Visual, Lexical, and Syntactic Effects on Failure to Notice Word Transpositions: Evidence from Behavioral and Eye Movement Data

Kuan-Jung Huang

University of Massachusetts Amherst

Follow this and additional works at: https://scholarworks.umass.edu/masters_theses_2

Part of the Cognitive Psychology Commons, and the Psycholinguistics and Neurolinguistics Commons

Recommended Citation
https://doi.org/10.7275/21410109.0 https://scholarworks.umass.edu/masters_theses_2/1055

This Open Access Thesis is brought to you for free and open access by the Dissertations and Theses at ScholarWorks@UMass Amherst. It has been accepted for inclusion in Masters Theses by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.
VISUAL, LEXICAL, AND SYNTACTIC EFFECTS ON FAILURE TO NOTICE WORD TRANSPOSITIONS: EVIDENCE FROM BEHAVIORAL AND EYE MOVEMENT DATA

A Thesis Presented

by

Kuan-Jung Huang

Submitted to the Graduate School of the University of Massachusetts Amherst in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

May 2021

Psychological and Brain Sciences
VISUAL, LEXICAL, AND SYNTACTIC EFFECTS ON FAILURE TO NOTICE WORD TRANSPOSITIONS: EVIDENCE FROM BEHAVIORAL AND EYE MOVEMENT DATA

A Thesis Presented

by

Kuan-Jung Huang

Approved as to style and content by:

________________________________
Adrian Staub, Chair

________________________________
Andrew Cohen, Member

________________________________
Brian Dillon, Member

________________________________
Caren Rotello, Department Chair
Psychological and Brain Sciences
ABSTRACT

Visual, lexical, and syntactic effects on failure to notice word transpositions: Evidence from behavioral and eye movement data

May 2021

Kuan-Jung Huang
B.A., National Cheng Kung University
M.S., University of Massachusetts Amherst
Directed by: Professor Adrian Staub

Evidence of systematic misreading has been taken to argue that language processing is noisy, and that readers take noise into consideration and therefore sometimes interpret sentences non-literally (rational inference over a noisy channel). The present study investigates one specific misreading phenomenon: failure to notice word transpositions in a sentence. While this phenomenon can be explained by rational inference, it also has been argued to arise due to parallel lexical processing. The study explored these two accounts. Visual, lexical, and syntactic properties of the two transposed words were manipulated in three experiments. Failure to notice the transposition was more likely when both words were short, and when readers' eyes skipped, rather than directly fixated, one of the two words. Failure to notice the transposition also occurred when one word was long. The position of ungrammaticality elicited by transposition (the first vs. second transposed word) influenced tendency to miss the error; the direction of the effect, however, depended on word classes of the transposed words. Failure of detection was not more likely when the second transposed word was easier to recognize than the first transposed word. Finally, readers’ eye movements on the transposed words revealed no disruption in those trials when they ultimately accepted the sentence to be grammatical. We consider the findings to be only partially supportive of parallel lexical processing and instead propose that word recognition is serial, but integration is not perfectly incremental, and that rational inference may take place before an ungrammatical representation is constructed.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>viii</td>
</tr>
<tr>
<td>CHAPTER</td>
<td></td>
</tr>
<tr>
<td>1. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1. The transposed-word effect and parallel word processing</td>
<td>3</td>
</tr>
<tr>
<td>1.2. Rational inference over a noisy channel</td>
<td>6</td>
</tr>
<tr>
<td>2. EXPERIMENT 1</td>
<td>9</td>
</tr>
<tr>
<td>2.1. Methods</td>
<td>9</td>
</tr>
<tr>
<td>2.2. Predictions</td>
<td>11</td>
</tr>
<tr>
<td>2.3. Analyses and results</td>
<td>13</td>
</tr>
<tr>
<td>2.3.1. Accuracy data</td>
<td>14</td>
</tr>
<tr>
<td>2.3.2. Eye movement data</td>
<td>16</td>
</tr>
<tr>
<td>2.4. Discussion</td>
<td>19</td>
</tr>
<tr>
<td>3. EXPERIMENT 2</td>
<td>22</td>
</tr>
<tr>
<td>3.1. Methods</td>
<td>22</td>
</tr>
<tr>
<td>3.2. Predictions</td>
<td>23</td>
</tr>
<tr>
<td>3.3. Analyses and results</td>
<td>23</td>
</tr>
<tr>
<td>3.3.1. Accuracy data</td>
<td>23</td>
</tr>
<tr>
<td>3.3.2. Eye movement data</td>
<td>25</td>
</tr>
<tr>
<td>3.4. Discussion</td>
<td>27</td>
</tr>
<tr>
<td>4. EXPERIMENT 3</td>
<td>34</td>
</tr>
<tr>
<td>4.1. Methods</td>
<td>34</td>
</tr>
<tr>
<td>4.2. Predictions</td>
<td>36</td>
</tr>
<tr>
<td>4.3. Analyses and results</td>
<td>37</td>
</tr>
<tr>
<td>4.4. Discussion</td>
<td>41</td>
</tr>
<tr>
<td>5. GENERAL DISCUSSION</td>
<td>47</td>
</tr>
</tbody>
</table>
5.1. The role of fixation status in sentence comprehension ........................................ 48

5.2. Serial or parallel processing ................................................................................ 49

5.3. On incremental integration, rational inference, and other misreading phenomena ...... 54

6. CONCLUSIONS ........................................................................................................... 58

REFERENCES .................................................................................................................. 60
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Example item in each transposed condition in Experiment 1</td>
<td>10</td>
</tr>
<tr>
<td>2. Mean accuracy for each condition in Experiment 1</td>
<td>14</td>
</tr>
<tr>
<td>3. Mean accuracy for each condition in Experiment 2</td>
<td>24</td>
</tr>
<tr>
<td>4. Example item in each condition in Experiment 3</td>
<td>35</td>
</tr>
<tr>
<td>5. Mean accuracy for each condition in Experiment 3</td>
<td>38</td>
</tr>
<tr>
<td>6. Mean reading time of the whole sentence by structure types and response types in Experiment 3</td>
<td>40</td>
</tr>
<tr>
<td>7. Mean response time to the detection question by structure types and response types in Experiment 3</td>
<td>41</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>1.</td>
<td>Histogram of error rates by subject in each experiment in Mirault et al. (2018) and the current study</td>
</tr>
<tr>
<td>2.</td>
<td>Accuracy for the transposed sentences, grouped by word class, word length, and fixation pattern in Experiment 1</td>
</tr>
<tr>
<td>3.</td>
<td>Model estimates of differences in eye movements between different response types on each region and each measure in Experiment 1</td>
</tr>
<tr>
<td>4.</td>
<td>Scatter plot of participants’ overall reading time on the filler sentences and their accuracy in detecting the transposition in Experiment 2</td>
</tr>
<tr>
<td>5.</td>
<td>Accuracy for the transposed sentences, grouped by word class and fixation pattern in Experiment 2</td>
</tr>
<tr>
<td>6.</td>
<td>Model estimates of differences in eye movements between different response types on each region and each measure in Experiment 2</td>
</tr>
<tr>
<td>7.</td>
<td>(a) Histogram of the by-subject error rates in Experiment 3</td>
</tr>
<tr>
<td></td>
<td>(b) Scatterplot of mean reading times on the filler sentences and their accuracy in detecting the transposition in Experiment 3</td>
</tr>
</tbody>
</table>
CHAPTER 1

INTRODUCTION

Written sentence comprehension models have traditionally viewed syntactic analysis as taking noise-free perceptual input into the parsing system (Frazier & Fodor, 1978; Jurafsky, 1996) by assuming that word recognition processes always turn out correct. Like word identity, word order also has been believed to be encoded without errors under serial-attention models (e.g., Reichle, Liversedge, Pollatsek, & Rayner, 2009; Reichle, Warren, & McConnell, 2009). However, there have been studies demonstrating that a word in a sentence can be misperceived (e.g., Slattery, 2009) or probabilistically perceived (Levy, Bicknell, Slattery, & Rayner 2009), and that interpretation of a sentence can be non-verdical (e.g., theta-role confusion in passive sentences; Ferreira, 2003, or in object-extracted relative clauses; Burnsky & Staub, in prep, and goal confusion in DO/PO structures; Gibson, Bergen, & Piantadosi, 2013). In light of these findings, it has been proposed that sentence processing is an interaction between top-down structural expectations and noisy bottom-up perceptual evidence, with the perception of word identities remaining uncertain and subject to linguistic knowledge not just prospectively but retrospectively (Futrell, Gibson, & Levy, 2020; Gibson et al., 2013; Levy et al., 2009; Levy, 2011).

In addition to semantics/plausibility-induced misinterpretation such as the examples above, misreading can be caused also by misrepresentation of word order. Most people have anecdotal experience of failing to detect a word transposition error while reading a sentence. Such a phenomenon has recently been formally examined by Mirault, Snell, and Grainger (2018). In this thesis, I focus on this specific type of misreading—the transposed-word effect—and ask how this phenomenon can help us advance models of sentence reading by further exploring when and why it happens.

Section 1.1 briefly summarizes the results from Mirault et al. (2018) and their proposed account for the findings. The account attributes failure to detect word transposition to noisy
positional encoding and parallel lexical processing for multiple words. A formal computational model that entertains these assumptions has been proposed by Snell, van Leipsig, Grainger, and Meeter (2018). The section also summarizes the key features and architecture of this model, OB1-Reader, which motivates the three experiments in the thesis. Section 1.2 introduces another possible account for the misreading of word order, namely rational inference over a noisy channel. Such a proposal has been demonstrated to be effective in explaining various types of misreading phenomena (e.g., Gibson et al., 2013; Poppels & Levy, 2016; Staub, Dodge, & Cohen, 2019). How this explanation can be applied to misreading word-transposition sentences will be described. The account implies that the veridical error is first perceived and later corrected via rational inference. The use of the online incremental measure of eye movements allows us to test this hypothesis.

Sections 2-4 report the three experiments designed to test out the two accounts on offer. There are four goals in Experiment 1: first, to replicate Mirault et al.’s findings in the same task while presenting the target sentences more naturally; second, to test whether word properties (i.e., word length and word class) of the transposed words influence error detection rates, as OB1-Reader has specific predictions for the two effects; third, to examine whether skipping at least one of the two transposed words makes detection failure more likely; and finally, to test whether eye movements on the transposed words are disrupted at all on those trials when participants make an incorrect response (i.e., accepting the word-transposition sentence to be good). A post-perceptual rational inference account predicts that disruption should emerge in all transposed-word sentences regardless of the ultimate response. Experiment 2 is mostly similar to Experiment 1 except that an even more natural task is adopted (detailed in Section 3) and that word length is not manipulated. Findings from the two experiments partially align with OB1-Reader while clearly arguing against a post-perceptual rational inference account. A new hypothesis is proposed that posits both serial word processing and rational inference, while assuming that post-lexical integration is not perfectly incremental but sometimes delayed such
that a syntactic misfit might not even be perceived in the first place. Experiment 3 further tests whether the point of ungrammaticality due to transposition (i.e., ungrammaticality right at the first transposed word vs. later at the second) has an effect on detection rates, which can potentially provide evidence against and for the two accounts. Experiment 3 also serves to provide empirical findings on how tolerant the parser is in the face of an outright grammatical error.

Section 5 provides general discussion of the three experiments. First, while skipping was associated with more failure of detection of transpositions, we argue that what skipping changed was the strength of belief in what has been read, not the veridical order of word recognition. Second, we argue that OB1-Reader was not fully supported by the findings under visual, lexical, and syntactic manipulations of the two transposed words. With results from the literature and results from the current study both taken into consideration, evidence for and against parallel lexical processing still remains equivocal. Finally, it was shown consistently that while there was significant disruption on the critical regions in those trials when participants explicitly reported an error, the processing of the problematic word in those trials when participants failed to report an error was essentially no more difficult than the processing of the same word in a grammatical position. This evidence argues against a post-perceptual rational inference account, and casts doubt on the popular view of highly incremental processing in reading. We argue, however, that rational inference might play a role here but that it must take place before or when integration begins. Such a proposal awaits further verification.

1.1. The transposed-word effect and parallel lexical processing

The question of how letters within a word are represented with respect to their order has evoked many studies in the visual word recognition literature for decades (e.g., Davis, 1999; Grainger & van Heuven, 2004; McClelland & Rumelhart, 1981; Whitney, 2001). One approach to address the question is the use of letter strings that stem from a word but whose two (adjacent)
letters are transposed. Transposed-letter effects refer to facilitatory priming provided by such strings (Perea & Lupker, 2003) or inhibition in recognizing a target word that has a word neighbor with the transposed letters (e.g., Andrews, 1996). Such findings have been taken as evidence that letters within a word are processed in parallel and letter position encoding is noisy (e.g., Davis, 2010).

Mirault, Snell, and Grainger (2018) suggested that, in addition to letter position, encoding of word order is also noisy. They tested this idea with sentences that contained a transposition error (e.g., The white was cat big). Participants were presented one sentence at a time and instructed to make a grammatical judgment as fast and accurately as possible while the sentence remained on the screen. They demonstrated that transposed-word sentences were incorrectly accepted more often than the control sentences (The white was cat slowly) and more often than grammatical sentences were incorrectly dismissed (The white cat was big). The mean response time for the transposed-word sentences were also significantly longer than that for the grammatical counterparts and for the control sentences. Analogous to the argumentation from transposed-letter effects, they posited that the activation of a word is spread across different positions. More importantly, with distributions overlapped across word positions, parallel processing of word recognition gives rise to the transposed-word effect. When cat is recognized prior to was, because the activation of cat extends from its veridical position to its surroundings, cat can be encoded as Word 3 rather than 4.

While parallel processing of words is one possibility to explain the misreading finding, whether or not readers can allocate attention simultaneously to multiple words to engage in lexical processing still remains a controversy among researchers (e.g., Brothers, Hoversten, & Traxler, 2017; Reichle, Liversedge, et al., 2009; Snell & Grainger, 2019; Veldre & Andrews, 2018; White, Palmer, Boynton, & Yeatman, 2019, see General Discussion). Given the ongoing debate, understanding the phenomenon of misreading transposed-word sentences is of theoretical importance. To do so, we test the most recent model of word recognition and eye
movements in text reading, OB1-Reader (Snell, van Leipsig, et al., 2018), which assumes parallel lexical processing and takes the transposed-word effect as supporting evidence.

The key features and architecture of the model are as follows. First, words are activated by open-bigram nodes which are formed by all possible combinations of any two letters in a word, with the relative position of the letters being respected (Grainger, Mathôt, & Vitu, 2014). Open-bigram nodes from one word can combine with those from the adjacent words, creating a pool of word candidates competing to be recognized, with activation of word nodes being determined by attention distribution and recognition threshold determined by frequency and predictability (Bicknell & Levy, 2010; Kliegl, Grabner, Rolfs, & Engbert, 2004). Unlike the E-Z Reader model, where words are always slotted into the correct, veridical order (because words are recognized serially one after another, from left to right), the word recognition process in OB1-reader includes both lexical activation and position mapping. OB1-Reader creates a sentence-level spatiotopic representation of several words (Snell, Meeter, & Grainger, 2017): upon fixating a word, a rough estimate of the shape and lengths of its neighboring words within the perceptual span will be created and stored in working memory; these are slots to be filled by words whose activation reaches their recognition threshold. For a word to be recognized and appended to a slot, its length and shape must match the low-level visual information in the spatiotopic representation. Furthermore, the sentence-level spatiotopic representation provides top-down high level syntactic and/or semantic feedback that guides word recognition. For instance, given a recognized article at Position 1 and a recognized adverb at Position 2, an adjective but not a noun is expected at Position 3. This syntactic constraint will facilitate activation of adjectives at Position 3 and might inhibit activation of nouns, influencing the word recognition process (Snell & Grainger, 2017; Wen, Snell, & Grainger, 2019; Wen, Mirault, & Grainger, 2020). Therefore, a transposed-word sentence like Dad caught big the fish this morning will sometimes be misread as a grammatical sentence since (a) the is likely to be recognized before big, (b) the two words are similar in length, and (c) an article is generally
more expected than an adjective, given the first two words. Such an account holds that failure
to notice transposition is due to failure at the level of perceptual encoding (i.e., during the
recognition stage). Several predictions can be derived from the model for different
manipulations on the two transposed words. The three experiments in the thesis test these
predictions.

1.2. Rational inference over a noisy channel

Mirault et al.’s finding could also be explained by a rational inference account (Gibson et
al., 2013; Levy et al., 2009; Ryskin, Futrell, Kiran, & Gibson, 2018). The premise is that
language communication processes are rife with noise, such that the intended message from
the producer might not be faithfully transmitted to the receiver. To recover the intended
message, the comprehender engages in optimal Bayesian decoding, formalized as (1):

$$ P(S_i | S_p) \propto P(S_i)P(S_i \rightarrow S_p) $$

The term on the left is the probability of the sentence actually carrying the intended message
by the speaker, given the perceived input. This probability is proportional to the prior
probability of that intended message (base rate), and the probability of that sentence being
corrupted to the perceived input (either due to the producer’s errors, the perceiver’s errors, or
environmental noise).

To illustrate rational inference, several studies have shown that readers sometimes attain
nonliteral interpretation of a sentence (e.g., Gibson et al., 2013, 2017). For instance, Gibson
and colleagues (2013) found that readers are more likely to interpret as plausible an
implausible sentence (e.g., *The mother gave the candle the daughter*) whose supposedly
intended message can be recovered by an insertion of *to*, than an implausible sentence (e.g.,
*The mother gave the daughter to the candle*) whose same intended message can be recovered
by a deletion of *to*. This difference was taken as evidence that readers, when interpreting a
sentence, consider the likelihood of noise: a deletion of a particular word is sampled only
from a small number of words within the sentence, while an insertion of a word is sampled
from a large number of words in the speaker’s lexicon. Therefore, the probability of a deletion having taken place due to the producer’s error is higher than that of an insertion. In the same study, it was also shown that readers take into consideration the local base rate of implausible sentences. If in a reading task most of the study sentences contained semantic errors (i.e., \( P(\text{implausible sentence}) \) was high), participants were more likely to interpret sentences containing a deletion/insertion error veridically without correction (e.g., a mother giving a baby to a bottle). In short, manipulations on the two terms on the right hand of the Bayesian formula did influence the probability of the intended message given a perceived input. Further studies demonstrated that readers even model the specific nature of noise (Poppels & Levy, 2016; Ryskin, et al., 2018). For instance, while implausible sentences like *The ball kicked the boy* and *The ball fell from the floor to the table* can both be a result of corruption from word swapping/exchange, participants were more likely to attain the nonliteral, plausible meaning of the sentence in the latter case as it involved swapping two prepositions while the former involved exchanging two nouns (Poppels & Levy, 2016).

Finally, later findings from Ryskin et al. (2020) showed that participants showed both a N400 and P600 when reading sentences with a word that does not fit semantically with the context, but whose word neighbor does (e.g., *The storyteller could turn any incident into an amusing antidote*). They suggested that the P600 is indicative of an online error correction. If so, it has to be true under this current noisy-channel account that rational inference occurs not only after the perception of the veridical stimulus but also after the integration of that word into the context, which led to the error signal (N400) as well as the correction signal (P600).

Consider applying a rational inference account to word-transposed sentences. While perceiving the sentence as *The white was cat big*, the reader might go through inference, computing how likely it is that the sentence was intended to be *The white cat was big* and how likely it is that it went through some corruption during the transmission of the signal (e.g., having fixated or paid attention to the words in the wrong order). If both the prior probability
and likelihood of the input given the intended message are high, the reader may attain the intended message as the ultimate interpretation, hence not reporting the transposition error.

Although not explicitly stated in any of their papers, Levy, Gibson, and colleagues’ rational inference account implies that the veridical signal is first observed/perceived as an error (Ryskin et al., 2020). One way to test this account is therefore to investigate whether there is evidence of incremental processing difficulty in perceiving the transposition error even when the reader ultimately responds that there was nothing wrong with the sentence. The first two experiments reported in this thesis adopted eye-tracking to test this post-perceptual rational inference hypothesis.
CHAPTER 2

EXPERIMENT 1

The first experiment was an attempt to replicate Mirault et al (2018)’s results and extend the investigation to the roles of word class, word length\(^1\), and fixation pattern while eye movements were being collected. All three factors, as explained in the Predictions section, were used to test the parallel processing account by Snell, van Leipsig, et al. (2018). Detailed analyses of eye-movements contingent on response type were further performed to test the rational inference account.

2.1. Methods

Fifty-three undergraduates from the University of Massachusetts Amherst were recruited either for payment or course credit. Participants were native speakers of English with normal or corrected-to-normal vision and without self-reported history of reading or language disorders.

To test OB1-Reader, we created different conditions of word length (both-short vs. one-long) and word class (different combinations of open-class and closed-class words) of the two transposed words; see Table 1 for examples. All sentences were 7 words long, presented in one line, and transposition occurred between the third and fourth word. Note that even after transposition, Word 3 itself is grammatical in context. Short words are of 2-4 letters. Long words are of 8-11 letters. For the both-short condition, the length difference between the two words is no greater than 1 letter. Open-class words are nouns, adjectives, or verbs. Closed-class words are pronouns, determiners, or prepositions. Trials were further grouped into two types based on the fixation pattern on the two transposed words: if a participant skipped at least one of the words on her first pass, the trial is coded as a skipping trial, and if she fixated both words on first-pass reading, a fixating-both-words trial.

\(^1\) We conveniently term this manipulation as word length when in fact it is a manipulation of word length difference of the two transposed words (see Discussion for Experiment 1).
Table 1. Example item in each transposed condition in Experiment 1. Transposed words are in bold; in grammatical conditions these words were presented in the opposite order. Word class labels reflect the order in transposed conditions.

We created 30 items for each cell of the design, with each having a grammatical version and a transposed counterpart. However, for the open-open, one-long condition, either the first or the second word could be long, so we created 60 items for this cell; in the analysis, we collapsed these sub-conditions, as post hoc analyses did not reveal significant differences in error rates between these two sub-conditions. Each participant saw either the grammatical or transposed version of each item. Fifty incomplete sentences served as ungrammatical fillers, which were included in order to encourage participants to read each sentence to the end. Each participant therefore read a total of 260 sentences, consisting of 105 transposed sentences, 105 grammatical sentences, and 50 incomplete sentence fillers.

The movement of the subject’s right eye was recorded using an EyeLink 1000 (SR Research, Toronto, ON, Canada) eye-tracker. The sampling rate was 1000 Hz. Subjects were seated 55 cm from a CRT monitor, with 1,024 × 768 resolution and a screen refresh rate of 120
Sentences were displayed on a single line in 11-point Monaco font, with between three and four characters subtending 1° of visual angle. Unlike Mirault et al. (2018), for every trial participants were first cued by a black dot to fixate on the left edge of the screen, after which the sentence appeared several spaces to the right of the dot position so that some saccades entering the beginning of the sentence must be made, as in natural reading. The sentence appeared on the screen until participants made a response or until 15000ms time-out. Participants were asked to make a judgement as quickly and accurately as possible whether the sentence was well-formed. The experiment took approximately 40 minutes.

2.2. Predictions

As detailed in Section 1.1., under OB1-Reader word recognition and appending must respect the shape of slots in the spatiotopic representation which are in turn estimated from the low-level visual information from multiple words within the perceptual span. The first prediction then is that readers are more likely to fail to notice transposition of similarly short transposed words compared to when one of the words is short and the other is rather long. For instance, while activation of her can reach its recognition threshold earlier than that of favorite in She repeated favorite her song ten times, its length is too different from that of favorite that it will not be able to fit in the slot for Position 3.

The second prediction concerns relative ease of recognition of the two transposed words. OB1-Reader states that if the second transposed word reaches its recognition threshold earlier than the first transposed word does (and when the lengths of the two are similar), the second transposed word would be erroneously appended to the first position (Snell, van Leipsig, et al., 2018). This is a prediction that is not specific to OB1-Reader but holds under parallel processing models in general (Reichle, Liversedge, et al., 2009). The manipulation of word class here provides a potential venue to test this hypothesis. As closed-class words are generally higher in frequency than open-class words, they should be likely to reach their recognition threshold earlier than open-class words (Inhoff & Rayner, 1986). A transposed sequence with
an open-class word followed by a closed-class word (i.e., the open-closed condition) is thus predicted to be more likely to elicit failure to notice the transposition error than a transposed sequence with a closed-class word followed by an open-class word (the closed-open condition). One thing to note here is that relative ease of recognition should ideally be directly manipulated by word frequency, rather than word class combination, with other factors controlled. Here we use word class combination mainly to follow the exploratory analysis by Mirault et al (2018), who found that accuracy was lower when at least one of the transposed words was closed-class.

Furthermore, if failing to notice a transposition is a result of the second word being recognized before the first word, this should predict that in trials when the participant judges the transposed sentence to be grammatical, the fixation pattern could not be sequential fixations on the two transposed words (or is at least less likely to be, when compared with correct-judgment trials): a fixation on the second word following a fixation on the first word would mean that the second word is not recognized before the first one, since activation of a recognized word falls back to zero and will not get selected as the saccade target (Engbert, Nuthmann, Richter, & Kliegl, 2005; Snell, van Leipsig, et al., 2018). Participants thus should have veridical perception of word order with this fixation pattern. On the other hand, according to E-Z Reader, because representation of word order by default follows the relative order of word recognition (Reichle, Liversedge, et al., 2009), and words are recognized serially even when skipped (e.g., Gordon, Plummer, Choi, 2013), it is predicted that skipping will not have an effect on detection rates.

In addition to the analysis of accuracy as a function of word length, word class, and fixation pattern, we compared eye movements (fixation durations and regression probabilities) contingent on the response. Trials were further grouped into three: when a participant correctly accepted a grammatical sentence (AG trials), when she incorrectly accepted a transposed-word sentence (AT trials), and when she correctly dismissed a transposed-word sentence (DT trials). The relatively small number of trials on which participants dismissed a grammatical sentence
were omitted. These analyses were to test the post-perceptual rational inference hypothesis, according to which misjudgment is a consequence of inference made based on perception of an erroneous input. If this is correct, even when the participant ends up incorrectly accepting the transposed-word sentence (AT trials), her incremental processing on the transposed words—here the second one especially, since it is the region where ungrammaticality emerges—as revealed by eye movements, should be as disrupted as that when she correctly dismissed the sentence (DT trials). Eye movements in both transposed trials should be inflated when compared to AG trials.

2.3. Analyses and results

Data from two participants were removed as their accuracy for grammatical and filler sentences was below 75%. One participant accepted all the transposed-word sentences so the data were also removed, leaving valid data from 50 participants in total.

We performed data analysis with mixed effects logistic regression models of judgment accuracy and regression probability and linear mixed effects models of reading time, implemented using the lme4 package (version 1.1-21; Bates et al, 2015) for the R statistical programming environment (version 3.5.3; R Core Team, 2019). Model comparison and selection followed the parsimonious principle by Bates, Kliegl, Vasishth, and Baayen (2015). Random slopes were first included for all possible contrasts without correlation parameters; if convergence failed, the contrast with the least variance was removed first. The models that successfully converged were compared with their random-intercept-only counterparts; only those showing significant improvement in the goodness of fit were kept. Models with correlation parameters would be ultimately adopted if they showed significant improvement in the goodness of fit from their no-correlation models.

---

Because word-class was a between-item variable, random item slopes did not include this effect. All generalized logistic linear mixed-effect models were random-intercept-only due to convergence issues. No interaction contrasts were entered in random slopes due to convergence issues.
2.3.1. Accuracy data

Accuracy as a function of word length, word class, and grammaticality is shown in Table 2. An initial model that includes grammaticality as the only fixed effect was constructed, with the transposed-word condition as reference against the grammatical and filler conditions respectively. Transposed sentences elicited significantly more errors (13.3% on average) than grammatical counterparts (6.7%, $z = 11.64$, $p < .001$) and fillers (incomplete sentences, 6.3%, $z = 6.76$, $p < .001$), thus replicating the transposed-word effect reported by Mirault et al. (2018).

Numerically, the error rates across the two experiments are also comparable. Figure 1 further shows the distributions of error rates by participants across their two experiments and the current two eye tracking experiments. In all the experiments, there was great individual difference in error detection, with many subjects correctly reporting the error on almost all the transposed-word trials while some failing to do so on a substantial proportion of trials.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Grammatical</th>
<th>Transposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>open-closed, both short</td>
<td>94.8 (1.2)</td>
<td>85.7 (2)</td>
</tr>
<tr>
<td>open-closed, one long</td>
<td>94.3 (0.9)</td>
<td>90.8 (1.9)</td>
</tr>
<tr>
<td>closed-open, both short</td>
<td>90.1 (1.2)</td>
<td>89.9 (1.6)</td>
</tr>
<tr>
<td>closed-open, one long</td>
<td>97.6 (0.6)</td>
<td>91.3 (1.7)</td>
</tr>
<tr>
<td>open-open, both short</td>
<td>90.9 (1.4)</td>
<td>75.9 (3)</td>
</tr>
<tr>
<td>open-open, one long</td>
<td>92.5 (1.5)</td>
<td>86.8 (2.2)</td>
</tr>
<tr>
<td>Filler</td>
<td></td>
<td>93.7 (1.3)</td>
</tr>
</tbody>
</table>

Table 2. Mean accuracy (%) for each condition in Experiment 1; standard error (by subjects) in parenthesis.
Fig. 1. Histogram of error rates by subject in each experiment in Mirault et al. (2018) and the current study.

For our main analyses, we included only the transposed-word trials. Fixed effects were word length, word class, and fixation pattern, and their interactions, with word class contrasts being treatment-coded using the open-closed condition as reference and the other two effects being sum-coded. The both-short condition was coded as 0.5 and the one-long condition as -0.5; The skipping-trial group was coded as 0.5 and the fixating-both-trial as -0.5. Figure 2 shows the accuracy for each. All three factors—word length, word class, and fixation pattern—
showed main effects (all $\chi^2$s > 23.8, $p < .001$). The both-short condition elicited more errors than the one-long condition. The open-open condition elicited many more errors than the open-closed and closed-open conditions, with the latter two not being significantly different from each other ($z = 1.43, p > .1$). There was no three-way interaction but only a two-way interaction between word class and fixation pattern ($\chi^2(2) = 10.13, p < .01$). Separating the data into three sets by word class, we found that whether both transposed words were fixated influenced error rates only for the open-open condition ($z = -5.03, p < .001$), with a numeric trend for the open-closed condition ($z = -1.7, p = .1$) and little sign of an effect for the closed-open condition ($z = -.91, p > .05$).

Fig. 2. Accuracy for the transposed sentences, grouped by word class, word length, and fixation pattern in Experiment 1. Error bars reflect 95% confidence interval, by subjects.

2.3.2. Eye movement data

Short fixations (<80ms) within 1 character position of another preceding or following fixation were automatically combined. Eye-movement data were analyzed separately on each of Words 3, 4, and 5 (spillover), and on the final region of each sentence. Word 4 (W4) is the

---

3 We reported $\chi^2$ from model comparison results because word-class had three levels and was treatment-coded which made the beta estimate of the slope of the other two factors reflect simple effects rather than main effects.
critical point at which a transposed sentence becomes ungrammatical. As described earlier, eye movement measures were analyzed contingent on response type: the AG, AT, and DT trials. In the statistical models, the response type variable is the only fixed effect, treatment-coded with AT as the reference level.

Four standard fixation time measures (Rayner, 1998) are reported: two early processing measures—first fixation duration (FFD; duration of the first fixation on a word) and gaze duration (the sum of the durations of all fixations on a word before leaving it) —and two later processing measures—go-past time (the sum of the durations of all fixations from the first on the word until the word is exited to the right, including any regressive re-reading) and total viewing time (the sum of the durations of all fixations on a word). All four measures included only observations of non-zero values; if a word was skipped on first-pass reading, that trial is not included in the calculation of FFD, gaze, or go-past, and if a word was not fixated at all, that trial is not included in the calculation of total time. We also analyzed two binary dependent measures: whether first pass reading of a word ended with a regression rather than a forward saccade, and whether there was a regression into a word from later in the sentence. Word 3 in AT and DT was compared with Word 4 in AG, and vice versa, so that the exact same words were compared.

Figure 3 shows the estimated difference in eye movement measures between different response-type trials (AG-AT and DT-AT) at different regions. A statistically negative AG-AT difference in first-pass reading will suggest that the ungrammatical, transposed sequence led to substantial disruption in the very first place and that the misjudgment was due to a later inference process. This will be further supported if it is also found that the DT-AT difference is not significant, such that ungrammatical, transposed input elicited the same extent of disruption (at least in the earliest stages). Surprisingly, however, it was revealed that the difference between AG and AT was not significant across the board (except differences in the total viewing time and regression-in rate on W3). On the other hand, the difference between DT and AT was
significant across the board. This difference was even as early as in the first-pass reading of W3, which theoretically should be grammatical at that point, hence suggesting a parafoveal-on-foveal effect of ungrammaticality (but see Discussion). At the final region, the direction of the difference became opposite, with the lowest reading time in DT and again barely no difference in reading time between AG and AT.

Fig. 3. Model estimates of differences in eye movements between different response types on each region and each measure in Experiment 1. Error bars reflect 95% confidence intervals.
AG=accept grammatical sentence; AT=accept transposed; DT=dismiss transposed. $p<.05$ *,
2.4. Discussion

Summarizing results in Experiment 1, first, we found significantly more judgment errors for transposed-word sentences than their grammatical counterparts and the fillers. This verified the transposed-word effect, even when target sentences were presented more naturally (cf. having the participants fixate at the center of the sentence as it appeared). It was also revealed that participants’ error rates ranged from zero to almost 50%, indicating substantial individual differences.

Second, word length had a reliable effect such that when the two transposed words differed greatly in length, the error was more easily detected. Third, we found a word class effect, albeit opposite in its direction to that found in Mirault et al. (2018): the error rate was the highest when the two words were both open-class words, with no difference between the open-closed and closed-open conditions. Fourth, skipping at least one of the transposed words was associated with a higher error rate, although this effect seemed to be dependent on the sentence structures. Finally, looking at the eye movement data, we found that participants’ reading pattern in AT trials was mostly indistinguishable from that in AG trials while differing significantly from that in DT trials, despite the fact that the perceptual inputs were actually the same ungrammatical sentences in AT and DT trials.

The first finding suggests that failure to notice word transposition was not due to a potential artifact in the previous design—fixation pattern might be more likely to be out of order due to the central presentation (i.e., first fixating the second transposed word and then the first transposed word)—or due to any other unnatural reading strategies. This issue was also recently address by Mirault, Guerre-Genton, Dufau, and Grainger (2020), although in a rather novel experimental setting (i.e., using virtual reality). They found the transposed-word effect to be replicable when sentences were presented to the right of the initial fixation cross. The next question to be asked is how prevalent this phenomenon is when neither speed nor
accuracy is emphasized. Experiment 2 thus adopted an even more natural reading task to tackle this.

Except for the word length manipulation, all the other three research questions (regarding word class and fixation pattern on accuracy and response type on eye movement data) will be re-assessed in Experiment 2. Given that the experimental setting is less natural in Experiment 1, it is also reasonable to hold some doubts regarding some of these findings. Therefore we wait to see if the results replicate in Experiment 2 and reserve detailed discussion of them (see Discussion for Experiment 2 and General Discussion).

In terms of the word length effect, the results provide some support for OB1-Reader. The model implements a sentence-level spatiotopic representation that can keep track of word order by first checking the low-level visual information from the perceptual input. This can prevent misrepresenting word order when the veridical sentence is in fact grammatical (e.g., *Her professor commented on* as *Her professor on commented*), which had been raised as a potential problem for parallel word processing (Reichle, Liversedge, et al., 2009). Here we indeed found that two transposed words that are similar in length were more likely to be read wrong. Two things must be noted, however. First, while the error rate in the both-short condition was relatively higher than that in the one-long condition, transposed-word sentences in the one-long condition still elicited more errors than their grammatical counterparts \((z = 4.35)\) and the fillers \((z = 4.58)\). OB1-Reader does not predict misreading when the two transposed words are saliently different in length (here 3-4 letters vs. 8-11 letters). Second, as word length conditions were not fully crossed (i.e., there was no both-long condition), it was not clear whether the effect here was a length difference or length effect. It could be the case that the one-long condition elicited fewer errors not because the lengths of the words differed but because misreading is common only when both words are extremely short. A stronger test for the visual constraint from a spatiotopic representation can be made by including both-long items (e.g., transposed words that are both 7-8 letters). If the error rates are similarly high for both-short
and both-long conditions, it will further support the assumption of spatiotopic representation in OB1-Reader.
CHAPTER 3

EXPERIMENT 2

Given that a speeded grammaticality judgment task, after all, does not resemble normal reading, it was not clear empirically how often readers fail to detect an error when little emphasis is placed on the detection process. Experiment 2 is aimed at answering this question and the other questions that motivated Experiment 1 under a more natural setting.

3.1. Methods

Sixty-five undergraduates from the University of Massachusetts Amherst were recruited either for payment or course credit. Participants were native speakers of English with normal or corrected-to-normal vision and without self-reported history of reading or language disorders.

We adopted the experiment design from Staub et al. (2019). Participants were instructed to read naturally and answer questions. Participants hit a proceed button once finishing reading the study sentence (or a 15000ms time-out), after which the sentence disappeared. Every study sentence was followed by a question which did not co-occur with the study sentence. While fillers were followed by a comprehension question (two-thirds of the trials), critical sentences were followed by an error-detection question, “Was there anything wrong with that sentence?” (one-third of the trials). Participants did not know what type of question would be asked while still in the study sentence phase. Critical sentences consisted of 30 transposed-word sentences and 30 of their grammatical counterparts; they were embedded in another 120 grammatical filler sentences. This reading-mainly-for-comprehension task therefore gets us closer to the everyday reading setting.

Critical sentences are subsets of the items taken from Experiment 1. Due to the need to include a great number of filler sentences, the number of critical sentences in each sub-condition was reduced from 15 to 10 for each experiment list. One-long items were also excluded due to this constraint. Therefore, only the word class manipulation was preserved.
The eye-tracker setting was the same as in Experiment 1. The experiment took about 40 minutes.

3.2. Predictions

Word processing strategies have been shown to differ between reading for comprehension and proofreading (Schotter et al., 2014). The first prediction thus is that, given the weaker emphasis on error detection, accuracy rate for the transposed-word condition should be even lower.

The predictions for word class and fixation pattern effects are the same as in Experiment 1. Given the results in Experiment 1, it is expected that accuracy rates for the open-closed and closed-open conditions will not differ. Whether or not open-open items will elicit more errors, however, is less clear, as results were not consistent across Experiment 1 and Mirault et al. (2018). It is also expected that skipping at least one of the transposed words will be associated with more errors.

As for the eye movement data, it is of interest whether we will see the same pattern in Experiment 1 (i.e., processing time: AG = AT < DT).

3.3. Analyses and results

Model construction and comparison followed the same principles as in Experiment 1. Data from one participant were excluded as the participant reported having ADHD. Another participant who did not complete the whole experiment was also removed, leaving 63 subjects in total for the analyses. All participants’ accuracy for the filler comprehension questions were above 75% (mean 95.5%).

3.3.1. Accuracy data

Table 3 shows the mean accuracy for grammatical and transposed versions of each word class condition. An initial model that used grammaticality as the only fixed effect showed that the error rate was significantly higher for the transposed-word condition than for the grammatical condition (14.4% vs. 7.7%, z = 6.82, p < .001). However, accuracy for the
grammatical open-open condition is particularly low. This could potentially mask the interpretation of any word class effect in the transposed-word condition (see Discussion of Experiment 3).

<table>
<thead>
<tr>
<th>Condition</th>
<th>Grammatical</th>
<th>Transposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>open-closed</td>
<td>95.2 (1.4)</td>
<td>86.2 (1.6)</td>
</tr>
<tr>
<td>closed-open</td>
<td>94.9 (1)</td>
<td>86.6 (1.8)</td>
</tr>
<tr>
<td>open-open</td>
<td>86.8 (2)</td>
<td>84.1 (1.6)</td>
</tr>
<tr>
<td>Filler, Comp. Question</td>
<td>95.5 (2.6)</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Mean accuracy (%) for each condition in Experiment 2; standard error (by-subjects) in parenthesis.

As shown in Figure 1 above, there was again considerable variability in error rates across participants. Since here we have a more reliable, ecologically valid measure of reading speed from the reading time of grammatical fillers, we attempted to capture the individual difference by correlating participants’ accuracy rate with their mean reading time on the fillers. There was a significant positive correlation between the two such that slower readers tended to detect the errors more ($r = 0.34$, $p < .01$, Figure 4).
For our main analyses we again included only the transposed-word trials. Figure 5 shows accuracy grouped by word class and fixation pattern. A generalized mixed effect model was constructed with word class and fixation pattern and their interaction as the fixed effects. The only significant effect was of fixation pattern ($\chi^2(1) = 10.19, p < .001$). When at least one of the transposed words were skipped, participants failed to notice the error 17.4% of the time. This rate was only 10.7% when both transposed words were fixated. Looking at the association the other way around, both words were more likely to be fixated when the transposition was detected (47%) than when the transposition was not detected (34%). Neither the main effect of word class nor the interaction was significant.

Fig. 5. Accuracy for the transposed sentences, grouped by word class and fixation pattern in Experiment 2. Error bars reflect 95% confidence interval (by subjects).

3.3.2. Eye movement data
Short fixations (<80ms) within 1 character position of another preceding or following fixation were automatically combined. As in Experiment 1, we compared eye movements on the critical regions (W3 and W4), and spillover region between trials of different response types. Response time to error detection questions was also compared. As shown in Figure 6, the results were qualitatively similar to those in Experiment 1. There was barely any statistical difference between AG and AT across the board, with the only exception in the regression rate out of W3 and a significant but relatively small difference in the regression rate into W3. The comparison between AT and DT, however, showed substantial disruption on eye movements for the latter. This was present across all the stages of W4 and the late stages of W3. Unlike Experiment 1, there was no disruption shown at the early stages (first-pass) of W3. For the RT to error detection questions, participants made substantially quicker responses in DT trials, indicating that they were explicitly aware of the error and ready to report it. RTs however were not distinguishable in AG and AT trials.

To compare Experiments 1 and 2 even better, we did the same analysis again, this time only on the subset of critical items in Experiment 1 (i.e., holding the material the same across the two experiments). This analysis on the subset items in Experiment 1 led to qualitatively identical patterns as that on the full set of items in Experiment 1. The one and only exception in this additional analysis is that the first fixation durations at Word 4 were statistically the same across three response types. The disruption in DT trials only appeared in gaze duration, and later measures, at Word 4. Critically, though, once again there was not any reliable hint that processing in AT trials was disrupted by the ungrammatical transposition across the board. Finally, as in the full-item-set analysis there was early disruption in DT trials at Word 3, different from Experiment 2. This difference might have arisen from the different tasks used (see the following discussion section).
Figure 6. Model estimates of differences in eye movements between different response types on each region and each measure in Experiment 2. Error bars reflect 95% confidence intervals. AG=accept grammatical sentence; AT=accept transposed; DT=dismiss transposed. $p<.05 \, *, \, p<.01 \, **, \, p<.001 \, ***$.

3.4. Discussion

Experiment 2 is similar in essence to Experiment 1, except that we attempted to extend the research to a more natural reading setting. The results were also fairly similar across the two experiments.
Summarizing Experiment 2, first, we found a numerically similar rate of not detecting transposition errors (14.4%, cf., 13.3 in Experiment 1). Second, there was no difference in error rates between the open-closed and closed-open conditions, as in Experiment 1. However, unlike Experiment 1, the error rate in the open-open condition was not significantly higher than those in the other two conditions. Third, we found again an effect of fixation pattern. The effect did not interact with word class. Finally, when it comes to eye movements, the patterns were similar between AG and AT trials, while they differed significantly from that in DT trials.

The first puzzle in Experiment 2 is that readers were not more likely to fail to notice transposition errors in a reading task that mainly focused on comprehension. Given that the instruction was to explicitly detect an error in Experiment 1, it was expected that failure to notice transposition errors should be less frequent there. One possible explanation for the current finding is that the speeded instruction made the participants in Experiment 1 more error-prone (Ratcliff & McKoon, 2008). The histogram of error rates for Experiment 1 (Figure 1) shows that while many subjects clustered in the left end bins, making very few errors, a few of them made many errors. The latter who focused more on the speeded instruction thus might have made the overall error rate higher. Second, while the task in Experiment 2 asked the participants to read naturally, it is possible that participants, knowing that on some trials there would be error detection questions, still implicitly attempted to detect errors, which made the overall error rate slightly lower. The task effect thus might have been attenuated and harder to find. It is conceivable that had we run an experiment that explicitly asks for error detection but does not emphasize speed and an experiment that has zero error-detection component (but in some way still gives us a measure of error detection rate), we would have found a reliable difference in error detection rates between the two. Regardless, the fact that we found a significant and reliable transposed-word effect in a rather natural reading task shows that readers do sometimes fail to notice a transposition error. Finally, while error detection rates did not differ between the two tasks, eye movement patterns differed notably such that participants
spent a lot more time on the problematic region when they noticed an error (the AT-DT contrast in Figure 6) in the reading-for-comprehension task, while they got stopped by ungrammaticality for only a short time in the speeded grammaticality judgment task both in the full-item-set analysis (Figure 3) and the subset analysis. This provides further evidence that readers’ eye movements are influenced by task demands (Schotter et al., 2014).

In terms of the effect of word class, we again failed to find a significant difference between the open-closed and closed-open condition. Different from Experiment 1, however, we did not see a significant difference between the open-open and the other two conditions either. Since word class manipulation will again be preserved in Experiment 3, we reserve the discussion for now.

As for the fixation pattern, in Experiment 1 we found a main effect of fixation status as well as an interaction between fixation status and word class such that the effect was not significant for the closed-open sentences (though in the same direction as in the other two word class conditions). In Experiment 2 we found a main effect of fixation pattern without an interaction with word class. Taken together, it seems appropriate to conclude that there is a true effect. When both transposed words were directly fixated, the readers were more likely to detect the error; similarly, there were more fixating-both-words trials when readers ultimately detected the error than when they failed to do so. Under OB1-Reader, a relationship between skipping and failure to notice a transposition is expected: When the second word is recognized before the first, the complete recognition of the second word should lead to a skip of this word. However, although we saw fewer fixating-both-words trials among the AT trials than among the DT trials, fixating-both-words trials still constituted a substantial proportion of the AT trials (45% in Experiment 1 and 34% in Experiment 2). It will be informative to see whether such a high proportion of direct fixation trials can be formally modelled by OB1-Reader under the proposal that the transposed-word effect arises because the second transposed word is recognized earlier than the first, and the assumption that once a word is recognized, it in
principle does not receive a fixation (Engbert et al., 2005).

On the other hand, under E-Z Reader, which posits that word recognition is always serial in the order as appearing on the page, word order representation need not be specially encoded but simply follows the relative order of recognition (Reichle, Liversedge, et al., 2009). This fundamental assumption is believed to be independent of fixation status (Rayner, Pollatsek, Liversedge, & Reichle, 2009). In E-Z Reader, cancellation of a saccade on Word N+1 is triggered when the L1 stage of Word N+1 is finished—when complete recognition of that word is imminent (Choi & Gordon, 2013, 2014; Gordon, et al., 2013; Reichle, Pollatsek, Fisher, & Rayner, 1998). The current version of E-Z Reader—which assumes that all words, regardless of fixation status, are processed serially from left to right and that word order is also constructed in this fashion—will thus predict no effect of skipping on misreading of word order. The effect of fixation pattern discovered here thus would seem at odds with the model. However, it could be the case that what skipping influenced here is not the word recognition process and order of word recognition (hence not compromising the seriality assumption) but strength of belief in what has been read. That is, perceptual evidence from direct fixations on the two words gives more confidence in interpreting the string as it is veridically, while skipping results in weaker perceptual evidence so that prior experience (linguistic knowledge about what sentences should look like) weighs more (Norris, 2006; Staub et al., 2019). Having skipped at least one of the words thus made the participants more likely to interpret the sentence non-literally, via inference which integrates both bottom-up input and top-down expectation (e.g., Gibson et al., 2013, Levy, 2008).

Regarding the eye movement data, we performed the same analyses as we did in Experiment 1. The patterns of eye movement data across different response types were qualitatively similar to those in Experiment 1. While significant disruption was present as early as the first-pass reading of W4, the point where the sequence no longer remains grammatical, in DT trials when participants explicitly reported an error in the sentence, there were no inflated
reading times found in AT trials in all, except one, eye movement measures on the critical regions. In those trials, it was as if the readers had been reading a grammatical sentence. In Experiment 1, the picture was less clear-cut than in Experiment 2: for the late measures of W3 (regress-in and total viewing time), there was significant disruption found in AT trials, when compared with eye movements in AG trials (this disruption, however, was still smaller than that found in DT trials, as the contrasts between AT and DT were also significant). For the early stages of W3, where the word sequence is still grammatical, there was also already indication of disruption. Such inflated processing time might have been due to reading strategies specific to an unnatural grammatical judgment task. While this speculation needs to be confirmed in future studies, it is important to note that in both Experiments 1 and 2, there was no disruption on W4, the very problematic region, in AT trials. That the disruption on W3 was not coupled with disruption on W4 suggested that the increased re-processing was not due to a perceived error from W4. Therefore, there is certainly no robust evidence of disruption in AT trials across the two experiments.

Our eye movement analyses separating and comparing trials based on the behavioral detection response were an attempt to connect failure to notice transposition errors and rational inference (Gibson et al., 2013; Staub et al., 2019). In the experiments, it was no surprise that in those trials when participants explicitly responded that there was something wrong with the sentence, their processing times on the problematic regions were longer. The result in the misreading trials (AT trials), however, provided insight as to when exactly inference takes place if non-literal interpretation is a result of rational inference. The current view of a noisy-channel framework suggests that it occurs fast enough to be captured in real time, yet slow enough such that a perception of an error due to the difficulty to integrate the word into the context has already emerged (Ryskin et al., 2020). Our eye-movement data however did not reveal any perception of an error, suggesting that if inference indeed took place, it must have been even faster, earlier than the high-level integration of the veridical word.
Before further presenting a revision of the rational inference account, it is necessary to first describe current views on the integration process in serial-attention models. E-Z Reader 10 (Reichle, Warren, et al., 2009) states that while only one word goes through lexical processing at one time, post-lexical integration of Word \( N-1 \) can run concurrently while lexical processing of Word \( N \) takes place. Forward eye movements will be stopped and a regression will occur when integration fails. The model specifies two scenarios when integration fails. One failure of integration is due to violation (e.g., syntactic misfit or implausibility) and the other type of failure of integration arises when integration finishes too late such that the recognition process of the next word is already complete. The model therefore assumes a perfect incremental integration mechanism. That is, integration of Word \( N \) always precedes that of Word \( N+1 \). In our current study, if sentence processing was highly incremental such that whenever a new word in a sentence is recognized, it immediately gets integrated into the context and forms a new representation, we should have seen processing difficulty on W4 on every transposed trials regardless of their response type: for instance in The boy on sat the school bus, once the W3, on, gets recognized and immediately integrated, a representation of [DP+PP] should be formed, and immediately following the recognition of W4, an integration of W4 [VP] into this [DP+PP] representation will fail, causing disruption to eye movements.

To account for our current findings of no processing cost in the AT trials, the assumption of perfect incremental processing should be relaxed. In fact, while many studies have shown early, immediate disruption of reading by a syntactically or semantically incompatible word (e.g., Frazier & Rayner, 1982; Rayner, Warren, Juhasz, & Liversedge, 2004), there has not been direct evidence for perfect incremental integration: by comparing mean reading times on anomalous and plausible/grammatical words, those studies only demonstrated that, on average, difficulty arises as soon as anomaly is encountered, not necessarily on every trial. It could be the case that, while on most trials failure to integrate immediately interrupts the processing flow, causing the mean to inflate, on a small proportion of trials a recognized word is not
incrementally integrated, as evidenced in our two experiments using analyses contingent on response types.

We propose that (in contrast to what E-Z Reader 10 prescribes) a stalling integration of Word $N$ does not always lead to disruption on attention and eye movement controls. Instead, when Word $N+1$ gets recognized before integration of Word $N$, the parser can entertain different ways to integrate multiple words (e.g., as $W3+W4$ or as $W4+W3$). It is during this stage that inference also takes place. The reader’s prior linguistic/word knowledge has a weight in determining in what order these words should be integrated, together with the bottom-up input, whose strength of certainty is influenced by fixation status (see above). Integration thus involves inference, and if inference takes place this early and alters the literal input, there will be no disruption at the ungrammaticality point, $W4$.

In short, our newly proposed account is in spirit similar to both E-Z Reader (Reichle, Warren, et al., 2009), which assumes seriality of word recognition, and rational inference over a noisy channel (Gibson et al., 2013), which allows nonliteral interpretation of a sentence, but ours relaxes the assumption of perfect incremental integration and specifies the timing of inference. While high-level information is integrated quickly and incrementally most of the time (e.g., Christiansen & Chater, 2016), the timing of integration is variable and sometimes it lags behind recognition of the next word. On such an occasion where more than one word is available to be integrated, inference takes place to decide which underlying message is more likely given the perceptual evidence and prior linguistic knowledge. If the resulting representation is the grammatical one, no difficulty of processing emerges, as in the AT trials. This possibility suggests that, in contrast to Mirault et al. (2018), misreading of word order does not necessarily entail that words are lexically processed in parallel, although the no-cost pattern in AT trials would also be consistent with Mirault et al., who attribute the misreading to mis-encoding of word order at the perceptual level (i.e., the words are misperceived as constituting a grammatical sequence in the first place, hence no processing difficulty).
CHAPTER 4
EXPERIMENT 3

So far we have obtained three findings robust across the two experiments. First, there existed a transposed-word effect (more judgment/detection errors when compared to grammatical or filler sentences) whether the task was a speeded grammaticality judgement or a relatively natural reading-for-comprehension task. Second, direct fixations on the two transposed words were associated with more accurate responses (less failure to notice transposition). Third, when the participants failed to correctly dismiss a transposed-word sentence, their eye movements, indistinguishable from those in a grammatical counterpart sentence, did not reveal disruption. The last finding motivated our proposal of a delayed-integration version of serial-attention models that incorporates inference over a noisy-channel (henceforth the delayed-integration-and-inference account), which undermines the necessity of OB1-Reader to explain the phenomenon. On the other hand, OB1-Reader is partly supported by some results in the two experiments. These include the effect of word length difference in Experiment 1 and the no-cost pattern of eye movements in AT trials. The word class effect (see later Discussion) and the non-trivial number of fixating-both-word trials among the AT trials, however, cast doubt on the model. Experiment 3 thus aims at providing evidence that can potentially further adjudicate the two accounts. Specifically, we manipulated the point of ungrammaticality stemming from transposition. Recall that in the previous two experiments, all items were specially designed such that transposition of the two words elicited a signal of ungrammaticality only at the second transposed position (i.e., W4). Here we added new conditions consisting of sentences in which transposition causes ungrammaticality as early as the first transposed position (i.e., W3). Hypotheses and predictions are further elaborated in Section 4.2.

4.1. Methods

Seventy-nine participants took part in the experiment online via Amazon Mechanical
Turk, among whom 7 failed to upload both sessions and 2 fell too much for the transposition sentences (correctly identified transposition errors below 24 trials out of 60, significantly below chance level). Nine participants whose accuracy on grammatical and filler sentences was below 75% were further excluded. This leaves 61 valid participants for the analyses. All participants self-identified as native American English speakers.

The procedure was the same as in Experiment 2, except that data were collected remotely without eye movements being tracked. Participants were instructed to read naturally and answer one question after each study sentence (Staub et al., 2019). Due to the additional inclusion of a number of new items (introduced below), the whole experiment was divided into two sessions. The experiment stopped half-way, and the participants determined on their own when to begin the second session. The experiment took about 50-60 minutes.

Table 4 exhibits an example for each condition in Experiment 3. We preserve the material used in Experiment 1 and 2 which elicits ungrammaticality at W4 in the transposed form (henceforth the Uat4 condition). We add their Uat3 counterparts (e.g., in They let see us, see (W3) is not a grammatical continuation for They let). This leads to a 3 (word class) x 2 (ungrammaticality point) design.

<table>
<thead>
<tr>
<th></th>
<th>Uat3</th>
<th>Uat4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open-closed</td>
<td>They let see us the entire document.</td>
<td>I walk fat my dog every morning.</td>
</tr>
<tr>
<td>Closed-open</td>
<td>We should him put to bed soon.</td>
<td>The boy on sat the school bus.</td>
</tr>
<tr>
<td>Open-open</td>
<td>The awfully lamp dim can be discarded.</td>
<td>The fragile cup red shattered into pieces.</td>
</tr>
</tbody>
</table>

4 Note that while word classes are held the same within a Uat3 and Uat4 pair, the specific parts of speech are not identical (e.g., in open-closed, verb-pronoun in Uat3 but adjective-determiner in Uat4), except for the open-open sequences (i.e., noun-adjective for both Uat3 and Uat4).
Table 4. Example item in each condition in Experiment 3. Word class labels reflect the order of W3 and W4 in the transposed form. In Uat3, ungrammaticality emerges at Position 3.

4.2. Predictions

The current design allows testing of the sentence-level spatiotopic representation, which plays an essential role in keeping track of word order, in OB1-Reader (Snell et al., 2017; Snell, van Leipsig et al., 2018). The model posits that the representation stores syntactic information of both the already-recognized words and the to-be-recognized words (Snell & Grainger, 2017; Wen et al., 2019, 2020). Given some already-recognized words (e.g., The boy), expected syntactic categories for the next word(s) (e.g., a verb or a preposition for Position 3) will be generated and attached to the to-be-filled slots in the spatiotopic representation. This information provides top-down influence for the word nodes activated by the bottom-up bigram nodes. Activation of verbs and prepositions will be facilitated and thus more likely to win the activation competition to be recognized and appended as Word 3. Activation of nouns might even be inhibited by this top-down regulation. Such a mechanism therefore makes it possible for words in later positions to be recognized earlier than those in earlier positions if the syntactic category of those words fits the expectation of the to-be-filled position. For Uat4 items, since both W3 and W4 can be a legitimate candidate for Position 3 given the preceding two words, facilitation will be given to words of both syntactic categories; both words therefore compete to be recognized as a word at Position 3. For Uat3 items, however, only Word 4, but not Word 3, fits Position 3 syntactically. Top-down expectation from the spatiotopic representation therefore strongly favors Word 4. For instance, given We should, a verb is highly expected, and activation of word nodes of verbs will be boosted strongly, making put more likely to reach its recognition threshold than him. Therefore, it is predicted, under OB1-Reader, that Uat3 items should elicit more failure to notice transposition than Uat4 items.

The delayed-integration-and-inference account has more equivalent predictions for Uat3 sentences. The earlier section has formulated how delayed-integration can lead to failure to
notice word transpositions without eliciting disruption: when the integration of the first transposed word stalls long enough that the recognition of the second word finishes earlier and intercepts the integration process of the first transposed word. However, it is possible that such a scenario only occurs when the first transposed word is a grammatical continuation for its preceding context. Recall that under E-Z Reader 10, besides stalling integration, there is another way that integration fails: direct violation of syntax or semantics. Again, given that it is widely assumed that reading is highly incremental and that not many studies have directly inspected incremental processing contingent on response type, it is an empirical question how tolerant the parser is in the face of an outright grammatical error on a trial-to-trial basis (cf., on average, Frazier & Rayner, 1982; Rayner et al., 2004). On the one hand, if our parser is highly sensitive to ungrammaticality—a word that is unable to be integrated to the previous context always and immediately crashes the parse and causes interruption—we would expect that Uat3 sentences lead to little failure to notice a transposition error. This is different from our proposed scenario in Experiment 1 and 2 where integration of Word 3 does not fail, but simply finishes so late that recognition of Word 4 is completed earlier, interfering with its integration process (via inference). On the other hand, it is possible that an outright syntactic error does not cause the parser to crash immediately, but instead prolongs the integration stage of that word (which might reflect the parser’s effort to work on integrating the word despite its ungrammaticality). This will then lead to a similar ‘delayed integration’ situation described earlier, where Word 4 gets recognized early enough to be considered as a candidate to be integrated as a word at Position 3. If so, error rates should be similar between Uat3 and Uat4 sentences.

4.3. Analyses and results

Model construction and comparison follow the same steps as in Experiment 1 and 2. Since Experiment 3 is not an eye-tracking experiment, we will mainly focus on the detection accuracy data and supplement it with analyses of whole sentence reading time and response time to the detection question.
Participants performed well on the filler questions (mean accuracy = 96.1%; Minimum = 84.1%), suggesting all of them were paying good attention during the task. An initial model that had grammaticality as the only fixed effect showed that participants were more likely to incorrectly miss the transposition error (17.6%) than falsely reported an error for the grammatical sentence (7.3%, $z = 7.44, p < .001$). Table 5 shows the mean accuracy by grammaticality, word class, and ungrammaticality point. It is obvious that GOO4 items, as in Experiment 2, were not as highly acceptable as items from the other conditions.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Grammatical</th>
<th>Transposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open-closed, Uat3 (OC3)</td>
<td>91.9% (1.72)</td>
<td>77.8% (2.23)</td>
</tr>
<tr>
<td>Closed-open, Uat3 (CO3)</td>
<td>95.8% (1.08)</td>
<td>72.1% (2.98)</td>
</tr>
<tr>
<td>Open-open, Uat3 (OO3)</td>
<td>93.2% (1.47)</td>
<td>89.6% (1.86)</td>
</tr>
<tr>
<td>Open-closed, Uat4 (OC4)</td>
<td>94.8% (1.25)</td>
<td>83.9% (2.04)</td>
</tr>
<tr>
<td>Closed-open, Uat4 (CO4)</td>
<td>92.4% (1.63)</td>
<td>84.5% (2.35)</td>
</tr>
<tr>
<td>Open-open, Uat4 (OO4)</td>
<td>88.3% (1.85)</td>
<td>86.5% (1.94)</td>
</tr>
<tr>
<td>Filler, Comp. Question</td>
<td>96.1% (0.46)</td>
<td></td>
</tr>
</tbody>
</table>

Table 5. Mean accuracy for each condition in Experiment 3. Standard error (by subject) in parenthesis. Word class labels reflect the order of W3 and W4 in the transposed form.

Figure 7a further shows the histogram of participants’ error rates collapsing across all transposed-word conditions. As in the previous experiments, the error rates ranged widely among participants. To capture this individual difference, we correlated participants’ error-detection accuracy with their mean reading time for the fillers. The correlation was significantly positive ($r = 0.37, p < .01$). Slower readers detected the errors more easily (Fig 7b).
When looking at only the transposed conditions, it appears that Uat3 sentences elicited more errors than Uat4 for the open-closed (UOC3 vs. UOC4) and closed-open (UCO3 vs. UCO4) conditions, while there was only a small difference between UOO3 and UOO4 in the opposite direction. This was confirmed by a generalized linear mixed effect model with word class and ungrammaticality point and their interaction as fixed effects. Given that the interest here is ungrammaticality point, this effect was treatment-coded, with Uat3 as the reference level. The three-level word class effect was deviation-coded. The interactive model was favored over the additive model ($\chi^2(2) = 8.25, p = .01$). The significant interaction hinted that the direction of the ungrammaticality point effect was opposite for UOC and UCO and UOO: when data were separated into their subgroups (OC, CO, and OO), it was shown that accuracy in Uat4 was significantly higher than Uat3 for CO ($z = 2.87$), numerically higher for OC ($z = 1.44$), but marginally lower for OO ($z = -1.67$). The interaction term also suggested that there was an effect of word class in the Uat3 condition (error rate: UCO = UOC, $z = 1.27$; UCO < UOO, $z = 4.4$) but no difference among the word class conditions within the Uat4 level.
While we did not have incremental processing data on each word, we did the same analyses of reading/response time contingent on response type as in the previous two experiments. Given that the experiment was conducted online, participants might on some trials leave the question unanswered for a long time. Trials were first removed if their response times exceeded 15000ms (n = 39, out of 7053 trials). Table 6 presents the mean reading time of the whole sentence for each condition (3 response types x 6 sentence structures), after outliers were removed.

<table>
<thead>
<tr>
<th>Structure labels</th>
<th>AG</th>
<th>AT</th>
<th>DT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open-closed, Uat3 (OC3)</td>
<td>3034 (136)</td>
<td>3178 (334)</td>
<td>3628 (197)</td>
</tr>
<tr>
<td>Closed-open, Uat3 (CO3)</td>
<td>2808 (138)</td>
<td>2322 (117)</td>
<td>3403 (168)</td>
</tr>
<tr>
<td>Open-open, Uat3 (OO3)</td>
<td>3513 (180)</td>
<td>3732 (245)</td>
<td>3953 (178)</td>
</tr>
<tr>
<td>Open-closed, Uat4 (OC4)</td>
<td>3053 (156)</td>
<td>2905 (301)</td>
<td>3415 (168)</td>
</tr>
<tr>
<td>Closed-open, Uat4 (CO4)</td>
<td>3037 (164)</td>
<td>3141 (284)</td>
<td>3369 (181)</td>
</tr>
<tr>
<td>Open-open, Uat4 (OO4)</td>
<td>3540 (192)</td>
<td>3055 (234)</td>
<td>3810 (195)</td>
</tr>
</tbody>
</table>

Table 6. Mean reading time of the whole sentence by structure types and response types. Standard errors (by subject) in parenthesis. Structure labels reflect the order of W3 and W4 in the transposed form. AG: accepting grammatical sentences; AT: accepting transposed-word sentences; DT: dismissing transposed-word sentences.

Visual inspection first revealed that when participants correctly detected the transposition error, those (DT) trials had longer whole-sentence reading times, than when they did not detect the error (AT). Reading times in DT were also longer than reading times in their grammatical counterparts when no error was falsely reported (AG). When collapsing across the six types of sentences and comparing only among the response types, the mixed-effect model showed that

---

5 This cutoff is somewhat arbitrary. We first used a criterion of 3 standard deviation above the mean, yet that still left us some trials with response time higher than 30000ms, which do not seem like normal, long responses. We thus used 15000ms, which was the longest response time in Experiment 2.
the difference between AT (mean = 2748 ms) and DT (mean = 3646 ms) was significant (t = 4.86, p < .001), while we failed to find significant difference between AT and AG (t = 0.47).

Table 7 presents the mean response time to the detection question for each condition. Similar to the findings from Experiment 2, participants reacted a lot faster in the DT trials than in AT and AG trials, arguably reflecting that they were explicitly aware of the error even before the question prompted them. For the AT trials and AG trials, there was only a numeric trend for the response times in the former to be faster. When collapsing across the six types of sentences and comparing only among the response types, the mixed-effect model showed that the difference between AT (mean = 1929 ms) and DT (mean = 1354 ms) was significant (t = -8.18, p < .001). The difference between AT and AG was not significant (t = 0.2).

<table>
<thead>
<tr>
<th>Structure Type</th>
<th>AG</th>
<th>AT</th>
<th>DT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open-closed, Uat3 (OC3)</td>
<td>2237 (167)</td>
<td>1886 (111)</td>
<td>1364 (75)</td>
</tr>
<tr>
<td>Closed-open, Uat3 (CO3)</td>
<td>2104 (140)</td>
<td>2058 (158)</td>
<td>1341 (76)</td>
</tr>
<tr>
<td>Open-open, Uat3 (OO3)</td>
<td>2319 (156)</td>
<td>2138 (142)</td>
<td>1341 (64)</td>
</tr>
<tr>
<td>Open-closed, Uat4 (OC4)</td>
<td>2103 (113)</td>
<td>2052 (120)</td>
<td>1295 (67)</td>
</tr>
<tr>
<td>Closed-open, Uat4 (CO4)</td>
<td>2160 (124)</td>
<td>2036 (140)</td>
<td>1314 (66)</td>
</tr>
<tr>
<td>Open-open, Uat4 (OO4)</td>
<td>2379 (172)</td>
<td>1899 (119)</td>
<td>1353 (72)</td>
</tr>
</tbody>
</table>

Table 7. Mean response time to the detection question by structure types and response types in Experiment 3. Standard errors (by subject) in parenthesis. Structure labels reflect the order of W3 and W4 in the transposed form. AG: accepting grammatical sentences; AT: accepting transposed-word sentences; DT: dismissing transposed-word sentences.

4.4. Discussion

We set out to explore if word-transposition that results in ungrammaticality at the first transposed position (W3, Uat3) would lead to more failure to detect the transposition when compared with word-transposition that resulting in ungrammaticality at the second transposed
position (W4, Uat4). The results revealed a complex picture, with an interaction between ungrammaticality point and word class. Detection accuracy was significantly lower in Uat3 than in Uat4, for both the open-closed and closed-open items, but the effect was opposite for the open-open items. Reading time analysis comparing AG, AT, and DT trials showed a very similar pattern as in Experiment 2: there was no indication of disruption when participants read a transposed-word sentence but failed to explicitly report the error. Analysis of response time to the question showed that readers reacted much faster in the DT trials than in the other two types of trials, mirroring the reading time analysis.

We first discuss the similarity between Experiment 2 and 3. First, there is no disruption in the AT trials. The measurement in Experiment 3 was not as fine-grained as that in Experiment 2, yet shows a consistent result that readers were not slowed down by the transposition in those trials when they ultimately reported no error. Another consistent picture across Experiment 2 and 3 was the response time to the error-detection question. This measure again showed that participants behaved similarly between the AT and AG trials and differently between the AT and DT trials, even though it was in the AT and DT trials that the stimuli were the same. Remarkably, the absolute response times and their magnitude of difference were numerically similar as in Experiment 2 (about 800ms). Although the findings in Experiment 3 by themselves cannot be taken as strong evidence due to its lack of fine-grained precision, taken together with Experiments 1 and 2, they are consistent with the view that there was no perception of an error at any stage for those trials when participants failed to report the error, and suggest that inference must have taken place early enough to preclude the encoding/integrating of an ungrammatical representation.

In terms of the accuracy data, for the Uat4 items that were included in both experiments, accuracy rates were comparable between the two experiments, all of which were around 85%. In addition, the differences among the open-closed, closed-open, and open-open conditions were not significant in either experiment. One specific prediction under OB1-Reader
motivating the current design is that if failure to notice word transposition is due to parallel
word processing such that the second transposed word gets recognized earlier than the first and
 appended to the first position (W3), a relatively easy second transposed word should make this
scenario more likely. Here however, in the two experiments (as well as Experiment 1), when
the second word was an easy closed-class word following an open-class word, the error rate
was not higher than when the second word was an open-class word following a closed-class
word. These findings might compromise the parallel processing account as an explanation for
the transposed-word effect. However, as mentioned earlier, the limitation here is that the ease
of recognition was not directly manipulated by word frequency but by word class, which made
the conditions differ in not only ease of recognition but also in sentence structure. In addition,
the open-class words here, due to their short length, tended to be rather high in frequency as
well so the difference in ease between open-class and closed-class words was small. The lack
of a difference might thus due to low power.

Putting aside the (null) difference between the open-closed and closed-open conditions,
the comparison between the open-open condition and the other two conditions has resulted in
a complex picture across the three experiments and the previous study. While in Mirault et al.
(2018), open-open items elicited less failure of detection, this effect was significant but in the
opposite direction in Experiment 1. For the Uat4 items in Experiment 2 and 3 there was no
significant differences between open-open and either of the other conditions. For the Uat3 items
in Experiment 3, the effect was significant and in the same direction as Mirault et al. (2018).
One thing to first note is that we examined the effect of word class by directly comparing the
accuracy for the transposed-word items of each condition. Ideally, we should have looked at
the effect of word class on the difference in accuracy between grammatical and transposed-
word items for each condition. We did not do so, however, under the assumption that the items
we created, in their grammatical version, should be extremely likely to be accepted as
grammatical such that all those instances when participants incorrectly rejected the
grammatical items were just due to random noise. As shown in Tables 3 and 5, however, it was clear that the GOO4 items we adopted were not always judged to be grammatical. This could be because the sentences here contain two consecutive adjectives, which in written English sometimes requires a comma in between. The effect of word class might have been obscured by the low acceptability of those items in their grammatical form.

To take this factor into account, here we did a signal detection theory analysis, calculating simple d-prime (d’; Macmillan & Creelman, 1990), comparing sensitivity in each word-class condition for each experiment. The sensitivity measure yielded a consistent picture across the three experiments, with OC4 and CO4 being comparable in Exp 1 (2.69 & 2.56), Exp 2 (2.75 & 2.74), and Exp 3 (2.62 & 2.45), both higher than OO4 in each experiment (2.03 in Exp 1⁶, 2.11 in Exp 2, & 2.29 in Exp 3). This relationship held in the one-long condition in Experiment 1, (2.91 vs. 3.34 vs. 2.56). For Uat3 conditions, the direction of difference changed, with higher sensitivity in OO3 (2.75) than in OC3 (2.16) and in CO3 (2.31). Taken together, when ungrammaticality emerged at Word 4, readers were less sensitive to the correct vs. transposed distinction than in the open-closed and closed-open items in the three experiments, opposite to Mirault et al’s post-hoc observational relationship between word class and detection accuracy. The pattern was obscured when only looking at the raw accuracy in the transposed conditions, either because OO in Experiment 1 had a pronounced ‘yes’-bias or because OO in Experiment 2 and 3 had a pronounced ‘no’-bias.

Mirault et al. (2018) tentatively suggested that their post-hoc exploratory findings (i.e., having at least one closed-class transposed word elicited more detection errors) as support for parallel processing. That is, since closed-class words serve grammatical functions, their syntactic information is easier to be grasped quickly and parafoveally (Koriat & Greenberg, 1994), even before full recognition of the closed-class word itself. The syntactic information

---

⁶ When including only items that were also used in Experiment 2 (the subset analysis), the sensitivities are 2.98, 2.87, and 1.79 respectively.
from these words becomes an anchor that rapidly generates a sentence-level frame, which does not necessarily reflect the literal sentence frame, and further guides word recognition in a parallel, cascaded fashion (Snell & Grainger, 2017; Wen et al., 2019, 2020). It remains unclear, however, why this only held for the Uat3 items used in our study but not for the Uat4 items, which even showed an opposite-direction effect.

Returning to our effect of interest, ungrammaticality point, it is clear that Uat3 sequences elicited substantially more errors than their Uat4 counterparts, for the OC and CO conditions. As elaborated in Section 4.2, such a pattern is predicted by OB1-Reader: a sequence in which only Word 4 but not Word 3 fits in with the preceding context should inhibit activation of Word 3 and facilitate activation of Word 4 and so should make the encoding/appending of Word 4 at Position 3 more likely. A serial-attention model, on the other hand, has less definite predictions for this manipulation. The finding that Uat3 sentences led to more failure to detect an error can suggest that an outright grammatical error does not always and immediately cause integration failure. Instead, the parser is tolerant of an error for a certain amount of time (which might reflect the parser’s effort to work on integrating the word despite its ungrammaticality) that is long enough for the next word to be recognized and to provide an alternative ordering of integration that will not lead to perception of a grammatical error. The empirical finding that readers on twenty to thirty percent of the trials failed to detect a transposition error even when the transposition resulted in outright ungrammaticality at the first word further casts doubt on a perfect incremental syntactic parser.

Finally, one caveat must be noted before a definitive conclusion is reached. As pointed out in Footnote 4, the exact sequences of parts of speech between Uat3 and Uat4 were not matched. In CO3 and OC3 where error rates were the highest, it turned out that the frequency of Word 2 in these conditions was extremely high (mean Zipf frequencies = 6.26, 6.16 respectively; cf. 5.11, 4.86 in CO4 and OC4 correspondingly). A possibility is that an extremely
frequent, easy word at Word 2, together with another frequent Word 3, prompted an effortless process that does not engage analytical and deliberate high-level processing, thus making readers more likely to overlook a syntactic error (Alter et al., 2007). Post-hoc analyses indeed revealed that if the effect of continuous frequency of Word 2 (or averaging frequency of Word 2 and Word 3) was first entered in the model, adding the effect of ungrammaticality point, the effect of word class, or both of them did not effectively reduce the deviance (all ps > .1), while the frequency effect mostly remained even after the two effects and their interaction have been accounted for ($\chi^2(1) = 3.23, p = .07$). Such possibility awaits further investigation by carefully controlling for word frequency and sentence structures. The OO3 and OO4 items, on the other hand, were highly comparable; the transposed words in both conditions involved a noun and adjective, and the only difference between the two was the second word (adverb vs. adjective), whose frequency on average was similar. In this well-controlled contrast, the OO3 items elicited marginally less failure to detect errors than the OO4. The occurrence of failure in the OO3 condition was only 10.5%. A possibility to explain the lack of transposed-word effect is that the beginning of the sentences here are a determiner and an adverb. The only word that this context allows next is an adjective. Such a strong contextual constraint might make detection of a syntactic category error extremely easy and fast, and the parse will fail immediately before the fourth word can make its way to interfere the integration process. On the other hand, from the OB1-Reader viewpoint, the strong contextual constraint (a determiner plus an adverb) should in fact facilitate activation of an adjective, making an adjective easier to be recognized as a word at Position 3. This turned out not to be the case.
This study was motivated by the recent study by Mirault et al. (2018) which reported that readers sometimes failed to detect a word-transposition error in a sentence. This phenomenon has been taken by the group as support for parallel lexical processing (e.g., Snell & Grainger, 2017; 2019, Snell et al., 2017; Snell, van Leipsig, et al., 2018, Wen et al., 2019, 2020). Our first goal was to see how prevalent the phenomenon is in a rather natural reading setting. Across experiments with different tasks we found quite similar occurrence of failure at about 15%, showing the robustness of this phenomenon.

Our second goal is to manipulate visual, lexical, and syntactic properties of the transposed words to test specific predictions on detection rates under the newly developed model of word recognition and eye movements in text reading, OB1-Reader. We found that transposed words of similar lengths were less likely to be detected, that relative ease of recognition between the two transposed words did not have an effect, and that the direction of ungrammaticality point effect depended on sentence structures (word class combinations of the transposed words). These findings provide only partial support for OB1-Reader. We discuss these results in the context of the controversial debate between serial-attention and parallel-processing models in the literature in Section 5.2.

Our third goal, with the use of an eye-tracker, is to test the relationship between fixation pattern and error detection. We found that skipping at least one of the transposed words is associated with more failure of detection. We discuss the role of fixation status in word recognition, word identity representation, and word order representation in Section 5.1.

Finally, with eye-tracking, which provides a measure of incremental processing difficulty, we tested whether participants’ eye movements showed disruption by the ungrammaticality at the transposition regions when participants ultimately reported no error on that trial. We found that they did not. This finding is not consistent with the current view of incremental processing
and the view of post-perceptual rational inference. We discuss this in Section 5.3 in the context of other misreading phenomena.

5.1. The role of fixation status in sentence comprehension

Under E-Z Reader (Reichle, Warren, et al., 2009), the target of a forward saccade is always the next unrecognized word in the text. Attention to the currently fixated word (Word $N$) can be shifted to the next word ($N+1$) before the eyes move, and if the shifted attention finishes the L1 stage (partial recognition, or the familiarity check) of Word $N+1$ early enough, preparation of the saccade will be re-programmed to target Word $N+2$, hence skipping Word $N+1$. The remaining processing of Word $N+1$ (L2 and integration), even when Word $N+1$ is skipped, still proceeds as it would when not skipped. Importantly, the recognition process of any word only starts after the recognition (L2 stage) of its preceding word has finished. In short, while the skipping decision is based on only the completion of L1, strict seriality of word recognition is preserved regardless of fixation status. Our finding that skipping influenced detection of errors, on the other hand, could suggest that order of recognition changes as readers skip words. Mirault et al. (2020) tested this hypothesis even more specifically by comparing the occurrence of failure to detect transposition between those trials when participants fixated the two transposed sequentially and those trials when participants first fixated the second transposed word and then regressed to the first transposed word. They tested an account which maintains strict serial word recognition but posits that the order of recognition is determined by the order of fixation. Such an account also should be able to explain the misreading phenomenon. However, they found that while readers made fewer errors in the in-order trials (14.1%) than in the out-of-order trials (17.3%), the error rate in the in-order trials was still nonnegligible. This suggested that order of recognition does not depend, or only minimally depends, on order of fixation. Indeed, previous studies by Gordon and colleagues have also provided evidence consistent with the view that even a skipped word is processed to a substantial extent: while word repetition and word frequency by themselves have an influence on skipping rates when
the preview is valid, the effects disappear when the preview is a nonword even if the nonword is orthographically similar to the target (Choi & Gordon, 2013; Gordon et al., 2013); whether or not orthographic or phonological similarity between the invalid preview and the target has an effect on skipping rates also depends on the lexicality of the invalid preview (Choi & Gordon, 2014).

Taken together, we do not believe that skipping influenced the detection rate by influencing the order of word recognition. Instead, we propose that what skipping influences is the strength of the reader’s belief in what has been read. Recognition of two words within one fixation event, rather than two separate fixation events, might make the representation of their relative order less definitive. When the bottom-up perceptual evidence is more equivocal, top-down knowledge weighs more and inference is more likely to override the literal input (Norris, 2006), even retrospectively (Futrell et al., 2020; Levy, 2011). Other evidence came from Staub et al. (2019) where failure to detect repetition of the, rather than transposition, was investigated. Fixating both the repeated thes was associated with increasing detection of the repetition error. Similarly, detection of omission of a the was also more likely when the two words that surrounded the omitted the were both fixated. Their findings can also be explained by the above hypothesis. For instance, in the skipping trials, while the two thes were still both veridically encoded, the representation of two thes was weaker due to the lack of direct, separate fixations, thus more amenable to inference into a nonliteral interpretation.

5.2. Serial or parallel processing

One theoretical motivation of this study is to provide further evidence that can potentially help adjudicate the long-lasting debate between serial and parallel word processing. The strongest version of a serial-attention model will not predict failure to notice word transposition since words are strictly recognized one at a time from left to right and word order representation
directly follows the order of recognition (Reichle, Liversedge, et al., 2009). Mirault et al. (2018) therefore took the finding of the transposed-word effect to undermine serial-attention models.

Other evidence for parallel processing came from the sentence-superiority effect (Snell & Grainger, 2017). Readers’ recognition of a word in a 4-word sentence that was rapidly presented for 200ms was higher when the whole sentence was grammatical than when its order was scrambled. This sentence-superiority effect was independent of the position of the to-be-reported-word, casting doubts on left-to-right seriality of word processing. Snell and Grainger (2017) argued that the better recognition rate of words arises because readers quickly generate a sentence-level syntactic frame from the grammatical sentence without fully recognizing all four words. The syntactic information guides recognition of the four words, facilitating correct identification of them. The fully scrambled sentence, however, does not allow a rapid construction of the sentence-level representation. Electrophysiological studies (Wen et al., 2019, 2020) showed that the sentence-superiority effect was associated with a smaller N400, which was interpreted as facilitation in the recognition process, echoing the view that quickly generated sentence-level syntactic frame guides word recognition (see also Pegado & Grainger, 2019; Snell et al., 2017).

On the other hand, one critical finding that has consistently been reported in many studies has been taken as strong evidence against parallel word processing. This is the lack of high-level parafoveal-on-foveal effect. If words are processed in parallel, the lexical variables of the word(s) in the parafovea should affect the ongoing processing of the fixated word. Such parafoveal-on-foveal effects only have been demonstrated either as a low-level effect (e.g., affected by preview of a repeated word, Angele, Tran, & Rayner, 2013) or in correlational, corpus studies (Kennedy & Pynte, 2005). When factors were more carefully manipulated in

---

7 We have shown, however, how serial-attention models, when working together with rational inference, can explain the transposed-word effect, see Section 5.3 for more.
experimental studies, no high-level parafoveal-on-foveal effects have been robustly found: in a series of experiments in Brothers et al. (2017), while manipulation of frequency had a clear effect on the target word, the first-pass reading of the pre-target word did not differ whether the word to its right as a preview (the target word) was high- or low-frequency. Using the boundary-change paradigm, Veldre and Andrews (2018) also found that invalid implausible or ungrammatical preview only had an effect on the target word itself but not the ongoing processing of the pre-target word (see also Snell, Deckerck, & Grainger, 2018, for no semantic parafoveal-on-foveal effects). In light of the elusive parafoveal-on-foveal effects, Snell and colleagues (Snell, Meeter, & Grainger, 2017; Snell, van Leipsig, et al., 2018; Snell & Grainger, 2019) proposed that a spatiotopic sentence representation in working memory enables readers to keep track of which information belongs to which word rather than integrating information from every word into a single mixture and thus prevents cross-word leakage of high-level information. This spatiotopic representation is thus an important feature in their model, and the current study tested some predictions it has in relation to the transposed-word effect.

Snell and colleagues’ account, as formally described in OB1-Reader model, has three fundamental assumptions: (1) association between word identities and locations is noisy, (2) words within the perceptual span are processed in parallel, and (3) a sentence-level spatiotopic representation in working memory keeps track of word order and provides top-down syntactic influence guiding on-going word recognition. While positional information is assumed to be noisy, the appending of an activated word node to a to-be-filled slot in the spatiotopic representation is constrained by their visual similarity: only transposed words of similar lengths should be read wrong. In Experiment 1 we tested this hypothesis. While both-short items elicited more errors than one-long items, the latter, when compared with their own grammatical counterparts, caused significantly more errors, still manifesting a transposed-word effect. These findings thus did not fully support OB1-Reader.

The second assumption of parallel word processing is tested here by manipulating the
relative ease of recognizing the first and second transposed words (Reichle, Liversedge, et al., 2009; Snell, van Leipsig, et al., 2018). If misreading of word order is due to parallel processing that allows the second transposed word to be recognized before the first transposed word, then the easier the second transposed word is the more likely misreading occurs. Across all three experiments, we saw no evidence that confirmed this prediction. When the second word was closed-class (the easiest kind of words in English) following an open-class word, the error rates were not higher than the other way around. Admittedly, our design did not maximize the frequency difference between open-class and closed-class words, potentially making the null effect due simply to a lack of power. However, for the Uat3 condition in Experiment 3, the direction of the difference was even numerically opposite to the prediction, which could not be attributed to lack of power. Finally, the sentence structures also differed among different ease conditions, adding another potential confound. We have recently conducted an experiment not formally reported here which varied frequency of the two transposed words within a fixed sentence structure. The frequency difference between conditions was extremely large (mean Zipf around 5 vs. 3, e.g., A painfully sound eerie came from the woods vs. The completely grits cold could not be eaten). With such large difference in ease of recognition, the error rates in the condition where a high-frequency second-transposed word followed a low-frequency first-transposed word turned out eliciting fewest errors. These results thus provide further evidence against the prediction that a word in a later position can be recognized before one in an earlier position.

The final assumption about the spatiotopic representation was tested by using transposition errors that introduced ungrammaticality at different positions (Ungrammaticality at Word 3 vs. 4). While a transposition error introduces ungrammaticality globally, sometimes it does not do so locally at the first transposed position. In this case, the syntactic expectation generated by the preceding two words will facilitate activation of both Word 3 and Word 4 at Position 3 (although the strength for each might differ). By contrast, a transposition error that
introduces ungrammaticality at the first transposed word means that the given context will not facilitate, but inhibit, Word 3 at Position 3 and facilitate Word 4 at Position 3, and therefore misreading should be more likely for Uat3 than for Uat4 items. In Experiment 3, we observed an interesting result in which this effect depended on sentence structures. While open-closed and closed-open sentences clearly showed evidence for OB1-Reader’s prediction, this effect could also have been explained by the frequency difference in Word 2. In a better-controlled contrast where both Word 2 frequency and sentence structures were similar, UOO3 vs. UOO4, it was the latter items that made participants err marginally more, opposite to OB1-Reader’s prediction. The results thus lend no strong support to the spatiotopic representation.

Overall, then, we believe the current version of OB1-Reader, or parallel processing models in general, falls short of explaining the results we have obtained. It is worth noting that some recent behavioral and neuronal data also demonstrated strong evidence against recognizing two words at the same time (White, Palmer, & Boynton, 2018; 2019; White et al., 2019). In their paradigms, participants performed either a lexical judgment, semantic categorization, or color detection task in either a single-task or dual-task condition. Two stimuli were presented simultaneously for about 50ms, preceded and followed by masks. The interval between the masks and the targets was adjusted for each participant to a threshold with which they could attain 80% accuracy in the single-task condition. Single- and dual-task conditions differed in whether there was a pre-cue signaling which side of the stimuli to attend to. The behavioral results showed that while participants’ accuracy was not lower in the dual-task condition when the task was to detect the color of the target, their accuracy decreased significantly in the dual lexical/semantic decision tasks. In these two dual-task conditions, their performance aligned more with the prediction under an all-or-none serial attention model than with the fixed-capacity or unlimited-capacity parallel model in the attention operating characteristic analysis. Furthermore, in an fMRI experiment using the same paradigm (White et al., 2019), it was found that while the BOLD responses in the anterior part of the left ventral
occipitotemporal cortex, termed VWFA-2, were not influenced by selective attention, they were influenced by lexical frequency of the attended word. This supports the hypothesis that there is a bottleneck preceding left VWFA-2 that prevents two words from entering this region through two separate parallel channels. These studies therefore reject the possibility that two words can be recognized simultaneously. Such findings are in line with earlier studies that employed a disappearing text paradigm, where processing remained normal when the word in the fovea disappeared 60ms after it was first fixated while processing became disrupted when the word right to the currently fixated word disappeared 60ms after the foveal word was fixated (Rayner, Liversedge, & White, 2006). Readers were able to visually encode the currently fixated word within 60ms and lexically and linguistically process it without visual input, yet they could not do the same thing simultaneously for the following word.

Finally, another experiment we conducted used the both-short items and fillers in Experiment 1 in a RSVP paradigm, where sentences were presented one word at a time every 250ms. We found that participants failed to report a transposition error about 20% of the time. These preliminary data indicate that even when the word is strictly serially encoded, misreading can still occur, arguing against a necessary role of parallel lexical processing in explaining failure to notice word transpositions.

**5.3. On incremental integration, rational inference, and other misreading phenomena**

Our second theoretical motivation is to see whether failure to notice word transposition, a phenomenon of nonliteral reading, can be explained by rational inference over a noisy channel (Gibson, 2013; Levy, 2008; Levy et al., 2009; Ryskin et al., 2018). Since the account emphasizes inference, whose output does not have to match the bottom-up input, and assumes noise can come from various sources (i.e., the producer, the perceiver, and the environment), in principle, it can explain why readers overlook any kind of errors as long as the literal input does not deviate substantially from the intended message, and these include transposition errors. Less known in the literature, however, is the timing of inference. It is currently assumed that
the error in the input is first perceived and then corrected/edited (e.g., Ryskin et al., 2020, a N400 followed by a P600). The common practice to look at processing difficulty in previous studies (Frazier & Rayner, 1982, Rayner et al., 2004; Ryskin et al., 2020) however, is to compare grammatical and ungrammatical/implausible sentences on average. Thanks to the error-detection questions in our design, we were able to separate trials of literal and nonliteral reading. By comparing the two, and with the use of eye-tracking, we could obtain temporal information about the inference process. Ryskin et al. (2020) argued that the inference process takes place fast enough to be captured online (cf. a deliberate process, Ryskin et al., 2018). However, as mentioned earlier, it is late enough for the literal grammatical error to be first perceived. Our results, on the other hand, revealed no disruption at all across the board on the two critical transposed regions, therefore suggesting that inference must take place before or during the integration of the ungrammatical word. This no-cost finding also poses problems to the assumption of incremental integration (Christianson & Chater, 2016; Reichle, Warren, et al. 2009). We therefore proposed a possibility of delayed integration, allowing high-level integration to occasionally even fall behind the recognition point of the next word. The newly recognized word can weigh in in the integration process. The decision of how to integrate these words involves inference and if the grammatical parse wins over the literal parse, no error detection ensues. Together with rapid inference and delayed integration, a serial-attention model like E-Z Reader is in principle able to explain misreading of transposed words.

To extend our current study, we looked back into the data in Staub et al. (2019), whose paradigm we followed in almost the exact same way, but which used repetition of the. When analyzing the eye movements contingent on response types (AG, AT, and DT), we found that participants’ eye movements were not disrupted at all for those trials when their ultimate response was “no error”. This supports the delayed-integration account: if the second the gets recognized quickly enough before the integration of the first the, no difficulty should emerge.

Still another finding that might be surprisingly incompatible with the perfect
incrementality assumption is the high error rates for the Uat3 items in Experiment 3. Even when the third word is outright ungrammatical, impossible to be integrated into the context, the participants failed to detect it on many trials (20–30%), which further suggests imperfectly-incremental integration. Future models of reading comprehension should thus take these findings into account.

While delayed integration seems compatible with our current findings about failure to notice transposition and repetition errors, it remains a question whether this holds for other misreading phenomena documented in the literature. One key feature of transposition and repetition is that the distance between the literal string and the intended string is close. Consider other examples, such as *The dog was bitten by the man* (Ferreira, 2003), it is implausible that the very first NP has not been integrated till the end of the sentence *the man*. Similarly, with the sentence *The mother gave the bottle the daughter*, *bottle* is slightly farther from *daughter* than two adjacent transposed/repeated words. Whether or not these phenomena arise because of offline inference (Ryskin et al., 2018) can be tested using the same paradigm as in our study.

Finally, it has to be noted, while the delayed-integration account is in theory able to explain the transposed-word effect itself, how it explains the effects of word length, word class, and ungrammaticality point is less clear. At face value, the delayed-integration account is not entirely compatible with our findings. For instance, if misreading of transposition occurs when the second word gets recognized quickly enough before the first word gets integrated, the absolute ease of the second word should influence the detection rate (i.e., the easier the latter word, the more quickly it gets recognized and the more likely it intercepts the integration process), which was not borne out here. It could be the case that the probabilities of delayed integration for W3 differed across word classes, which confounded the results. Similarly, in

---

8 It will be interesting to see whether for these items there is also no incremental processing difficulty at Word 3. This will require an eye-tracking setting. Our analyses on the whole-sentence reading time and response time, however, did suggest this to be the case (i.e., no cost for the AT trials).
Experiment 3 we saw the Uat3 items elicited more errors than the Uat4 items (arguably because of the immediate failure to integrate W3), but only for the open-open condition. Like OB1-Reader, the delayed integration account also falls short to explain the interaction pattern in Experiment 3. Future studies with even better control and with specific predictions by the delayed-integration account are needed to verify and refine this theory.
CHAPTER 6
CONCLUSIONS

Summarizing our study, we furthered the investigation of the recently reported phenomenon, failure to notice transposition errors, the results of which provide us more insight into reading comprehension processes. Our results first showed robustness of the phenomenon, which occurs in both grammaticality judgment and reading for comprehension tasks. Second, we found that transposed words of similar lengths were more likely to be read wrong, although those of two strikingly different lengths also got read wrong to a certain extent. Third, there was no difference across experiments between transposed words consisting of an open-class word followed by a closed-class word and those consisting of a closed-class word and open-class word. Having at least one closed-class word in the transposed sequence and having both open-class words, however, showed a difference, although the direction depended on whether the transposition introduced ungrammaticality at the first or second word. Similarly, the ungrammaticality point had opposite effects, depending on the word-class combination of the two transposed words. Fourth, directly fixating both transposed words in first-pass reading was associated with more error detection. Finally, for those trials when participants failed to explicitly report the error, their eye movements did not show disruption on the problematic regions, as if readers had been reading a grammatical sentence.

We are most confident in the last finding, based on which we argue that rational inference can fit with the results. However, we maintain that inference must take place during the integration process itself, not after. We also highlight that the no-cost pattern does not align with a perfect incremental integration view. Instead, we propose that the integration of a recognized word occasionally lags behind the recognition of the next word. Our study thus furthers our understanding of incremental language processing and rational inference over a noisy channel.

All results considered, we believe that the newly developed model, OB1-Reader, is only
partly supported by our data; hence our doubts of parallel word processing remain. On the other hand, our current proposal of serial word recognition with delayed-integration also might not be entirely consistent with the results. Both accounts require refining in light of these challenges.
REFERENCES


Staub, A., & Goddard, K. (2019). The role of preview validity in predictability and frequency


