during scene viewing

A fundamental goal in the study of gaze control during scene viewing is to understand the factors that determine where fixation will be placed. Two general hypotheses have been advanced to explain fixation locations in scenes. According to what we will call the visual saliency hypothesis, fixation sites are selected based on image properties generated in a bottom-up manner from the current scene. On this hypothesis, gaze control is, to a large degree, a reaction to the visual properties of the stimulus confronting the viewer. In contrast, according to what we will call the cognitive control hypothesis, fixation sites are selected based on the needs of the cognitive system in relation to the current task. On this hypothesis, eye movements are primarily controlled by task goals interacting with a semantic interpretation of the scene and memory for similar viewing episodes (Hayhoe & Ballard, 2005; Henderson & Ferreira, 2004a). On the cognitive control hypothesis, the visual stimulus is, of course, still relevant: The eyes are typically directed to objects and features rather than to uniform scene areas (Henderson & Hollingworth, 1999); however, the relevance of a particular object or feature in the stimulus is determined by cognitive information-gathering needs rather than inherent visual salience.

The visual saliency hypothesis has generated a good deal of interest over the past several years, and in many ways has become the dominant view in the computational vision literature. This hypothesis has received primary support from two lines of investigation. First, computational models have been developed that use known properties of the visual system to generate a saliency map or landscape of visual salience across an image (Itti & Koch, 2000, 2001; Koch & Ullman, 1985). In these models, the visual properties present in an image give rise to a 2D map that explicitly marks regions that are different from their surround on image dimensions such as color, intensity, contrast, and edge orientation (Itti & Koch, 2000; Koch & Ullman, 1985; Parkhurst, Law, & Niebur, 2002; Torralba, 2003), contour junctions, termination of edges, stereo disparity, and shading (Koch & Ullman, 1985), and dynamic factors such as motion (Koch & Ullman, 1985; Rosenholtz, 1999). The maps are generated for each image dimension over multiple spatial scales and are then combined to create a single saliency map. Regions that are uniform along some image dimension are considered uninformative, whereas those that differ from neighboring regions across spatial scales are taken to be potentially informative and worthy of fixation. The visual saliency map approach serves an important heuristic function in the study of gaze control because it provides an explicit model that generates precise quantitative predictions about fixation locations and their sequences, and these predictions have been found to correlate with observed human fixations under some conditions (e.g., Parkhurst et al., 2002).

Second, using a scene statistics approach, local scene patches surrounding fixation points have been analyzed to determine whether fixated regions differ in some image
properties from regions that are not fixated. For example, high spatial frequency content and edge density have been found to be somewhat greater at fixated than non-fixated locations (Mannan, Ruddock, & Wooding, 1996, 1997b). Furthermore, local contrast (the standard deviation of intensity in a patch) is higher and two-point intensity correlation (intensity of the fixated point and nearby points) is lower for fixated scene patches than control patches (Krieger, Rentschler, Hauske, Schill, & Zetzsche, 2000; Parkhurst & Neibur, 2003; Reinagel & Zador, 1999).

Modulating the evidence supporting the visual saliency hypothesis, recent evidence suggests that fixation sites are tied less strongly to saliency when meaningful scenes are viewed during active viewing tasks (Land & Hayhoe, 2001; Turano, Geruschat, & Baker 2003). According to one hypothesis, the modulation of visual salience by knowledge-driven control may increase over time within a scene-viewing episode as more knowledge is acquired about the identities and meanings of previously fixated objects and their relationships to each other and to the scene (Henderson, Weeks, & Hollingworth, 1999). However, even the very first saccade in a scene can often take the eyes in the likely direction of a search target, whether or not the target is present, presumably because the global scene gist and spatial layout acquired from the first fixation provide important information about where a particular object is likely to be found (Antes, 1974; Brockmole & Henderson, 2006b; Castelhano & Henderson, 2003; Henderson et al., 1999; Mackworth & Morandi, 1967).

Henderson and Ferreira (2004a) sorted the knowledge available to the human gaze control system into several general categories. Information about a specific scene can be learned over the short term from the current perceptual encounter (short-term episodic scene knowledge) and over the longer term across multiple encounters (long-term episodic scene knowledge). Short-term knowledge underlies a viewer’s tendency to refixate areas of the current scene that are semantically interesting or informative (Buswell, 1935; Henderson et al., 1999; Loftus & Mackworth, 1978; Yarbus, 1967), enables the prioritization of newly appearing or disappearing objects from a scene (Brockmole & Henderson, 2005a, 2005b), and ensures that objects are fixated when needed during motor interaction with the environment (Land & Hayhoe, 2001). Long-term episodic knowledge involves information about a particular scene acquired and retained over time. Recent evidence suggests that good memory for the visual detail of fixated regions of a viewed scene is preserved over relatively long periods of time (Castelhano & Henderson, 2005; Henderson & Hollingworth, 2003; Hollingworth, 2004; Hollingworth & Henderson, 2002; Williams, Henderson, & Zacks, 2005; for review see Henderson & Castelhano, 2005). The contextual cueing phenomenon shows that perceptual learning of complex visual images can take place relatively rapidly over multiple encounters (Chun & Jiang, 1998), and this effect has been shown to influence eye movements (Peterson & Kramer, 2001). We have recently found that this same type of learning can take place even more rapidly for real-world scenes (Brockmole & Henderson, 2006a). Furthermore, we have shown that these learned representations can facilitate eye movements during search in real-world scenes (Brockmole & Henderson, 2006b). Another interesting example of the influence of episodic scene knowledge on gaze control is the finding that viewers will often
fixate an empty scene region when that region previously contained a task-relevant object (Altmann, 2004; Richardson & Spivey, 2000).

A second source of information that can guide gaze is scene schema knowledge, the generic semantic and spatial knowledge about a particular category of scene (Biederman, Mezzanotte, & Rabinowitz, 1982; Friedman, 1979; Mandler & Johnson, 1977). Schema knowledge includes information about the objects likely to be found in a specific type of scene (e.g., bedrooms contain beds) and spatial regularities associated with a scene category (e.g., pillows are typically found on beds), as well as generic world knowledge about scenes (e.g., beds do not float in the air). Scene identity can be apprehended and a scene schema retrieved very rapidly (Potter, 1976; Schyns & Oliva, 1994), and schema knowledge can then be used to limit initial fixations to scene areas likely to contain an object relevant to the current task (Henderson et al., 1999).

A third source of information important in gaze control is task-related knowledge (Buswell, 1935; Yarbus, 1967). Task-related knowledge can involve a general gaze control policy or strategy relevant to a given task, such as periodically fixating the reflection in the rear-view mirror while driving, and moment-to-moment control decisions based on ongoing perceptual and cognitive needs. Gaze control differs during complex and well-learned activities such as reading (Rayner, 1998), tea and sandwich making (Land & Hayhoe, 2001), and driving (Land & Lee, 1994). The distribution of fixations over a given scene changes depending on whether a viewer is searching for an object or trying to memorize that scene (Henderson et al., 1999). Gaze is also strongly influenced by moment-to-moment cognitive processes related to spoken language comprehension and production (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; see Henderson & Ferreira, 2004b).

2. Present study

As reviewed above, there is abundant evidence that fixation placement during scene viewing is strongly affected by cognitive factors. Most proponents of the visual saliency hypothesis acknowledge that cognitive factors play a role in gaze control, but they tend to focus on the adequacy of a saliency-based approach to account for much of the data on fixation placement (e.g., Parkhurst et al., 2002).

In the present study, we investigated further the degree to which fixation location is related to image properties during scene viewing. First, we collected eye movement data from participants who viewed full-color photographs of real-world outdoor scenes while engaged in a visual search task in which they counted the number of people who appeared in each scene. We then analyzed the fixation data in three ways to investigate the adequacy of the visual saliency hypothesis. First, we compared the fixation data against the predictions generated from an established visual saliency model. Second, we conducted an image statistics analysis to determine whether image properties differed at fixated and non-fixated locations. Third, we tested whether any observed correlations
between fixation locations and image statistics might be due to the meaning of the fixated locations. Our conclusion is that the evidence supporting the visual saliency hypothesis is weak, and that the existing evidence is consistent with the hypothesis that cognitive factors play the dominant role in gaze control.

2.1. Method

The eye movements of 8 Michigan State University undergraduates were monitored as they viewed 36 full-color photographs of real-world outdoor scenes displayed on a computer monitor (see, e.g., Figure 1). The photographs were shown at a resolution of 800 × 600 pixels and subtended 16 deg. horizontally by 12 deg. vertically at a viewing distance of 113 cm. Eye position was sampled at a rate of 1000 Hz from a Fourward Technologies Generation 5.5 Dual Purkinje Image Eyetracker, and raw eye-tracking data were parsed into fixations and saccades using velocity and distance criteria.
The subject’s head was held steady with an anchored bite-bar made of dental impression compound. Prior to the first trial, subjects completed a procedure to calibrate the output of the eyetracker against spatial position on the display screen. This procedure was repeated regularly throughout the experiment. Observers were instructed to count the number of people in each photograph. Each participant saw all 36 scenes in a different random order. Each photograph contained between 0 and 6 people and was presented until the participant responded or for 10 s. maximum. Across all search photographs, accuracy on the counting task was 82%, with greater accuracy for scenes with fewer targets present. Accuracy was below 100% because some targets were well hidden and difficult to find in the scenes.

3. Analysis 1: Comparing saliency model predictions to human fixations

A benchmark for the visual saliency hypothesis is the saliency map model of Itti and Koch (2000, 2001). This model produces explicit predictions about where viewers should fixate in complex images. The Itti and Koch model has been shown to predict human fixations reasonably well under some conditions (e.g., Parkhurst et al., 2002), though Turano et al. (2003) demonstrated that the correlations between the model and human fixations were eliminated when the viewing task was active. However, one could argue that this latter result was a consequence of the dynamic interaction between a moving viewer and the real world, a situation for which the model was not specifically developed. In Analysis 1, we examined the degree to which the Itti and Koch saliency map model is able to predict fixation locations in static scenes (the situation for which it was developed) during an active visual search task.

3.1. Do Human Fixations Fall on the Most Visually Salient Regions?

In a first analysis we compared the number of saccadic entries into, and the number of discrete fixations in, the scene regions that the saliency map model specified as most salient. For this analysis, the Itti and Koch bottom-up saliency model posted on the website [http://ilab.usc.edu/toolkit/downloads.shtml] on 16 May 2005, was used to determine the visually salient regions in each of our test scenes. The model “viewed” the scenes for 10 s each (the same amount of time given to the participants) with a foveal radius of 1°. While viewing each scene, the model generated a cumulative saliency map showing the scene regions that it found most salient over the 10 s of viewing. Any region in these cumulative saliency maps that held a value greater than zero was defined as a salient region in the eye movement analysis. To better understand the relationship between the regions the Itti and Koch model found salient and the regions participants fixated while viewing the scenes, two measures of the participants’ eye movements were examined, Salient Region Entries and Salient Region Fixations. “Salient Region Entries” was defined as the proportion of all participant-generated saccades that started outside of a given salient region and landed in that region. This measure captures the degree to which the eyes tended to
Ch. 25: Visual Saliency Does Not Account for Eye Movements During Visual Search

![Participant fixation distribution](image)

Figure 2. Spatial distribution of all participant-generated fixations across all scenes. The figure shows an overall bias for fixations to be placed along a lower central horizontal band.

saccade to salient regions. “Salient Region Fixations” was defined as the proportion of all participant-generated fixations that fell in a given salient region. This measure reflects all fixations in a salient region regardless of whether the fixation was due to a saccade from beyond that region or within that region. As control contrasts, random fixations of equal number to the participants’ were generated by two random models. The first model (pure random) simply sampled from all possible fixation locations. To control the participants’ bias to fixate in the lower central regions of the scenes (see Figure 2), a second control contrast (biased random) used randomly generated fixations based on the probability distribution of fixation locations from the participants’ eye movement behavior across all the scenes.

Salient Region Entries. The first analysis was carried out by taking the proportion of saccades that entered a salient region (see Figure 1 for an example). All saccades from the participant trials and an equal number of saccades from the pure- and biased-random models were included in the analysis. A higher proportion of salient region entries means that a greater number of saccades were made into salient regions. If the saliency map model is able to identify regions that capture attention better than chance, the proportion of salient region entries made by participants should be higher than the proportion of salient region entries made by the random models. If the proportion of salient region entries does not differ between participants and the random models, participants are no
J. M. Henderson et al.

Figure 3. (a) Mean proportion (with standard error) of all participant-generated saccades that moved from outside to inside salient regions, compared with those generated by a pure random model and a biased random model. (b) Mean proportion (with standard error) of all participant-generated fixations that fell within salient regions, compared with those generated by a pure random model and a biased random model.

More likely to saccade to regions identified by the model than they are by chance. As can be seen in Figure 3a, although the saliency map model predicted entries better than did a pure random model, $t(34) = 3.64, p < 0.001$, it did not predict entries better than a biased-random model that took into account participants’ general tendency to fixate the lower and central regions of all scenes, $t(34) < 1$. Contrary to the visual saliency hypothesis, participants’ tendency to move their eyes to specific scene regions was not accounted for by the saliency model.
**Salient Region Fixations.** The second analysis was carried out by taking the proportion of all fixations that landed in a salient region (see Figure 1 for an example). All fixations from the participant trials and an equal number of fixations from the two random models were included in the analysis. Once again, if participants were to have a higher proportion of salient region fixations than the random model, this would suggest that the model is finding regions that capture attention better than chance. On the other hand, if the proportion of salient region fixations does not differ between participants and the random models, this would suggest that the model is finding regions that are no more likely to be fixated than by chance. Observers fixated salient regions identified by the model more often predicted by the pure-random model, \( t(34) = 6.81, p < 0.001 \), and the biased-random model, \( t(34) = 4.02, p < 0.001 \), indicating that the saliency model predicted the number of fixations in salient regions more accurately than models based on chance (see Figure 3b).

**Summary.** The Salient Region Entries analysis demonstrates that viewers were no more likely to saccade to a salient scene region (as identified by the saliency map model) than they were by chance. On the other hand, the Salient Region Fixations analysis shows that viewers fixated salient regions more often than would be expected by chance. Together, these data suggest that although the eyes are not specifically attracted to salient regions, they do tend to stick to them once there. The latter result might be taken as at least partial support for the saliency control hypothesis. However, because this hypothesis is supposed to account for the movement of the eyes through a scene rather than the tendency to dwell in a given region, the support is weak. Furthermore, as detailed below, the latter result is also consistent with the possibility that saliency is correlated with “object-ness”, and that viewers tend to gaze at objects.

### 3.2. Do Human Fixations Correspond with Model-Generated Fixation Predictions?

In addition to generating a map of salient scene regions, the saliency model also produces a set of fixations. Therefore, a second way to test the ability of the model to predict human fixations is to compare directly the human- and model-generated fixation locations. We quantified the distance between these fixation locations in two ways, one based on a similarity metric devised by Mannan, Ruddock, & Wooding (1995) and a second using the one that we developed as an extension of this metric.

**Mannan, Ruddock, & Wooding (1995) Similarity Metric.** The fixation location similarity metric introduced by Mannan et al. (1995) compares the spatial proximity of fixations derived from two unique fixation sets (e.g. model generated and observer generated). The location similarity metric compares the linear distance from one set of fixation locations to the closest fixation in the other set, and vice versa. A high score indicates high similarity. As a control, we also computed the same similarity metric for all pairwise comparisons among participants. If the saliency map model is able to predict the locations of human fixations, its similarity to human observers should be comparable to the similarity of one human viewer to another.
The index of similarity \( (I_s) \) introduced by Mannan et al. is based on the squared distances between corresponding fixations in two gaze patterns \( (D_m \text{ and } D_{mr}) \) and is defined in the following manner:

\[
I_s = 100 \left[ 1 - \frac{D_m}{D_{mr}} \right],
\]

with

\[
D_m^2 = \frac{n_1 \sum_{j=1}^{n_2} d_{2j}^2 + n_2 \sum_{i=1}^{n_1} d_{1i}^2}{2n_1n_2(w^2 + h^2)},
\]

where \( n_1 \) and \( n_2 \) are the number of fixations in the two gaze patterns, \( d_{1i} \) is the distance between the \( i \)th fixation in the first gaze pattern and its nearest neighbor fixation in the second gaze pattern, \( d_{2j} \) is the distance between the \( j \)th fixation in the second gaze pattern and its nearest neighbor fixation in the first gaze pattern, and \( w \) and \( h \) are the width and height of the image of the scene. The calculation of \( D_{mr} \) is the same as \( D_m \) but with randomly generated gaze patterns of the same size being compared. Similar to a correlation, identical gaze patterns produce an \( I_s \) score of 100, random gaze patterns produce an \( I_s \) score of 0, and systematically different gaze patterns generate a negative score (Mannan et al., 1995). For our analysis, we examined the first seven fixations each participant produced when viewing each scene and compared them against the first seven fixations produced by the saliency model.

Figure 4a shows the mean similarity score \( I_s \) for each participant against all other participants (left bar) and all participants against the model (right bar). As can be seen in the figure, the participants’ fixations were significantly less similar to those generated by the saliency model than they were to each other, \( t(35) = 7.87, p < 0.001 \).

A Unique Assignment Variant of the Mannan et al. (1995) Metric. A potential concern with the Mannan et al. (1995) similarity metric is that it does not take into account the overall spatial variability in the distribution of fixations over an image. For example, if all of the fixations in Set 1 are clustered in one small region of a scene, and there is at least one fixation in that same region in comparison Set 2, all the Set 1 fixations will be compared against that single Set 2 fixation. Another way to compute similarity in the same spirit as the Mannan et al. method that corrects for this issue is to require that each fixation in each set be assigned to a unique fixation in the other set. A metric can then be computed based on the distance of each point in Set 1 to its assigned point in Set 2. Intuitively, this unique-assignment metric better takes into account the overall spatial distributions of fixations. (Unlike the Mannan et al. analysis, this method requires that there be an identical number of fixations in each set.) In our unique-assignment analysis, all possible assignments of each fixation in Set 1 to a unique fixation in Set 2 were examined to find the single assignment that produced the smallest average deviation. This assignment was then used to compute the similarity metric, which is the squared deviation of each fixation point in Set 1 to its mate in Set 2.
Figure 4. Similarity of participant fixation locations to model-generated locations (left bars) and to each other (right bars) for the Mannan et al. Index of Similarity (a), our Unique-Assignment “Warping” similarity index (b), and the Levenshtein Distance metric for sequence similarity (c).
More precisely, unique-assignment distance \(W_s\) between two gaze patterns \(D_m\) and \(D_{mr}\) was defined as:

\[
W_s = 100 \left[ 1 - \frac{D_w}{D_{wr}} \right],
\]

with

\[
D_w = \frac{1}{n} \sum_{j=1}^{n} p_j^2,
\]

where \(n\) is the number of fixations in the gaze patterns, \(p_j\) is the distance between the \(j\)th unique pair of one fixation from the first set and one fixation from the second set. The calculation of \(D_{wr}\) is the same as \(D_w\) except that randomly generated gaze patterns of the same size are compared. Identical gaze patterns produce a \(W_s\) score of 100, random gaze patterns produce a \(W_s\) score of 0, and systematically different patterns generate negative scores.

Again, as a contrast, we also computed the unique-assignment similarity metric for all participants against all other participants. If the saliency model is able to predict human fixations, its similarity to human observers should be comparable to the similarity for all pairwise comparisons of participants. As above, we restricted the analysis to the first seven fixations each participant produced when viewing each scene and the first seven fixations produced by the saliency model.

Figure 4b shows the mean similarity score \(W_s\) for each participant against all other participants (left bar) and for all participants against the model (right bar). As can be seen in the figure, as with the first similarity metric, the fixations generated by the saliency model were significantly less similar to those of the participants than were those of the participants to each other, \(t(35) = 5.27, p < 0.001\).

### 3.3. Are Human Fixation Sequences Predicted by Model Fixation Sequences?

Both the original Mannan et al. (1995) similarity metric and our unique-assignment variant of it ignore information about fixation sequence. In the case of the Mannan et al. (1995) metric, there is no requirement that fixations be assigned in a one-to-one correspondence across sets, and in the unique-assignment variant, the correspondence is based purely on spatial proximity and so does not take into account the temporal order in which the fixations were produced. It could be that the saliency model does a better job of predicting fixation order (scan pattern) than it does the exact locations of fixations. To investigate this possibility, we computed the Levenshtein Distance, a similarity metric specifically designed to capture sequence. The analysis uses a set of basic transformations
to determine the minimum number of steps (character insertion, deletion, and substitution) that would be required to transform one character string into another. This general method is used in a variety of situations including protein sequencing in genetics (Levenshtein, 1966; Sankhoff & Kruskal, 1983). To conduct the analysis, we divided each scene into a grid of 48 regions of about 2° by 2° each. This division allowed some noise in fixation location so that minor deviations from the model would not disadvantage it. Each of the 48 regions was assigned a unique symbol. Each fixation was coded with the symbol assigned to the region in which it fell. We again analyzed the first seven fixations, so each fixation sequence produced a 7-character string. The similarity metric between two strings was the number of steps required to transform one string into another. Identical strings generated a value of 0, and the maximum value was 7. As in the first two analyses, we computed the similarity of each subject’s fixation sequence for each scene to the sequence generated for that scene by the model. Again, as a control, we also computed the string metric for all participants against all other participants for each scene. If the saliency model is able to predict human fixations, its similarity to human observers should be comparable to the similarity of the human participants to each other. Figure 4c shows the mean distance score for each participant against all other participants (left bar) and for all participants against the model (right bar). The fixation sequences generated by the saliency model were significantly less similar to those of the participants than were those of the participants to each other, $t(35) = 10.2, p < 0.001$.

### 3.4. Saliency Map Model Comparison Summary

In a first set of analyses, we tested the ability of an implemented saliency map model to predict human fixation locations during an active viewing task. Overall, the results suggested that the model does not do a particularly good job. Human fixations did not land in regions of a scene that the model considered to be visually salient, and the similarity of the participants’ fixations to each other was much greater than the similarity of the participants’ fixations to model-generated fixations. Of course, the ability of a given model to predict human performance is a function both of those aspects of the model that are theory-inspired and other incidental modeling decisions required for the implementation. One could argue that the spirit of the model is correct, but not the implementation. Similarly, one might argue that the implementation is correct, but not the specific parameter choices. However, it is important to remember that this version of the model has been reported to predict human fixation locations reasonably well under other conditions (Parkhurst et al., 2002; Parkhurst & Neibur, 2003, 2004). The model seems to do a particularly good job with meaningless patterns (such as fractals) and in relatively unstructured viewing tasks. In this context, the present results can be taken to suggest that whereas visual salience (as instantiated in the Itti and Koch saliency map model) does a reasonable job of accounting for fixation locations under some circumstances, it does a poor job when the viewing task involves active search and the image is a real-world scene.
4. Analysis 2: Measuring local image statistics at fixated locations

Several studies have demonstrated that the image properties of fixated regions tend to differ in systematic ways from regions that are not fixated (Krieger et al., 2000; Mannan et al., 1995, 1996; Mannan, Ruddock, & Wooding, 1997a; Parkhurst & Niebur, 2003; Reinagel & Zador, 1999). Specifically, fixated scene regions tend to be lower in intensity but higher in edge density and local contrast, and are more likely to contain third-order spatial relationships such as T-junctions and curves, than non-fixated regions. These results have been taken to suggest that such regions act as “targets” for fixations. Do these results generalize to an active visual task with real-world scenes?

4.1. Scene Statistics Method

In the present study, we measured the local image statistics associated with the fixations generated by our viewers, and compared those values to the values associated with randomly selected scene locations (see Parkhurst et al., 2002). For each scene image, ensembles of image patches were created. These patches had a radius of 1° of visual angle, approximating the spatial extent of foveal vision. Three different types of ensembles were created. In the subject ensemble, patches were defined by the subject-selected fixation positions within each image. That is, the center of each patch was defined by the \((x, y)\) coordinates of each fixation. Thus, the subject ensemble was completely constrained by subject behavior. In the random ensemble, patches were centered on randomly selected positions within each scene. Thus, the patches in the random ensemble were completely unconstrained and every point in the image was equally likely to be selected. In the shuffled ensemble, patches were derived by “shuffling” subject-selected fixation locations from one image onto a different, randomly selected image. Like the biased random control condition in Analysis 1, this shuffled ensemble was used to account for the participants’ bias to fixate more centrally in an image (see Parkhurst & Niebur, 2003).

For each ensemble, several measures of local image statistics were calculated. Analyses then focused on evaluating the similarity of the image statistics within each type of ensemble. Image statistics within the subject ensemble are characteristic of those image properties that are fixated. Since it is a random sampling, image statistics within the random ensemble are characteristic of the image properties in the scenes overall. Image statistics within the shuffled ensemble are characteristic of the image properties in those scene regions that tend to be fixated across scenes, such as the lower scene center (see Figure 2). The degree to which the scene statistics of the subject ensembles differ from the random and shuffled ensembles indicates the extent to which fixation location is correlated with particular image statistics.

Three common measures of local image statistics were examined: intensity, contrast, and edge density. These image statistics characterize different properties of image luminance. The luminance of each scene was extracted by converting the scene’s RGB values to the CIE \(L^*a^*b^*\) colorspace (Oliva & Schyns, 2000) which separates the luminance information of an image into a distinct dimension \((L^*)\). The chromatic information in the \(a^*\) and \(b^*\) dimensions
was discarded, and analyses focused on the values in the $L^*$ dimension. Intensity was defined as the average luminance value of the pixels within a patch (see Mannan et al., 1995). Greater intensity is associated with higher luminance, or a higher degree of subjectively perceived brightness. Local contrast was defined as the standard deviation of luminance within a patch (see Parkhurst & Niebur, 2003; Reinagel & Zador, 1999). Local contrast, then, is a measure of how much the luminance values of pixels within a patch vary from each other. More uniform patches have less contrast. Edge density was defined as the proportion of edge pixels within an image patch. Edge pixels were found by filtering the scenes with a Sobel operator that responds to contours in scenes represented by steep gradients in luminance (see Mannan et al., 1995, 1996). Greater edge density is associated with image patches containing a greater number of contours.

**Results.** Representative patches from the subject, shuffled, and random ensembles are depicted in Figure 5. Quantitative analyses of the local image statistics available in patches from each ensemble are summarized in Figure 6. For all analyses, the local image
Figure 6. Mean intensity, contrast and edge density (with standard errors) for the subject, shuffled, and random ensembles.
statistics observed at fixation (the subject ensemble) were tested against the random and shuffled ensembles using one-sample $t$-tests.

Consistent with the prior findings in the literature cited earlier, patches derived from the subject ensembles were reliably different from those from the shuffled and random ensembles for all three local image statistics. Intensity within the subject ensemble patches was 6% lower than that within the shuffled patches ($t(284) = -3.88, p < 0.001$), and 8% lower than that in the random patches ($t(284) = -6.55, p < 0.001$). Local contrast within the subject ensemble patches was 14% higher than that within the shuffled patches ($t(284) = 6.72, p < 0.001$) and 9% higher than that in the random patches ($t(284) = 7.59, p < 0.001$). Edge density within the subject ensemble patches was 19% higher than that within the shuffled patches ($t(284) = 8.03, p < 0.001$), and 29% higher than that in the random patches ($t(284) = 15.5, p < 0.001$).

Summary. Replicating prior results, observers fixated regions that were lower in intensity and higher in local contrast and edge density than either control regions selected randomly or based on fixations from another image. On the face of it, these data could be taken to suggest that regions marked by differences in local image properties compared to the remainder of the scene act as “targets” for fixation, irrespective of the semantic nature of the information contained in those regions (Parkhurst et al., 2002; Parkhurst & Neibur, 2003). However, because these analyses only establish a correlation between fixation locations and image properties, it is also possible that the relationship is due to other factors. In the following section we explore the hypothesis that region meaning is such a factor. Specifically, we measured the semantic informativeness of the subject, shuffled, and random ensembles to determine whether meaning was also correlated with fixation location.

5. Analysis 3: Are fixated scene regions more semantically informative?

The purpose of this analysis was to determine whether fixated regions that have been shown to differ from non-fixated regions in intensity, contrast, and edge density, also differ in semantic informativeness. To investigate this question, an independent group of observers rated the degree to which patches from each ensemble were semantically informative (Antes, 1974; Mackworth & Morandi, 1967).

One hundred patches were selected from each of the subject, shuffled, and random ensembles generated from the scene statistics analysis reported above. These patches met two constraints. First, patches had to be representative of their ensemble supersets (subject, shuffled, random) in terms of local image statistics (as determined above) so that the reliable statistical differences observed would be preserved. Second, a minimum distance of 2° of visual angle was established between the center points of any two patches originating from the same scene so that selected patches could not spatially overlap. Within these constraints, patches were chosen randomly. Patches from all scenes and subjects were represented in the final subset used in Analysis 3.
Seven Michigan State University undergraduates viewed all 300 selected patches on a computer monitor. Stimuli for presentation were created by placing each patch in the center of a uniform gray background that subtended 16° horizontally and 12° vertically. Each individual patch subtended 2° horizontally and vertically. Presentation order of patches was randomly determined. Using a 7-point Likert-type scale, observers were instructed to rate how well they thought they could determine the overall content of the scene from the small view they were shown.

**Results.** Mean ratings for each patch type are illustrated in Figure 7. Patches from the subject, shuffled, and random ensembles received mean ratings of 4.65, 4.25, and 4.11, respectively. A one-way repeated-measures ANOVA demonstrated a reliable effect of patch type, $F(2, 12) = 19.4, p < 0.001$, with all pairwise comparisons reliable. Critically, patches from the subject ensemble were judged to be more informative of scene identity than those from the shuffled and random ensembles. This analysis demonstrates that observers in the eyetracking experiment fixated scene regions that were more likely to provide meaningful information about the scene. These results challenge the hypothesis that local scene statistics and semantic informativeness are independent.

**Summary.** Image statistics of areas selected for eye fixation within scenes differ in systematic ways from areas that are not fixated. A possible interpretation of these results is that fixation position can be accounted for by low-level image statistics (Krieger et al., 2000; Mannan et al., 1995, 1996, 1997a; Parkhurst & Niebur, 2003; Reinagel & Zador, 1999). The present results, however, call this interpretation into question. We conclude that examining the relationship between image statistics and fixation location without also measuring the semantic content of fixated regions can provide a partial or even misleading characterization of bottom-up influences on gaze control. Though it is possible that image properties in a scene directly influence gaze control, the results from scene statistics analyses cannot be taken as strong support for it.
6. General discussion

Gaze control during scene perception is critical for timely acquisition of task-relevant visual information. In this study, we tested the hypothesis that the selection of locations for fixation during complex scene viewing is primarily driven by visual salience derived from a bottom-up analysis of image properties. To test this visual saliency hypothesis, we collected eye movement data from participants who viewed full-color photographs of real-world scenes during an active visual search task. We then analyzed the eye movement data in a variety of ways to test the visual saliency hypothesis. In an initial set of analyses, we compared the fixation data against the predictions generated from what is arguably the standard among computational visual saliency models. We found that visual salience, as instantiated by the model, did a poor job of predicting either fixation locations or sequences. In a second set of analyses, we examined whether image properties differ at fixated and non-fixated locations. Consistent with other reports in the literature, we found clear differences in intensity, contrast, and edge density at fixated scene regions compared to regions that were not fixated. However, in a third analysis, we showed that fixated regions also differ in rated meaning compared to regions not fixated. Therefore, any observed correlations between fixation locations and image statistics could be due to the informativeness of fixated locations rather than to differences in the image statistics themselves. Our conclusion is that the evidence supporting the visual saliency hypothesis is weak, and that the existing evidence is consistent with the hypothesis that cognitive factors play the dominant role in gaze control.

6.1. Visual saliency or cognitive control?

To what extent is there strong evidence that gaze is primarily controlled by visual salience? As we have shown, the main sources of evidence, correlation of fixation positions with model-determined visual saliency, and differences in scene statistics at fixated and non-fixated locations, are both problematic. In the case of correlations with saliency model output, there is very little evidence that for active viewing tasks in the real world, existing saliency models do a good job of predicting fixation locations (Turano et al., 2003). In the present study, we showed that an existing saliency model also does a poor job of predicting fixation locations during active visual search in static images.

In the case of analyses showing that image statistics differ at fixated and non-fixated locations, our results suggest that previously reported effects may just as well be due to differences in region meaning as to differences in the image statistics themselves. This confound is probably unavoidable: meaningful objects differ from scene background in their image properties. The important conclusion is that showing differences in image properties at fixated and non-fixated regions cannot be used as unambiguous support for the hypothesis that those properties are driving eye movements. It is at least as likely that the meaning of an image region is responsible for the fact that it was fixated.
The few prior attempts to investigate a causal link between image statistics and fixation point selection have produced mixed results. Mannan and colleagues (1995, 1996) demonstrated that fixation locations in normal and low-pass filtered scenes are similar over the first 1.5s of viewing. Because the objects in many of the low-pass filtered scenes were not identifiable, these data suggest that human gaze control does not initially select fixation sites based on object identity information. However, Einhauser and König (2003) demonstrated that perceptually detectable modifications to contrast have no effect on fixation point selection, suggesting that contrast does not contribute causally to fixation point selection, though this study has been criticized on methodological grounds (Parkhurst & Niebur, 2004). Nevertheless, Einhauser and König (2003) concluded that top-down, rather than bottom-up, factors determined attentional allocation in natural scenes.

Given that local image statistics associated with semantically informative regions such as objects undoubtedly differ systematically from those of backgrounds, and given our demonstration that such relationships exist for fixated scene regions, the results obtained by investigations linking local image statistics and gaze are entirely consistent with the conclusion that cognitive factors guide eye movements through scenes.

6.2. The special case of sudden onsets

Are there any conditions in which stimulus-based factors have priority over cognitive factors in controlling fixation location during scene viewing? We know of only one definitive case: The top-down direction of the eyes can be disrupted by the abrupt appearance of a new but task-irrelevant object, a phenomenon called oculomotor capture (Irwin, Colcombe, Kramer, & Hahn, 2000; Theeuwes, Kramer, Hahn, & Irwin, 1998). We have recently found that during real-world scene viewing, the transient motion signal that accompanies an abruptly appearing new object attracts attention and gaze quickly and reliably: up to 60% of fixations immediately following the onset are located on the new object (Brockmole & Henderson, 2005a, 2005b). This effect is just as strong when the new object is unexpected as when the observer’s goal is to search for and identify new objects, suggesting that the allocation of attention to transient onsets is automatic. Thus, visually salient scene regions marked by low-level transient motion signals introduced by sudden changes to a scene can influence gaze in a manner divorced from cognitive control.

6.3. How should stimulus-based and knowledge-based information be combined?

The fact that gaze control draws on stored knowledge implies that image properties about potential fixation targets must somehow be combined with top-down constraints. How is this accomplished? One approach is to construct the initial stimulus-based saliency map taking relevant knowledge (e.g., visual properties of a search target) into account from the outset (Findlay & Walker, 1999; Rao, Zelinsky, Hayhoe, & Ballard, 2002).
A second approach is to combine independently computed stimulus-based and knowledge-based saliency maps so that only salient locations within knowledge-relevant regions are considered for fixation. For example, Oliva, Torralba, Castelhano, & Henderson (2003) filtered an image-based saliency map using a separate knowledge-based map of scene regions likely to contain a specific target. A yet more radical suggestion would be to move away from the concept of an image-based saliency map altogether, and to place primary emphasis on knowledge-based control. For example, in what we might call a Full Cognitive Control model, objects would be selected for fixation primarily based on the types of knowledge discussed in the Introduction, such as episodic and schema-based scene knowledge. The visual image in this type of model would still need to be parsed to provide potential saccade targets, but unlike the assumption of the salience hypothesis, the objects and regions would not be ranked according to their inherent visual saliency, but rather would be ranked based on criteria generated from the cognitive knowledge base. For example, if I am searching for the time of the day, and I know I have a clock on the wall in my office, I would rank non-uniform regions in the known location on the wall highly as a saccade target. In this view, visual saliency itself would play no direct role in saccade location selection. Image properties would only be directly relevant to the extent that they support processes needed to segregate potential targets from background. (And of course they would be necessary as input for determining that I’m in my office, how I’m oriented in my office, where the wall is, and so on.) We call this idea that inherent visual saliency plays no role the Flat Landscape Assumption and contrast it with the differentially peaked visual saliency landscapes assumed by the saliency hypothesis.

7. Conclusion

What drives eye movements through real-world scenes during active viewing tasks? Despite the recent popularity of the visual saliency hypothesis as an explanation of gaze control, the evidence supporting it is relatively weak. Cognitive factors are a critical and likely dominant determinant of fixation locations in the active viewing of scenes.

Acknowledgements

This research was supported by the National Science Foundation (BCS-0094433 and IGERT ECS-9874541) and the Army Research Office (W911NF-04-1-0078) awarded to John M. Henderson. Affiliations of co-authors are now as follows: James Brockmole, University of Edinburgh; Monica Castelhano, University of Massachusetts at Amherst; Michael Mack, Vanderbilt University. We thank John Findlay and the volume editors for comments on an earlier draft of this chapter. Please address correspondence to John M. Henderson, Psychology, 7 George Square, University of Edinburgh, EH8 9JZ, United Kingdom, or to John. M. Henderson@ed.ac.uk.
References


Gareze, & Findlay (this volume).


Hooge, Vlaskamp, & Over (this section).


Shen, & Reingold (this section).


Underwood (this volume).


<table>
<thead>
<tr>
<th>Target</th>
<th>Cohort</th>
<th>Unrelated</th>
</tr>
</thead>
<tbody>
<tr>
<td>beaker</td>
<td>beetle</td>
<td>carriage</td>
</tr>
</tbody>
</table>

Fixation proportions over time

![Fixation proportions over time](image)

**Color Plate 2.** Schematic illustrating proportion of fixation curves. *(See Figure 1, Chapter 20, p. 449.)*

![Color Plate 2](image)

**Color Plate 3.** Top left: Original scene. Top middle: Model-determined salient regions in the scene. Top right: Fixation locations from all participants. Bottom: Scene with salient regions and participant fixations overlaid. Red dots show participant fixations within a salient region. Red tails mark saccade paths that originated in a non-salient region. Green dots denote participant fixations outside of the salient regions. *(See Figure 1, Chapter 25, p. 543.)*

![Color Plate 3](image)