The Roles of Thematic Knowledge in Sentence Comprehension

Kazunaga Matsuki
The University of Western Ontario

Supervisor
Ken McRae
The University of Western Ontario

Graduate Program in Psychology

A thesis submitted in partial fulfillment of the requirements for the degree in Doctor of Philosophy

© Kazunaga Matsuki 2013

Follow this and additional works at: http://ir.lib.uwo.ca/etd

Part of the Cognitive Psychology Commons

Recommended Citation
http://ir.lib.uwo.ca/etd/1661

This Dissertation/Thesis is brought to you for free and open access by Scholarship@Western. It has been accepted for inclusion in Electronic Thesis and Dissertation Repository by an authorized administrator of Scholarship@Western. For more information, please contact tadam@uwo.ca.
The Roles of Thematic Knowledge in Sentence Comprehension

(Thesis format: Integrated Article)

by

Kazunaga Matsuki

Graduate Program in Psychology

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

The School of Graduate and Postdoctoral Studies
The University of Western Ontario
London, Ontario, Canada

© Kazunaga Matsuki 2013
Abstract

People possess a great deal of knowledge concerning what commonly happens in various types of events. This knowledge, specifically with respect to event participants and their relations within an event (thematic knowledge), is an important component of how people understand language. A number of studies have shown a rapid influence of thematic knowledge during moment-to-moment sentence processing, in both investigations of lexical processing in sentential contexts, and temporary syntactic ambiguity resolution. The main goal of this dissertation is to further our understanding about the roles of thematic knowledge during sentence processing and sentence understanding by examining two critical unresolved issues. Chapter 2 investigated whether manipulation of thematic knowledge can lead to processing disruption in sentences that are otherwise assumed to be free of processing difficulty. That is, I investigated whether simple sentences can be made more difficult. This issue is particularly important for adjudicating among two major theories of sentence comprehension, two-stage and constraint-based theories. I found that main clause sentences could be made more difficult by manipulating thematic fit of the initial noun phrase, as in *The host invited* versus *The guest invited*. Because the influence of thematic fit was found at the earliest point at which it could be expected, the results strongly support constraint-based models. Chapter 3 investigated how thematic knowledge affects the construction of sentential meaning representations, and how misinterpretations can occur during that process. Specifically, the study evaluated several possibilities regarding how misanalyses of thematic roles might occur in full passive sentences that varied in plausibility. Participants’ understanding was determined by asking them to recall the agent or patient of each target sentence. The novel aspects of this study involved in-depth analyses of the types of errors that participants make, and using ERPs to investigate on-line processing differences. A major result was that the N400 ERP component was smaller for trials on which participants made an error versus when they responded correctly, indicating that errors occurred when readers were not sensitive to thematic implausibility. In summary, the studies reported in this dissertation provide novel and important theoretical insights into thematic role processing during sentence comprehension.
Keywords

Sentence Comprehension, Event Knowledge, Thematic Roles, Constraint-based Models, Good-enough Processing, Eye-tracking, ERPs.
Co-Authorship Statement

All projects in this dissertation were conducted under the supervision of my advisor, Dr. Ken McRae.

Chapter 2 has been written in preparation for publication, authored by Kazunaga Matsuki, Lindsay Donaghy, Michael J. Spivey, Michael K. Tanenhaus, and Ken McRae. Lindsay Donaghy assisted in collection of the data for the norming studies in Chapter 2. I designed, implemented, analyzed, and interpreted the eye-tracking experiment. Michael Spivey and Michael Tanenhaus participated in the initial conceptual development of the studies.

Chapter 3 is entirely my own work. It has been written in preparation for publication, authored by Kazunaga Matsuki and Ken McRae. I conceptualized, designed, implemented, analyzed, and interpreted all of the research presented in Chapter 3.

The written material in this thesis is my own work, which was edited by my advisor, Dr. Ken McRae.
Acknowledgements

First and foremost, I would like to thank my supervisor Ken McRae for all his support, guidance, and patience in carrying out my dissertation work, as well as for his generosity and kindness in and outside of the laboratory. My great appreciation also goes to the members of my supervisory and examining committees, Marc Joanisse, John Paul Minda, Albert Katz, J. B. Orange, and Craig Chambers for their thoughtful questions and suggestions on the final draft. I am grateful to my colleagues, Lisa King, Peter Nguyen, and Kelsey Santerre for their helpful comments and to Layla Amer for her help in transcribing and coding the hours of audio recordings. Finally, I thank Eriko Ando and my family for their care and support.
# Table of Contents

Abstract ............................................................................................................................... ii

Co-Authorship Statement ................................................................................................. iv

Acknowledgements ........................................................................................................... v

Table of Contents ............................................................................................................. vi

List of Tables ..................................................................................................................... ix

List of Figures ..................................................................................................................... x

List of Appendices ........................................................................................................... xii

Chapter 1 ........................................................................................................................... 1

1 General Introduction ....................................................................................................... 1

1.1 Event Knowledge and lexical processing ................................................................. 1

1.2 Event Knowledge and Syntactic Ambiguity ............................................................. 4

1.3 The Role of Event Knowledge ............................................................................... 6

1.4 Overview of present research ................................................................................ 7

1.5 References ............................................................................................................... 10

Chapter 2 .......................................................................................................................... 16

2 Thematic Fit Induces Competition Effects in Sentences that are Resolved as a Main
Clause ............................................................................................................................... 16

2.1 Introduction .............................................................................................................. 16

2.1.1 Making Simple Sentences Harder .................................................................... 21

2.1.2 Biasing Main Clauses toward Sentence Complements ..................................... 22

2.1.3 Biasing Main Clauses toward Reduced Relatives ............................................. 23

2.1.4 The Present Research ...................................................................................... 25

2.2 Norming Study 1a ................................................................................................... 26

2.2.1 Method ............................................................................................................... 26
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.2.2</td>
<td>Results and Discussion</td>
<td>28</td>
</tr>
<tr>
<td>2.3</td>
<td>Norming Study 1b</td>
<td>30</td>
</tr>
<tr>
<td>2.3.1</td>
<td>Methods</td>
<td>30</td>
</tr>
<tr>
<td>2.3.2</td>
<td>Results and Discussion</td>
<td>30</td>
</tr>
<tr>
<td>2.4</td>
<td>Norming Study 2a</td>
<td>31</td>
</tr>
<tr>
<td>2.4.1</td>
<td>Methods</td>
<td>33</td>
</tr>
<tr>
<td>2.4.2</td>
<td>Results</td>
<td>33</td>
</tr>
<tr>
<td>2.5</td>
<td>Norming Study 2b</td>
<td>33</td>
</tr>
<tr>
<td>2.5.1</td>
<td>Method</td>
<td>33</td>
</tr>
<tr>
<td>2.5.2</td>
<td>Results and Discussion</td>
<td>35</td>
</tr>
<tr>
<td>2.6</td>
<td>Experiment</td>
<td>37</td>
</tr>
<tr>
<td>2.6.1</td>
<td>Method</td>
<td>38</td>
</tr>
<tr>
<td>2.6.2</td>
<td>Results and Discussion</td>
<td>41</td>
</tr>
<tr>
<td>2.7</td>
<td>General Discussion</td>
<td>45</td>
</tr>
<tr>
<td>2.8</td>
<td>References</td>
<td>47</td>
</tr>
<tr>
<td>Chapter 3</td>
<td>Inducing Misinterpretation of Simple Passive Sentences</td>
<td>52</td>
</tr>
<tr>
<td>3.1</td>
<td>Introduction</td>
<td>52</td>
</tr>
<tr>
<td>3.1.1</td>
<td>“Good Enough” Sentence Processing</td>
<td>53</td>
</tr>
<tr>
<td>3.1.2</td>
<td>Temporal Coordination of Heuristic-based and Syntax-based Processing</td>
<td>57</td>
</tr>
<tr>
<td>3.1.3</td>
<td>The Locus of Misinterpretation</td>
<td>59</td>
</tr>
<tr>
<td>3.1.4</td>
<td>The Present Study</td>
<td>65</td>
</tr>
<tr>
<td>3.2</td>
<td>Methods</td>
<td>67</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Participants</td>
<td>67</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Materials</td>
<td>68</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Procedure</td>
<td>69</td>
</tr>
</tbody>
</table>
List of Tables

Table 2-1. Means (and ranges) of the agenthood and patienthood ratings for Binder et al.’s (2001) Experiment 1 balanced and biased items. Means (and ranges) of McRae et al.’s (1998) good patient and good agent items are also presented for comparison................................. 29

Table 2-2. Mean completion percentages for sentence fragments ending at verb and at the determiner (or her/his) for Binder et al.’s (2001) Experiment 1 items. Means as reported in Binder et al. (2001) are also presented for comparison................................................................. 32

Table 2-3. Means (and ranges) of the agenthood and patienthood ratings for the good agent and good patient items selected for Experiment 1. .............................................................................. 34

Table 2-4. Mean (and standard deviation) of sentence fragment completion percentages obtained at the verb and at the prepositional phrase (PP) for the 20 selected target sentences used in Experiment 1. .............................................................................................................. 36

Table 2-5. Mean first-pass reading times (in ms), first-pass regression proportions, and total reading times (in ms) by condition, for each region of the sentences................................................................. 43

Table 3-1. Percent accuracy for each sentence type in each question condition .................. 76

Table 3-2. The fixed effect statistics for the final model for overall accuracy, including the estimated coefficient, standard error, Wald Z-statistics, and p-values. ............................................ 78

Table 3-3. Mean confidence ratings for each condition, separated by correct versus incorrect response and error type. ................................................................................................................ 87

Table 3-4. The t-values and their statistical significance for each contrast in each ROI. . 97
List of Figures

Figure 3-1. Electrode montage. Each group of electrodes connected by lines formed a region of interest. ............................................................ 72

Figure 3-2. Mean accuracy for each sentence type and question type as a function of time. The upper panel is based on observed accuracy for each block and the lower panel is based on the model’s predicted probability for varying trial positions. ................................. 80

Figure 3-3. Mean accuracy and predicted probability for each sentence type as a function of time. The upper panel is based on observed accuracy for each block and the lower panel is based on the model’s predicted probability for varying trial positions. ................................. 81

Figure 3-4. Mean accuracy and predicted probability for each N2 plausibility level for question type as a function of time. ........................................................ 82

Figure 3-5. Mean frequency count per block for each error type. .................................................. 86

Figure 3-6. Mean accuracy for each level of inattentiveness rating. .............................................. 90

Figure 3-7. Grand average ERP waveforms for the local plausibility contrast comparing the P+A- to the P+A+ sentences (e.g., The patient was treated by the client vs. doctor). For each ROI, ERPs for sets of electrodes are averaged. The x-axis represents the time from -200 ms pre-stimulus onset to 800 ms post-stimulus onset, ticked every 100 ms. ........................................ 92

Figure 3-8. Grand average ERP waveforms for the distant plausibility contrast comparing P-A+ and P+A+ sentences (e.g., The clinician vs. patient was treated by the doctor). For each ROI, ERPs for sets of electrodes are averaged. The x-axis represents the time from -200 ms pre-stimulus onset to 800 ms post-stimulus onset, ticked every 100 ms. ........................................ 93

Figure 3-9. Grand average ERP waveforms for the global plausibility contrast comparing the P-A- to the P+A+ sentences (The clinician was treated by the client vs. The patient was treated by the doctor). For each ROI, ERPs for sets of electrodes are averaged. The x-axis represents the time from -200 ms pre-stimulus onset to 800 ms post-stimulus onset, ticked every 100 ms. .................................................................................. 94
Figure 3-10. Voltage maps across each time window for each plausibility contrast showing difference in ERPs to each sentence type (local = P+A-, distant = P-A+, and global = P-A-) relative to ERPs to the fully plausible sentences (P+A+). Scale is -2μV (blue) to +2μV(yellow).

Figure 3-11. Grand average ERP waveforms for plausible-correct condition and implausible-correct condition.

Figure 3-12. Grand average ERP waveforms for implausible-correct condition and implausible-incorrect condition.

Figure 3-13. Grand average ERP waveforms for plausible-correct condition and implausible-incorrect condition.
List of Appendices

Appendix A: Sentences used in Experiment of Chapter 2.................................................. 121

Appendix B: Verbs and noun phrases used in Experiment of Chapter 3.............................. 124

Appendix C: Ethics Approval................................................................................................. 127

Appendix D: Ethics Approval................................................................................................. 128
Chapter 1

1 General Introduction

One important step in successful comprehension of single sentences is to establish thematic roles and relations, that is, basically who did what to whom. One way to do so is to rely on syntactic information. Syntax determines that “a boy kissed a girl” has a meaning distinct from “a girl kissed a boy,” and that “a girl boy a kissed” is agrammatical. Another way is to infer thematic relations from event-based world knowledge (i.e., knowledge about how an event in the world commonly takes place; henceforth event knowledge). As language is often used by people to express what they are doing, what they have done in the past, and what they are planning to do in the future, it is necessary to bring to bear event knowledge to communicate with one another in a sensible and sufficiently conventionalized way. Although there are no disputes that both syntactic and event knowledge are important for single sentence comprehension, there have been intense debates regarding how and when the two streams of information are used, how they interact, and how they are integrated.

1.1 Event Knowledge and lexical processing

Some theories of sentence processing make a clear distinction between linguistically relevant knowledge, which is encoded in the lexicon, and event knowledge, which is represented outside of the linguistic system (Bornkessel & Schlesewsky, 2006; Chomsky, 1975; Katz, 1972; Schlesinger, 1995; Sperber & Wilson, 1986; Warren and McConnell 2007). That is, event knowledge is assumed to be part of the cognitive knowledge system, and not part of lexical/linguistically relevant knowledge. Given this architecture, when
people encounter a word, activation of linguistically relevant knowledge is privileged and automatic, whereas activation of event knowledge is not direct, and is subject to a time delay. An alternative view claims that event-based world knowledge is activated and used immediately during on-line sentence comprehension without a built-in delay (see McRae & Matsuki, 2009, for a review). According to the latter view, when reading or listening to a sentence, event knowledge not only becomes incorporated immediately when constructing an interpretation, but it also influences the ways in which readers or listeners process words and sentences.

The two views make contrasting predictions particularly regarding how people process words or phrases within a sentence. For example, according to accounts that assume delayed access of event knowledge, the word *carrot* in *John used a knife to chop a large carrot* versus *John used an axe to chop a large carrot* is not processed differently, at least not initially. No initial differences exist because knowledge about whether a *carrot* is a common thing to be cut with a *knife* versus with an *axe* is not part of linguistically relevant knowledge. In contrast, the alternative account suggests that *carrot* in the above example should be processed faster following *a knife to chop* than *an axe to chop* because knowledge activated from combining the instrument and the verb produces different expectations for the upcoming word *carrot*. While several studies have shown evidence in support of the former account (Rayner, Warren, Juhasz, & Liversedge, 2004; Warren & McConnell, 2007), others have shown the immediate use of event knowledge, supporting the latter account (Matsuki, et al., 2011; see also Bicknell et al., 2010; Kamide, Altmann, & Haywood, 2003; Morris, 1994).
The notion that people use their event knowledge to generate expectations for upcoming concepts in the linguistic stream has been examined extensively by measuring event-related potentials (ERPs) during reading or listening (Camblin et al., 2007; Federmeier & Kutas 1999; Hagoort, Hald, Bastiaansen, & Petersson, 2004; Otten & Van Berkum 2007; Van Berkum et al., 2005). For example, DeLong, Urbach, and Kutas (2005) demonstrated that anticipation of upcoming words occurs during reading of sentences such as *The day was breezy so the boy went outside to fly a kite/an airplane*. In their study, the target words that were less predictable from contexts (*airplane*) elicited a larger N400 compared to highly predictable words (*kite*). The N400 ERP component is a negativity that peaks approximately 400 ms after a word’s onset, and is considered to generally reflect semantic processing. Although contextual predictability effects on the N400 have been shown by others as well (Federmeier & Kutas, 1999; Federmeier, McLennan, De Ochoa, & Kutas, 2002; Kutas & Hillyard, 1984), DeLong et al.’s study was unique in that related effects also were found at the articles preceding the target words (*a/an*). More specifically, the N400 was larger for the article that agreed with the less predictable target (*an* for *airplane*) than with the more predicable target (*a* for *kite*). Such data suggest that expectancies generated from contextual information (based on knowledge about *what a boy would go outside to fly on a breezy day*) was constraining enough to allow anticipation of upcoming words’ phonological form, because it was the target nouns’ phonological form that determined whether or not the article is expected. In summary, event knowledge does indeed rapidly influence semantic and lexical processing during sentence comprehension, strongly suggesting that there is no delay of its access and use.
1.2 Event Knowledge and Syntactic Ambiguity

A related debate exists with regard to the interaction between syntactic knowledge and processing on the one hand, and event knowledge on the other. There have been numerous investigations of the nature of this interaction that have used sentences containing a temporary syntactic ambiguity as a tool. For example, the sentence *The officer examined by the lawyer turned out to be unreliable* is temporarily syntactically ambiguous at the verb *examined* because *examined* could be a part of a main clause (MC; e.g., *The officer examined the evidence very carefully*), or part of a reduced relative (RR) clause (**examined by the lawyer**, as in the first example). It has been shown frequently that when reading sentences containing reduced relative clauses, readers experience processing disruption following the initial verb (**examined**) as though they were expecting the sentence to continue as a syntactically simpler MC. A number of studies have investigated the conditions under which this ambiguity effect can be observed, and several models have been proposed. An aspect of this ambiguity that is particularly relevant to this dissertation is that the alternative interpretations of the MC/RR ambiguity not only differ in terms of the structural configuration of the verb, but also in terms of the thematic roles of the entity denoted by the initial noun phrase in the event denoted by the verb (**the officer** is the agent of **examining** in a MC, the person doing the examining, and is the patient in a RR interpretation, the person being examined). A key issue that has been debated over the years is whether people’s knowledge about who commonly does what to whom (i.e., event knowledge or more specifically **thematic knowledge**) can help to resolve such an ambiguity, and if so, how rapidly thematic knowledge exerts its
influence during sentence processing. The debate is best exemplified by contrasting two classes of models.

The Garden-Path model (Frazier, 1987; Frazier & Rayner, 1982) claims that the influence of thematic knowledge is delayed because the model includes two separate and serially ordered stages of processing, namely, syntactic parsing and reanalysis. For example, when reading sentences with a MC/RR ambiguity, the syntactic parser initially prefers the MC interpretation because it is syntactically simpler than the alternative RR interpretation, and is in accordance with the Garden-Path model’s syntactically-governed principles, irrespective of other sources of information (Ferreira & Clifton, 1986). When the initial parse turns out to be incorrect (e.g., the RR interpretation is correct), the incorrect parse must be reanalyzed using information from non-syntactic sources. Processing disruption is assumed to reflect this second stage, and it always occurs when a syntactically more complex interpretation such as the RR is the correct one.

Alternatively, in constraint-based models (MacDonald, Pearlmutter, & Seidenberg, 1994; Trueswell & Tanenhaus, 1994), there are assumed to be rapid influences of event knowledge as well as other information from both linguistic and non-linguistic sources. Constraint-based models assume that syntactic ambiguities are resolved based on the relative likelihood or probability of each interpretation given the various constraints that are based on multiple sources of information. Processing difficulty arises when the resulting probabilities favour the incorrect interpretation, and thus have to be altered. Although readers are generally biased toward syntactically simpler interpretations because they occur much more frequently, ambiguity effects can be reduced or eliminated if other constraints strongly bias readers toward the correct alternative.
interpretation. For example, in a sentence containing a MC/RR ambiguity, if the entity denoted by the initial noun is much more likely to be a patient than an agent (e.g., The evidence examined, or The crook arrested), readers should be biased toward the RR interpretation. This prediction has been supported in a few studies (Clifton et al., 2003; McRae, Spivey-Knowlton, and Tanenhaus, 1998; Trueswell, Tanenhaus, & Garnsey, 1994). In addition to the constraint based on thematic knowledge, several other non-syntactic constraints have been tested and shown to affect the early stage of syntactic ambiguity resolution (see McRae & Matsuki, 2013, for review).

The proponents of constraint-based models often have argued that one potential mechanism by which event knowledge influences syntactic ambiguity resolution is through expectancy generation (Altmann, van Nice, Garnham, & Henstra, 1998; Hare, Elman, Tabaczynski, & McRae, 2009; Hare, McRae, & Elman, 2003; McRae & Matsuki, 2009; Spivey, 2006; Taraban & McClelland, 1990; Trueswell et al., 1993). This is similar to studies of lexical processing in sentence contexts in which conceptual combination of contextual information generates expectancies for upcoming words and concepts. In this case however, knowledge about the roles that are likely to be played by the entities in a denoted event generates expectations for unfolding syntactic structures that could satisfy the likely roles and relations.

1.3 The Role of Event Knowledge

Evidence thus far from these two streams of research jointly suggests that the influence of event knowledge during on-line sentence processing is rapid, and such influence may occur through expectancy generation. In terms of the roles that event knowledge plays during sentence processing, there are two ways to interpret such evidence. On the one
hand, the rapid, expectation-driven use of event knowledge during sentence comprehension can be viewed as beneficial to the sentence processing system because it can speed-up and/or ease the processing of upcoming linguistic input. On the other hand, the use of event knowledge, and particularly the predictive processing based on such knowledge, can be seen as having detrimental effects, specifically in cases where the upcoming linguistic inputs disconfirms expectations. From this point of view, what the current evidence does not yet sufficiently provide is the understanding about the extent to which the use of event knowledge can cause disruption. In addition, if a theory is to focus on the role of event knowledge during sentence processing, there is an essential need to understand how the use of event knowledge influences the final interpretation of a sentence. For example, one issue that has not been addressed in detail concerns how possible disruptions of sentence comprehension through the use of event knowledge affect the final interpretation of sentences.

1.4 Overview of present research

The main goal of this dissertation is to examine further the roles of thematic knowledge during sentence processing and sentence understanding. To do so, the influence of event knowledge was examined in two ways. The first study (Chapter 2) investigated whether manipulation of thematic knowledge can lead to processing disruption during reading of sentences that are otherwise free of processing difficulties. The second study (Chapter 3) investigated how thematic knowledge affects the construction of sentential meaning representations and how misinterpretations can occur during that process.

In Chapter 2, I re-examined constraint-based models with regard to the role of thematic fit constraints. As described earlier, constraint-based models assume that the relative fit
of the initial NP to the event denoted by the verb (e.g., *The cop vs. The crook arrested* …) should bias readers toward one of the alternative interpretations. Yet, the evidence discussed thus far has focused primarily on biasing readers in a specific direction. That is, studies have shown that readers can be biased toward a more syntactically complex interpretation (i.e., a RR interpretation in the case of a MC/RR ambiguity), so that processing difficulty associated with temporary syntactic ambiguity can be alleviated. In other words, difficult sentences can be made easier. However, given the assumption underlying constraint-based models that all relevant constraints are used immediately to influence interpretation probabilities, it should be possible to make easy sentences harder. That is, manipulating thematic fit should induce processing disruption in syntactically ambiguous sentences that are otherwise free of processing difficulties (i.e., making main clause sentences more difficult to process). In fact, this has been a major challenge of constraint based models, originally put forth by Frazier (1995).

To investigate this, I used sentences containing the MC/RR ambiguity that are resolved in favor of the MC interpretation, which should be easy to process from a purely syntactic point of view. Thematic fit can be manipulated to bias readers toward the RR interpretation, and should result in processing disruption due to the unexpected continuation of a sentence as a MC. Finding processing difficulties in such sentences as a result of a thematic fit manipulation would strongly support the predictions of the constraint-based models.

In Chapter 3, I investigated the role of thematic knowledge in relation to sentence interpretation and misinterpretation. In many psycholinguistic theories of sentence processing, it is commonly assumed that people extract deep and rich representations of
the meaning and structure of sentences, and the goal of empirical work is to uncover the underlying mechanism that enables this to occur. For example, in studies of temporary syntactic ambiguity resolution, the underlying essential question of “How is the ambiguity resolved?” presupposes the idea that, with or without temporal processing disruptions, people in the end arrive at a correct interpretation of sentences every time. Indeed, all current models of sentence processing discussed above assume that people generate complete, detailed, and accurate representations of the linguistic input. Despite the differences in the models’ assumptions regarding how and when constraints are used, both constraint-based and two-stage models assume that ambiguities are resolved eventually by incorporating all relevant constraints. However, there is a body of evidence to question, if not undermine, this very assumption.

A growing number of studies have shown that comprehenders do not always arrive at an accurate interpretation of a sentence (see Ferreira & Patson, 2007; Sanford & Sturt, 2002, for reviews). Instead, under certain circumstances, comprehenders seem to occasionally generate inaccurate or distorted representations. The good enough approach to language comprehension proposed by Ferreira and colleagues (Ferreira, Bailey, & Ferraro 2002; Ferreira & Patson, 2007) suggests that the comprehension system occasionally generates incomplete interpretations for the sake of cognitive economy. That is, because complete and accurate analyses of linguistic inputs based on syntactic information are sometimes computationally costly, particularly in situations in which the input is noisy due to perceptual noise, disfluencies, grammatical errors, and so on, the comprehension system may rely on more “quick and dirty” heuristics that are based on probabilistic knowledge...
about language and events. Although heuristic-based processing is efficient and effective most of the time, it can lead to occasional misinterpretation.

Currently, there are a number of studies supporting the notion that heuristic-based processes indeed operate during sentence processing. There is also evidence suggesting the potential time course of heuristic-based processes in relation to syntax-based processing (Kim & Osterhout, 2005). However, it is not entirely clear how the online (i.e., moment-to-moment during reading) use of heuristics directly relates to occasional misinterpretation. To examine the relation between online processing and final sentence interpretation, I extended the findings of Ferreira (2003) in which participants were more likely to make comprehension errors when sentences mismatch with event knowledge (i.e., what can be expected on the basis of probabilistic knowledge about events). In particular, as in Chapter 2, I focused on thematic knowledge. Using a paradigm similar to Ferreira, the study in Chapter 3 measured ERPs and obtained additional behavioral measures to examine and evaluate how the online use of heuristic-based processing relates to the final interpretation of sentences.

1.5 References


McRae, Ken, & Matsuki, K. (2009). People use their knowledge of common events to understand language, and do so as quickly as possible. *Language and Linguistics Compass, 3*, 1417-1429.


Chapter 2

2 Thematic Fit Induces Competition Effects in Sentences that are Resolved as a Main Clause

2.1 Introduction

Sentences containing a temporary syntactic ambiguity have played an important role in adjudicating among theories of on-line sentence processing. There are several types of temporary syntactic ambiguities that have been studied extensively in English, and two of them are exemplified in the following:

1. The student found the book was written poorly.
2. The officer examined by the lawyer turned out to be unreliable.

The first type is known as the Direct Object / Sentential Complement (DO/SC) ambiguity. Sentence (1) is temporarily ambiguous at the post-verbal noun phrase (the book) because it can be either the DO of the verb (found; as in The student found the book in the library) or the subject of a SC (as in 1). The ambiguity is present because the complementizer that following the verb can be dropped in English. The second type is called the Main Clause / Reduced Relative (MC/RR) ambiguity. Sentence (2) is temporarily ambiguous at the verb (examined) because the verb could be in past tense active form and thus part of a main clause (e.g., The officer examined the evidence very carefully), or it could be in past-participle passive form and thus part of a reduced relative clause (as in 2). The MC/RR ambiguity occurs because, in English, a large number of verbs have identical past tense and passive participle forms, and because who was or that was can be omitted. A common finding is that when reading or listening to these types of
ambiguous sentences, comprehenders experience a processing disruption (a syntactic ambiguity effect, also known as a garden-path effect) whereby they initially misinterpret the sentence as a syntactically simpler structure (an active main clause structure, in these cases). A number of studies have investigated the conditions under which these temporary syntactic ambiguity effects are observed and several models have been proposed.

Two types of models have most often been contrasted regarding their predictions for how people deal with this type of ambiguity in moment-to-moment processing. One class of models consists of serial, two-stage models, which are often realized as the Garden-Path model (Frazier, 1987; Frazier & Rayner, 1982; see Crocker, 1994; Inoue & Fodor, 1995; Kimball, 1973; Pickering, 1994 for other variants). The key assumption is that privileged (primarily syntactic) information is processed and analyzed before any other types of information are combined to refine the initial analysis. For example, according to the Garden-Path model, during processing of sentences with a MC/RR ambiguity, the human parsing mechanism initially prefers the MC interpretation on the basis of syntactic parsimony. Only upon reaching the region of structural conflict (e.g., turned out in 2) does the parser reanalyze and construct an alternative RR interpretation. Reanalysis is assumed to be cognitively costly and results in longer reading times. In these models, the first analysis is considered to be fast and automatic because inputs are analyzed according to syntactically-governed principles (Ferreira & Clifton, 1986). Reanalysis, on the other hand, is optional but is necessary for correctly interpreting syntactically complex sentences. Contextual (semantic, thematic, and discourse-based) information exerts its influence on syntactic parsing only during a second reanalysis stage. As such, two-stage
models predict that ambiguity effects occur purely on the basis of syntactic misanalysis, and irrespective of meaning or contextual cues that may be useful for initial correct interpretation, so that contextual influences are delayed.

Two-stage models typically have been contrasted with parallel, constraint-based models (MacDonald, Pearlmutter, & Seidenberg, 1994; Trueswell & Tanenhaus, 1994). In constraint-based models, sentence processing is statistical or probabilistic in nature, and occurs by way of computing, combining, and integrating various types of information rapidly from both linguistic and non-linguistic sources. That is, there is no assumption of an architecturally-determined delay of information. Multiple partial interpretations are entertained in parallel when processing ambiguous sentences, although types of information can be weighted differentially. According to constraint-based models, ambiguity effects can be reduced or eliminated if constraints are manipulated in certain ways, and a number of studies have focused on determining the constraints that matter in various circumstances, and how rapidly they can reduce or eliminate ambiguity effects.

One constraint, the likelihood that a certain verb is used with certain syntactic structures (verb-bias), has been shown to rapidly influence syntactic ambiguity resolution. For example, Trueswell, Tanenhaus, and Kello (1993) investigated how reading times are influenced by people’s knowledge of the relative frequencies with which verbs take a DO versus a SC. The relative structural frequencies of verbs were measured in a separate sentence fragment completion study in which participants completed a fragment such as *John insisted ___*, and the relative proportion with which the fragments were continued with a DO versus a SC were calculated. They contrasted sentences that differed in terms of the verbs’ structural biases, as in (3) and (4).
3. The waiter confirmed the reservation was made yesterday. (DO-biased)

4. The waiter insisted the reservation was made yesterday. (SC-biased)

The verb confirm is often followed by a DO and rarely by a SC (100% vs. 0% completions, respectively), whereas insisted is more frequently followed by a SC and rarely by a DO (71% vs. 0%, respectively). Thus, according to constraint-based models, if readers use their knowledge of verb bias, they should be less likely to interpret the reservation as a DO when reading (4) than when reading (3), and therefore should be less likely to show slower reading times, as measured against an unambiguous baseline in which that follows the initial verb. Alternatively, according to two-stage models, because information such as verb bias is not used during initial syntactic parsing, syntactic ambiguity effects should be observed for both (3) and (4), and an influence of verb-bias should appear only later (Ferreira & Henderson, 1990). Trueswell et al. confirmed the prediction of constraint-based models in self-paced reading and eye-tracking experiments. Several other studies have found similar rapid influences of verb-bias during the reading of sentences containing a DO/SC ambiguity (Garnsey, Pearlmutter, Myers, & Lotocky, 1997; Hare, McRae, & Elman, 2003; but see, Kennison, 2001). Hare, McRae, and Elman (2004) present evidence for, and a discussion of, the conditions that determine rapid influences of verb bias.

Another well-studied constraint is thematic fit, which concerns the degree of match between an entity or object, and a specific thematic role in an event denoted by a verb. McRae, Spivey-Knowlton and Tanenhaus (1998) investigated whether ambiguity effects can be reduced when thematic fit is manipulated to bias a reader toward one interpretation over another in sentences containing a MC/RR ambiguity. McRae et al.
compared sentences with a reduced relative clause such as those below that differ in terms of thematic fit of the initial noun.

5. The cop arrested by the detective was guilty of taking bribes. (good agent, poor patient)

6. The crook arrested by the detective was guilty of taking bribes. (good patient, poor agent)

In an arresting event, a cop is more commonly an agent (i.e., the one who performs the action of arresting) whereas a crook is more often a patient (i.e., the one who gets arrested). In (6), the fact that crook is a good patient but a poor agent of an arresting event makes the RR interpretation more likely. Therefore, constraint-based models predict that a reader should be less likely to experience processing disruption in sentences like (6) relative to (5). Note that reading times are compared to an unambiguous baseline in which who was or that was follows the initial verb. Two-stage models instead predict that thematic fit does not influence initial syntactic parsing. McRae et al. showed using self-paced reading that thematic fit modulated ambiguity effects at arrested by and was guilty, supporting constraint-based models. Several other studies have manipulated thematic fit in terms of the animacy of the initial noun phrase (e.g., The defendant [vs. evidence] examined …). Using eye-tracking, Trueswell, Tanenhaus, and Garnsey (1994) found a rapid influence of animacy in eliminating processing difficulty in the inanimate condition, which is otherwise present in the animate condition. Clifton et al. (2003) showed similar but somewhat modest effects of an animacy manipulation, in that it reduced, rather than eliminated, the ambiguity effect, despite using the same materials as Trueswell et al. Although it is apparent that the manner in which all other constraints
support one or the other interpretation can determine whether the influence of thematic fit is sufficiently strong to eliminate versus reduce the processing difficulty related to syntactic ambiguity, it is clear that the influence of thematic fit is rapid.

2.1.1 Making Simple Sentences Harder

At this point in time, a number of studies have shown immediate influences of multiple constraints on syntactic ambiguity resolution, favoring constraint-based models. However, in a critical review of constraint-based models, Frazier (1995) argued that both constraint-based and two-stage serial models can account for evidence showing the reduction or elimination of syntactic ambiguity effects in temporarily ambiguous sentences that are resolved as complex syntactic structures via manipulation of various constraints. Frazier urged that a more rigorous test is to investigate whether a syntactically simple structure could be made harder to understand through the manipulation of constraints. Frazier challenged researchers to demonstrate that a normally-preferred, syntactically simpler sentence can be made harder to process.

At the time Frazier (1995) was published, there was one study by Pearlmutter and MacDonald (1995) showing that a simple sentence could indeed be made more difficult. However, Frazier argued that this evidence was insufficient. Pearlmutter and MacDonald showed that high, but not low, reading span participants experienced difficulty comprehending main clause transitive sentences containing a MC/RR ambiguity. They compared sentences such as *The soup cooked in the pot but was not ready to eat* to unambiguous intransitive counterparts such as *The soup bubbled in the pot but was not ready to eat*. Difficulty with the ambiguous sentences was attributed to differential on-line (but equal off-line) sensitivity of high- versus low-span readers to plausibility
information, as established by off-line norms and regression analyses of reading times. Frazier argued that Pearlmutter and MacDonald’s study suffered from post hoc theoretical predictions, the lack of clear explanations of their hypothesis and regression results, and the absence of ambiguity effects in a large number of participants (there were twice as many low-span readers as high-span readers). Frazier therefore concluded that Pearlmutter and MacDonald’s results are inconclusive at best.

Although still scarce, a few investigations of this issue have taken place in the last decade or so (Binder, Duffy, & Rayner, 2001; Kennison, 2001; van Berkum, Brown, & Hagoort, 1999; Wilson & Garnsey, 2009). In particular, the two most recent studies, Wilson and Garnsey and Binder et al., investigated the two aforementioned ambiguity types, DO/SC and MC/RR respectively, and reached contrasting conclusions. I review those studies in turn.

2.1.2 Biasing Main Clauses toward Sentence Complements

Wilson and Garnsey (2009) investigated whether manipulating verb bias produces processing difficulty during the reading of simple DO sentences containing a DO/SC ambiguity. Based on offline proper name-verb fragment completion norms, they identified verbs that are biased toward direct object (76% DO vs. 13% SC) and embedded clause completions (11% DO vs. 58% SC). They constructed ambiguous DO sentences, ambiguous SC sentences, and unambiguous SC sentences with a that-complementizer, as in below.

DO-biased verbs

7. The CIA director confirmed the rumor when he testified before Congress.
8. The CIA director confirmed (that) the rumor could mean a security leak.

SC-biased verbs

9. The ticket agent admitted the mistake because she had been caught.

10. The ticket agent admitted (that) the mistake might not have been caught.

In both word-by-word self-paced reading and eye-tracking experiments, Wilson and Garnsey (2009) found effects of verb bias. First, they replicated previous studies showing reduced ambiguity effects at the disambiguation region (the two words following the ambiguous noun phrase, as in *could mean or might not*) in SC sentences when embedded clauses followed SC-biased verbs (8) than they followed DO-biased verbs (10). Second, they found an interaction between verb-bias and sentence type such that longer reading times were associated with the bias-continuation mismatch. That is, readers were garden-pathed (there was greater competition) when reading SC-structures with DO-biased verbs and critically, when reading DO-structures (i.e., a main clause structure) with SC-biased-verbs. Thus, in a DO/SC ambiguity, a normally-preferred, syntactically simpler DO sentence was made harder.

2.1.3 Biasing Main Clauses toward Reduced Relatives

In contrast, Binder et al. (2001; Experiment 1) conducted an eye-tracking study in which they manipulated a thematic fit/verb bias (as one rather than two separate factors) to investigate whether processing difficulty emerges in main clause sentences with a MC/RR ambiguity, as well as whether processing difficulty is reduced in ambiguous RR sentences. They used off-line NP-verb fragment completion norms to create two conditions for each sentence type (RR and MC). The balanced condition consisted of
sentences that were completed approximately equally often as a MC or a RR (55% vs. 45%, respectively). The biased condition consisted of sentences that were primarily completed as a MC (93% MC vs. 7% RR). Examples are shown below.

Main Clause Biased

11. The wife (had) deserted her unfaithful husband and moved to another country.

12. The wife (who was) deserted by her unfaithful husband moved to another country.

Balanced

13. The patient (had) cured the inexperienced doctor and became famous.

14. The patient (who was) cured by the inexperienced doctor became famous.

Contrary to previous results, comprehension difficulties occurred irrespective of the type of initial noun-verb combination when reading the ambiguous RR sentences (12 and 14). More germane to the present study, Binder et al. (2001) found no evidence of processing disruption in the MC sentences (11 and 13). The balanced-biased manipulation did not modulate reading times at the ambiguous verb region (*cured*) or at the disambiguation region (which in their study, corresponded to *the inexperienced*). Binder et al. concluded that thematic fit information does not influence ambiguity resolution and therefore is not used in the initial stage of sentence comprehension.

Given the opposite conclusions of Wilson and Garnsey (2009) and Binder et al. (2001), and their theoretical importance, it is critical to reconcile them. The studies featured different types of ambiguities and constraints. Indeed, Wilson and Garnsey pointed out the possibility that the kind of constraints used in Binder et al. are computationally more complex than is verb bias, and thus their influence might be slower and weaker.
However, there do exist alternative accounts of Binder et al.’s results. First, although Binder et al. stated that they manipulated thematic fit, they did not actually measure it. Therefore, it is possible that there might be some items for which the thematic fit manipulation is not sufficiently strong to show reliable effects. Second, the initial verbs in Binder et al.’s MC sentences were followed by the definite article the for most of their items, and possessive determiners (his or her) for the remainder. When these words appear directly following an initial noun-verb segment, this produces an extremely strongly bias toward a MC interpretation (henceforth referred to as the the-bias). It is possible that this the-bias masked potential effects of thematic fit. In other words, if thematic fit is manipulated strongly, and if the influence of the the-bias is eliminated, it may be possible to use thematic fit to make simple sentences harder.

2.1.4 The Present Research

The purpose of Chapter 2 was to examine whether thematic fit can induce processing difficulty in temporarily ambiguous MC/RR sentences that are resolved in favor of the structurally simpler MC. Constraint-based models predict that main clause sentences can be made harder via manipulation of thematic fit, insofar as the manipulation is sufficiently strong and no other, potentially stronger biases such as the the-bias counteracts its effects. Two-stage models, on the other hand, predict that because the comprehension system by default always constructs a MC interpretation upon encountering this ambiguity, it is not possible to show a processing disruption in MC sentences.

To empirically examine our speculations regarding Binder et al.’s (2001) findings, we began by conducting two norming studies. In Norming Study 1a, we obtained thematic fit...
ratings on Binder et al.’s items. In Norming Study 1b, we used fragment completion norms to assess Binder et al.’s items, in particular to investigate whether the presence of the determiner changes how the fragments are continued. We then conducted thematic fit ratings (Norming Study 2a) and fragment completion norms (Norming Study 2b) to construct items to be used in an eye-tracking experiment. To avoid the potential the-bias, we used a construction known as heavy NP shift in which the post-verbal NP (direct object) occurs after an intervening phrase, as shown in (15).

15. The host invited to the luncheon three old buddies who really had no business being there.

In the eye-tracking experiment, we examined reading times in temporarily ambiguous MC sentences that are initially biased toward a RR interpretation, compared to those that are initially biased toward a MC interpretation.

2.2 Norming Study 1a
To test the strength of Binder et al.’s (2001) manipulation, we conducted thematic role/filler typicality norms using Binder et al.’s items. This norming task was originally used by McRae et al. (1997), and has been used in a number of studies since that time to establish thematic fit.

2.2.1 Method
2.2.1.1 Participants
Thirty-five undergraduates from the University of Western Ontario, who did not participate in any of the other studies, participated for course credit. Twenty participants completed the agentialhood ratings, and 15 rated patienthood.
2.2.1.2 Materials and Procedure

We used the 48 verbs and the subject NP associated with each verb from Binder et al.’s (2001) Experiment 1. The object NPs were also included along with 3 filler nouns, in order to encourage participants to use the entire scale. The task was administered in a paper and pencil format. Participants rated how common it is for some type of object or person in the world to play a specific role in an event on a 7-point scale (1 corresponded to very uncommon and 7 to very common). Examples of an agenthood and a patienthood item are given below.

Agenthood
How common is it for a/an
- soldier
- inexperienced doctor
- business woman
- physician
- patient
to cure someone/something?

Patienthood
How common is it for a/an
- soldier
- inexperienced doctor
- business woman
- physician
- patient
to be cured by someone/something?

Examples were provided before participants began the task. No time limit was imposed.
2.2.2 Results and Discussion

Table 2-1 presents the mean agenthood and patienthood ratings on the initial NPs used in Binder et al.’s (2001) biased and balanced conditions. For comparison, Table 2-1 also includes the mean agenthood and patienthood ratings reported in McRae et al. (1998), who found an immediate effect of thematic fit. A strong manipulation consists of agenthood and patienthood ratings that are highly polarized such that Binder et al.’s biased items should have high agenthood ratings and low patienthood ratings, whereas their balanced items should show the opposite pattern. In contrast, the mean patienthood ratings for Binder et al.’s items were greater than the mean agenthood ratings in both conditions. In addition, the agenthood and patienthood ratings overlapped in both conditions. This sharply contrasts with McRae et al.’s ratings which show large differences in opposite directions for good agents and good patients. Therefore, Binder et al.’s weak thematic fit manipulation suggests a clear reason why they did not observe an early effect.
Table 2-1. Means (and ranges) of the agenthood and patienthood ratings for Binder et al.’s (2001) Experiment 1 balanced and biased items. Means (and ranges) of McRae et al.’s (1998) good patient and good agent items are also presented for comparison.

<table>
<thead>
<tr>
<th></th>
<th>Agenthood</th>
<th>Patienthood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Norming Study 1a: Binder et al.’s Items</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Balanced</td>
<td>3.0 (1.7 - 6.6)</td>
<td>5.5 (3.2 - 6.9)</td>
</tr>
<tr>
<td>Biased</td>
<td>4.0 (1.6 - 6.7)</td>
<td>4.5 (1.9 - 6.9)</td>
</tr>
<tr>
<td>From McRae et al. (1998)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Good Patient</td>
<td>2.0 (1.0 - 3.8)</td>
<td>6.0 (4.0 - 7.0)</td>
</tr>
<tr>
<td>Good Agent</td>
<td>6.3 (4.4 - 7.0)</td>
<td>2.5 (1.2 - 4.1)</td>
</tr>
</tbody>
</table>
2.3 Norming Study 1b

To investigate our second speculation regarding Binder et al.’s (2001) findings, we conducted two sets of sentence completion norms in which participants were asked to complete sentence fragments that were either truncated at the verb (e.g., The patient cured _____) or at the post-verbal determiner (e.g., The patient cured the _____).

2.3.1 Methods

2.3.1.1 Participants

Twenty-five undergraduates from the University of Western Ontario participated for course credits. Of these, 20 completed the post-verb completion and five completed fragments ending in a determiner or pronoun.

2.3.1.2 Materials and Procedure

We used the 48 sentences (24 balanced ambiguous and 24 biased ambiguous) from Binder et al.’s (2001) Experiment 1. The target sentences, which were truncated either at the verb or at the post-verbal determiner in separate lists, were intermixed with 116 fillers so that target fragments were separated by at least 2 fillers. Participants were asked to write down a sensible and grammatical continuation for each fragment.

2.3.2 Results and Discussion

Responses were categorized in terms of a main clause (MC), a reduced relative (RR), a self-continuation (as in The runner timed herself...), an embedded clause, or a nonsensical response or blank (less than 1% of responses; excluded when calculating percentages). Self-continuations were distinguished from MC-continuations because although they are syntactically the same, self-continuations are thematically ambiguous
because the entity denoted by the pre-verbal NP (e.g., runner) is used as both the agent and patient of the action denoted by the verb (timing). Table 2-2 presents the mean percentage of each continuation type. Table 2-2 also presents the values reported in Binder et al. (2001) for comparison.

For the completions at the verb, although the percentages for the biased items are similar to Binder et al.’s, the percentages for the balanced items are not. Given completion percentages observed in other studies that have investigated the MC/RR ambiguity, the results we obtained appear to fall within a reasonable range (McRae et al, 1998). That is, in no study other than Binder et al. have RR completions been as high as 44%. The results for the completions at the post-verbal determiner were extremely straightforward. Regardless of condition, not a single item was completed as a RR. Thus, the the-bias is an extremely strong cue for a MC interpretation. This observation implies that any potential effects of a moderate thematic fit manipulation were masked by the post-verbal the-bias in Binder et al. One implication of this result is that the strong influence of the the-bias needs to be avoided by using, for example, sentences containing an NP-shift.

2.4 Norming Study 2a

To construct heavy NP shift sentences that contain a MC/RR ambiguity and are initially biased either toward MC or RR interpretations, we conducted two norming studies similar to Norming Study 1a and 1b. In Norming Study 2a, we obtained thematic fit ratings, and selected a subset of items that showed clear polarization of agenthood and patienthood ratings across conditions. The results of Norming Study 2a served as the basis for item selection for Norming Study 2b and the ensuing eye-tracking experiment.
Table 2-2. Mean completion percentages for sentence fragments ending at verb and at the determiner (or her/his) for Binder et al.’s (2001) Experiment 1 items. Means as reported in Binder et al. (2001) are also presented for comparison.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Continuation Type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Main clause</td>
</tr>
<tr>
<td><strong>Norming Study 1b</strong></td>
<td></td>
</tr>
<tr>
<td>At the Verb</td>
<td></td>
</tr>
<tr>
<td>Biased</td>
<td>91</td>
</tr>
<tr>
<td>Balanced</td>
<td>68</td>
</tr>
<tr>
<td>At the Determiner</td>
<td></td>
</tr>
<tr>
<td>Biased</td>
<td>100</td>
</tr>
<tr>
<td>Balanced</td>
<td>100</td>
</tr>
<tr>
<td><strong>From Binder et al. (2001)</strong></td>
<td></td>
</tr>
<tr>
<td>At the Verb</td>
<td></td>
</tr>
<tr>
<td>Biased</td>
<td>93</td>
</tr>
<tr>
<td>Balanced</td>
<td>55</td>
</tr>
</tbody>
</table>

*Note:* Self completions are included in the main clause percentages.
2.4.1 Methods

2.4.1.1 Participants

Twenty undergraduates from the University of Western Ontario participated for course credit.

2.4.1.2 Materials and Procedure

Eighty verbs and several associated nouns (the number of nouns ranging from 3 to 7) were selected. The Procedure was identical to that of Norming Study 1a.

2.4.2 Results

We selected 40 verbs based on the criteria that there were two animate nouns with highly polarized ratings (i.e., with a high patienthood rating and a low agenthood rating or vice versa) with minimal overlap in range. The means and ranges of the agenthood and patienthood ratings for the sentences used in the eye-tracking experiment are presented in Table 2-3.

2.5 Norming Study 2b

The 40 verbs, each paired with a good agent/poor patient and a poor agent/good patient were presented in sentence fragments that were truncated either at the verb or at the end of the postverbal prepositional phrase. The proportions of various types of sentence completions were measured and used to assess the items for the eye-tracking experiment.

2.5.1 Method

2.5.1.1 Participants

Forty undergraduate from University of Western Ontario participated for course credit.
Table 2-3. Means (and ranges) of the agenthood and patienthood ratings for the good agent and good patient items selected for Experiment 1.

<table>
<thead>
<tr>
<th></th>
<th>Agenthood</th>
<th>Patienthood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good Agents</td>
<td>2.8 (1.5 - 4.1)</td>
<td>6.3 (3.9 - 7.0)</td>
</tr>
<tr>
<td>Good Patients</td>
<td>6.1 (4.6 - 7.0)</td>
<td>2.0 (1.0 - 3.8)</td>
</tr>
</tbody>
</table>
### 2.5.1.2 Materials and Procedure

From the 40 verbs selected from Norming Study 2a, 40 sentence pairs were constructed. Each sentence pair was identical except for the initial noun. Each sentence was truncated at the verb (e.g., *The host invited _____*), or at the end of a prepositional phrase (PP; e.g., *The host invited to the luncheon _____*). For continuations from the verb and PP, 2 lists were created to ensure that no participants encountered the same verb twice. For each list, 112 fillers were included so that target items were separated by at least one filler. Participants completed one of the lists by writing sensible and grammatical continuations for each fragment.

### 2.5.2 Results and Discussion

We scored responses by counting the frequency with which the continuations were in one of following forms: main clause (MC, intransitive MC in the case of PP completions), reduced relative (RR), heavy NP shift (NPS), self-completions (e.g., *The patient cured himself*), and nonsensical responses or blanks (less than 1% of data; excluded when calculating percentages).

From the 40 items used in this study, 20 verbs and their associated nouns were chosen for the eye-tracking experiment. Item selection was based on the proportion of RR completions at the verb, proportion of RR completions and NP shift completions at the PP, and the agenthood and patienthood ratings obtained in Norming Study 2a. The completion percentages for the selected 20 sentences are presented in Table 2-4.

The data from the completions at the verb suggest that, as with Binder et al.’s materials, there was an overall bias toward MC interpretations than RR interpretations, although the
Table 2-4. Mean (and standard deviation) of sentence fragment completion percentages obtained at the verb and at the prepositional phrase (PP) for the 20 selected target sentences used in Experiment 1.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Main clause</th>
<th>Noun-phrase shift</th>
<th>Reduced Relative</th>
<th>Self</th>
</tr>
</thead>
<tbody>
<tr>
<td>At the Verb</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Good Agent</td>
<td>99.0 (2.6)</td>
<td>---</td>
<td>0.1 (0.8)</td>
<td>0.9 (2.2)</td>
</tr>
<tr>
<td>Good Patient</td>
<td>79.6 (14.6)</td>
<td>---</td>
<td>14.4 (10.0)</td>
<td>6.0 (7.5)</td>
</tr>
<tr>
<td>At the PP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Good Agent</td>
<td>28.4 (25.8)</td>
<td>34.7 (18.8)</td>
<td>36.9 (25.2)</td>
<td>---</td>
</tr>
<tr>
<td>Good Patient</td>
<td>14.9 (19.8)</td>
<td>4.6 (6.7)</td>
<td>80.6 (22.2)</td>
<td>---</td>
</tr>
</tbody>
</table>

*Note: Main clause continuation includes both transitive and intransitive uses of the verb.*
RR completions did increase for the good patient condition. In contrast, the data from the completions at the PP showed that RR completions increased overall, but critically much more so for the good patient condition than for the good agent condition. Importantly, there is a crossover of completion biases at the PP, such that there is a stronger MC than RR bias for the good agent condition, but a stronger RR than MC bias for the good patient items. Finally, the 35% heavy NP shift completions at the PP for the good agent items demonstrate that this type of structure could be produced by participants.

2.6 Experiment

The goal of the main experiment of Chapter 2 was to determine whether thematic fit, which is presumed to influence second stage processing in two-stage models, but initial comprehension processes in constraint-based models, can rapidly influence the resolution of temporary syntactic ambiguities in sentences that are eventually resolved as syntactically-simpler main clause sentences. Participants read sentences while their eye movements were monitored. Thematic fit (good-agent vs. good-patient) and ambiguity (ambiguous vs. unambiguous; i.e., without vs. with had) were manipulated. Examples of good-agent and good-patient sentences are shown below.

16. The organizer (had) invited to the luncheon three old buddies who really had no business being there. [Good Agent/Poor Patient]

17. The guest (had) invited to the luncheon three old buddies who really had no business being there. [Poor Agent/Good Patient]

First-pass reading times were examined to measure immediate processing, and first-pass regressions and total-reading times were indices of late processing. Constraint-based
models predict an interaction in first-pass reading times such that an ambiguity effect should emerge early in the good-patient condition, but not in the good agent condition. This interaction should occur at the disambiguating post-PP NP region (*three old buddies*). Two-stage models predict no such effect.

2.6.1 Method

2.6.1.1 Participants

Forty-eight native English-speaking undergraduates from the University of Western Ontario participated either for course credit or for $10 CAD compensation. They did not participate in any of the other studies. All participants had normal or corrected-to-normal visual acuity.

2.6.1.2 Materials

Twenty sets of experimental sentences were used in four conditions: good-agent ambiguous, good-agent unambiguous, good-patient ambiguous and good-patient unambiguous. As in Binder et al. (2001), unambiguous sentences included *had* prior to the initial verb, whereas ambiguous sentences did not. All target sentences were followed by a second sentence to increase the meaningfulness of each short narrative. For example, (16) and (17) were followed by, *They ended up drinking too much and getting him in a lot of trouble*. All items are presented in Appendix A. There were four lists so that each participant saw each sentence in only one of its four forms. Each list contained 20 target sentences randomly ordered and intermixed with 64 two-sentence fillers to ensure that the targets never occurred adjacently and that each experimental session began with six filler trials. No filler items contained a heavy NP shift.
2.6.1.3 Procedure

The experiment was conducted using a SR-Research EyeLink 1000 desk-mounted eye-tracker (spatial resolution: 0.01 degrees; sampling rate: 1000) with a chin-forehead rest. Stimuli presentation was implemented using the EyeTrack software (Version 0.7.9) developed at University of Massachusetts, Amherst. Sentences were presented in black 18 pt bold Courier New font on a light grey background (RGB 232.232.232). Participants were seated approximately 70 cm apart from a 21 inch CRT monitor (refresh rate: 100 Hz; resolution: 1280 × 1024); 3 characters equaled one degree of visual angle. Viewing was binocular, but only the participant’s dominant eye was tracked (the right eye for approximately 60% of the participants). The participants’ dominant eye was determined prior to the experiment by a variant of the Miles test (Miles, 1930). In the test, participants form with their hands a triangle with a small opening, and view a distant object through the opening with both eyes open. Then, they close one eye at a time. The eye that can view the object is determined as the dominant eye. Button responses were collected using a hand-held game pad.

Each participant was assigned to one of four lists. At the start of the experiment, the experimenter performed the EyeLink's standard 9-point calibration and validation procedure. The procedure was repeated at least once every 25 trials, or if the experimenter noticed a decline in measurement accuracy. Each trial began with the presentation of a central fixation dot, followed by a small black rectangle in the first character position of the upcoming text display (five characters from the left in the middle row of the screen). When participants had fixated on the target box for at least 200 ms, the two-sentence discourse appeared. Participants were instructed to read at a
normal pace and to press the right-hand button on the game pad when they were finished reading. Following that, either the subsequent trial began, or a comprehension question was administered. Comprehension questions appeared on 36% of trials pseudo-randomly, and participants responded by pressing either the left ('no') or the right ('yes') response button. An experimental session lasted approximately 40 minutes.

2.6.1.4 Data Analyses

Fixation coordinates were mapped onto character positions using EyeDoctor (Version 0.6.3) developed at University of Massachusetts, Amherst. Prior to analyses, trials with poor vertical accuracy or with blinks occurring while reading critical portions of text for the first time were excluded. A total of 48 trials were removed on these bases. Fixations less than 80 ms in duration were pooled with the neighbouring fixations if they were within one character space of the short fixations. Fixations less than 40 ms in duration were excluded if they were within three character spaces of adjacent fixations. Other fixations shorter than 80 ms were excluded, as they are thought to reflect oculomotor programming time rather than cognitive processes (Morrison, 1984). Likewise, all fixations longer than 800 ms were excluded as they are likely due to track losses.

For the purpose of analyses, each sentence was divided into four regions as shown below.

Verb  PP  Modifier  Noun

The organizer (had) / invited / to the luncheon / three old / buddies / who really had…

Fixations landing on the space between two adjacent regions were counted as part of the region to the right of the space (Rayner & Pollatsek, 1989). For each region, fixation data
were summarized in terms of three commonly reported eye-tracking measures: (1) first-pass reading time (the time from fixating a region for the first time until leaving the region either to the left or to the right; (2) first-pass regressions (the probability of making a regressive eye movement after fixating the region for the first time, i.e., first-pass fixation); and (3) total reading time (the sum of all fixations on a region, including re-reading).

2.6.2 Results and Discussion

All participants scored at least 75% accuracy on the comprehension questions ($M = 92\%; SD = 6\%$). Therefore, all participants were included in data analyses. Table 2-5 summarizes the results by measure and region. Each reading time measure was analyzed by fitting linear mixed effects models (Baayen, Davidson, & Bates, 2008), whereas regression proportion was analyzed by fitting mixed logit models (Jaeger, 2008) with the same random and fixed effects. All analyses were conducted using the lme4 package in R (Bates, Maechler, & Bolker, 2011), following the procedure outlined in Levy, Bicknell, Slattery, and Rayner (2009, Supporting Information Appendix). The estimated coefficients ($bs$ indicating effect size in milliseconds for reading times and change in log odds for first-pass regressions), standard errors, $t$-statistics (for reading times), $z$-statistics (for first-pass regressions), and $p$-values are reported. In the cases in which the intercept-only models were fit, the $p$-values for the linear mixed effects models were estimated based on Markov Chain Monte Carlo (MCMC) simulations (using the function `pvals.fnc();` Baayen, 2008). Otherwise, they are based on the normal approximation to the $t$-statistic. The $p$-values for logit models are based on the $z$-distribution. Following the significant ambiguity by thematic fit interaction, the simple main effects of ambiguity in
each thematic fit condition were assessed by fitting separate models for each level of thematic fit.

2.6.2.1 First Pass Reading Times

In the modifier region (*three old*), thematic fit interacted with ambiguity ($b = -53.96$, $SE = 24.96$, $t = -2.16$, $p < .04$). There were also main effects of ambiguity ($b = -31.15$, $SE = 12.48$, $t = -2.50$, $p < .01$) and thematic fit ($b = 33.28$, $SE = 12.51$, $t = 2.66$, $p < .02$).

Simple main effects analyses yielded a reliable 60 ms ambiguity effect in the good patient condition ($b = -60.87$, $SE = 18.67$, $t = -3.26$, $p < .003$), but a non-significant ambiguity effect for good agents ($b = -7.37$, $SE = 16.19$, $t = -0.46$, $p > .6$). There were no main effects or interactions in any other region.

Thus, there are clear and immediate effects of thematic fit. Strong thematic bias against the simpler MC interpretation in the good-patient condition induced processing difficulty at the disambiguating modifier region. When thematic fit aligned with a MC interpretation (good agent but poor patient), no processing difficulty resulted. The results strongly support constraint-based models, in that readers can be garden-pathed during the processing of main clause sentences. Note that this effect was obtained even prior to the post-PP noun (*buddies*), and thus is clearly a rapid one.
Table 2-5. Mean first-pass reading times (in ms), first-pass regression proportions, and total reading times (in ms) by condition, for each region of the sentences.

<table>
<thead>
<tr>
<th>Region</th>
<th>Good Agent</th>
<th>Good Patient</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ambiguous</td>
<td>Unambiguous</td>
</tr>
<tr>
<td><strong>Verb Region</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First-pass reading times</td>
<td>285</td>
<td>277</td>
</tr>
<tr>
<td>First-pass regressions</td>
<td>0.13</td>
<td>0.16</td>
</tr>
<tr>
<td>Total reading times</td>
<td>418</td>
<td>429</td>
</tr>
<tr>
<td><strong>Prepositional Phrase Region</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First-pass reading times</td>
<td>552</td>
<td>533</td>
</tr>
<tr>
<td>First-pass regressions</td>
<td>0.13</td>
<td>0.25</td>
</tr>
<tr>
<td>Total reading times</td>
<td>766</td>
<td>764</td>
</tr>
<tr>
<td><strong>Modifier (the + adjective) Region</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First-pass reading times</td>
<td>432</td>
<td>414</td>
</tr>
<tr>
<td>First-pass regressions</td>
<td>0.15</td>
<td>0.12</td>
</tr>
<tr>
<td>Total reading times</td>
<td>608</td>
<td>525</td>
</tr>
<tr>
<td><strong>Post-verbal Noun Region</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First-pass reading times</td>
<td>289</td>
<td>287</td>
</tr>
<tr>
<td>First-pass regressions</td>
<td>0.13</td>
<td>0.09</td>
</tr>
<tr>
<td>Total reading times</td>
<td>350</td>
<td>320</td>
</tr>
</tbody>
</table>
2.6.2.2 First-pass regressions

In the PP region, participants were more likely to regress back to earlier regions in the unambiguous than in ambiguous sentences \((b = 1.24, SE = 0.21, z = 6.01, p < .001)\). In contrast, in the noun region, there were more regressions in the ambiguous sentences \((b = -0.86, SE = 0.23, z = -3.72, p < .001)\). Additionally in the noun region, more regressions occurred in the good patient than in the good agent condition \((b = 0.55, SE = 0.23, z = 2.39, p < 0.02)\). No other main effects or interactions were significant. This pattern suggests that the presence of the auxiliary verb *had* strongly biased readers toward a main clause interpretation, possibly generating an expectancy for an upcoming noun phrase, and thus resulting in re-reading of earlier materials upon encountering the prepositional phrase. The readers were less committed to a main clause interpretation, or to a noun phrase expectation, when the auxiliary verb *had* was absent and the sentence was temporarily syntactically ambiguous.

2.6.2.3 Total Reading Times

Overall, participants’ total reading times were longer in the good patient condition than in the good agent condition at all four regions: at the verb \((b= 58.45, SE = 20.28, t = 2.88, p<0.01)\), prepositional phrase \((b = 88.50, SE = 25.10, t = 3.53, p < .001)\), modifier \((b = 134.90, SE = 26.23, t = 5.14, p < .001)\), and at the noun region \((b = 45.64, SE = 11.75, t = 3.89, p < .001)\). In addition, total reading times were longer in the ambiguous than in the unambiguous sentences at the modifier \((b = -92.80, SE = 21.57, t = -4.28, p < .001)\), and noun regions \((b = -25.39, SE = 11.72, t = -2.17, p < .04)\). No other main effects or interactions were significant.
2.7 General Discussion

The purpose of Chapter 2 was to examine whether constraints such as thematic fit can immediately influence online resolution of temporary syntactic ambiguity in such a way that processing difficulty is induced in sentences that are resolved in favor of the structurally simpler interpretation. According to Frazier (1995), this is the ultimate test of two-stage versus constraint-based models. In two-stage models, non-syntactic constraints such as thematic fit cannot influence the processing of simple main clauses because no alternative complex structures are considered. In contrast, constraint-based models predict processing disruption or competition whenever the constraints are sufficiently strongly biased toward an incorrect alternative interpretation. Prior to the current study, although Wilson and Garnsey (2009) showed such an influence using the DO/SC ambiguity, Binder et al. (2001) found no influence of a combined verb bias/thematic fit manipulation when investigating the MC/RR ambiguity. The present study reconciles the two conflicting results in two ways. First, we provided an empirically-based explanation of Binder et al.’s null results. Second, in an eye-tracking experiment that avoided potential issues in Binder et al.’s items, thematic fit clearly influenced the comprehension of syntactically simple sentences.

In Norming Studies 1a and 1b, we showed that Binder et al.’s (2001) items may not have been optimally constructed to uncover an influence of thematic fit during the comprehension of ambiguous sentences that are resolved as main clauses. The results of our thematic fit and completion norms indicated a less than optimal manipulation between their balanced and biased items. The completion norms also suggested that any potential effects of their thematic fit manipulation may have been overwhelmed by the
post-verbal *the*-bias. In Norming Study 2a and 2b, we avoided the extremely strong post-verbal *the*-bias, and created items that showed clear differences in thematic fit. Using those items, in the eye-tracking experiment, there was a rapid influence of thematic fit in which readers experienced difficulties during the processing of main clause sentences. Thus, the present results are consistent with those of Wilson and Garnsey (2009), but inconsistent with Binder et al. (2001). Given that our norming studies demonstrate why Binder et al. did not find an influence of thematic fit, the evidence now clearly supports constraint-based models, and provides additional support for the view that the influence of thematic fit is rapid and strong given an appropriate balance of other constraints.

In summary, there has been a long-standing debate between two-stage and constraint-based models of sentence comprehension. An outstanding and theoretically important issue concerned whether there exist results that can be explained only by constraint-based models, and Chapter 2 provides such results. Therefore, as constraint-based models claim, sentences are processed by rapidly computing, combining and integrating information from all relevant sources.

An important question remains with regard to how or whether readers constructed the final correct interpretation of the sentences that are syntactically simple yet difficult to process, such as those used in Chapter 2. In this study, and in other similar studies investigating issues regarding the time-course of syntactic ambiguity resolution, the comprehension questions almost invariably serve the secondary purpose of checking whether participants paid attention during the experiment. Comprehension questions typically concern content that is unrelated to the main manipulation. This is done purposely to avoid drawing participants’ attention to the ambiguities in the materials.
Therefore, it is not completely clear whether participants actually have understood the content of the sentences correctly, for example, in terms of correct thematic assignments.

Although it is often assumed on the basis of reading time data that participants spend time to revise incorrect expectations, a recent series of studies by Ferreira and colleagues (see Ferreira & Patson, 2007 for review) suggest that such an assumption may not be adequate. Instead, their research suggests that readers often misinterpret sentences because they retain the initial incorrect interpretation along with, or as a part of, the final correct interpretation. Thus, according to Ferreira and colleagues, neither the two-stage nor constraint-based models (at least a version of them, e.g., MacDonald, et al., 1994) are correct with regard to readers’ final interpretations of sentences because those models assume that temporary syntactic ambiguities are resolved one way or another by the end of a sentence. Whether or not such misinterpretations occur, the type of incorrect interpretation a reader may construct, and whether or not constraint-based models generally can account for such findings, are interesting topics for further research. The study in Chapter 3 addresses some of these issues.

2.8 References


Chapter 3

3 Inducing Misinterpretation of Simple Passive Sentences

3.1 Introduction

In psycholinguistic studies of sentence processing, there has been an increasing awareness that when reading or listening to sentences, people do not always arrive at an accurate interpretation of a sentence every time. Instead, readers or listeners sometimes generate interpretations that are inaccurate and distorted representations of the linguistic input, although these misrepresentations tend to be systematic and meaningful. In other words, people’s representations of a sentence are often “shallow” (Sanford & Sturt, 2002), but are “good-enough” (Ferreira, Bailey, & Ferraro 2002; Ferreira & Patson, 2007).

Fillenbaum (1971; 1974) provided a classic empirical demonstration by asking participants to paraphrase sentences such as “Get a move on or you will catch the bus.” Although the sentence should be paraphrased as “if you hurry up, then you will NOT catch the bus,” participants tended to normalize it to make it sensible and conventional, as in “if you hurry up, then you will catch the bus.” Similarly, Erickson and Mattson’s (1981) Moses’ Illusion shows that, when asked “How many animals of each kind did Moses put on the ark?”, a large number of participants quickly answered "two," when in fact the answer is zero, because it was Noah, not Moses, who put animals on the ark. Numerous other examples of misinterpretations are reviewed elsewhere (Sanford & Sturt, 2002). The common theme is that people often overlook what linguistic inputs specify
and instead rely on their world knowledge to (mis-)understand sentences, particularly when sentences are complex, semantically strange, or anomalous.

### 3.1.1 “Good Enough” Sentence Processing

Recently, Ferreira and colleagues (Christianson, Hollingworth, Halliwell, & Ferreira, 2001; Christianson, Williams, Zacks, & Ferreira, 2006; Ferreira, 2003; Ferreira, Bailey, & Ferraro, 2002; Ferreira, Christianson, & Hollingworth, 2001; Patson, Darowski, Moon, & Ferreira, 2009) claimed that sentence processing systems often generate representations that are *good enough* rather than completely detailed and accurate. They argued that underspecified representations are common in various areas of human cognition due to environmental demands and/or limitations of cognitive resources, and language processing is no exception. To examine the extent to which people’s interpretations are faithful to a sentence’s content, Ferreira and colleagues have conducted a series of experiments investigating the contents of people’s sentence representations.

For example, Christianson et al. (2001) had participants read sentences such as *While the man hunted the deer ran into the woods*. In the absence of a comma after *hunted*, the sentence is temporarily syntactically ambiguous because the noun phrase *the deer* is usually treated initially as the direct object of *hunted* but later turns out to be the subject of the main clause (when *ran* is encountered). It has been shown that such ambiguity leads to processing disruption, often measured in terms of elevated reading times for the main clause verb *ran* (Adams, Clifton, & Mitchell, 1998; Mitchell, 1987; Staub 2007; van Gompel & Pickering, 2001). Although it is commonly assumed that these elevated reading times reflect reanalysis, and that readers eventually reanalyze these sentences
successfully (or fail to do so completely, e.g., by deeming it to be ungrammatical), Christianson et al. examined these assumptions by asking participants comprehension questions such as “Did the deer run into the woods?” and “Did the man hunt the deer?” If reanalysis produced the correct interpretation, the answers should be ‘yes’ and ‘no’ respectively. Although participants answered the first question correctly nearly 90% of the time, they responded incorrectly approximately 50% of the time to the second question, and did so with relatively high levels of confidence (as indicated by confidence ratings). To test whether overall plausibility played a role in making incorrect interpretations, Christianson et al. used sentences such as While the man hunted the deer paced in the zoo. It is unlikely that the man was hunting the deer because two events presumably are taking place in different locations. Participants were less likely to incorrectly say ‘yes’ to “Did the man hunt the deer?” Notably, however, even with those sentences, participants were inaccurate 20% of the time (and 43% when the NP was longer, as in the deer that was brown and graceful). They also made errors 21% of the time (and 31% with a longer NP) when the sentence was not syntactically ambiguous, as in, While Bill hunted the pheasant the deer ran into the woods.

Christianson et al.’s (2001) results suggest that people’s representations of sentences are indeed partially incorrect, possibly because complete reanalysis of the sentence did not take place, or the initial incorrect interpretation was not completely abandoned. One alternative explanation suggests that the pattern is specific to the task. That is, it is not the representations that readers constructed during reading that are often incorrect; instead, the comprehension questions and the reevaluation process lured people into forming the incorrect representations. To test this explanation, Patson, Darowski, Moon, and Ferreira
(2009) replicated the study using a methodology in which participants were asked to paraphrase the sentences they read. Analyses indicated that participants retained partially incorrect analyses of the sentences (e.g., the man hunted the deer and it ran into the woods), more so when the sentences were ambiguous, but also when the sentence were unambiguous (69% and 38%, respectively). Thus, readers do seem to attempt to generate representations of sentences that are good enough.

Even more intriguing, a study by Ferreira (2003) suggests that people sometimes misinterpret sentences that are syntactically much simpler and unambiguous. One focus of her study was to investigate whether heuristics are used during comprehension. Specifically, Ferreira considered two kinds of heuristics, one based on plausibility, with the other being the so-called NVN strategy (Bever, 1970; Townsend & Bever, 2001). The NVN strategy is a general bias for any noun-verb-noun sentence structure to be interpreted as actor-action-object (i.e., agent-action-patient). In Ferreira (2003), participants listened to a number of simple active and passive transitive sentences and were asked to recall either the agent or patient (Experiment 1). Whereas participants were accurate with active sentences about 97% of the time, they were less so with passives (85%), particularly when the sentences described implausible events (80%, as in The dog was bitten by the man). They also took longer to provide responses following passives than actives (2071 vs. 1807 ms). Although higher error rates and longer response latencies with passives can be attributed to an atypical order of thematic roles (i.e., the NVN strategy yields an incorrect outcome), it is also possible that the results are due to the fact that passives are in general less frequent and more complex than actives. To explore this possibility, Ferreira used subject-clefts (e.g., It was the dog that bit the man;
Experiments 2 & 3) and object-clefts (e.g., *It was the man the dog bit; Experiment 3*). Both constructions are less frequent than passives and are complex, but maintain the same word order as do actives and passives respectively. Overall, passives and object-clefts were processed more slowly and less accurately than actives and subject-clefts, suggesting that sentences with atypical thematic role ordering are more difficult to process. Another important finding was that there was an overall tendency across experiments for implausible sentences to be more difficult, and this was particularly evident when the sentences contained atypical thematic role ordering. This suggests that people sometimes interpret implausible sentences according to what is more sensible in terms of their event-based knowledge (i.e., a man biting a dog is an unlikely event, whereas a dog biting a man happens more often). Given these results, Ferreira (2003) suggested that heuristic-based processes that rely on statistical probabilities such as structural biases (e.g., NVN strategy) and event knowledge (e.g., plausibility) are used and co-exist with syntax-based algorithmic processes. Whereas syntactic processes guarantee the correct solution eventually, they are often cognitively costly both in terms of time and resources under certain situations. Particularly given the noisiness of linguistic inputs in natural environments due to disfluencies, grammatical errors, and so on, complete reliance on algorithmic processes is not always optimal. Heuristics, on the other hand, are fast and are expected to generate the solutions with a reasonable margin of error. Interplay between the two produces the *good enough* representations that work most, but not all, of the time.

Although the results and interpretations of Ferreira’s (2003) findings are relatively straightforward, several critical issues remain. For example, assuming operation of the
two qualitatively different processes (i.e., syntax-based and heuristic-based), one important issue concerns their relative time courses. As pointed out by Ferreira, it is unclear “whether heuristics and algorithms are applied in parallel, or if one system is used when the other fails,” and “it will be important to determine how the outputs of the two systems are coordinated” (p. 197). In addition, it is not entirely clear whether or how heuristics are applied on a moment-to-moment basis during sentence processing. Another important question concerns the locus of misinterpretation. It is unclear how the use of heuristics can directly or indirectly cause occasional, but not too frequent, misinterpretations, and what other conditions or factors may lead to misinterpretation. It is also unclear whether misinterpretation occurs, for instance, on-line during incremental sentence processing (e.g., incorrect construction of partial representations, inattentiveness to important cues), or off-line at the time of answering questions. In the next section, these two issues are addressed in more detail, and several possible accounts are outlined.

3.1.2 Temporal Coordination of Heuristic-based and Syntax-based Processing

There are several possibilities regarding how heuristic-based processing operates in relation to syntax-based interpretation. First, the heuristic-based and syntax-based processes may operate in a serial manner such that syntactic processes first generate an interpretation and then heuristics support or compete with it. For example, heuristics may be employed only when syntax-based processing leads to difficulty generating interpretations. Second, the two processes may operate serially in the opposite order: “quick-and-dirty” heuristics operate first to generate a pseudo-parse that is later checked against the slower but more accurate syntax-based interpretation. This second possibility
is suggested by the Late Assignment of Syntax (LAST) model proposed by Townsend and Bever (2001). Alternatively, the heuristic-based and syntax-based processes operate in parallel, such that the interpretation that receives the strongest support, or the one that is completed first, is taken as the final interpretation.

Several ERP studies provide insight into this issue. Traditionally, in ERP studies of sentence comprehension, two ERP components are associated with semantic and syntactic processing. A negative-going component that usually peaks approximately 400 ms after the onset of a target word, known as the N400, is associated with semantic processing in that it is sensitive to violations of semantic congruency, expectation, or plausibility (see Kutas & Federmeier, 2011, for a review). A positive-going component that usually peaks approximately 600 ms post-stimulus onset, known as the P600, is associated with syntactic processing in that it is sensitive to violations of syntactic expectations. However, more recent studies have shown an unusual coupling of the type of violations and elicited ERP components, namely what is sometimes called the semantic or thematic P600 (Kim & Osterhout, 2005; Kuperberg, Sitnikova, Caplan, Holcomb, 2003; Kolk et al., 2003; Hoeks, Stowe, Doedens, 2004). For example, Kim and Osterhout (2005) investigated the ERP component generated at the verb in sentences such as The hearty meal was devouring the kids. In this type of sentence, syntax-based and plausibility-based interpretations conflict in that, whereas syntax unambiguously signals that the hearty meal is a subject (thus agent) of an active sentence, plausibility-based analyses suggest the hearty meal is an unlikely agent and likely patient of "devour." If syntax-based processing precedes semantic processing, an N400 is expected because the syntax-based interpretation is semantically implausible. On the other hand, if semantic
processing precedes syntax-based processing, then a P600 is expected because the input syntactically mismatches what is expected based on event-knowledge (i.e., *devouring* is the incorrect form of the verb). Kim and Osterhout found that a P600 rather than a N400 was elicited, as compared to control sentences such as *The hungry boy was devouring the cookies* and *The hearty meal was devoured by the kids*. They suggested that semantic-based processing is "in control" of initial sentence interpretation under certain circumstances.

One way to interpret these findings is that heuristic- (or semantic-) based processing operates prior to syntax-based processing, supporting the serial heuristic-first account as in the LAST model. Such an interpretation may be problematic for a few reasons. First, although the account assumes a straightforward linking of the N400 and P600 to semantic and syntactic processing respectively, this is a matter of ongoing debate (see Brouwer, Fitz, & Hoeks, 2012, for a review). Second, even if the linking assumption is correct, the results may be viewed as one instance of parallel operation in which heuristics provided an interpretation more quickly. Nevertheless, ERPs provide useful insight into the relative time course of syntax-based and heuristic-based processes.

### 3.1.3 The Locus of Misinterpretation

Although the studies discussed above illustrate the possibility that heuristics indeed operate, and how they may operate in relation to syntax-based processing, it is not clear how occasional misinterpretation might result from such an architecture. The key here is the word *occasional*, because it is the case that the use of heuristics does not always lead to misinterpretation. Take for example Ferreira’s (2003) Experiment 1 in which participants were less accurate with implausible reversible passive sentences (e.g., *The
dog was bitten by the man; 20% errors) than plausible reversible passive sentences (e.g., The man was bitten by the dog; 10% errors). Although the use of plausibility heuristics is a reasonable explanation for the difference between conditions, it does not explain why error rates increase by only 10%. It could be the case that people used heuristics 10% of the time, or it could be the case that people always used heuristics but there are other reasons that their representations occasionally are inaccurate. There are at least three possibilities of how misinterpretations may result. I discuss these possibilities below, particularly in relation to Ferreira (2003).

The first possibility is that misinterpretations result from the absence of conflict detection or a faulty or nonexistent monitoring process. In sentences such as The hearty meal was devouring the kids (Kim & Osterhout, 2004), if semantic-based processing initiates the interpretation that the hearty meal is the patient of the devour event, but the anomaly of such an interpretation based on syntax (was devouring rather than was devoured) is detected (as shown by a P600), it should result in correct interpretation. Logically, then, misinterpretations may occur when the syntactic anomaly was not detected (e.g., a lack of a P600). More generally, misinterpretation may result from the absence of conflict detection (whether implicit or explicit).

Some indirect support for this possibility comes from studies comparing persons with Broca’s aphasia and age-matched normal controls. Persons with Broca’s aphasia in many instances have syntactic comprehension difficulties. For example, given The girl kissed the boy and a semantically equivalent but syntactically different sentence such as The boy was kissed by the girl, some persons with agrammatic Broca’s aphasia have difficulty comprehending the latter, often misinterpreting it to mean the boy kissed the girl (Kearns,
In an ERP study, Wassenaar and Hagoort (2007) found that persons with Broca's aphasia make a greater number of picture-sentence matching errors than do normal controls due to a lack of on-line sensitivity to thematic role assignment, which is reflected in the absence of a signature ERP profile (larger P600 to the verb) that normal controls show. Although these are between-group differences, it is conceivable that similar differences in cortical activity may exist within participants on trials on which they correctly versus incorrectly interpret sentences.

In the context of Ferreira’s (2003) plausible and implausible passives, the fragments up to the verb (the man was bitten … and the dog was bitten) should be processed without conflict between syntax-based and plausibility-based processing (a man or a dog can be bitten). When the post-verbal noun phrase (the dog or the man) is encountered, semantic conflict should occur in the implausible sentences. Such semantic conflict is likely to produce a larger N400 for the implausible than plausible sentences as shown in a previous study using simple passives (Paczynski & Kuperberg, 2012). However, if the detection or resolution of such conflict does not occur, possibly because people glossed over or ignored the actual syntax, misinterpretations should result. In these cases, a larger N400 amplitude reflecting sensitivity to an atypical situation may be absent during on-line listening or reading of the implausible sentences.

The second possibility is that misinterpretation results due to faulty working memory. More specifically, multiple interpretations based on heuristics and faithful processing of the syntax are entertained, weighted, and retained in memory. Whether or not the correct interpretation can be retrieved when it is required (i.e., when answering questions or identifying the thematic role-fillers) is subject to working term memory, which is not
consistently reliable due to, for example, decay and/or interference. A similar account of how misinterpretation emerges has been suggested by Slattery, Sturt, Christianson, Yoshida, and Ferreira (2013). In their study, participants read sentences containing a temporary syntactic ambiguity, as in *While Frank dried off the truck/glass that was dark green was peed on by a stray dog*. Subsequently, they read a sentence that contained the same referent, as in *Frank quickly finished drying himself off then...* Whereas reading time measures on first sentences suggested that the initially constructed incorrect interpretation (i.e., Frank dried off the truck) was revised, reading time on second sentences was affected by the ambiguity (presence vs. absence of a comma) and plausibility (implausible: *the glass* vs. plausible: *the truck*) of the first sentence. This suggests that readers retained the incorrect interpretation of the first sentence and it interfered with the reading of the subsequent text. A few other studies (Kaschak & Glenberg, 2004; van Gompel, Pickering, Pearson, Jacob, 2006) also suggest that in the case of reading syntactically ambiguous sentences, initially constructed incorrect analyses can remain active in working memory.

Third, misinterpretations may occur as a result of inattentiveness to cues and/or to the task. For instance, the aforementioned Moses illusion can be seen as an example of this. Several follow-up studies of the illusion indicate that people are more likely to notice the distortion when the salience of the distortion is emphasized. This has been accomplished using it-clefts (*It was Noah who …*, Bredart & Modolo, 1988) and by printing NOAH in upper case (Kamas, Reder, & Ayers, 1996). Interestingly, in a self-paced reading study, Reder and Kusbit (1991) found that readers spent longer on the distortion word when they made an error (i.e., failed to notice the distortion) than when they were accurate.
Such findings at first seem counterintuitive because reading times typically are thought to reflect the amount of time spent encoding a word. However, more recent studies show an association between longer reading times and inattentiveness in terms of mind-wandering. Mind-wandering is a state in which people's attention is decoupled or zoned out from the external world, and often times, people engage in task-irrelevant thought (Smallwood & Schooler, 2006). In the case of reading texts, the phenomenon is known as mindless reading, in which the reader's mind is zoned out while her eyes continue to move across the text (Rayner & Fischer, 1996; Reichle, Reineberg, & Schooler, 2010; Sayette, Reichle, & Schooler, 2009; Schooler, Reichle, & Halpern, 2004; Smallwood, 2011; Vitu, O’Regan, Inhoff, & Topolski, 1995). Whether and when participants zone out is often determined by self-reports (e.g., press a button whenever they zoned out) or by answers to periodic probes (e.g., by asking if they have zoned out every 2-4 minutes).

An interesting finding is that mindless reading is negatively correlated with comprehension performance (Schooler, Reichle, & Halpern, 2004; Smallwood, McSpadden, & Schooler, 2008; but see, Sayette, Reichle, & Schooler, 2009), and is associated with more blinks and fewer fixations (Smilek, Carriere, & Cheyne, 2010). Feng, D’Mello, and Graesser (2013) reported that mind-wandering disproportionally affects comprehension performance on texts that differ in difficulty, where increases in errors during mind-wandering episodes were found only for difficult texts. They also showed that people are more like to engage in mindless reading when reading difficult texts. In term of ERP evidence, several studies in other areas have reported reduction of other ERP components during mind-wandering episodes (Smallwood, Beach, Schooler, Handy, 2008; Barron, Riby, Greer, & Smallwood, 2011). An ERP study by Sanford,
Leuthold, Bohan, and Sanford (2010) showed N400 effects that usually occur in response to semantic anomaly do not occur during comprehension of sentences with difficult-to-detect anomalies (e.g., the Moses’ illusion), which may be viewed as an instance of mindless reading.

In Ferreira (2003), it is possible that participants were more likely to engage in mindless reading of passive implausible sentences than of passive plausible or active sentences. Because there are greater conflicts between heuristics and syntax-based interpretations in passive implausible sentences (i.e., both NVN strategy and plausibility heuristic suggest the same incorrect outcome) compared to the other sentence types, reading passive implausible sentences could be more difficult or cognitively demanding. The higher cognitive demand or difficulty of passive implausible sentences make it more likely that people engage in mindless reading and make errors.

It should be noted that the three possibilities listed above are not exhaustive, and more importantly they are not necessarily mutually exclusive. For example, the third explanation can be seen as supplemental to the first because it is possible that occasional absence of conflict detection/resolution during online reading is due to mindless reading. It is also probable that different possibilities apply to different individuals or to the same individual at different time points. Thus, to uncover the source of misinterpretation, it is important to obtain multiple measurements that can be used to evaluate possible explanations. The experiment in this chapter was designed to address this.
3.1.4 The Present Study

To investigate the factors that contribute to misinterpretation of simple unambiguous sentences and whether misinterpretation results from on-line or off-line processes, the current experiment adopted Ferreira’s (2003) approach in which participants listened to simple transitive sentences and were tested on recall accuracy of either the agent or the patient. The current experiment differs from Ferreira (2003) in four major respects. Sentences were read rather than heard, only passives were used and thematic fit was manipulated more systematically, there were additional behavioral analyses, and ERPs were used to measure on-line sensitivity to implausibility.

First, the current study focused only on reversible passives. In Ferreira (2003), reversible passives produced a 21% error rate, and there was a clear effect of plausibility. In addition to swapping the two arguments, thematic fit of the agent and patient was manipulated more systematically to investigate whether the position at which the thematic-syntactic conflict occurs affects performance on the thematic role decision task, as in the following example set of sentences.

1. The patient was treated by the doctor before the ambulance arrived.
2. The patient was treated by the client before the ambulance arrived.
3. The clinician was treated by the doctor before the ambulance arrived.
4. The clinician was treated by the client before the ambulance arrived.

In the type of treating event that occurs in a hospital or a clinic, both a clinician and a doctor are good agents/poor patients, and both a patient and client are good patients/poor agents. Sentences (1) and (4) correspond to Ferreira’s (2003) plausible and implausible
reversible passives respectively. Sentences (2) and (3) are new, and are particularly interesting in that it cannot be determined on the basis of global plausibility which noun is a better agent versus patient of the specified event. If the thematic fit between the first argument and the verb is the key to misinterpretation, then (3) and (4) should be equally error-prone and result in a greater number of errors than (1) and (2). In contrast, if the thematic fit between the second argument and the verb is vital, then (2) and (4) should be equally more error-prone than (1) and (3). If thematic fit of both arguments and the verb are equally important, then (1) should be the easiest, (4) should be the most difficult, and (2) and (3) should be intermediary. It is also possible that relative difficulty of these sentence types depends on which argument participants are asked to recall.

Second, in addition to collecting response accuracy, several other behavioral measures were collected. Because participants’ vocal responses were recorded, I coded incorrect responses by error type to examine the characteristics of the types of errors people make. In addition, participants’ response confidence was measured using a 5-point confidence rating scale. If participants are as confident in making erroneous responses as they are in making correct responses, that would imply that they have constructed a single incorrect representation. If their confidence varies between correct and incorrect trials, that would indicate less complete or consistent representations. Similarly, mind-wandering probes were included at pseudo-random intervals to investigate how mindless reading relates to comprehension performance.

Finally, electroencephalogram (EEG) was recorded during word-by-word reading of each sentence to examine how sentences are processed on-line. Because the sentences containing implausible arguments (2-4) become implausible once the post-verbal noun
(N2) is read, the ERPs time-locked to the N2 were analyzed. Two sets of analysis were performed on the ERP data. The first set examined the effect of plausibility by comparing differences in mean amplitude at different time windows. This provides information regarding how each sentence type was processed. It is expected that the three types of sentences that contains implausible arguments (2-4) would elicit a larger N400 compared to the fully plausible sentence (1), although there might be a slight differences in size and timing of the component or existence of additional components. The second set of analyses examined the relation between the on-line sentence profile and off-line comprehension performance by comparing the ERPs from correct trials to incorrect trials. If the ERPs between two sets of trials are identical, it can be inferred that sentences are processed the same way in both sets of trials, and thus errors occur due to processes involved in responding to the memory probes.

### 3.2 Methods

#### 3.2.1 Participants

Forty-eight native English-speaking students from the University of Western Ontario participated for either course credit or for $20 compensation. The data from one participant was excluded from all analyses because accuracy was below 50%. In addition, three participants were excluded because of technical problems in recording the EEG. Thus, the data from the remaining 44 participants were included in all analyses. No participants reported a history of neurological or other mental issues.
3.2.2 Materials

Sixty sets of one verb and four nouns (e.g., treated, doctor, clinician, patient, client) were generated so that two nouns were typical agents (and poor patients) and the other two were typical patients (and poor agents) of the event denoted by the verb, as determined by the experimenter’s intuitions (see Appendix B for a complete list). For each set, four types of simple passive sentences were constructed as in (1) to (4) above. In P+A+ sentences (1), the initial NP denotes a typical patient of the event, and the second, post-verbal NP denotes a typical agent. A ‘+’ refers to the fact that the syntax and thematic fit are congruent. In P+A- (2), both NPs denote typical patients, but the syntactic cues signal the second NP to be the agent (thus the ‘A-’). In P-A+ (3), both NPs denote typical agents, but the syntactic cues signal the initial NP to be the patient. Finally, in P-A- (4), the initial NP denotes a typical agent and the second NP denotes a typical patient of the event, whereas the syntactic cues signal the opposite interpretations.

Eight sentences (two for each sentence type) were created from each word set. A total of 480 sentences (120 per type) were divided into four lists such that each verb appeared twice in each list but with pairs of non-overlapping nouns. In each list, agent probes ("DO-ER") followed half of the experimental items, and patient probes ("ACTED-ON") followed the other half. To counterbalance the type of item and probe, eight lists were created. Each participant saw only one of the lists. In each list, experimental stimuli were combined with 160 filler sentences (120 active sentences and 40 passives) that were followed by four types of probes: 40 asked about the main action in the sentence ("ACTION"), 40 asked about the location in which the described event took place ("LOCATION"), 40 asked about temporal information of the event ("WHEN"), and
remaining 40 asked about the color name that appeared in the sentence ("COLOR"). The resulting 280 items were presented in a pseudo-random order.

3.2.3 Procedure

Participants were told that they would read sentences and make decisions regarding them. They were informed using a sample sentence that there would be six types of decisions: DO-ER, ACTED-ON, ACTION, LOCATION, WHEN, and COLOR. Given *An old lady ordered a green tea Frappuccino at Starbucks this afternoon, an old lady* is the do-er, *a green tea Frappuccino* is the acted-on, *ordering* (or any other conjugation: *order* or *ordered*) is the action, *Starbucks* is the location, *this afternoon* is when the event took place, and *green* is the color. Once the participants understood the probe types and the rationale for each response, they were given 12 practice items (two for each of the six probe types) on a CRT monitor. If a participant answered more than four incorrectly, that person was re-informed about the probe types and retested on some of the practice items.

Each experimental session began with six practice items to familiarize participants with rapid serial visual presentation (RSVP) of a sentence. Each trial started with the presentation of a central fixation cross (+), followed by RSVP of a sentence in which each word/phrase of the sentence was presented one at a time in the center of the screen for 350ms, separated by a blank screen for 100 ms. Following each sentence, participants responded to the probe by speaking into two microphones: one recorded response latency (in ms) via an E-prime button box and another recorded the participant’s actual vocal responses for offline scoring. Participants also were asked to rate their confidence in making the judgment using a scale from 1 (not confident) to 5 (highly confident). For one third of the trials, Participants also were asked to rate their degree of inattentiveness to
the immediately preceding task using the scale of 1 (on-task) to 5 (off-task). The entire session took less than 2 hours (including approximately 40 min of EEG set-up and clean-up).

3.2.4 EEG Recording and Preprocessing

The continuous EEG signal was recorded using the BioSemi Active-Two system from 32 Ag-AgCl active electrodes mounted in an elastic electrode cap (Electrocap International) in accordance with the standard 10-20 system (see Figure 3-1). To record eye movements and blinks, four additional electrodes were used (above, beneath, and to the left of the left eye, and to the right of the right eye). As per the system’s design, two additional electrodes (Common Mode Sense active electrode and Driven Right Leg passive electrode) were used as reference and ground electrodes, respectively. Recordings were digitized on a PC using ActiView software (BioSemi) with a sampling rate of 512 Hz (a bandwidth of DC to 104 Hz, 3 dB/octave). The recorded EEG signals were re-referenced off-line to linked mastoids, and were band-pass filtered offline using an IIR filter (0.1–30 Hz, 24 dB/octave).

For each trial, ERPs were segmented in epochs time-locked to the onset of the N2, spanning from 200 ms pre-stimulus to 800 ms post-stimulus onset, and baseline-corrected relative to the 200 pre-stimulus baselines. Epochs containing ocular artifacts or excessive noise were identified and removed offline using a maximum voltage criterion of ±75 μV at any electrode. For the statistical analyses, the mean amplitudes from three

---

1 To be more precise, given the sampling rate of 512 Hz, the epoch spanned from 199.22 ms pre-stimulus onset to 800.78 ms post-stimulus onset.
time windows (100-300, 300-500, 500-700 ms)² were extracted. The three time windows were chosen based on previous studies using similar materials (Kim & Osterhout, 2005; Paczynski & Kuperberg, 2012), as well as on visual inspection of the data.

### 3.2.5 Statistical Analyses

Linear mixed effect (LME) models were used to analyze continuous data (e.g., ERP amplitudes) using the function *lmer()* from the *lme4* library in R version 3.0.1. Generalized linear mixed effect (GLME) models with a logit link function (Jaeger, 2008) were used to analyze binomial data (i.e., question-answer accuracy) using the *glmer()* functions from the *lme4* library. Both LME and GLME are variants of general linear models that include both fixed and random effects. These analysis methods were chosen over more traditional ANOVA for several reasons. For example, LME and GLME can be more robust than ANOVAs particularly when the data involves multiple and/or crossed random effects (e.g., both by-participant and by-item variability can be modeled simultaneously; Baayen, Davidson, & Bates, 2008). In addition, when analyzing binomial data such as question answering accuracy, GLME has been suggested as a better alternative to ANOVAs because the use of ANOVAs on such data is not only theoretically inappropriate as the data violate the assumptions of ANOVAs, but also has the danger of providing spurious results, whereas GLME does not (Jaeger, 2008).

---

² Given the sampling rate of 512 Hz, the precise windows were 99.61-300.78, 300.78-500.0, and 500.00-699.22 ms, respectively.
Figure 3-1. Electrode montage. Each group of electrodes connected by lines formed a region of interest.
Fitting of LME and GLME models generally followed a procedure in which, first, all the factors of interest and their interaction terms were entered into the model as fixed predictors. Then, the optimal random effect was determined through a series of iterative comparison procedures in which, starting from random by-subject (and by-item intercepts where applicable), progressively more complex models are compared against simpler models using likelihood ratio tests. For analyses of ERP data, electrodes were grouped into nine regions of interest (ROI; see Figure 3-1) based on a 3 laterality gradient (left, middle, and right) by 3 anteriority gradient (anterior, central, and posterior) grid over the scalp to simplify the model fitting procedure (Newman et al., 2012). The models included a by-participant intercept, condition, and ROI as crossed random effects. Inclusion of ROI as random effects enabled the model to capture individual variability in the effects across the scalp, as well as individual corrections among ROIs. Likelihood ratio tests were again used to justify the inclusion of ROI as random effects.

Statistics reported for each model vary and depend on how the categorical factors were coded. Whenever the models included only two-level categorical variables, the categorical variables were coded using contrast coding (i.e., one level was coded as -1 and the other as 1). When the model is fitted in this way, the coefficient of a fixed effect predictor corresponds to its main effect. In these cases, Wald-Z statistics (for GLME) or t statistics (for LME) are reported. When one of the predictors involved a categorical factor with more than two levels, the categorical variables were coded using treatment coding (i.e., one level as 1 and all the other level as 0). In these cases, F ratios associated with each factor as a whole are reported. In the LME models, determining the exact denominator degree of freedom (df) is complex (Bates, 2005). Thus, I adopted the
estimation of the \( df \) based on upper-bound value (number of data points minus the sum of all fixed effect numerator \( df \)s) and lower-bound value (upper-bound value minus the number of random effects), which are anti-conservative and conservative, respectively (Newman, et al., 2012; Tremblay & Ransijn, 2013). Only the lower-bound \( p \)-value is reported because the two values are similar in most cases in which the number of data points is large (as in the current experiment).

### 3.3 Results and Discussion

#### 3.3.1 Behavioral Data

For each trial, the following data were collected: vocal response to the question, time taken to respond to the question, and confidence (5 point scale). In addition, for one third of the trials, inattentiveness (i.e., mind-wandering) ratings (5 point scale) were collected.

The participants’ vocal responses were coded for accuracy by the experimenter as either correct or incorrect. Because the questions are open-ended, prior to providing their final response, participants often uttered their thought process loudly enough that it triggered the voice-activated relay in the button box. Because the precision of this measurement is of questionable quality, and because the decision time data are of secondary importance, this dependent variable was not analyzed.

#### 3.3.1.1 Overall Accuracy

Table 3-1 presents the percent accuracy for each sentence type in each question condition. Participants had the least difficulty with the fully plausible sentences (i.e., \(P+A+: 84\%\)) and the most difficulty with the fully implausible sentences (\(P-A-: 76\%\)). Accuracy on the two semi-implausible sentences fell in between (\(P+A-: 80\%; P-A+: \))
The overall accuracy across the four types of sentences was the same for patient decisions (80%) and agent decisions (80%).

To determine the influences of both sentence and question type on accuracy, GLME models were fit. The model included the following fixed factors: plausibility of first noun (N1 plausibility; plausible vs. implausible), plausibility of second noun (N2 plausibility; plausible vs. implausible), question type (ACTED-ON vs. DO-ER), and the corresponding two-way and three-way interaction terms. The model included random intercepts for participants and items. The random slopes parameters were entered into the model in a forward stepwise manner. The final model included random by-participants and by-item slopes for question type, as well as random by-item slopes for N2 plausibility. No other random effects or interactions significantly improved the fit.

Table 3-2 lists the fixed effects of the final model, including the estimated coefficient (b’s indicating change in log odds), standard error, Wald z-statistics, and p-values.

Participants were significantly more accurate when the initial noun of the sentence was plausible with respect to the verb (e.g., *The patient was treated ...*; 82%) than when it was implausible (e.g., *The clinician was treated... *; 77%). Likewise, participants were more accurate when the post-verbal noun was plausible (e.g., *... was treated by the doctor*; 81%) than when it was implausible (e.g., *... was treated by the client;* 78%). However, the effect of N2 plausibility interacted with question type. Simple main effects were investigated by refitting the models without the question type effects on each half of the data with different question types. There was no effect of N2 plausibility when the question asked about the initial noun (i.e., ACTED-ON). In contrast, when the question concerned the post-verbal noun (i.e., DO-ER), participants were significantly more
Table 3-1. Percent accuracy for each sentence type in each question condition.

<table>
<thead>
<tr>
<th>Sentence Type</th>
<th>Plausibility</th>
<th>Question Type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N1</td>
<td>N2</td>
</tr>
<tr>
<td>P+A+</td>
<td>Plausible</td>
<td>Plausible</td>
</tr>
<tr>
<td>P+A-</td>
<td>Plausible</td>
<td>Implausible</td>
</tr>
<tr>
<td>P-A+</td>
<td>Implausible</td>
<td>Plausible</td>
</tr>
<tr>
<td>P-A-</td>
<td>Implausible</td>
<td>Implausible</td>
</tr>
</tbody>
</table>
accurate when it was plausible. In other words, the plausibility of the post-verbal noun influenced performance only when the question concerned that noun. On the other hand, the lack of such an interaction between N1 plausibility and question type indicates that N1 plausibility affected performance on both questions regarding both nouns.

One of the notable differences between the results of the current study and Ferreira’s (2003) is that the overall accuracy in the current experiment (80%) was somewhat lower in comparison to Ferreira’s reversible passive sentences (85%). One possibility for this discrepancy is that the task in the current experiment was generally more difficult because the sentences were generally longer. In addition, there were slightly more items (120 experiment item in the current study vs. 72 items in Ferreira, 2003), which may have led to fatigue effects. Thus, sentence and experiment length may have lowered the overall accuracy. To investigate this possibility, sentence-length (specified as the number of separately presented phrases) and trial position were entered in a forward stepwise manner into the final GLME model described above (see Table 3-3). Sentence-length significantly affected performance in that the longer the sentences, the lower the accuracy. Sentence-length did not interact with any other variable as indicated by the lack of significant improvement in the models, $\chi^2(10) = 3.83, p > .95$, and thus the interaction terms were removed. This confirms the notion that the lower overall accuracy in the current experiment is at least partly due to sentence length.
Table 3-2. The fixed effect statistics for the final model for overall accuracy, including the estimated coefficient, standard error, Wald Z-statistics, and p-values.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>b</th>
<th>SE</th>
<th>Wald Z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Model with N1 plausibility, N2 plausibility and Question type</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Intercept)</td>
<td>1.593</td>
<td>0.1356</td>
<td>11.75</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>N1</td>
<td>-0.1751</td>
<td>0.0382</td>
<td>-4.59</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>N2</td>
<td>-0.1303</td>
<td>0.0381</td>
<td>-3.42</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Question</td>
<td>0.0079</td>
<td>0.0527</td>
<td>0.15</td>
<td>ns</td>
</tr>
<tr>
<td>N1:N2</td>
<td>0.0203</td>
<td>0.0382</td>
<td>0.53</td>
<td>ns</td>
</tr>
<tr>
<td>N1:Question</td>
<td>-0.0327</td>
<td>0.0381</td>
<td>-0.86</td>
<td>ns</td>
</tr>
<tr>
<td>N2:Question</td>
<td>0.1056</td>
<td>0.0381</td>
<td>2.77</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>N1:N2:Question</td>
<td>-0.0001</td>
<td>0.0381</td>
<td>-0.002</td>
<td>ns</td>
</tr>
<tr>
<td>B. Model A plus Trial position and Sentence Length</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Intercept)</td>
<td>1.6620</td>
<td>0.1387</td>
<td>11.98</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>N1</td>
<td>-0.1747</td>
<td>0.0399</td>
<td>-4.38</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>N2</td>
<td>-0.1133</td>
<td>0.0399</td>
<td>-2.84</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Question</td>
<td>-0.0102</td>
<td>0.0555</td>
<td>-0.18</td>
<td>ns</td>
</tr>
<tr>
<td>Trial</td>
<td>0.0054</td>
<td>0.0005</td>
<td>10.84</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Length</td>
<td>-0.2500</td>
<td>0.0491</td>
<td>-5.09</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>N1:N2</td>
<td>0.0379</td>
<td>0.0399</td>
<td>0.95</td>
<td>ns</td>
</tr>
<tr>
<td>N1:Question</td>
<td>-0.0425</td>
<td>0.0399</td>
<td>-1.07</td>
<td>ns</td>
</tr>
<tr>
<td>N2:Question</td>
<td>0.0785</td>
<td>0.0399</td>
<td>1.97</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>N1:Trial</td>
<td>0.0009</td>
<td>0.0005</td>
<td>1.72</td>
<td>= .09</td>
</tr>
<tr>
<td>N2:Trial</td>
<td>0.0007</td>
<td>0.0005</td>
<td>1.47</td>
<td>ns</td>
</tr>
<tr>
<td>Question:Trial</td>
<td>-0.0004</td>
<td>0.0007</td>
<td>-0.57</td>
<td>ns</td>
</tr>
<tr>
<td>N1:N2:Question</td>
<td>0.0105</td>
<td>0.0398</td>
<td>0.26</td>
<td>ns</td>
</tr>
<tr>
<td>N1:N2:Trial</td>
<td>0.0011</td>
<td>0.0005</td>
<td>2.30</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>N1:Question:Trial</td>
<td>0.0001</td>
<td>0.0005</td>
<td>0.10</td>
<td>ns</td>
</tr>
<tr>
<td>N2:Question:Trial</td>
<td>-0.0017</td>
<td>0.0005</td>
<td>-3.38</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>N1:N2:Question:Trial</td>
<td>0.0007</td>
<td>0.0005</td>
<td>1.32</td>
<td>ns</td>
</tr>
</tbody>
</table>
Trial position was a significant predictor of the performance. However, the pattern is the opposite of fatigue effects in that participants’ accuracy improved as the experiment progressed (see Figure 3-2). More interestingly, there was a significant three-way interaction among N1 plausibility, N2 plausibility and trial position (see Figure 3-3), as well as among N2 plausibility, question type, and trial position (see Figure 3-4). The former interaction is due to the fact that participants’ performance improved over time in such a way that the difference in performance due to N1 and N2 plausibility becomes minimal. The latter interaction suggests that the two-way interaction between N2 plausibility and question type discussed earlier - that the effect of N2 is present only when the question was on the second noun - diminished as the experiment progressed.

The main effect and interactions regarding trial position suggest not only that participants adapted to the task as trials progressed, but also that there are several ways in which they adapted. First, it is likely that the improvement for fully plausible sentences (P+A+) is due to adaptation of syntactic-thematic expectation (i.e., NVN strategy). Passive sentences are generally less common in English (less than 10 % in several corpora; Roland & Jurafsky, 2002), as is object-action-actor ordering. However, because slightly more than half of the items were passive constructions, participants may have updated their syntactic-thematic expectation to match the probabilities that exist in the experiment. Second, whereas such a syntactic-thematic adaptation may be present equally for all sentence types, the difference in improvement (i.e., slope) between conditions seems to vary as a function of plausibility. Thus, it is possible that the improvement on those sentences is due not only to syntactic adaptation but also to adaptation or desensitization of event-knowledge-based expectations.
Figure 3-2. Mean accuracy for each sentence type and question type as a function of time. The upper panel is based on observed accuracy for each block and the lower panel is based on the model’s predicted probability for varying trial positions.
Figure 3-3. Mean accuracy and predicted probability for each sentence type as a function of time. The upper panel is based on observed accuracy for each block and the lower panel is based on the model’s predicted probability for varying trial positions.
Figure 3-4. Mean accuracy and predicted probability for each N2 plausibility level for question type as a function of time.
Another notable difference between the current study and that of Ferreira (2003) is that whereas Ferreira found that participants performed better on patient questions (89%) than on agent questions (81%), there was no main effect of question type in the current experiment. On the other hand, question type interacted with N2 plausibility. The presence of this interaction (and the absence of N1 plausibility × question type interaction) indicates that whereas N1 plausibility affected the performance regardless of the question type, N2 plausibility affected the performance on only the questions relevant to it. This suggests that the memory representations of the first and second arguments have a different status. More specifically, it is possible that N1 may play a more dominant role in the construction of a sentential representation, which is essentially the explanation that Ferreira provided for the effect of question type in her study. Thus, the current study and Ferreira’s captured slightly different outcomes of the same underlying process.

3.3.1.2 Errors by Type

Because the thematic role decision task uses open-ended questions, participants could make errors in several ways. To investigate whether error type provides interesting insights, the experimenter and a research assistant transcribed each response and then coded each incorrect response as one of four error types: thematic-analysis errors, schema-driven errors, memory-retrieval errors, and miscellaneous. A thematic-analysis error occurred when participants provided the agent rather than the patient, and vice versa (e.g., identifying the clinician as DO-ER of sentence The clinician was treated by the client...). A schema-driven error occurred when participants produced a noun that was not in the sentence, but is related to the described event (e.g., responding with “guard” given
The prisoner was punished by the officer...). A memory-retrieval error refers to participants either being unable to provide the answer within the time limit, or providing “I don’t know” or a similar response. Finally, when participants seemed to misread the cues in some way, as in reading ACTED-ON as ACTION, and produced the verb instead of a noun, this was coded as miscellaneous. Because there were only 13 miscellaneous errors in total, they were not included in the analyses. There are two issues of interest. The first concerns whether error type differs as a function of plausibility and question type. The second concerns whether the frequency of each error type decreases over time.

The most common type of error in terms of raw frequency counts was thematic-analysis (\(M = 11.84, SD = 9.73, median = 9.0\)), followed by schema-driven (\(M = 6.68, SD = 4.14, median = 5.5\)), and then memory-retrieval errors (\(M = 5.57, SD = 6.78, median = 2.0\)). Thus, approximately half of all errors were thematic-analysis errors (46%), followed by schema-driven (30%) and memory retrieval errors (23%). To investigate whether the occurrence of each error type varied as a function of plausibility and question type, mixed-effects Poisson regression models were fit separately for each error type count. In all three cases, N1 and N2 plausibility, question type, and the two-way and three-way interaction terms as fixed factors, as well as a random by-participant intercept, were entered into the models. Only thematic-analysis errors varied significantly across conditions. They were more frequent when N1 was implausible than when it was plausible, Wald \(Z = 4.13, p < .001\), and when N2 was implausible than when it was plausible, Wald \(Z = 5.13, p < .001\).

To test whether the frequency of each error type changed throughout experiment, average frequency counts per block (of 70 trials including fillers) for each error type was
calculated (see Figure 3-5). Although the frequency of thematic-analysis and schema-driven errors decreased over time, memory retrieval errors remained relatively constant.

### 3.3.1.3 Confidence Ratings

Participants were overall highly confident in their responses, with a mean confidence rating of 4.04 ($SD = 0.53$) on a 5-point scale, where 1 corresponds to *not at all confident* and 5 correspond to *highly confident*. Clearly, participants were more confident after responding correctly ($M = 4.38$, $SD = 0.47$) than incorrectly ($M = 2.61$, $SD = 0.70$). This finding contrasts with those of Christianson et al. (2001, 2006), in which participants were highly confident in both their correct and incorrect responses. One possibility for this difference is that the error trials in the current experiment include those in which participants failed to provide responses (e.g., “I don’t know/remember” responses), for which their confidence is quite low. Table 3-3 presents mean confidence ratings for each condition, split into correct and incorrect trials, and further into each error type.

Confidence ratings for the memory errors are much lower ($M = 1.26$, $SD = 0.43$, $n = 31$) than for thematic-analysis errors ($M = 3.06$, $SD = 0.81$, $n = 39$) or schema-driven errors ($M = 2.83$, $SD = 0.84$, $n = 38$). Nevertheless, there is clear difference in confidence between correct and incorrect trials, suggesting that participants were at least partially aware when their answers were wrong.
Figure 3-5. Mean frequency count per block for each error type.
Table 3-3. Mean confidence ratings for each condition, separated by correct versus incorrect response and error type.

<table>
<thead>
<tr>
<th>Question Type</th>
<th>Overall</th>
<th>By Accuracy</th>
<th>By Error Type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Correct</td>
<td>Incorrect</td>
</tr>
<tr>
<td>Sentence Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ACTED-ON</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P+A+</td>
<td>4.12</td>
<td>4.41</td>
<td>2.52</td>
</tr>
<tr>
<td>P+A-</td>
<td>3.99</td>
<td>4.23</td>
<td>2.69</td>
</tr>
<tr>
<td>P-A+</td>
<td>3.89</td>
<td>4.27</td>
<td>2.65</td>
</tr>
<tr>
<td>P-A-</td>
<td>4.08</td>
<td>4.38</td>
<td>2.97</td>
</tr>
<tr>
<td><strong>DO-ER</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P+A+</td>
<td>4.21</td>
<td>4.49</td>
<td>2.64</td>
</tr>
<tr>
<td>P+A-</td>
<td>3.93</td>
<td>4.29</td>
<td>2.32</td>
</tr>
<tr>
<td>P-A+</td>
<td>4.09</td>
<td>4.50</td>
<td>2.43</td>
</tr>
<tr>
<td>P-A-</td>
<td>4.04</td>
<td>4.39</td>
<td>2.83</td>
</tr>
</tbody>
</table>
3.3.1.4 Mind-wandering

Due to technical problems, 17 out of 44 participants’ inattentiveness ratings were not recorded. Thus, the results are based on 27 participants, which amounts to 1080 data points because there were 40 randomly chosen trials for each participant that were followed by mind-wandering probes. On average, participants were well-focused, with mean inattentiveness ratings of 1.98 (SD = 0.87) where 1 corresponds to attentional focus being on-task and 5 correspond to off-task. Linear mixed effect models with a full factorial design of N1 plausibility, N2 plausibility, and Question type, and by-participant and by-item random intercept showed that the inattentiveness ratings did not differ across conditions, |t|s| < 1.7.

There was a tendency for participants to be less accurate as their focus became off-task (Figure 3-6). To examine whether the degree of attentional focus on the task can predict errors, GLME models were fit. First, to test whether the pattern of effects in the smaller subset (i.e., the 1080 trial on which the inattentiveness ratings were obtained) deviated from the overall pattern discussed previously, the model with the same structure and procedures described in the overall accuracy section were fit to the smaller subset of the data. In this model, the significant predictors were N1 plausibility, trial position, sentence-length, and the interaction between N2 plausibility and trial position. Then, the mean centered inattentiveness ratings were entered into the model as an additional fixed effect. Its addition significantly improved the model’s fit, $\chi^2(1) = 52.08, p < .001$, without changing the overall pattern of the other predictors’ effects. The inattentiveness ratings, however, did not interact with other factors, $\chi^2(15) = 9.45, p > .85$. The model predicts
that accuracy is lower as the attentional focus becomes off task, \( b = -0.71, SE = .09, \) Wald \( Z = -7.43, p < .001. \)

The results thus confirm that misinterpretations occur partly as a result of inattentiveness. However, inattentiveness does not seem to explain differential error patterns observed across sentence types, as seen by the lack of interactions. Thus, unlike previous studies showing the relation between mind-wandering and sentence comprehension (Feng, D’Mello, & Graesser, 2013), task difficulty or complexity of items in the current experiment did not vary sufficiently to induce differing frequency of inattentiveness or differing comprehension accuracy across conditions.

### 3.3.2 ERP Data

Two sets of analyses were conducted on the ERP data. The first set involved assessing plausibility effects at the post-verbal NP, in order to examine whether readers detected implausibility online during reading. The second set of analyses examined the difference in online processing between correct and incorrect trials. If online processing relates to misinterpretation, the pattern of activity should differ between correct and erroneous trials. More specifically, it is possible that the N400 is smaller for error than for correct trials.
Figure 3-6. Mean accuracy for each level of inattentiveness rating.
3.3.2.1 The Effects of Plausibility

Based on previous studies showing plausibility-based N400 effects in passive sentences (Paczynski & Kuperberg, 2012), it was expected that sentences containing implausible role fillers (i.e., P+A-, P-A+, P-A-) would elicit larger negativities at the post-verbal noun in the 300-500 ms window and possibly in the earlier 100-300 ms window, compared to the fully plausible sentences (i.e., P+A+). Following any significant effect of sentence type or a sentence type by ROI interaction, three comparisons were conducted by contrasting the fully plausible sentences (P+A+) against each of the remaining three sentence types. First, a local plausibility contrast investigated the effect of N2 plausibility by comparing the P+A- to the P+A+ sentences (e.g., The patient was treated by the _client vs. doctor; see Figure 3-7). Second, a distant plausibility contrast investigated the effect of N1 plausibility at the N2 by comparing P-A+ and P+A+ sentences (e.g., The _clinician vs. patient was treated by the _doctor; see Figure 3-8). Third, a global plausibility contrast compared the fully implausible versus fully plausible sentences (P-A- vs. P+A+; _The clinician was treated by the _client vs. _The patient was treated by the _doctor; see Figure 3-9). In all cases, because the target word mismatched with the earlier part of the sentences semantically, a larger N400 amplitude relative to the P+A+ conditions was expected. The voltage maps showing each contrast for each time window are presented in Figure 3-10.
Figure 3-7. Grand average ERP waveforms for the local plausibility contrast comparing the P+A- to the P+A+ sentences (e.g., *The patient was treated by the client* vs. *doctor*). For each ROI, ERPs for sets of electrodes are averaged. The x-axis represents the time from -200 ms pre-stimulus onset to 800 ms post-stimulus onset, ticked every 100 ms.
Figure 3-8. Grand average ERP waveforms for the distant plausibility contrast comparing P-A+ and P+A+ sentences (e.g., The clinician vs. patient was treated by the doctor). For each ROI, ERPs for sets of electrodes are averaged. The x-axis represents the time from -200 ms pre-stimulus onset to 800 ms post-stimulus onset, ticked every 100 ms.
Figure 3-9. Grand average ERP waveforms for the global plausibility contrast comparing the P-A- to the P+A+ sentences (*The clinician was treated by the client* vs. *The patient was treated by the doctor*). For each ROI, ERPs for sets of electrodes are averaged. The x-axis represents the time from -200 ms pre-stimulus onset to 800 ms post-stimulus onset, ticked every 100 ms.
Figure 3-10. Voltage maps across each time window for each plausibility contrast showing difference in ERPs to each sentence type (local = P+A-, distant = P-A+, and global = P-A-) relative to ERPs to the fully plausible sentences (P+A+). Scale is -2μV (blue) to +2μV(yellow).
In all three time windows, there was a significant sentence type × ROI interaction: 100-300 ms, \( F(24, 5024) = 2.33, p < .001 \); 300-500 ms, \( F(24, 5024) = 4.42, p < .001 \); and 500-700 ms, \( F(24, 5024) = 1.90, p < .001 \). There was also a significant main effect of sentence type in all time windows: 100-300 ms, \( F(3, 5024) = 3.05, p < .05 \); 300-500 ms, \( F(3, 5024) = 3.14, p < .05 \); and 500-700 ms, \( F(3, 5024) = 3.12, p < .05 \). Following the significant interaction, three planned comparisons were conducted in each time window (see Table 3-4 for the statistics).

In the local plausibility contrast (P+A- vs. P+A+; e.g., *The patient was treated by the client* vs. *doctor*; see Figure 3-7), there were clear N400 differences in the middle central to posterior regions, which is consistent with previous studies showing plausibility/semantic congruency effects. The P+A- sentences showed significantly greater negativities than the P+A+ sentences at all posterior regions as well as the middle central region at 300-500 ms. In addition there were long lasting positivities in right posterior regions, which were marginal at the 100-300 ms window, and significant at the 300-500 and 500-700 ms windows.

In the distant plausibility contrast (P-A+ vs. P+A+; e.g., *The clinician vs. patient was treated by the doctor*; see Figure 3-8), there were again clear N400 differences, although they seemed to start earlier and last longer. The P-A+ sentences showed significantly greater negativities than did the P+A+ sentences in the middle central-posterior regions at the 100-300 ms window, in all central to posterior regions at 300-500 ms, and middle posterior regions at 500-700 ms. There were marginal differences in middle anterior region at 300-500 ms, and in middle central and left posterior regions at 500-700 ms.
Table 3-4. The t-values and their statistical significance for each contrast in each ROI.

<table>
<thead>
<tr>
<th>Contrast</th>
<th>Time Windows</th>
<th>ROI</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>100-300</td>
<td>300-500</td>
<td>500-700</td>
<td></td>
</tr>
<tr>
<td><strong>Local Plausibility Contrast:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P+A- vs. P+A+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>left anterior</td>
<td>-0.02</td>
<td>-0.75</td>
<td>-1.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>left central</td>
<td>-2.06*</td>
<td>-2.52*</td>
<td>-1.36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>left posterior</td>
<td>-1.96*</td>
<td>-3.00*</td>
<td>-0.99</td>
<td></td>
<td></td>
</tr>
<tr>
<td>middle anterior</td>
<td>-0.80</td>
<td>-1.15</td>
<td>-1.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>middle central</td>
<td>-3.06*</td>
<td>-3.84*</td>
<td>-1.74†</td>
<td></td>
<td></td>
</tr>
<tr>
<td>middle posterior</td>
<td>-2.79*</td>
<td>-3.98*</td>
<td>-1.87†</td>
<td></td>
<td></td>
</tr>
<tr>
<td>right anterior</td>
<td>0.32</td>
<td>0.34</td>
<td>-0.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>right central</td>
<td>-1.14</td>
<td>-1.42</td>
<td>-0.82</td>
<td></td>
<td></td>
</tr>
<tr>
<td>right posterior</td>
<td>-1.90†</td>
<td>-2.67*</td>
<td>-1.38</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Distant Plausibility Contrast:</strong></td>
<td>P+A- vs. P+A+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>left anterior</td>
<td>-0.38</td>
<td>-1.44</td>
<td>-0.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>left central</td>
<td>-1.88†</td>
<td>-2.95*</td>
<td>-1.18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>left posterior</td>
<td>-1.80†</td>
<td>-2.96*</td>
<td>-1.85†</td>
<td></td>
<td></td>
</tr>
<tr>
<td>middle anterior</td>
<td>-1.01</td>
<td>-1.88†</td>
<td>-0.57</td>
<td></td>
<td></td>
</tr>
<tr>
<td>middle central</td>
<td>-2.58*</td>
<td>-3.62*</td>
<td>-1.87†</td>
<td></td>
<td></td>
</tr>
<tr>
<td>middle posterior</td>
<td>-2.50*</td>
<td>-3.20*</td>
<td>-2.50*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>right anterior</td>
<td>-0.04</td>
<td>-0.59</td>
<td>0.93</td>
<td></td>
<td></td>
</tr>
<tr>
<td>right central</td>
<td>-1.43</td>
<td>-2.07*</td>
<td>-0.47</td>
<td></td>
<td></td>
</tr>
<tr>
<td>right posterior</td>
<td>-2.46*</td>
<td>-2.37*</td>
<td>-1.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Global Plausibility Contrast:</strong></td>
<td>P-A- vs. P+A+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>left anterior</td>
<td>0.64</td>
<td>-0.40</td>
<td>1.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>left central</td>
<td>-0.38</td>
<td>-1.92†</td>
<td>0.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>left posterior</td>
<td>-0.32</td>
<td>-2.44*</td>
<td>0.11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>middle anterior</td>
<td>0.92</td>
<td>-0.23</td>
<td>1.64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>middle central</td>
<td>-0.21</td>
<td>-2.22*</td>
<td>0.95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>middle posterior</td>
<td>-0.69</td>
<td>-2.92*</td>
<td>-0.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>right anterior</td>
<td>1.94†</td>
<td>1.87†</td>
<td>2.98*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>right central</td>
<td>0.46</td>
<td>-0.76</td>
<td>0.88</td>
<td></td>
<td></td>
</tr>
<tr>
<td>right posterior</td>
<td>-0.80</td>
<td>-1.72†</td>
<td>0.29</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: * = p < .05, † = p < .09
Finally, the global plausibility contrast (P-A- vs. P+A+; e.g., *The clinician was treated by the client* vs. *The patient was treated by the doctor*; see Figure 3-9) also showed clear N400 differences, this time starting early but not lasting long. The P-A- sentences elicited significantly greater negativities than did the P+A+ sentences in the central-posterior regions on the left and the midline, as well as right posterior regions at 100-300 ms and 300-500 ms. At 500-700 ms, there were only marginal negative differences in middle central-posterior regions.

These three contrasts indicated that when participants processed sentences correctly, they detected the implausible pairing of the role fillers in the implausible sentences, and this elicited a larger N400 to the implausible sentences compared to the plausible sentences. One observation that is worth noting is that in the local plausibility contrast (i.e., P+A- vs. P+A+ as in *The patient was treated by the client* vs. *doctor*), but not in other two contrasts, there was an anterior positivity that was significant in the right anterior region in 300-500 and 500-700 ms time windows. Several previous studies using similar materials have shown positive differences between implausible and plausible materials, but the effects were usually distributed in central parietal regions in later time windows than N400, corresponding to P600 (Kim & Osterhout, 2005; Paczynski & Kuperberg, 2012; van Herten, Chwilla, & Kolk, 2006). Because the positivity found here differs both in terms of timing and scalp distribution, and also because the items in current study do not involve animacy violations, it is unlikely that the positivity found here reflects the same process.

A pattern of EPRs more similar to the pattern in the local plausibility contrast has been observed, for example, by Delong, Urbach, Groppe, and Kutas (2011). They found a
larger N400 amplitude for less predictable targets compared to highly predictable targets given the same contexts leading up to the target words. The N400 difference was distributed widely but particularly pronounced in central-parietal regions. In addition, they found greater positivity to low predictability than high predictability items in anterior sites, starting in the same time window as N400 and extending. They also observed that both the N400 amplitudes and the amplitude of positivity were correlated with predictability (i.e., cloze probability) of items, although each correlation was maximal at slightly different time points (N400 were earlier). Delong et al. suggest that the positivity indexes sentential constraint violation and/or expectancy violation, whereas N400 index semantic activation and/or integration.

Although the current study did not manipulate cloze probability or predictability in a systematic way, it is conceivable that the combination of the plausible patient and verb (e.g., \textit{The patient was treated}) resulted in expectations that mismatched the implausible agent (\textit{client}), resulting in greater anterior positivity in the P+A- sentences. In other contrasts, however, the combination of an implausible patient and verb (e.g., \textit{The clinician was treated}) is unlikely to provide sufficiently constraining expectations (because clinicians can indeed be treated), therefore no differences in the positivity. Indeed, the P-A+ sentences and the P-A- sentences differed in N400 amplitudes (in middle parietal regions, $t=-2.10$, $p < .05$, and a marginal difference in the middle central region, $t = -1.7$, $p < .09$, when the window were narrowed to 350-450 ms; not significant in the 300-500 ms window) but not in positivities. Thus, although participants may have found the P-A+ sentences to be more acceptable than the P-A- ones (i.e., \textit{doctor} is a more reasonable continuation than \textit{client} of \textit{The clinician was treated by}), the two sentences
were not sufficiently constraining to generate strong expectancy violations as indexed by the positivity.

To summarize, when participants processed sentences correctly, they were sensitive to the implausibility of the target words with respect to the earlier part of sentences, and they also seem to show a sign of expectancy-driven processing during reading.

### 3.3.2.2 Differences between Correct and Incorrect Trials

To investigate whether on-line processing differs between correct and incorrect trials, the ERPs at the post-verbal nouns (N2) were compared among three conditions: plausible-correct, implausible-correct, and implausible-incorrect. The plausible-correct condition consists of the trials with correct responses for the fully plausible sentences (P+A+) only. The implausible-correct condition consists of the trials with correct responses for the remaining three sentence types (i.e., P-A-, P+A-, P-A+), and the implausible-incorrect condition consists of the trials with incorrect responses for P-A-, P+A-, and P-A+. The three sentence types were pooled for two reasons. First, the previous analyses clearly showed greater negativity in all three implausible sentence types compared to the fully plausible sentences. Thus, the differences in the ERP amplitude between the implausible sentences averaged together and the plausible sentences should reflect the common underlying processes associated with the implausibility. Second, incorrect trials for individual sentence types did not occur with sufficient frequently to provide interpretable ERP signals for individual conditions, but pooling the three sentence types did. That is, because a relatively large number of trials are required for ERP analyses, it was not possible to analyze ERPs by sentence type for error trials only.
Visual inspection of the averaged ERPs suggests that there is a clear difference between the plausible-correct and the implausible-correct conditions in central-posterior sites (see Figure 3-11). The implausible-correct condition shows greater negativity in central parietal sites compared to the implausible-incorrect condition (see Figure 3-12), and the ERPs for the implausible-incorrect condition are similar to the plausible-correct condition (see Figure 3-13). To confirm these observations, LME models were fit for each comparison on mean amplitudes at each of the three time windows (100-300, 300-500, 500-700 ms). Significant interactions between the condition effect and ROI were followed up with post-hoc comparisons of the condition effect in each ROI.

A comparison between the implausible-correct and plausible-correct conditions revealed a significant plausibility × ROI interaction in all three time windows: 100-300 ms, $F(8, 2314) = 5.28, p < .001$, 300-500 ms, $F(8, 2314) = 9.44, p < .001$, and 500-700 ms, $F(8, 2314) = 3.16, p < .001$. There was also a significant main effect of plausibility at 300-500 ms, $F(1, 2314) = 6.95, p < .001$. The interaction at 100-300 ms indicated that implausible-correct sentences elicited significantly greater negativity than plausible-correct ones in the middle central, middle posterior and right posterior regions ($t = -2.16, -2.38, -2.04$, respectively, all $p < .05$). The interaction in the 300-500 ms window also showed significantly greater negativity in implausible-correct sentences in the same regions, as well as in the left central and posterior regions ($t = -3.87, -4.09, -2.74, -2.97, -3.36$, respectively, all $p < .05$). The interaction at 500-700 ms was due to a non-significant positive difference in the right anterior sites whereas all other sites showed non-significant negative differences.
Figure 3-11. Grand average ERP waveforms for plausible-correct condition and implausible-correct condition.
Figure 3-12. Grand average ERP waveforms for implausible-correct condition and implausible-incorrect condition.
Figure 3-13. Grand average ERP waveforms for plausible-correct condition and implausible-incorrect condition.
Comparisons between the implausible-correct and implausible-incorrect conditions showed a significant main effect of plausibility at 100-300 ms, $F(1, 2314) = 5.26, p < .05$, and a significant plausibility × ROI interaction at 300-500 ms, $F(8, 2314) = 7.17, p < .001$, and 500-700 ms, $F(8, 2314) = 2.61, p < .001$. The significant main effect at 100-300 ms suggests that there was greater negativity throughout the scalp for implausible sentences paired with correct responses. The interaction at 300-500 ms indicated significantly more negativity for correct response items in the left-to-right posterior regions ($t = -2.52, -2.71, -2.49$, respectively, all $p < .05$), and marginally in the middle central region ($t = -1.88, p = .06$). The interaction at 500-700 ms was due to non-significant differences between conditions varying in polarity across ROI (positive difference in right anterior-central region and negative elsewhere).

Finally, implausible-incorrect condition did not differ from plausible-correct condition in any of the time windows. That is, the ERP patterns suggest that when participants made recall errors following implausible sentences, they had processed the sentences in the same way as fully plausible sentences. Thus, this lack of differences suggests that participants did not register the implausibility, which resulted in errors on the probe task. On the other hand, there were numerous ERP differences between implausible sentences for correct versus error trials.

These patterns of results confirm the idea that one important source of misinterpretation can be found during on-line sentence processing, and it is reflected in the absence of a differential N400. One way to interpret the results is in terms of expectancy generation. In the correct implausible trials, participants generated expectations for upcoming concepts which were subsequently disconfirmed by the implausible second (actual agent)
noun phrases. Such disconfirmation and subsequent updating of sentential interpretation may have strengthened the memory representation of the sentences, resulting in better accuracy in identifying the thematic role fillers of the sentences. On the other hand, there was no such disconfirmation in the incorrect implausible trials, presumably because readers did not successfully combine syntactic and semantic information to generate expectancies.

3.4 Summary and Conclusion

The goal of Chapter 3 was to examine the notion that people’s representations of sentences are often good enough, such that people sometimes misinterpret sentences that are as simple as the dog was bit by the man (Ferreira, 2003). More importantly, the novel contribution of the current study lies in its detailed investigation of the factors or processes that separate successful from unsuccessful comprehension, and whether such factors or processes can be revealed during online processing of sentences, using ERPs.

Adopting the experimental paradigm of Ferreira (2003) in which participants listened to (read in the current study) sentences and later recalled one of the thematic role fillers, the current study found that implausibility of sentences (or arguments) increased the probability of incorrect recall. In addition, how well one can recall the particular role filler of a sentence in part depended on the plausibility of that argument with respect to the verb. Furthermore, the plausibility of the first argument influenced recall performance for both the first and second arguments, suggesting that a sentential representation is constructed incrementally with the first argument and the verb playing a more dominant role in memory formation. These results are in accord with those of Ferreira (2003), and suggest that people do use plausibility heuristic-based processing.
The analyses of errors by type were informative in that three major types of errors were identified. The schema-based errors (recalling unmentioned nouns that are related to the described event) likely result from a strong reliance on heuristics and memory retrieval problems regarding the lexical concepts that actually appeared in the sentences. Retrieval errors (unable to recall) likely resulted from memory retrieval problems (interference and/or decay). Participants were aware that they were making both types of errors as indicated by their confidence ratings. The thematic-analysis errors (providing the other, incorrect noun from the sentence) occurred most frequently, with relatively higher confidence than the other two types of errors, but lower than when making correct responses. The frequency of thematic-analysis errors varied across sentence types as a function of both N1 plausibility and N2 plausibility, suggesting that people do have problems with thematic role assignments when sentences contain unlikely role-fillers.

The analyses of behavioral data including trial position as a factor indicated that there was improvement of performance over time, and that there are possibly two types of learning that occurred during the experiment. The first type is adaptation of syntactic-thematic expectation (i.e., NVN strategy). Because all experimental items in the current study were simple full passives (although there were a large number of fillers of different syntactic structures), and because this construction is relatively rare, people started with low expectations for such a construction, leading to higher error rates in the beginning, and later learned to adjust their expectations. Another type of learning is adaptation or desensitization of expectations based on semantic or event-based knowledge. Because greater than three quarters of experimental items expressed somewhat unusual events,
and because reliance on knowledge of typical events may indeed impair performance in the current task, participants may have learned to place less weight on such expectations.

The inattentiveness ratings suggest that participants do occasionally engage in mindless reading at the cost of accuracy. This supports the notion that mind-wandering is a potential cause of misinterpretation. The results however do not speak to the mind-wandering-based explanation of differential difficulties across sentence types. It is possible that the sample size may not be sufficiently large, or that the materials used in the current experiment may not be well suited for uncovering the previously reported effects of mind wandering that are specific to task-demands or stimulus-complexity.

The most important contributions arise from the ERP data. First, when sentences are processed correctly, those that contained implausible arguments elicited a larger N400. Second, this N400 effect is absent in implausible misinterpreted sentences. Furthermore, the implausible sentences that lead to errors were processed similarly to plausible, correctly interpreted sentences. These results suggest that one important source of misinterpretation occurs during online sentence processing as reflected in the absence of a differential N400. In contrast to previous studies in which the authors suggested that reanalysis occurs and the resulting representation competes with the incorrect one (Ferreira, 2003; Slattery, et al., 2003), the present results suggest an insensitivity to implausibility during sentence comprehension. Thus, this is the first clear evidence that on-line processing differs systematically between sentences following which an error is made on a thematic probe task, and those on which it is not.
Finally, it should be noted that the lack of an N400 does not explain all errors. Performance on entirely plausible sentences was not perfect, and there should be no violation of expectancy for sentences in which both the agent and patient are typical. These errors may be due to a NVN heuristic that is incompatible with full passive sentences (as evidenced by their decrease in frequency across the experiment). In addition, as would be expected, some errors presumably are due to memory requirements of this task, in terms of holding the surface structure in memory during the interval between sentence reading and making a response. However, the present results make it clear that it is certainly not the case that all errors are due to processes that occur between sentence comprehension and the eventual response.

3.5 References


Chapter 4

4 General Discussion

The main goal of my dissertation was to provide insight into our understanding the role of event knowledge during sentence processing, particularly with respect to issues that previous studies have not fully addressed. The first study (Chapter 2) demonstrated that a strong manipulation of thematic fit can bias readers toward a normally non-preferred structural interpretation of sentences containing a temporary syntactic ambiguity. In combination with previous research that has demonstrated effects of thematic fit in the other direction (making complex sentences easier), Chapter 2 provides strong support for the conclusion that thematic knowledge rapidly constrains expectations for upcoming sentence structure. The second study (Chapter 3) explored the role of event-based thematic knowledge in constructing interpretations of sentences, along with other potential causes of misinterpretation. A key result was that misinterpretation is associated with a lack of sensitivity to the implausibility of an event described by a sentence.

4.1 Implications for Models of Language Comprehension

Chapter 2 was targeted at adjudicating between two major models of sentence processing: the garden-path model (and other two-stage models), and the constraint-based model. The models differ regarding the time-course over which information based on thematic knowledge can influence syntactic ambiguity resolution, and whether a general bias toward simpler syntactic structures can be overridden by a sufficiently strong manipulation of thematic fit (and other constraints). The present results suggest that thematic constraints can rapidly overcome general structural biases, supporting the
predictions of constraint-based models. By adding the study presented in Chapter 2 to the literature, the evidence is now overwhelmingly in favour of constraint-based models.

Chapter 3 examined claims that counter the commonly held assumption that people’s interpretations of sentences are always correct in the end (at least in the absence of a global ambiguity), which underlies many models of language compression including the two-stage and constraint-based models. Consistent with Ferreira’s (2003) results, the current evidence shows that people make relatively frequent interpretation errors regarding thematic role assignment even in relatively simple sentences, and the occurrence of errors can be predicted by the plausibility or thematic fit of the arguments. Moreover, the current evidence suggested that misinterpretation of sentences containing implausible arguments is associated with a lack of sensitivity to implausibility during sentence reading. This was reflected by the absence of a significantly larger N400 to implausible versus plausible thematic role fillers at the final noun phrase in the misinterpreted passive sentences. To account for these results, models of sentence comprehension need to incorporate a mechanism whereby the weighting of information based on event-based plausibility should be adjusted to allow an occasional disproportionate contribution of such information. Alternatively, or additionally, as discussed below, models should incorporate the mechanism of expectation-driven processing.

4.2 Expectation-driven Processing

As noted briefly in Chapter 1, one possible mechanism by which event knowledge can exert its influence rapidly during the resolution of temporary syntactic ambiguity is through generation of expectations (implicit prediction) about upcoming concepts and
structures. Based on this view, the Chapter 2 results can be considered as arising from a situation in which syntactic expectations were influenced by thematic knowledge, but those expectations were incongruent with the subsequent input when the initial noun was a good patient for the event denoted by the verb. The sentences continued a main clause rather than a reduced relative structure, resulting in temporary processing disruption relative to the unambiguous sentences that included had. Interestingly, the Chapter 3 results also suggest the possibility of expectancy-driven processing during reading. I argued that expectancy-driven processing has a potential role in correctly interpreting sentences, based on the observation of a larger N400 for implausible than for plausible sentences when participants responded correctly, but not when they responded incorrectly.

On the surface at least, expectancy generation is related to Ferreira’s (2003) construct of plausibility heuristics, which were described in Chapter 3 as the tendency to choose the analysis that is most consistent with a person’s event knowledge. Thus, both expectation-driven processing and plausibility heuristics are based on people’s knowledge about what is likely in the real world. It may seem contradictory to argue on the one hand that the use of plausibility-heuristics can cause misinterpretations, and on the other hand that implicit prediction is beneficial in constructing correct interpretations. It has long been shown that expectation-driven processing results in processing disruption when the upcoming input disconfirms predictions. Indeed, longer reading times in Chapter 2 and larger N400s in Chapter 3 for unexpected continuations show that to be the case.

I argue that this type of disruption reflects comprehenders’ sensitivity to violation of expectancies, as has been assumed by some models of sentence processing. This
sensitivity (which is assumed to be implicit, Lau, Holcomb, & Kuperberg, 2013) may strengthen the memory trace of the updated interpretation. The lack of an N400 for the implausible misinterpreted sentences suggests the lack of an alteration of expectancies in those sentences, possibly because no expectations were generated in the first place. To generate expectations for upcoming inputs, global and local syntactic and semantic information must be integrated. In the absence of such processing, no expectations will be generated or altered. In this case, information based on event-based plausibility and local coherence (integrating nouns and verbs piecemeal without a correct overall incorporation of syntactic structure) may take precedence, resulting in faulty comprehension, incorrect sentence memory representations, and recall errors.

This argument may seem to contradict the findings within the framework of good enough processing proposed by Ferreira and colleagues (Christianson, Hollingworth, Halliwell, & Ferreira, 2001; Christianson, Williams, Zacks, & Ferreira, 2006; Ferreira, 2003; Ferreira, Christianson, & Hollingworth, 2001; Slattery, Sturt, Christianson, Yoshida, & Ferreira, 2013; Patson, Darowski, Moon, & Ferreira, 2009). These researchers have argued that incorrect interpretations of syntactically ambiguous sentences can occur despite assumed reanalyses indicated by longer reading times. However, what they have not shown is the direct link between the reading time measures and the probability of making interpretation errors. That is, they have not separately analyzes reading times for sentences followed by correct versus incorrect responses. Although the materials in Chapter 3 did not involve syntactic ambiguity, the results indicated that there is a relation between the amplitude of the N400 to the implausible sentences and whether the sentences were interpreted correctly or incorrectly. Thus, my account predicts that there
is such a relation, and I argue that future studies should be directed to understanding this relation.

4.3 References


Appendices

Appendix A: Sentences used in Experiment of Chapter 2

The suspect (had) interrogated for three weeks the crooked detectives that he had captured.

The inspector (had) interrogated for three weeks the crooked detectives that he had captured.

The slave (had) tortured for two hours the wicked man who had mistreated him for years.

The kidnapper (had) tortured for two hours the wicked man who had mistreated him for years.

The suspect (had) arrested near the station the notorious murderer who had been eluding the police for months.

The cop (had) arrested near the station the notorious murderer who had been eluding the police for months.

The gangster (had) investigated for three months the young punks who had been rumored to be working in his territory.

The auditor (had) investigated for three months the young punks who had been laundering money in Chicago.

The audience (had) entertained in the club some visiting businessmen who were out for a night on the town.

The strippers (had) entertained in the club some visiting businessmen who were out for a night on the town.

The patient (had) cured with the treatment numerous cancer patients since she had recovered and began using it on others.

The doctor (had) cured with the treatment numerous cancer patients since she had started using it.

The baby (had) frightened during the night three teenage hikers who heard him scream loudly in the dark.

The monster (had) frightened during the night three teenage hikers who heard him grunt loudly in the dark.

The student (had) graded by the new criteria all first-year undergraduates who were enrolled in the lab that she TAed.
The teacher (had) graded by the new criteria all first-year undergraduates who had been enrolled in her class.

The hostage (had) rescued from the dungeon three frightened peasants who were being held captive with him.

The knight (had) rescued from the dungeon three frightened peasants who were being held captive with him.

The child (had) punished with excessive force the irritating kids who had been making fun of her.

The babysitter (had) punished with excessive force the irritating kids who had been left in her care.

The victim (had) terrorized for two weeks the rich landowners who had treated him as a slave for years.

The pirate (had) terrorized for two weeks the rich landowners who had lived on the island for years.

The goddess (had) worshipped for many years the divine Creator and all of the earth's creatures.

The priest (had) worshipped for many years the divine Creator and all of the earth's creatures.

The victim (had) killed near the border the immigration officials who had been after him.

The assassin (had) killed near the border the immigration officials who had been after him.

The witness (had) questioned during the trial the prosecution's account of the murder.

The lawyer (had) questioned during the trial the prosecution's account of the murder.

The fugitive (had) trapped near the river two small animals that he cooked over an open fire.

The hunter (had) trapped near the river two small animals that he cooked over an open fire.

The child (had) abandoned in the forest the screaming baby he had abducted from a car.

The kidnapper (had) abandoned in the forest the screaming baby he had abducted from a car.

The fugitive (had) captured in the forest two large rabbits that had been living only a few hundred yards from his hut.
The hunter (had) captured in the forest two large rabbits that had been living only a few hundred yards from his hut.

The guest (had) invited to the luncheon three old buddies who really had no business being there.

The organizer (had) invited to the luncheon three old buddies who really had no business being there.

The candidate (had) interviewed on the evening news three civil servants who had worked for the previous administration.

The reporter (had) interviewed on the evening news three civil servants who had worked for the previous administration.

The patient (had) examined in the laboratory the official report that detailed his doctor's mistakes.

The scientist (had) examined in the laboratory the official report that detailed his mistakes.
**Appendix B: Verbs and noun phrases used in Experiment of Chapter 3**

<table>
<thead>
<tr>
<th>verb</th>
<th>good agents</th>
<th>good patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>accused</td>
<td>witness</td>
<td>criminal</td>
</tr>
<tr>
<td>adopted</td>
<td>family</td>
<td>orphan</td>
</tr>
<tr>
<td>advised</td>
<td>supervisor</td>
<td>student</td>
</tr>
<tr>
<td>anesthetized</td>
<td>surgeon</td>
<td>donor</td>
</tr>
<tr>
<td>applauded</td>
<td>critic</td>
<td>musician</td>
</tr>
<tr>
<td>appointed</td>
<td>president</td>
<td>secretary</td>
</tr>
<tr>
<td>arrested</td>
<td>policeman</td>
<td>robber</td>
</tr>
<tr>
<td>assassinated</td>
<td>hitman</td>
<td>governor</td>
</tr>
<tr>
<td>beaten</td>
<td>tyrant</td>
<td>slave</td>
</tr>
<tr>
<td>betrayed</td>
<td>politician</td>
<td>supporter</td>
</tr>
<tr>
<td>bit</td>
<td>snake</td>
<td>man</td>
</tr>
<tr>
<td>bullied</td>
<td>manager</td>
<td>trainee</td>
</tr>
<tr>
<td>calmed</td>
<td>therapist</td>
<td>schizophrenic</td>
</tr>
<tr>
<td>cheered</td>
<td>spectator</td>
<td>athlete</td>
</tr>
<tr>
<td>complimented</td>
<td>columnist</td>
<td>ballerina</td>
</tr>
<tr>
<td>criticized</td>
<td>food columnist</td>
<td>pastry chef</td>
</tr>
<tr>
<td>cursed</td>
<td>wizard</td>
<td>prince</td>
</tr>
<tr>
<td>deceived</td>
<td>politician</td>
<td>taxpayer</td>
</tr>
<tr>
<td>diagnosed</td>
<td>physician</td>
<td>miner</td>
</tr>
</tbody>
</table>


disciplined coach commandant trainee cadet
dressed stylist designer model actress
entertained singer comedian fan audience
expelled principal dean rebel thug
fed social worker community worker drug addict homeless person
hired manager administrator intern applicant
hypnotized psychologist magician smoker insomniac
idolized teenager groupie celebrity popstar
interviewed newswriter journalist linebacker quarterback
invited organizer host guest visitor
kicked out doorman bouncer minor drunk
lectured professor instructor undergrad student
measured tailor seamstress businessman bride
mocked bully brat nerd geek
molested pedophile pervert kindergartener tot
mugged thief hoodlum pedestrian victim
oppressed king queen peasant maid
overcharged telemarketer salesman elderly housewife
persecuted warden attorney inmate offender
petted zookeeper veterinarian orangutan chimp
praised mother father kid toddler
protected troop soldier refugee villager
<table>
<thead>
<tr>
<th>Verb</th>
<th>Officer</th>
<th>Guard</th>
<th>Prisoner</th>
<th>Convict</th>
</tr>
</thead>
<tbody>
<tr>
<td>Punished</td>
<td>Officer</td>
<td>Guard</td>
<td>Prisoner</td>
<td>Convict</td>
</tr>
<tr>
<td>Pursued</td>
<td>Sheriff</td>
<td>Detective</td>
<td>Mugger</td>
<td>Robber</td>
</tr>
<tr>
<td>Questioned</td>
<td>Investigator</td>
<td>Detective</td>
<td>Informant</td>
<td>Suspect</td>
</tr>
<tr>
<td>Quizzed</td>
<td>Tutor</td>
<td>Teacher</td>
<td>Student</td>
<td>Sophomore</td>
</tr>
<tr>
<td>Recruited</td>
<td>Scout</td>
<td>Coach</td>
<td>Rookie</td>
<td>Freshman</td>
</tr>
<tr>
<td>Rescued</td>
<td>Policewoman</td>
<td>Fireman</td>
<td>Dog</td>
<td>Kitty</td>
</tr>
<tr>
<td>Ripped Off</td>
<td>Owner</td>
<td>Landlord</td>
<td>Renter</td>
<td>Tenant</td>
</tr>
<tr>
<td>Scammed</td>
<td>Hustler</td>
<td>Gangster</td>
<td>Addict</td>
<td>Junkie</td>
</tr>
<tr>
<td>Scared</td>
<td>Large Monster</td>
<td>Spooky Ghost</td>
<td>Little Girl</td>
<td>Little Boy</td>
</tr>
<tr>
<td>Scolded</td>
<td>Nanny</td>
<td>Babysitter</td>
<td>Troublemaker</td>
<td>Juvenile</td>
</tr>
<tr>
<td>Served</td>
<td>Waitress</td>
<td>Waiter</td>
<td>Customer</td>
<td>Patron</td>
</tr>
<tr>
<td>Stopped</td>
<td>Umpire</td>
<td>Referee</td>
<td>Player</td>
<td>Competitor</td>
</tr>
<tr>
<td>Supervised</td>
<td>Scientist</td>
<td>Professor</td>
<td>Assistant</td>
<td>TA</td>
</tr>
<tr>
<td>Tasered</td>
<td>Security Guard</td>
<td>Police Officer</td>
<td>Demonstrator</td>
<td>Protester</td>
</tr>
<tr>
<td>Threatened</td>
<td>Persistent Stalker</td>
<td>Determined Kidnapper</td>
<td>Vulnerable Woman</td>
<td>Helpless Family</td>
</tr>
<tr>
<td>Treated</td>
<td>Doctor</td>
<td>Clinician</td>
<td>Patient</td>
<td>Client</td>
</tr>
<tr>
<td>Washed</td>
<td>Mother</td>
<td>Nanny</td>
<td>Infant</td>
<td>Baby</td>
</tr>
<tr>
<td>Whipped</td>
<td>Farmer</td>
<td>Cowboy</td>
<td>Cow</td>
<td>Donkey</td>
</tr>
<tr>
<td>Worshipped</td>
<td>Priest</td>
<td>Monk</td>
<td>Creator</td>
<td>Goddess</td>
</tr>
</tbody>
</table>
Appendix C: Ethics Approval

Department of Psychology  The University of Western Ontario

Western

Use of Human Subjects - Ethics Approval Notice

<table>
<thead>
<tr>
<th>Review Number</th>
<th>10 08 07</th>
<th>Approval Date</th>
<th>10 08 23</th>
</tr>
</thead>
<tbody>
<tr>
<td>Principal Investigator</td>
<td>Ken McRae/Kazunaga Matsuki</td>
<td>End Date</td>
<td>11 04 30</td>
</tr>
<tr>
<td>Protocol Title</td>
<td>Thematic roles and eye-movements during reading</td>
<td>Sponsor</td>
<td>n/a</td>
</tr>
</tbody>
</table>

This is to notify you that The University of Western Ontario Department of Psychology Research Ethics Board (PREB) has granted expedited ethics approval to the above named research study on the date noted above.

The PREB is a sub-REB of The University of Western Ontario’s Research Ethics Board for Non-Medical Research Involving Human Subjects (NMREB) which is organized and operates according to the Tri-Council Policy Statement and the applicable laws and regulations of Ontario. (See Office of Research Ethics web site: http://www.uwo.ca/research/ethics/)

This approval shall remain valid until end date noted above assuming timely and acceptable responses to the University’s periodic requests for surveillance and monitoring information.

During the course of the research, no deviations from, or changes to, the protocol or consent form may be initiated without prior written approval from the PREB except when necessary to eliminate immediate hazards to the subject or when the change(s) involve only logistical or administrative aspects of the study (e.g. change of research assistant, telephone number etc.). Subjects must receive a copy of the information/consent documentation.

Investigators must promptly also report to the PREB:

a) changes increasing the risk to the participant(s) and/or affecting significantly the conduct of the study;
b) all adverse and unexpected experiences or events that are both serious and unexpected;
c) new information that may adversely affect the safety of the subjects or the conduct of the study.

If those changes/adverse events require a change to the information/consent documentation, and/or recruitment advertisement, the newly revised information/consent documentation, and/or advertisement, must be submitted to the PREB for approval.

Members of the PREB who are named as investigators in research studies, or declare a conflict of interest, do not participate in discussion related to, nor vote on, such studies when they are presented to the PREB.

Clive Seligman Ph.D.
Chair, Psychology Expedited Research Ethics Board (PREB)

The other members of the 2009-2010 PREB are: David Donoix, Bill Fisher, Riley Hinson and Steve Lupker

CC: UWO Office of Research Ethics

This is an official document. Please retain the original in your files.
Appendix D: Ethics Approval

Use of Human Subjects - Ethics Approval Notice

<table>
<thead>
<tr>
<th>Review Number</th>
<th>Approval Date</th>
<th>Principal Investigator</th>
<th>Protocol Title</th>
<th>Sponsor</th>
</tr>
</thead>
<tbody>
<tr>
<td>12 09 21</td>
<td>12 09 21</td>
<td>Ken McRae/Kazuhisa Matsuki</td>
<td>Brain-wave pattern during sentence understanding</td>
<td>n/a</td>
</tr>
<tr>
<td></td>
<td>13 04 30</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This is to notify you that The University of Western Ontario Department of Psychology Research Ethics Board (PREB) has granted expedited ethics approval to the above named research study on the date noted above.

The PREB is a sub-REB of The University of Western Ontario’s Research Ethics Board for Non-Medical Research Involving Human Subjects (NMREB) which is organized and operates according to the Tri-Council Policy Statement and the applicable laws and regulations of Ontario. (See Office of Research Ethics web site: http://www.uwo.ca/research/ethics/)

This approval shall remain valid until end date noted above assuming timely and acceptable responses to the University’s periodic requests for surveillance and monitoring information.

During the course of the research, no deviations from, or changes to, the protocol or consent form may be initiated without prior written approval from the PREB except when necessary to eliminate immediate hazards to the subject or when the change(s) involve only logistical or administrative aspects of the study (e.g., change of research assistant, telephone number etc.). Subjects must receive a copy of the information/consent documentation.

Investigators must promptly also report to the PREB:
   a) changes increasing the risk to the participant(s) and/or affecting significantly the conduct of the study;
   b) all adverse and unexpected experiences or events that are both serious and unexpected;
   c) new information that may adversely affect the safety of the subjects or the conduct of the study.

If these changes/adverse events require a change to the information/consent documentation, and/or recruitment advertisement, the newly revised information/consent documentation, and/or advertisement, must be submitted to the PREB for approval.

Members of the PREB who are named as investigators in research studies, or declare a conflict of interest, do not participate in discussion related to, nor vote on, such studies when they are presented to the PREB.

Clive Seligman Ph.D.
Chair, Psychology Expedited Research Ethics Board (PREB)

The other members of the 2011-2012 PREB are: Mike Atkinson (Introductory Psychology Coordinator), Rick Goffin, Riley Hilton, Albert Katz (Department Chair), Steve Laplak, and Karen DiSloven (Graduate Student Representative)

CC: UWO Office of Research Ethics
This is an official document. Please retain the original in your files
Curriculum Vitae

Name: Kazunaga Matsuki

Post-secondary Education and Degrees:

University of Wisconsin - Platteville
Platteville, Wisconsin, U.S.A.
2002-2003 A.A.

University of Wisconsin - Madison
Madison, Wisconsin, U.S.A.
2004-2005, B.S.

The University of Western Ontario
London, Ontario, Canada
2006-2008 M.Sc.

The University of Western Ontario
London, Ontario, Canada
2008-2013 Ph.D.

Honours and Awards:

G. Keith Humphrey Memorial Award
2012

Western Graduate Research Scholarship
University of Western Ontario
2006-2012

Related Work Experience:

Teaching Assistant
The University of Western Ontario
2006-2012

Publications:

Peer Reviewed


Submitted Manuscript


Invited Talks

Matsuki, K. (2013, February). *The roles of thematic knowledge in sentence comprehension*. Colloquium at Universität Stuttgart, Stuttgart, Germany

McRae, K., Matsuki, K., Elman, J. L., & Hare, M. (2008, June). *Semantic and syntactic event-based expectancy generation in on-line language comprehension*. In A. Myachykov (Chair), General Cognition and Language Processing. Symposium conducted at the Third International Conference on Cognitive Science. Moscow, Russia


Conference Presentations


