

Doctoral Dissertation

Activation of Color Information in
Second Language Comprehension

(第二言語理解における色情報の活性化)

Faculty of Humanities,
Graduate School of Humanities,

Nagoya University

TERAI Masato

February 2023

Abstract

Studies of second language vocabulary have attempted to uncover the mechanisms of processing and developing their knowledge. The mechanisms have been described and studied with various models. Some of the models explained that second language learners use their first language to understand the words in their second language, especially at lower levels. They also speculate that the reliance on their first language decreases as second language proficiencies increase. However, these models cannot fully explain concepts assumed to be represented after orthographic or phonological processing. Nor do they elucidate the changes in representation with increasing second language proficiency. Embodied cognition studies have investigated conceptual processing during language comprehension. Some studies suppose that linguistic and non-linguistic processing, such as visual aspects of objects, play an essential role in language comprehension. Although several empirical research studies have observed the use of both linguistic and non-linguistic knowledge during first language comprehension, very little is known about second language comprehension.

The study investigated whether readers activated non-linguistic information, specifically object color, during second language vocabulary processing and whether second language proficiency affects them. A semantic Stroop task was conducted with 35 native English speakers and 72 native Japanese speakers. Thirty-six native Japanese speakers performed the task in Japanese, and the remainder performed it in English. In the task, a sentence was presented (e.g., *Joe was excited to see a bear in the woods*) to the participants. They were then presented with a word from the sentence (*bear*) in a colored font. There were three conditions with the color: the typical (brown), the atypical (white: a polar bear), and

the unrelated color (green) of the object to which the word refers. They must answer the color of the font with a keyboard as quickly as possible. All target words were nouns representing concrete objects. Each word had two conditions that differed in typicality implied by the sentence (typical/atypical). The typicality of the experimental materials was determined by the results of two pilot studies conducted with 26 native Japanese speakers, none of whom participated in the main studies.

Additionally, typicality rating tasks were conducted with the participants after they completed the semantic Stroop task. This procedure confirmed that the typicality of the experimental items selected in the pilot studies was consistent with that of the participants. Finally, the English proficiency of the Japanese participants was measured with a vocabulary test (Meara & Miralpeix, 2016).

Here the results showed that readers activated the object colors during language comprehension in their first language. This finding was consistent with previous studies. Specifically, when native English and Japanese participants performed the semantic Stroop task in their first language, they responded to typical color words faster than atypical and unrelated colors, regardless of how typical the sentence was. The second language task results showed that as participants' second language proficiency increased, they responded significantly faster to a typical object color than an object in atypical or unrelated colors. Furthermore, even learners with lower language proficiency responded significantly faster to typical color words than unrelated ones when the typical color was red. These results imply that readers activated non-linguistic knowledge during second language processing at higher proficiency levels, as they do in their first languages. The results also imply that word knowledge development differs depending on the relationship between an object's color and its typicality. The author expects the findings to provide a novel explanation for existing

vocabulary processing and knowledge models.

Acknowledgments

I would like to thank my thesis advisor, Prof. Junko Yamashita, who supported me with critical, helpful comments and warm encouragement. The door to Prof. Yamashita's office was always open when I had problems with my research.

My gratitude also goes to my sub-supervisors Prof. Sugiura Masatoshi and Ass. Prof. Koji Miwa. I had attended Prof. Sugiura Masatoshi's seminar for five years since I started the master's program. Prof. Sugiura Masatoshi gave me insightful comments every time and encouraged me when I talked about my research. I would also like to thank Ass. Prof. Koji Miwa, who always welcomed and advised me in a friendly and enthusiastic manner. Ass. Prof. Koji Miwa often provided me with research articles related to my research.

I am extremely thankful to Mr. Ryouzuke Mikami for his kind help and discussion about this thesis. In particular, the discussion of the theoretical background and statistical methods of my research has opened new perspectives for my research. Whenever I have faced difficulties not only in this research but also in my life, he has always encouraged me to overcome the challenges.

It is my privilege to thank my parents for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching for and writing this thesis.

Last but not least, this work was financially supported by JST SPRING, Grant Number JPMJSP2125. I would like to take this opportunity to thank the "Interdisciplinary Frontier Next-Generation Researcher Program of the Tokai Higher Education and Research System." Also, I would like to thank *Editage* (www.editage.com) for English language editing.

Table of Contents

Abstract.....	i
Acknowledgments	iv
Table of Contents.....	v
List of Tables	xiii
List of Figures.....	xv
Chapter 1: Introduction.....	1
Background.....	1
Aim of the Present Study	2
Outline of the Dissertation.....	4
Chapter 2: Literature Review.....	7
Background.....	7
Processing and Representation of L2 Vocabulary	7
Models of Bilingual Mental Lexicons.....	7
The Development of Mental Lexicon.....	9
Embodied Cognition.....	12
Empirical Support for the Embodiment Approach.....	16
The Quality of Simulation and Mental Images.....	19
Embodied Cognition Research in L2.....	21
The Present Study	25
Research Questions and Hypotheses	29
Chapter 3: Pilot Study.....	32
Aim of the Pilot Studies.....	32
Pilot Study 1.....	32
Participants.....	32
Experimental Items	32
Rating Tasks	33
Word Typicality Rating Task.....	33
Sentence Typicality Rating Task.....	34
Procedure	36
Results.....	36
Pilot Study 2.....	36
Participants.....	36
Experimental Items	37
Procedure	38
Results.....	38
Determining the Sample Size	39
Experiment 1	39
Experiment 2.....	39
Chapter 4: Experiment 1	41

Aim of Experiment 1	41
Method	41
Participants.....	41
Native English Speakers.	41
Native Japanese Speakers.	42
Tasks	43
Semantic Stroop Task.	43
Experimental Sentences.....	47
Critical Sentences.	47
Filler Sentences.....	49
Word Typicality Rating Task.	49
Sentence Typicality Rating Task.	50
Vocabulary Size Test.	50
Procedure	51
Analysis	54
Results and Discussion	60
Word Typicality Rating Task	60
Native English Speakers.	60
Native Japanese Speakers.	61
Sentence Typicality Rating Task	61
Native English Speakers.	61
Native Japanese Speakers.	61
Semantic Stroop Task	61
Native English Speakers.	61
Descriptive Statistics.....	61
Modeling Results.	66
Native Japanese Speakers.	72
Descriptive Statistics.....	72
Modeling Results.	79
Summary of Experiment 1	85
Chapter 5: Experiment 2	86
Aim of Experiment 2	86
Method	86
Participants.....	86
Tasks and Materials	87
Procedure	88
Analysis	88
Results and Discussion	92
Rating Tasks	92
Word Typicality Rating Task.	92
Sentence Typicality Rating Task.	92
Semantic Stroop Task	92

Descriptive Statistics.....	92
Modeling Results.....	98
Summary of Experiment 2.....	110
Chapter 6: General Discussion.....	111
Summary of the Results.....	111
Experiment 1.....	111
Experiment 2.....	111
Simulation of Object Colors.....	112
Color Simulation in L1 Processing.....	112
Color Simulation in L2 Processing.....	113
The Representation of L2 Mental Lexicon.....	115
Limitations and Directions for Future Research.....	117
Chapter 7: Conclusions.....	127
References.....	130
Appendix A: The Results of the Rating Tasks (Pilot Study 1).....	141
Word.....	141
Rating Scores.....	141
Stacked Bar Chart.....	141
Sentence.....	142
Agreement Rates.....	142
Entirely.....	143
Each Sentence.....	143
Balloon Plot.....	143
Appendix B: The Results of the Rating Tasks (Pilot Study 2).....	145
Word.....	145
Rating Scores.....	145
Stacked Bar Chart.....	145
Sentence.....	147
Agreement Rates.....	147
Entirely.....	147
Each Sentence.....	148
Balloon Plot.....	148
Appendix C: Power Analysis.....	150
Experiment 1.....	150
Native English and Native Japanese Speakers.....	150
Results.....	150
The Required Sample Sizes.....	151
Experiment 2.....	151
Native Japanese Speakers Learning English.....	151
Results.....	151
The Required Sample Sizes.....	152
Appendix D: The Experimental Items Used in Experiment 1 (Native English Speakers).....	153

Practice Session	153
Critical Sentences (Three Sentences)	153
Filler Sentences (Two Sentences).....	153
Main Session.....	153
Critical Sentences (Before) (Ninety Sentences)	153
Critical Sentences (After) (Ninety Sentences).....	155
Filler Sentences (Before) (Ninety Sentences).....	156
Filler Sentences (After) (Ninety Sentences)	162
Appendix E: The Experimental Items Used in Experiment 1 (Native Japanese Speakers)	
.....	170
Practice Session	170
Critical Sentences (Three Sentences)	170
Filler Sentences (Two Sentences).....	170
Main Session.....	170
Critical Sentences (Before) (Ninety Sentences)	170
Critical Sentences (After) (Ninety Sentences).....	171
Filler Sentences (Before) (Ninety Sentences).....	172
Filler Sentences (After) (Ninety Sentences)	179
Appendix F: The Japanese Version of the Instructions	187
Practice Session	187
Main Session.....	187
Appendix G: Rating Task (Native English Speakers)	188
Word	188
Rating Scores	188
Cleveland Dot Plot.....	190
Sentence	192
Agreement Rates.....	192
Entirely.....	192
Each Sentence.....	193
Balloon Plot	194
Appendix H: Rating Task (Native Japanese Speakers)	195
Word	195
Rating Scores	195
Cleveland Dot Plot.....	197
Sentence	198
Agreement Rates.....	198
Entirely.....	199
Each Sentence.....	199
Balloon Plot	200
Appendix I: Statistical Modeling (Native English Speakers).....	202
List of Variables.....	202
Change Coding of the Typicality.....	202

Change Coding of the Categorical Variables	203
Sentence	203
Word	203
Position	203
Scaling the Continuous Variables	203
Sentence Reading Time	203
Choose Probabilistic Distributions for the Observed Data	203
Possible Covariates	204
Specification of the Best Random-Effects Structure	206
Maximal Model.....	206
Output.	206
Random-Effects Principal Components Analysis.....	206
Maximal Model (Zero-Correlation-Parameter)	206
Output.	207
Random-Effects Principal Components Analysis.....	207
Dropping Variance Components.	207
Output.	207
Random-Effects Principal Components Analysis.....	207
Dropping Variance Components.	208
Output.	208
Random-Effects Principal Components Analysis.....	208
Dropping Variance Components.	208
Output.	208
Random-Effects Principal Components Analysis.....	208
Dropping Variance Components.	209
Output.	209
Random-Effects Principal Components Analysis.....	209
Dropping Variance Components.	209
Output.	209
Random-Effects Principal Components Analysis.....	209
Dropping Variance Components.	210
Output.	210
Random-Effects Principal Components Analysis.....	210
Dropping Variance Components.	210
Output.	210
Random-Effects Principal Components Analysis.....	212
Dropping Variance Components.	212
Output.	212
Random-Effects Principal Components Analysis.....	214
Dropping Variance Components.	214
Model Comparisons.....	214
Checking If Including Correlation Parameter Increases the Goodness-of-Fit.....	215

Model Comparisons.....	215
Results of the Final Model.....	216
Summary of the Final Model.....	216
Variance Inflation Factors (VIF).....	217
Model Diagnosis.....	218
The Model That Only Includes Significant Predictors.....	218
Appendix J: Statistical Modeling (Native Japanese Speakers).....	221
List of Variables.....	221
Change Coding of the Categorical Variables.....	221
Sentence.....	221
Word.....	221
Position.....	222
Scaling the Continuous Variables.....	222
Sentence Reading Time.....	222
Choose Probabilistic Distributions for the Observed Data.....	222
Possible Covariates.....	223
Specification of the Best Random-Effects Structure.....	225
Maximal model.....	225
Output.....	225
Random-Effects Principal Components Analysis.....	225
Maximal Model (Zero-Correlation-Parameter).....	225
Output.....	226
Random-Effects Principal Components Analysis.....	226
Dropping Variance Components.....	226
Output.....	226
Random-effects Principal Components Analysis.....	226
Dropping Variance Components.....	226
Output.....	227
Random-Effects Principal Components Analysis.....	227
Dropping Variance Components.....	227
Output.....	227
Random-Effects Principal Components Analysis.....	227
Dropping Variance Components.....	227
Output.....	228
Random-Effects Principal Components Analysis.....	228
Dropping Variance Components.....	228
Output.....	228
Random-Effects Principal Components Analysis.....	228
Dropping Variance Components.....	228
Output.....	229
Random-Effects Principal Components Analysis.....	229
Dropping Variance Components.....	229

Output.....	229
Random-Effects Principal Components Analysis.....	229
Dropping Variance Components.....	229
Output.....	230
Random-Effects Principal Components Analysis.....	231
Dropping Variance Components.....	231
Output.....	232
Random-Effects Principal Components Analysis.....	232
Dropping Variance Components.....	232
Model Comparisons.....	232
Checking If Including Correlation Parameter Increases the Goodness-of-Fit.....	233
Model Comparisons.....	233
Results of the Final Model.....	234
Summary of the Final Model.....	234
Variance Inflation Factors (VIF).....	235
Model Diagnosis.....	236
The Model That Only Includes Significant Predictors.....	236
Appendix K: List of Correction for the English Sentences.....	238
Appendix L: Rating Task (Native Japanese Speakers Learning English).....	240
Word.....	240
Rating Scores.....	240
Cleveland Dot Plot.....	242
Sentence.....	243
Agreement Rates.....	243
Entirely.....	244
Each Sentence.....	244
Balloon Plot.....	245
Appendix M: Statistical Modeling (Japanese Learners of English).....	247
List of Variables.....	247
Change Coding of the Categorical Variables.....	247
Sentence.....	247
Scaling the Continuous Variables.....	248
Sentence Reading Time.....	248
Scores of the Vocabulary Size Test.....	248
Choose Probabilistic Distributions for the Observed Data.....	248
Possible Covariates.....	249
Specification of the Best Random-Effects Structure.....	251
Maximal Model.....	251
Output.....	251
Random-Effects Principal Components Analysis.....	251
Maximal Model (Zero-Correlation-Parameter).....	251
Output.....	251

Random-Effects Principal Components Analysis.....	252
Dropping Variance Components.	252
Output.	252
Random-Effects Principal Components Analysis.....	252
Dropping Variance Components.	252
Output.	253
Random-Effects Principal Components Analysis.....	253
Dropping Variance Components.	253
Output.	253
Random-Effects Principal Components Analysis.....	253
Dropping Variance Components.	253
Output.	254
Random-Effects Principal Components Analysis.....	254
Dropping Variance Components.	254
Output.	254
Random-Effects Principal Components Analysis.....	254
Dropping Variance Components.	254
Output.	255
Random-Effects Principal Components Analysis.....	255
Dropping Variance Components.	255
Output.	255
Random-Effects Principal Components Analysis.....	255
Dropping Variance Components.	256
Output.	256
Random-Effects Principal Components Analysis.....	256
Dropping Variance Components.	256
Output.	256
Random-Effects Principal Components Analysis.....	259
Dropping Variance Components.	259
Output.	259
Model Comparisons.....	261
Dropping Variance Components.	262
Output.	262
Model Comparisons.....	264
Checking If Including Correlation Parameter Increases the Goodness-of-Fit.....	265
Model Comparisons.....	265
Results of the Final Model.....	266
Summary of the Final Model.....	266
Variance Inflation Factors (VIF).	268
Model Diagnosis.....	269
The Model That Only Includes Significant Predictors	269

List of Tables

Table 1 <i>Native English Speakers' Age</i>	41
Table 2 <i>Native Japanese Speakers' Descriptive Statistics</i>	42
Table 3 <i>An Example of Each Condition (bear)</i>	48
Table 4 <i>The Number of Trials in Each Condition</i>	49
Table 5 <i>The Goodness-of-Fit Statistics and Information Criterion (Native English Speakers)</i>	57
Table 6 <i>The Goodness-of-Fit Statistics and Information Criterion (Native Japanese Speakers)</i>	57
Table 7 <i>Dependent Variables and Their Assigned Codes</i>	58
Table 8 <i>Descriptive Statistics of the Reaction Times of the Semantic Stroop Task (Native English Speakers)</i>	62
Table 9 <i>Descriptive Statistics of the Reaction Times of the Semantic Stroop Task (Native English Speakers: Pre-Context Condition)</i>	64
Table 10 <i>Descriptive Statistics of the Reaction Times of the Semantic Stroop Task (Native English Speakers: Post-Context Condition)</i>	64
Table 11 <i>Results of Mixed-Effects of the Native English Speakers</i>	67
Table 12 <i>Results of Mixed-Effects of the Native English Speakers (Only Significant Variables)</i>	71
Table 13 <i>Descriptive Statistics of the Reaction Times of the Semantic Stroop Task (Native Japanese Speakers)</i>	73
Table 14 <i>Descriptive Statistics of the Reaction Times of the Semantic Stroop Task (Native Japanese Speakers: Pre-Context Condition)</i>	75
Table 15 <i>Descriptive Statistics of the Reaction Times of the Semantic Stroop Task (Native Japanese Speakers: Post-Context Condition)</i>	75
Table 16 <i>Results of Mixed-Effects of the Native Japanese Speakers</i>	81
Table 17 <i>Results of Mixed-Effects of the Native Japanese Speakers (Only Significant Variables)</i>	85
Table 18 <i>Descriptive Statistics of the Japanese Learners of English</i>	87
Table 19 <i>Goodness-of-Fit Statistics and Information Criterion</i>	90
Table 20 <i>Dependent Variables and Their Assigned Codes</i>	91
Table 21 <i>Descriptive Statistics of the Reaction Times of the Semantic Stroop Task (Japanese Learners of English)</i>	93
Table 22 <i>Descriptive Statistics of the Reaction Times of the Semantic Stroop Task (Japanese Learners of English: Pre-Context Condition)</i>	94
Table 23 <i>Descriptive Statistics of the Reaction Times of the Semantic Stroop Task (Japanese Learners of English: Post-Context Condition)</i>	95
Table 24 <i>Results of Mixed-Effects of the Japanese Learners of English</i>	99
Table 25 <i>Results of Mixed-Effects of the Japanese Learners of English (Only Significant Variables)</i>	104

Table 26 <i>Reaction Times of the Semantic Stroop Task for Individual Words</i>	106
Table 27 <i>Mean and Standard Deviations of Scores of Word Typicality Rating Task of Each Word</i>	109
Table 28 <i>Frequencies of the Experimental Items</i>	118
Table 29 <i>Spearman's Rank Correlation Coefficients and 95% Confidence Interval Between Frequency and the Reaction Times of the Semantic Stroop Task</i>	119
Table 30 <i>The Number of Letters in English and Japanese Experimental Items</i>	120
Table 31 <i>Descriptive Statistics of the Reaction Times of the Semantic Stroop Task (Native English Speakers: Filler Items)</i>	124
Table 32 <i>Descriptive Statistics of the Reaction Times of the Semantic Stroop Task (Native Japanese Speakers: Filler Items)</i>	125
Table 33 <i>Descriptive Statistics of the Reaction Times of the Semantic Stroop Task (Japanese Learners of English: Filler Items)</i>	125

List of Figures

Figure 1 <i>Word Typicality Rating Task Example</i>	34
Figure 2 <i>Sentence Typicality Rating Task Example</i>	35
Figure 3 <i>Diagram of the Practice Program</i>	44
Figure 4 <i>Diagram of the Semantic Stroop Task (Critical Items)</i>	46
Figure 5 <i>Diagram of the Semantic Stroop Task (Filler Items)</i>	47
Figure 6 <i>Word Typicality Rating Task Example (English Version)</i>	50
Figure 7 <i>The Positions of the Fingers on the Keys</i>	52
Figure 8 <i>The Distribution of the Reaction Times of the Semantic Stroop Task (Native English Speakers)</i>	55
Figure 9 <i>The Distribution of the Reaction Times of the Semantic Stroop Task (Native Japanese Speakers)</i>	56
Figure 10 <i>Mean Reaction Times of the Semantic Stroop Task (Native English Speakers)</i>	63
Figure 11 <i>Mean Reaction Times of the Semantic Stroop Task (Native English Speakers: Pre-Context Condition)</i>	65
Figure 12 <i>Mean Reaction Times of the Semantic Stroop Task (Native English Speakers: Post-Context Condition)</i>	65
Figure 13 <i>Effects of Sentence Typicality and Word Typicality on the Reaction Times of the Semantic Stroop Task (Native English Speakers)</i>	68
Figure 14 <i>The Scaled Reading Time of Each Sentence and the Presentation Order Variable Included in the Final Model (Native English Speakers)</i>	70
Figure 15 <i>Mean Reaction Times of the Semantic Stroop Task (Native Japanese Speakers)</i>	74
Figure 16 <i>Mean Reaction Times of the Semantic Stroop Task (Native Japanese Speakers: Pre-Context Condition)</i>	76
Figure 17 <i>Mean Reaction Times of the Semantic Stroop Task (Native Japanese Speakers: Post-Context Condition)</i>	76
Figure 18 <i>The Distributions of Native English and Japanese Participants' Age</i>	78
Figure 19 <i>The Relation of Native English and Japanese Participants' Age and Reaction Times of the Semantic Stroop Task</i>	79
Figure 20 <i>Effects of Sentence Typicality and Word Typicality on the Reaction Times of the Semantic Stroop Task (Native Japanese Speakers)</i>	82
Figure 21 <i>The Scaled Reading Time of Each Sentence and the Presentation Order Variable Included in the Final Model (Native Japanese Speakers)</i>	83
Figure 22 <i>The Distribution of the Reaction Times of the Semantic Stroop Task (Japanese Learners of English)</i>	89
Figure 23 <i>Mean Reaction Times of the Semantic Stroop Task (Japanese Learners of English)</i>	94

Figure 24 <i>Mean Reaction Times of the Semantic Stroop Task (Japanese Learners of English: Pre-Context Condition)</i>	95
Figure 25 <i>Mean Reaction Times of the Semantic Stroop Task (Japanese Learners of English: Post-Context Condition)</i>	96
Figure 26 <i>The Distributions of Native English Speakers and English Learner' Age</i> ..	97
Figure 27 <i>The Relation of Native English Speakers and L2 Learners' Age and Reaction Times of the Semantic Stroop Task</i>	98
Figure 28 <i>Effects of Word Typicality and Scaled Vocabulary Size Test Scores on the Reaction Times of the Semantic Stroop Task</i>	102
Figure 29 <i>The Scaled Reading Time of Each Sentence and Presentation Order Variable Included in the Final Model (Japanese Learners of English)</i>	103
Figure 30 <i>Reaction Times of the Semantic Stroop Task in Each Color (Critical Items)</i>	107
Figure 31 <i>Reaction Times in Each Color (Filler Items)</i>	108
Figure 32 <i>Reaction Times of the Semantic Stroop Task in Each Color (Native English Speakers: Critical Items)</i>	122
Figure 33 <i>Reaction Times of the Semantic Stroop Task in Each Color (Native Japanese Speakers: Critical Items)</i>	123

Chapter 1: Introduction

Background

Many people who speak more than one language are likely to find that their proficiency in one or the other varies. Learners of a second language (L2) often find it more difficult to understand verbal information in their L2 than in their first language (L1). Investigating the causes of this difficulty will provide useful information for creating effective teaching and learning materials for L2 learners, as well as provide clues to the mechanisms of language processing and how L1 and L2 knowledges are stored.

Studies of L2 lexical processing have investigated the mechanisms of L1 and L2 processing and the development of their knowledge. These studies hypothesize that our L1 interacts with the L2 to understand the words in the L2 (e.g., Dijkstra et al., 2019; Dijkstra & Van Heuven, 2002; Jiang, 2000; Kroll & Stewart, 1994). These mechanisms of vocabulary processing or the structures of vocabulary knowledge have been described and studied using various models. Jiang (2000) explained that when we encounter a new word in the L2, we must first connect it to the equivalent translation in our L1. For instance, when a Japanese learner of English first encounters the word *dog*, they have to find out what the string composed of the letters “d,” “o,” and “g” means in Japanese (*inu* is the corresponding word in this example). Jiang (2000) assumed that our L1 already has a direct connection to the concept. Therefore, we can understand the L2 word through our L1. L2 proficiency changes the relational structure between L1, L2, and the concept. Some models assume that as L2 proficiency increases, the learner can access the concept without relying on translation equivalents (e.g., Jiang, 2000; Kroll & Stewart, 1994). The models have been contributing to revealing how we understand the words in the L1 and L2. The

results of these studies are also used to explain why some methods of vocabulary learning are more effective than others (e.g., Terai, 2019; Terai et al., 2021).

However, these models need to be further explored to provide a more comprehensive account of our use of language. For example, the Revised Hierarchical Model (Kroll & Stewart, 1994), the BIA+ model (Dijkstra & Van Heuven, 2002), and the Three-Stage Model (Jiang, 2000) cannot fully explain what we understand after orthographic or phonological processing is complete; more specifically, what process takes place after we access the concept. The specific nature of the concept has not been the main target of studies on L2 vocabulary processing. However, recent findings have shown that it is not just about linguistic processing. Previous studies have reported that L2 learners imagine the shape of the word referent less than in their L1 (Hayakawa & Keysar, 2018). In addition, some studies have reported that affective processing was reduced in a foreign language (Pavlenko, 2017 for a review). Thus, this research suggests that there might be a difference in understanding conceptual details between L1 and L2. Current models of L2 lexical processing need to be reconsidered to explain the processing of non-linguistic aspects.

Aim of the Present Study

The studies of L2 vocabulary processing have yet to reveal what we understand from verbal information. Studies of *embodied cognition* help fill this gap. They have studied the interplay between linguistic and non-linguistic processes to understand the meaning of words (e.g., Barsalou, et al., 2008). For example, in one empirical study, it was reported that the corresponding picture was recognized more quickly after the sentence was presented (Zwaan et al., 2002). Readers recognized a picture of an eagle with outstretched wings faster than a picture of an eagle with folded wings after reading the sentence *the*

ranger saw the eagle in the sky, and the difference was reversed after they read *the ranger saw the eagle in the nest*. This match effect implied that readers mentally simulate the eagle that stretches its wings when they read the sentence *the ranger saw the eagle in the sky*; they mentally simulate the eagle that folds its wings when they read *the ranger saw the eagle in the nest*. The results show that readers mentally represent what the verbal information implies without seeing the actual object. Empirical research on embodied cognition has focused mainly on L1 processing, whereas very little is known about L2 processing. Therefore, research on embodied cognition can contribute to studies of vocabulary processing by examining conceptual representation in the L2.

The present study aims to determine whether L2 learners can mentally represent the detailed images of words and what influence their L2 proficiency has. Among the various elements of images, the study focuses on color. There are three reasons for this. First, previous studies have reported that color plays an important role alongside other visual aspects of an image (e.g., shape, size, and orientation) (e.g., de Koning et al., 2017; Zwaan & Pecher, 2012). Second, studies suggest a strong correlation between visual aspects (de Koning et al., 2017). This suggests that the activation of color is associated with the activation of other aspects, such as the shape and size of the object. Third, methodologically, color makes it easier to manipulate experimental items. There are some objects that are the same size but different in color, such as bears. There are brown, black, and white bears, and the color-referent combination might be different in terms of typicality. We can compare the results by changing colors while keeping other elements (e.g., shape and size) the same.

Study 1 investigated whether readers activate object colors while reading words in their L1. This task was implemented for native English and Japanese speakers. This study

was meant to take baseline L1 data. The semantic Stroop task used in Connell and Lynott (2009) was employed with some modifications. In the original task, participants read a sentence, then are presented with a word from the sentence in a colored font. They must name the color of the font as quickly as possible. If a participant mentally represents the object color, the naming speed of the font color that matches its representation will be significantly faster than that of the non-matching color. With this task, we can find out which color readers mentally represent when they read words in their L1 (therefore, baseline data).

Study 2 was conducted with L2 English learners, using the same task as in Study 1. Study 2 examined the activation of color in L2 processing, comparing the results with those of L1 speakers. Subsequently, it was also investigated whether L2 proficiency modulated representational patterns. The models of L2 vocabulary assume that learners with lower language proficiency can also access the concepts through their translation equivalents in L1 (Jiang, 2000; Kroll & Stewart, 1994). If they are correct, there should be no impact of L2 proficiency. However, if there is an L2 proficiency effect, the existing L2 models cannot account for developmental changes in learners' conceptual understanding. Therefore, studying L2 proficiency is important to test existing L2 models.

Outline of the Dissertation

Chapter 2 begins with a literature review of L2 vocabulary studies. There are several models of L2 vocabulary processing and representation. Models that are relevant to the present research are explained in detail, such as the Revised Hierarchical Model by Judith Kroll and Erika Stewart (1994) and the Three-Stage Model by Nan Jiang (2000). These models have made an important contribution to studies on L2 vocabulary acquisition and are used as a theoretical background for the research area. Although both models

include a conceptual representation, the specific role of the concept has not yet been clarified. To address this problem, the next section introduces the idea of *embodied cognition*. In this research paradigm, we assume that language processing involves not only the manipulation of symbols, but also the activation of the mental image of the object to which the symbols refer. We can conclude that language processing involves generating rich images from artificial symbols. First, an overview of the paradigm is given, and the following section describes empirical research that supports the hypothesis of *embodied cognition* in language processing in both L1 and L2. The final section of the chapter identifies the limitations of previous studies, considering both L1 and L2 research. Chapter 2 concludes with the research questions and hypotheses of the study.

Chapter 3 describes two pilot studies. The pilot studies were conducted to create the experimental items. The chapter begins with how each material was created based on previous studies and experiments. The conditions and criteria for the materials are also explained. The chapter ends with a section on how the research determined the required sample size. Chapter 4 and Chapter 5 discussed the experiments with L1 and L2 participants, respectively. In Experiment 1 (Chapter 4), L1 speakers performed the semantic Stroop task in their L1, followed by the word and sentence rating tasks. In Experiment 2 (Chapter 5), Japanese learners of English performed the same tasks (the Stroop task and the rating tasks) in their L2. They also participated in a vocabulary test to measure their L2 proficiency. Both chapters begin with the details of the experimental designs, procedures, and analyses. The final section explains the results and discusses them with the research questions and hypotheses. Chapter 6 begins with a summary of both the L1 and L2 studies. The remainder of the chapter describes the more general discussion of language processing based on the research findings. The chapter ends with the limitations

and future directions of the study. Chapter 7 contains the summary and conclusions of the study.

Chapter 2: Literature Review

Background

Processing and Representation of L2 Vocabulary

Models of Bilingual Mental Lexicons. Studies of bilingual mental lexicon have examined how bilinguals, including L2 learners, process words and how their knowledge of L2 and L1 words is represented. Previous studies have proposed models to understand the complex interaction between L1, L2, and their referents (concepts) (e.g., de Groot, 1992; Dijkstra et al., 2019; Dijkstra & Van Heuven, 2002; Jiang, 2000; Kroll & Stewart, 1994; Paivio et al., 1988; Pavlenko, 2009). This approach was expressed in the title of Brysbaert et al. (2010): “Models as hypothesis generators and models as roadmaps.” For example, some models, such as the Multilink Model (Dijkstra et al., 2019), are used to simulate the activation patterns of vocabulary knowledge computationally. On the other hand, other models, such as BIA+ model, the Revised Hierarchical Model (Kroll & Stewart, 1994) and the Three-Stage Model (Jiang, 2000), are used to understand the L2 vocabulary processing conceptually. This difference does not mean that computational models are used only to simulate patterns of L2 vocabulary processing and that conceptual models are not used for simulation research. The BIA+ Model is also considered as a computational model (e.g., Chuang et al., 2021; Li & Xu, 2022). In the following, the assumptions of the models are compared. However, the review focuses only on the models related to the current research, the Revised Hierarchical Model and the Three-Stage Model, since validation of the other models is not the main objective of this study.

The Revised Hierarchical Model (Kroll & Stewart, 1994) is one of the most influential models in the research of bilingual lexical processing. The model consists of separate L1 and L2 mental lexicons and concepts. It is a “revised hierarchy” because it

includes the assumptions of the two other models of the mental lexicon: word association and concept mediation (Potter et al., 1984). The word association model assumes that newly learned L2 words are directly associated with words in the L1 when the L2 is weaker than the L1. In contrast, the concept mediation model states that L2 words are linked to non-linguistic concepts but not directly to L1 words. The Revised Hierarchical Model highlights the asymmetry in the strength of lexical connections from either L2 to L1 or L1 to L2. The lexical connection from L2 to L1 is stronger than from L1 to L2. This assumption follows from the fact that even relatively fluent bilingual speakers know more words in their L1 than in their L2. The translation speed was faster in L2 to L1 than in L1 to L2 (e.g., Kroll & Stewart, 1994). The model includes the developmental hypothesis that the structure of lexical knowledge changes as a speaker's L2 proficiency increases. This point is discussed in more detail in the following section.

As for the connections between the L1 or L2 lexicon and the concepts, it is assumed that L1 has stronger connections than L2. The L2 has weaker connections because learning of L2 words usually begins with the mapping of L2 words to their translation equivalents in L1. Speakers are more likely to activate concepts when translating from L1 to L2 than when translating from L2 to L1 because the link to concepts is stronger. This assumption was also supported by the results of their experiments (Kroll & Stewart, 1994). They compared the translation speed of L1 to L2 and L2 to L1. The participants were asked to translate the list of words whose categories were identical (e.g., *dress, suit, shoes*) or randomized (e.g., *orange, lion, ambulance*). They found that the translation speed was slower under the identical conditions. In addition, the trend was much more pronounced when translating from L1 to L2. This category interference in translation from L1 to L2 was evidence of stronger concept activation during translation from L1 to L2. In contrast,

weaker or absent interference during translation from L2 to L1 was interpreted as evidence for more lexically mediated translation in that direction. However, this does not mean that translation from L2 to L1 is never mediated by concepts. Brysbaert and Duyck (2010) argued that not only learners with very high L2 proficiency but also learners with lower proficiency are influenced by concepts in the translation from L2 to L1.

The Revised Hierarchical Model was originally proposed to account primarily for the translation asymmetry occurring in production tasks (e.g., Kroll et al., 2010); it was not clearly mentioned that the model was for word production when they introduced the model (Kroll & Stewart, 1994). However, some studies focusing on L2 word recognition use the model as a theoretical background or investigate its validity using translation recognition tasks (e.g., Poarch et al., 2015; Talamas et al., 1999; Terai et al., 2021; Wu & Juffs, 2019). Brysbaert et al. (2010) reported that 83 studies were about perception out of 166 studies that cited the Revised Hierarchical Model between 1994 and 2009 (the other 82 dealt with production, and one research was unable to be classified).

The Development of Mental Lexicon. Learners' L2 proficiency must be taken into account when understanding L2 language processing and representation. The structure of the bilingual mental lexicon may not be stable; consequently, the processing of L2 words depends on the learner's L2 proficiency. Some models assume that the relationship between the L1 and L2 mental lexicon changes as the learner's L2 proficiency increases (e.g., Jiang, 2000; Kroll et al., 2002; Kroll & Sunderman, 2003). The Revised Hierarchical Model involves a developmental hypothesis about the strength of connectivity among L1 lexicon, L2 lexicon, and concept. The Revised Hierarchical Model hypothesizes that the lexical connection from L2 to L1 is stronger than from L1 to L2. However, the asymmetry between the two connections decreases as the speaker's L2 proficiency increases (Kroll et

al., 2002; Kroll & Sunderman, 2003). For example, Kroll et al. (2002) compared the translation speed of the participants whose L2 proficiencies varied. In the experiment, participants translated L2 words into L1 words or L1 words into L2 words. The results showed that participants were faster when translating L2 words into L1 than when translating L1 into L2, consistent with the Revised Hierarchical Model. Interestingly, the speed difference was greater for the less proficient participants than for the more proficient participants.

Jiang (2000) proposed the Three-Stage Model of bilingual mental lexicon. The uniqueness of this model is that it explains how each word develops during the learning process. It is possible to estimate where L2 learners are among the stages because most L2 words correspond to their L2 proficiency level. Jiang (2000) suggested that not all words in an L2 learner's lexicon are at the same levels. The Three-Stage Model assumes that the L1 lexical representation includes four components: two lexeme information (phonology/orthography and syntax) and two lemma information (semantics and morphology). In contrast, the lexical representation of the L2 in the initial stage contains only the L2 phonology and orthography. At this stage, the use of the L2 word relies on the L1 translation equivalents because the L2 lexical representation does not contain semantics, morphology, and syntax. Thus, accessing the concept requires L1 translation equivalents. Repeated activation of the L2 word leads to the strength of the connection between L2 words and their L1 translation equivalents, which develops the representation. In the second stage, the L1 lemma of the translation equivalents is copied into the lexical representation of the L2. In contrast to the first stage, the L2 word can access the concept directly (with the copied lemma) and indirectly (via L1 translation equivalents). The L2 word provides a direct link to the concept in this stage; however, the direct link is weak

and the asymmetric assumption is similar to the Revised Hierarchical Model. In the third stage, the lexical representation in the L2 is fully anchored in the L2 information. This is the complete development of the L2 word. Access to concepts does not require the use of L1 information.

Jiang notes that most words stop developing in the second stage because of two constraints: first, the absence of strongly contextualized L2 input; second, the presence of the L1 lemma in the L2 representation. This makes it more difficult for L2 learners to form L2 lemmas in their L2 lexical representation (Jiang, 2000, 2002), and the tendency was called *lexical fossilization*. L2 collocational research has been contributing to investigating lexical fossilization (e.g., Wolter & Gyllstad, 2011, 2013; Wolter & Yamashita, 2015, 2018; Yamashita & Jiang, 2010). By the time a learner reaches the final stage, there should be no more L1 influence because the words at that stage are directly linked to the concepts. L1 influence on online lexical processing means that most words in the learner stop at the second or first developmental stage. Previous research has shown that even advanced learners were under the influence of the L1 (e.g., Wolter & Gyllstad, 2013; Wolter & Yamashita, 2018, except for Yamashita & Jiang, 2010). For example, Terai et al. (under review) investigated whether an increase in learners' L2 proficiency reduced the L1 influence on on-line L2 collocational processing. They conducted an acceptability judgment task with Japanese learners of English. The results showed that even with high level learners who were assumed to have mastered the CEFR C1 level, most of the words in their L2 stop at the second developmental stage.

To summarize, L2 models of the bilingual mental lexicon admit the importance of L2 proficiency in revealing the structure and process of L2 words. However, even for advanced L2 learners, most of their words are before the final stage of lexical

development. More specifically, although they can make a direct connection from the L2 to their concept, their lexical processing in the L2 is still influenced by their L1. Thus, both direct and indirect connections (via L1 translation equivalents) were present in any L2 lexical information.

The Revised Hierarchical Model (Kroll & Stewart, 1994) and the Three-Stage Model (Jiang, 2000) have contributed significantly to studying lexical processing and representation in the L2. They serve as a theoretical background not only for studying lexical processing and representation in the L2, but also for vocabulary learning in the L2 (e.g., Terai et al., 2021). However, these models do not fully account for speakers' understanding of word meaning. More specifically, what do people represent when they process a word? More recently, Pavlenko (2009) argued that lexical concepts are not amodal but rather multimodal mental representations. The concepts include visual, auditory, perceptual, and kinesthetic information as implicit memory. Although Pavlenko (2009) does not provide a detailed explanation for the multimodality of lexical concepts, this view is consistent with recent studies in cognitive psychology (e.g., Barsalou, 1999; Barsalou et al., 2008). The studies in language processing have provided evidence that people activate not only linguistic but also non-linguistic knowledge during language processing. These areas of research will shed new light on the conceptual knowledge of L2 learners. The details of the activation of non-linguistic knowledge will be discussed in the next section.

Embodied Cognition

Earlier theories of human cognition assume that cognitive representations are not inherently perceptual and are referred to as the amodal view of cognition (e.g., Fodor, 1975/1979). Although perceptual states arise in sensory-motor systems, they are

transformed into an entirely new representation of language that is not perceptually related (Barsalou, 1999). Knowledge thus exists separately from the modal perceptual systems of the brain, such as vision and hearing (Barsalou, 2008). Models of amodal systems theory speculate that language comprehension depends on abstract and amodal symbols that are arbitrarily mapped onto referents. Amodal theory has remained dominant in the study of language comprehension since the onset of the cognitive revolution in the 1950s (e.g., Horchak et al., 2014). However, the assumption of the amodal view has been criticized (Barsalou, 1999; Barsalou et al., 2008). One of the most famous criticisms of the purely symbolic model is *the symbol grounding problem* (Harnad, 1990, and also see Searle's [1980] *Chinese Room Argument* for the original proposal of the criticism about the amodal view). *The symbol grounding problem* is that the symbols that are not grounded cannot have meaning. One of the best-known examples of this problem is the confusion in Chinese dictionaries. It is impossible to learn Chinese when the only source of information we have is a Chinese-Chinese dictionary. The reason is that the information in the dictionary consists of symbols that are meaningless to the learner who does not know Chinese. In addition, Barsalou (1999) introduced the following other problems of amodal theories: lacking direct empirical evidence that amodal symbols exist, being challenged by neuroscience research, and failure to provide a satisfactory account of the transduction process.

In contrast to Amodal theory, more recent theories propose that knowledge is embodied in the modal system of the brain. Embodied theories of cognition suggest that symbols are grounded in their references to the environment and challenge the amodal views (e.g., Horchak et al., 2014). The embodied view assumes that language comprehenders create simulations to represent the meaning of the texts (e.g., Barsalou,

1999, 2008; Barsalou et al., 2008). Barsalou (2008) defined simulations as “the reenactment of perceptual, motor, and introspective states acquired during the experience with the world, body, and mind” (p. 618). For example, readers mentally represent an image of the dog and his owner when they read the sentence, “*John takes his dog for a walk.*” The simulation also includes the color and the orientation of the dog and this simulation is updated as readers proceed with the sentences (Barsalou, 1999). The theory of perceptual symbols postulates that our simulation is limitless and generated by a simulator. In perceptual theory, a simulator is equivalent to a concept. A simulator is organized with related perceptual symbols (Barsalou, 1999). It is important to note that mental simulation and mental imagery are not equivalent in the theory (Barsalou, 1999, 2008). Mental simulations are thought to occur automatically and unconsciously outside of working memory. In contrast, mental images are consciously constructed in working memory (Barsalou, 2008).

Theories that support the embodied view have been proposed in the past few decades, such as Perceptual Symbol Theory (Barsalou, 1999), Indexical Hypothesis (Glenberg & Robertson, 1999), Action-Based Language (Glenberg & Gallese, 2012), Immersed Experienced Framework (Zwaan, 2004), Language and Situated Simulation (Barsalou et al., 2008), and Symbol Interdependency System (Louwerse & Connell, 2011) (see Horchak et al., 2014 and Mochizuki, 2015 for summaries of the aforementioned theories). These theories are categorized into a strong embodied view or a moderate embodied view (e.g., Horchak et al., 2014; Mochizuki, 2015). Theories that support a strong embodied view are Perceptual Symbol Theory (Barsalou, 1999), Indexical Hypothesis (Glenberg & Robertson, 1999), Action-Based Language (Glenberg & Gallese, 2012), and Immersed Experienced Framework (Zwaan, 2004). Theories supporting a

strongly embodied view claim that human cognitive processing, including language processing, always activates sensory-motor knowledge (e.g., Horchak et al., 2014; Mochizuki, 2021). In contrast, more recently proposed models such as Language and Situated Simulation (Barsalou et al., 2008) and Symbol Interdependency System (Louwerse & Connell, 2011) support a moderate embodied view. In this view, the theories assume that both the linguistic system and the simulation system are involved in the cognitive process (e.g., Horchak et al., 2014; Mochizuki, 2021).

Barsalou's language and situated simulation (LASS) theory assumes that both language systems (linguistic forms) and simulation systems (situated simulations) represent knowledge. They interact continuously to produce conceptual processing (Barsalou et al., 2008). LASS Theory assumes that both the linguistic and simulation systems are activated when a word is perceived. Still, at the initial conceptual processing, the linguistic system plays a central role (i.e., more active than the simulation system). This linguistic processing is considered more superficial than situated simulation. Recognition of the word leads to activation of the associated simulations, which is often automatic and rapid. Therefore, the simulations are more situated. These simulations are allowed due to the presence of *simulators*. Barsalou (1999) assumed that "simulators have two levels of structure: (1) an underlying frame that integrates perceptual symbols across category instances, and (2) the potentially infinite set of simulations that can be constructed from the frame" (p. 586). Barsalou et al. (2008) mentioned the distinction between a linguistic and a simulation system. However, it is only for simplifications to focus on mechanisms of research interest; he does not mean that the two systems are unrelated.

The ideas of the situated simulations are applied to both "concrete" concepts and "abstract" concepts (e.g., Barsalou et al., 2008; Barsalou et al., 2018; Connell & Lynott,

2012). Some may think that it is impossible to create a simulator for abstract concepts like love because the concept is invisible. Certainly, the types of concepts have been represented differently because the foci of their situations are different. Nevertheless, both concrete and abstract concepts represent situations. The framework assumes that concepts emerged from processing situations. Concrete concepts focus more on objects and settings (perceptual and movement information). Abstract concepts focus more on mental states and events (social, introspective, and affective information). Humans focus on the focal content, whether it is concrete or abstract. For example, love and fear relate to a person's internal mental state (affective information), and abstract concepts may be more integrated than concrete concepts. However, love and fear also integrate stimuli that elicit these emotions, such as parents and spiders (Barsalou et al., 2018). Therefore, situated simulations can explain the processing of words that refer to both concrete and abstract concepts (see Barsalou [1999] for more discussions of abstract concepts such as negation).

In summary, human cognition studies, including language processing research and amodal views, have been challenged by embodied views. The embodied views assume that cognition is based on modal simulations. In addition, theories that support a moderate view, such as Language and Situated Simulation (Barsalou et al., 2008) and Symbol Interdependency System (Louwerse & Connell, 2011), are more supported than theories that support a strong embodied view.

Empirical Support for the Embodiment Approach

Some studies empirically investigated the theory of embodied cognition (e.g., Bergen et al., 2007; Connell & Lynott, 2011; de Koning et al., 2017; Garofalo & Riggio, 2022; Pecher et al., 2009; Richter & Zwaan, 2009; Rommers et al., 2013; Yaxley & Zwaan, 2007). Previous studies have examined the relationship between language

processing and various aspects of embodied representations (e.g., visual, motor, and affective). Because this study aims to examine the visual aspects of representation, the review focuses on the studies of visual aspects from behavioral studies. For instance, Zwaan et al. (2002) investigated the activation of an object's shape during sentence processing. They conducted a sentence-picture verification (SPV) task. In the task, participants first read a sentence and were then asked to judge whether the object was mentioned in the previous sentence. They hypothesized that when readers represented the visual aspects of the objects mentioned, they also represented what the sentence implied. Thus, readers checked more quickly whether the shape of the objects matched the shape of the image presented after reading the sentence. For example, the reaction times to a picture of an eagle with outstretched wings would be faster after they read "*The ranger saw the eagle in the sky*" than after reading "*The ranger saw the eagle in its nest.*" The results supported their hypothesis that readers represent the shape of the mentioned object during sentence processing.

Connell (2007) conducted an SPV task to investigate whether readers simulate the color of an object. Participants were asked to judge whether a presented picture matched the preceding sentence. Each sentence was paired with a sentence that implied the same object but with a different color. For example, "*John looked at the steak on his plate*" represents a brown steak because the sentence implies that the steak is cooked. Thus, the picture of a brown steak is the matched condition. A red steak was used as the mismatched condition. This sentence was paired with, "*John looked at the steak in the butcher's window.*" In this case, the picture of the red steak is the matched condition and the picture of the brown steak is the mismatched condition. In contrast to the matching advantage observed when trials targeted other visual aspects (e.g., Zwaan et al., 2002), the matched

condition was significantly slower than the mismatch condition. Although Connell (2007) concluded based on the results that color is simulated during language processing, the representation of color might be different from other visual objects, such as shapes, because color can only be perceived by one sense.

Zwaan & Pecher (2012) conducted large-scale replication studies. They replicated studies of orientation (Stanfield & Zwaan, 2001), shape (Zwaan et al., 2002), and color (Connell, 2005, 2007) via Amazon's Mechanical Turk to recruit participants who were varied in age and educational background. In the study, each visual trait (orientation, shape, and color) was examined separately, and each trait was replicated with two experiments, resulting in six experiments. In total, data from 992 participants were statistically analyzed after the elimination process. They used the same experimental items as in the original studies, except for the filler items and the comprehension questions. The study replicated the benefits of matching orientation and shape. The results showed that color also showed a matching benefit, i.e., reaction time was shorter in matching conditions than in mismatching conditions, which contradicts the results of Connell (2005, 2007). Furthermore, the effect size of the color was as large as the shape and was bigger than the orientation (Color: Bayes Factor (BF_{01}) = 0.01, Shape: BF_{01} = 0.01, Orientation: BF_{01} = 0.04).

De Koning et al. (2017) have reported that color showed the strongest match effect among visual aspects (shape, size, color, and orientation). They performed SPV tasks with a within-subjects design. In a single session, participants saw a sentence that implied one of the aspects and pictures. The results showed that color had the strongest matching advantage, followed by shape and size. Orientation, however, showed no matching advantage. They also examined the relationship between all visual aspects. A correlation

analysis showed that color, shape, and size were significantly correlated; however, orientation did not correlate with the three aspects. Therefore, the results contrast with the results from Connell (2007), that reported opposite trends of color match advantage but agreed with the findings of Zwaan & Pecher (2012). They suggested that L1 readers mentally simulate the shape, size, and color of the indicated object, but not necessarily the orientation of the object. The simulation of the three visual properties was an interesting result. We can infer that the other properties were also simulated when we found activation of one of the properties. Bai et al. (2022) investigated this point more directly with an SPV task. The results showed that both shape and color properties were integrated into the simulation incrementally.

The Quality of Simulation and Mental Images

Previous studies have reported that embodied knowledge is activated during language comprehension. However, it is also important to reveal how much the simulation is sophisticated to understand the mechanisms of language processing. For example, Zwaan et al. (2002) revealed that readers activate the shape of an object; however, the results did not shed light on how vivid the shapes were. Hoeben Mannaert et al. (2017) investigated how much visual information is included in a mental simulation. They performed an SPV task with materials that differed in color saturation. They hypothesized that the mismatch condition would have a greater discrepancy at full color saturation than at reduced color saturation when readers vividly simulated the color. The first study conceptually replicated Zwaan and Pecher (2012) with the full-color items. Results showed that reaction times were faster when the color of a presented image matched the color implied in the preceding sentence than when the image and implied color did not match. In a second study using items with reduced saturation, it was found that the matching

advantage became smaller when saturation was reduced. Thus, the degree of saturation influences mental simulation.

It is also important to know whether readers can simulate multiple images during language comprehension. Connell and Lynott (2009) also examined one of the visual aspects: color, but from a different perspective. They investigated whether sentence comprehension involves the activation of color and whether we can simulate two objects during sentence comprehension. Fifty-four native English speakers were recruited for the study. They performed a semantic Stroop task. In this task, they were first presented with a sentence. After the participants read the sentence, they were presented with a colored target word. They named the color of the target word as quickly and accurately as possible. When a reader activated the color of the word, the response to the ink matching the color in the mental simulation was expected to be faster than in the mismatch condition. There were two conditions with sentences implying the typical or atypical color of the object and three conditions with the ink of the target word: the typical, the atypical, or the unrelated color of the object. Although there was no significant interaction between the typicality of the sentence and the color of the ink, the congruence effect between the implied color and the color of the ink was observed. When a sentence implied the typical color of the object, participants responded faster to the word colored with the typical color. Interestingly, a similar tendency was observed in an atypical condition; however, response times were similar for typical and atypical words, which were much faster than the unrelated color. Thus, they found that color representation is activated during the processing of L1 sentences. When the sentence implied an atypical color of the object, readers activated both typical and atypical colors.

Embodied Cognition Research in L2

In the previous section, the author reviewed studies on the question of whether embodied knowledge is activated during L1 language processing. Many studies have provided evidence for the activation of embodied knowledge in various aspects. Do readers show the same tendency during L2 processing? Or is the extent of embodied knowledge activation lower during L2 processing? Jiang (2000) argued that there are practical constraints in L2 learning that lead to a fundamental difference between L1 and L2 lexical development. One of the limitations is that in L2 learning, there is already an established conceptual/semantic system of the L1. Thus, when L2 learners learn new vocabulary, they tend to rely on their L1, especially if they are adult learners. In Jiang (2000)'s model, L2 learners need to activate L1 to access its concepts at the initial stage of development. However, as learners' L2 proficiency increases, they no longer need to rely on L1 to understand the concepts. Thus, L2 proficiency can be expected to influence the relationship between L2 processing and simulations.

There is another important difference between L1 acquisition and L2 learning. In English as a foreign language, such as in Japan, L2 learning takes place mainly in the classroom. L1 acquisition involves more experience with the world than in a EFL context. Therefore, unlike L1 acquisition, which involves more sensory-motor experiences, L2 learning may not lead to a strong connection between L2 forms and embodied knowledge, as L2 learning usually occurs through symbol manipulation, such as translation from L1 to L2 or vice versa, as Kühne & Gianelli (2019) argue. Even if we are able to acquire embodied knowledge for L2 from L1 equivalents, there might be a difference in the extent of embodiment. For example, L2 learners might acquire more embodied knowledge with familiar concepts than with unfamiliar concepts. In summary, investigating whether

embodied knowledge can be activated during both L1 and L2 processing is important not only for research on L2 acquisition but also for research on embodied cognition. Moreover, it is important to investigate the influence of L2 proficiency and the context of L2 acquisition to uncover the relationship between language processing and embodied knowledge.

Contrary to research in L1, there is far less research targeting L2 processing (e.g., Ahn & Jiang, 2018; Athlberg et al., 2018; Awazu & Suzuki, 2020; Buccino et al., 2017; Dudschig et al., 2014; Norman & Peleg, 2021; Patterson, 2021; Vukovic & Williams, 2014; Monaco et al., 2021). Buccino et al. (2017) examined whether L2 learners have active motor representations during L2 processing in a go-no go paradigm. They recruited native Italian speakers whose English proficiency corresponded to reference level C1 of the Common European Framework of Reference for Languages (CEFR). Participants judged whether the stimulus presented referred to a real object or not. The stimulus was either a word (a noun or a pseudoword) or a picture (which referred to a real object or a scrambled picture that made no sense). If the stimulus referred to a real object (a noun or a picture that referred to a real object), participants were asked to press the button. If the stimulus was a pseudoword or scrambled picture, they were instructed not to respond. Results showed that participants responded more slowly when the stimulus (nouns and pictures) was graspable (e.g., ear, leaf) than when the stimulus was non-graspable (e.g., air, thunder). According to the researchers, the slowing of reaction times reflects the cognitive cost of simultaneously activating the motor system, which is activated in two routes (the physical movement of pressing keys and the activation by seeing the graspable stimuli). They concluded that the difference in reaction times reflects the activation of the motor representation in L2.

In line with the study, Ahn and Jiang (2018) and Vukovic and Williams (2014) reported that highly proficient L2 learners activated visual aspects of embodied knowledge during L2 processing. The activation of embodied knowledge in L2 processing was also observed in the study targeting learners with much lower L2 proficiency. Awazu and Suzuki (2020) investigated whether sensory-motor representation is activated during sentence processing by Japanese learners of English whose L2 proficiency ranged between A2 to B1 of CEFR. They performed a sensible judgment task on both L1 and L2. In the task, participants judged whether the presented sentence was acceptable. They found that even low L2 proficiency learners activate their sensory-motor representation during L2 sentence processing.

Kogan et al. (2020) reviewed 29 articles (34 experiments) that examined whether embodied knowledge is activated in action-related words. Based on this review, they concluded that embodied knowledge is activated in both early and late learned languages and that early language exposure may not be necessary for embodiment to occur (but see Monaco et al., 2021 for results suggesting differences in embodiment between L1 and L2).

The studies above found that L2 learners activate their embodied knowledge during L2 processing. Moreover, it may not depend on L2 learners' proficiency levels. However, some studies did not find the activation of embodiment knowledge. For example, Norman and Peleg (2021) investigated whether L2 comprehension includes perceptual visual simulations. They performed an SPV task for 80 late Hebrew-English bilinguals. They performed the task in L1 and L2, with language order counterbalanced between participants. They found that the match/mismatch condition affected reaction times only in L1. Moreover, the significant difference occurred only when participants performed the

task in L1 before the task in L2. They concluded that embodied knowledge might not be activated during the processing of L2 sentences.

Chen et al. (2020) reported a similar result. They investigated whether shape information was simulated during L2 processing with delayed SPV tasks. In the task, participants listened to sentences and decided whether each sentence made sense. About 10 minutes later, they were presented with pictures and decided whether they had been mentioned in the sentences they had heard in the previous phase. This task allowed the researchers to determine whether the participants' embodied knowledge was strong or durable. They compared the influence of learners' L2 proficiency as within-subject factors. The participants speak Cantonese as their L1, Mandarin as their L2, and English as their L3. The results showed that the match effect occurred only in the participants' L1 but not in their L2 or L3, regardless of their proficiency (within-subject factor).

Thus, previous research has not concluded that L2 processing involves the activation of embodied knowledge. Moreover, the influence of L2 proficiency and learning context, such as EFL vs. ESL, needs further investigation. Some studies that have found evidence of simulation in the L2 have not comprehensively considered L2 skills. For example, the effects of L2 proficiency were not statistically examined (e.g., Buccino et al., 2017; Vukovic & Williams, 2014) or learners with relatively high L2 proficiency were recruited. To illustrate, Buccino et al. (2017) reported that the participants were at the C1 level of the CEFR. Vukovic and Williams (2014) reported that the mean of the self-report L2 proficiency (7-point Likert scale) was 6.4 in writing ($SD = 0.59$) and 6.3 in speaking ($SD = 0.65$). Further research is needed to reveal whether lower proficiency learners simulate during L2 processing.

The Present Study

Previous studies have shown that different aspects of embodied knowledge are activated during L1 language processing. This study focused primarily on color, as color is thought to play one of the most important roles along with other visual aspects (shape, size, orientation) (de Koning et al., 2017). Furthermore, color facilitates the manipulation of experimental objects to reveal the content of the representation (e.g., Connell & Lynott, 2009; Hoeben Mannaert et al., 2017).

The relation between object color-typicality and simulations, which was investigated by Connell and Lynott (2009), is intriguing in terms of the interaction of conceptual knowledge and language processing. However, some issues should have been further examined in Connell and Lynott (2009). First, the significant interaction of sentence color and word color was not observed in a strict sense ($p = .057$). Nevertheless, they performed a simple effects analysis to compare speed between sentence levels. In the atypical sentence condition, they found no significant difference between reaction times for typical and atypical word colors, and reaction times for these colors were significantly faster than for unrelated colors. In the typical sentence condition, reaction times for typical color words were faster than for atypical and unrelated colors (there was no significant difference between atypical and unrelated colors). They took these results as evidence for dual activation of typical and atypical colors when reading atypical sentences. However, there was no main effect of sentence typicality; only the effect of word typicality was significant. Thus, their results only suggest that typical colors alone were activated, regardless of sentence typicality.

Second, their results show that the activation of both typical and atypical colors of the object might depend on their experimental items. Eighty percent of the target sentences

were context-end, in which a phrase that determines the typicality of the target's color was placed after the target word (e.g., *Joe was excited to see a bear at the North Pole*). Thus, participants in their study may have responded more quickly to typical and atypical words because they activated the typical color when they read the target word (*bear*) and then changed the image of the color after they read the contextualizing phrase. Some studies have reported that readers simulate an object in the middle of the sentence and update the image based on the implied information (e.g., Sato et al., 2013).

Sato et al. (2013) reported that readers activate the representation of object shape before they finish reading the sentence. The representation of the shape can be changed quickly depending on the meaning of the sentence. Further, Kang et al. (2020) investigated whether grammatical tense markers influenced the simulation of object-state change during reading sentences in an L1. They performed an SPV task. The experimental sentences differed in either past or future tense in two experiments (Experiment 2 and Experiment 3) (e.g., *The woman dropped/will drop the ice cream*). They found that participants responded more quickly when the sentences were written in the past tense and the object state of the picture (e.g., a dropped ice cream) matched what the sentences implied (*The woman dropped the ice cream*). However, when the sentences were written in the future tense, the match advantage was replicated only when the sentences implied the change (*The woman will drop the ice cream*). In contrast, the reaction times to pictures of both the original and changed form of the objects were not significantly different after reading the original sentence that did not imply the change (e.g., *The woman will choose the ice cream*). They argued that in the future tense, both the original and the changed state could be simulated since the sentence does not express the end of the state of the objects.

Horchak and Garrido (2021) also investigated object state change during the reading L1 sentences with an SPV task. They compared the two possibilities of simulating an object state change: a constant scenario and a competing scenario. A constant scenario predicts that “only the consequences of the described action will be encoded” (Horchak & Garrido, 2021, p. 4). A competing scenario predicts that “both the initial canonical and the end non-canonical states of the object would be equally integrated into the mental model” (Horchak & Garrido, 2021, p. 5). Based on their seven experiments with different conditions, they found that both the initial and changed states are equally accessible in simulations when a context implies a change implicitly. According to their results, for example, L1 English readers simulate both the unsquashed tomato and squashed tomato when they read, “*A bowling ball fell on a tomato.*” Consequently, they supported a competing scenario.

Based on these results, we can infer that L1 readers simulate atypical and typical colors when they read sentences that imply atypical colors, which is consistent with the arguments of Connell and Lynott (2009). L1 readers simulate the typical color of the object (e.g., *bear* in brown) when reading the word (*bear*), then update the image or also simulate a different image of a bear (*bear* in white) when reading *at the North Pole*. However, previous studies have not shown whether L1 readers simulate both typical and atypical colors when reading sentences that determine color typicality (e.g., *at the North Pole*) before reading the key words (*bear*). Providing the context could prevent the simulation of typical color.

Thus, the finding of multiple activation of color representation in Connell and Lynott (2009) needs further investigation because there is no statistical evidence and the possibility of item dependencies exists. This brings us to the new research questions: Do

readers activate typical and atypical colors when the context is introduced before the target word? Testing this hypothesis sheds new light on the activation mechanism, as it would provide further evidence as to whether color typicality can be altered by context.

In L2 research, there is a growing body of research that uses embodied cognition as a theoretical framework, but it does not yet represent the majority of research targets. The embodied paradigm will provide details about what we understand from verbal information, apart from the activation of orthographic and phonological information, and provide new insights into the difference between L1 and L2 acquisition and processing. Previous studies that have investigated L2 simulation are not yet in agreement about the influence of readers' L2 proficiency on L2 simulation. The influence of L2 proficiency is important when considering the role of non-linguistic information in language processing. L2 lexical models such as the Revised Hierarchical Model and the Three-Stage Model are concerned with the development of the relationship between L1 and L2 lexicons and their concepts. However, these models do not assume that L2 proficiency may modulate content understanding. Jiang (2000), for example, assumes three stages of lexical development. The stages are different regarding the connection between L2 forms and their concepts. As mentioned, in the initial stages, the L2 accesses its concepts through its L1 translation equivalents.

The question here is whether there is a difference in understanding concepts between direct and indirect access. If so, can learners simulate non-linguistic information through indirect access? None of the models explain this question. Therefore, it is critical to understand whether learners with lower language proficiency, most of whose vocabulary is thought to be in the beginning indirect stage, can simulate colors during L2 processing. Therefore, L2 proficiency is an important factor in L2 embodiment research.

Research Questions and Hypotheses

The purpose of this study is to investigate whether color is simulated during L2 sentence processing and whether learners' L2 proficiency affects color activation. Previous studies have shown that L1 processing involves non-linguistic processes (e.g., Connell & Lynott, 2009). The current study aims to replicate Connell and Lynott's (2009) investigation of whether object colors are simulated during L1 processing using the material developed by the author. These data will serve as the basis for comparison with the results of the L2 data. This study addresses three research questions for the L1 conceptual processing (Chapter 4) and four questions for the L2 conceptual processing (Chapter 5):

1. Are the objects' colors simulated during L1 processing?
2. Does the simulation of color depend on the color that is implied by the L1 sentence?
3. Does the position of the context phrases change the simulation in L1?
4. Are the objects' colors simulated during L2 processing?
5. Does the simulation of color depend on the color that is implied by the L2 sentence?
6. Does the position of the context phrases change the simulation in L2?
7. Does L2 proficiency affect the degree of simulation of objects' color?

The following hypotheses for each of the research questions were based on the results of previous studies:

1. The colors of the objects are simulated in L1. Reaction time shortens when the color of the target words (object) matches the color implied in the sentences. The typical colors have a faster reaction time than atypical colors.
2. The simulation of color depends on the color implied by the L1 sentence. That is, if a sentence implies a typical color of the object, the reaction times for the typical color are faster than for the atypical or unrelated colors. In contrast, when a sentence implies an atypical color, there is no significant difference in reaction times between the typical and atypical colors. However, they are faster than for the unrelated color.
3. The position of the context phrases changes the simulation in L1. When the context phrases are introduced before the keywords, reaction times are faster whenever the contextual color and the color of the word match. On the other hand, if the context phrases are inserted after the keywords, two phenomena occur: (1) the reaction times of typical colors are faster than those of atypical colors when the sentence implies typical colors, and (2) the reaction times of typical and atypical colors do not differ when the sentence implies atypical colors.
4. When learners' L2 proficiency is not considered, no differences in reaction times are found between match and mismatch conditions in L2.
5. When learners' L2 proficiency is not considered, colors implied by sentences do not affect reaction times in L2.
6. When learners' L2 proficiency is not considered, the position of the context phrases does not influence on reaction times.
7. L2 proficiency affects the degree of simulation of objects' color. Higher L2 proficiency leads to a similar pattern as L1 results. Therefore, research hypotheses

1 through 3 hold true for higher proficiency L2 learners. In contrast, lower proficiency learners will not show a similar pattern to the results from L1.

Chapter 3: Pilot Study

Aim of the Pilot Studies

Two pilot studies were conducted to (1) construct the experimental items in both English and Japanese and (2) calculate the target sample size for the main study (Experiments 1 and 2). This study used items from Connell and Lynott (2009), but the number of items used in the study was small ($N = 10$). Therefore, the author created new experimental items. Results of Pilot Study 1 and 2 were used to calculate the target sample sizes of the main experiments.

Pilot Study 1

The main aim of Pilot Study 1 was to validate ten experimental items used by Connell and Lynott (2009), as well as the items that the author constructed. Because Experiment 1 was conducted in both English and Japanese, the English and Japanese versions of the experimental material were validated using word- and sentence-level rating tasks. In addition, about half of the participants performed the semantic Stroop task using the computer program created by the author. This task was performed to test whether the computer program would work as intended.

Participants

Twenty-six participants were recruited for the study. They were all native Japanese speakers learning English. Thirteen participants were assigned to the word typicality rating task, and 25 participants were assigned to the sentence typicality rating task. In addition, 12 participants performed the semantic Stroop task before the rating tasks.

Experimental Items

In the Stroop task, there were 20 experimental sentences and ten critical words (*bear, chameleon, hair, horse, leaf, steak, strawberry, tea, tomato, and tree*). Each test

sentence has two different versions in terms of typicality. For example, the noun *bear* was embedded in “Joe was excited to see a bear *in the woods*” or “Joe was excited to see a bear *at the North Pole*.” Each test word was colored with three different colors corresponding to the typicality of the noun: typical, atypical, and unrelated. Four colors were selected: Red, Brown, White, and Green. More details about the procedure for determining the colors are described in Pilot Study 2. The Japanese version of the experimental material was created by translating the English material.

Rating Tasks

Word Typicality Rating Task. Connell and Lynott (2009) used two rating tools (sentence-level and word-level) to determine the typicality of the experimental materials. The author used the same sentence typicality task but modified the word typicality task for the following reasons. While they did not provide detailed criteria, they decided on the typicality of colors by reviewing photographs of the corresponding objects and then asking participants to choose which color pair (e.g., bear-brown, bear-white) was more typical. The color chosen was considered valid if most participants matched the typicality determined by the researchers. However, it is not certain whether participants actually believed that the color the researcher considered represented atypical colors. Another weakness of their method is that participants did not rate the typicality of unrelated colors. Therefore, the author modified the word typicality rating task to more rigorously determine the typicality of object colors.

The task was created by the author using Google Forms (Figure 1). Participants were asked to evaluate which colors correspond to typical, atypical, and unrelated colors. In the case of the *bear*, it was presented with three colors: brown, white, and green (see Figure 1). Participants simply clicked on the radio button to determine the color that

corresponded to typicality. There was no restriction on the choice. For example, they could select the same box for all colors by not selecting the other two boxes (e.g., white-typical, brown-typical, and green-typical).

Figure 1

Word Typicality Rating Task Example

クマ (bear) *	代表的な色	ありえるが代表的でない色	あり得ない色
白	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
茶	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
緑	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Note. The first column had three colors that differed in typicality. The title of the second, third, and fourth columns were “typical color,” “possible but not typical color,” and “unrelated color.”

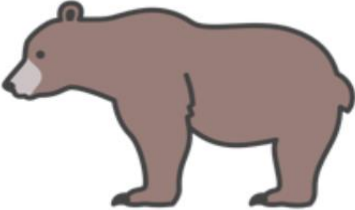
Sentence Typicality Rating Task. The sentence typicality rating task was created to investigate whether the typicality implied by the test sentences matches the author’s intended typicality. The author created the task using Google Forms (Figure 2). A test sentence was presented with two images (free images downloaded from websites) and four forced-choice alternatives; best matched by the first picture, best matched by the second picture, matched by both pictures equally, and matched by neither picture. One of the images represented a typical color of the keyword (e.g., a brown bear), and the other image represented an atypical color of the object (e.g., a white bear). Participants were asked to

choose one option regarding the match between the sentence and the pictures.

Figure 2

Sentence Typicality Rating Task Example

Joe was excited to see a bear in the woods. *

1. 

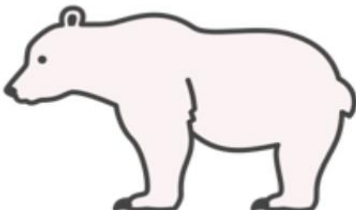
2. 

写真1と合う

写真2と合う

どちらも合う

どちらも合わない

Note. The choices are “best matched by the first picture,” “best matched by the second picture,” “matched by both pictures equally,” and “matched by neither picture.”

There were two versions of the task to balance the typicality of the sentences. Each task consisted of half of the sentences implying the typical color of the object and the other half of the sentences implying the atypical color of the object. Each participant worked on only one of the two versions.

Procedure

Participants were given two rating tasks on Google Forms. They could take as much time as they needed. They were not allowed to use the Internet or a dictionary during the testing period.

Results

The results of the word typicality rating task showed that for most of the items, all participants agreed with the typicality of colors. For example, all the participants considered red to be typical of *tomato*. More than 90 percent of the participants considered green to be atypical and white to be unrelated. However, some items were problematic. For example, most of the participants did not agree with the typicality of *tree*. For the first screening for item construction, words that received less than 50 percent agreement were dropped. Three items (*hair*, *tea*, and *tree*) that did not reach the criterion were deleted.

As for sentence typicality rating task, Connell and Lynott (2009) used a 25% match as the criterion for validating the typicality of each sentence—the present study also followed this criterion. All items met this criterion. The detailed results can be found in Appendix A.

Pilot Study 2

Based on Pilot Study 1, the author created 14 new test items. The purpose of Pilot Study 2 was to validate them.

Participants

The participants were 24 students from Pilot Study 1. None saw the new 14 items. Twelve participants were assigned to the word typicality rating task, and 24 participants were assigned to the sentence typicality rating task. In addition, 12 participants performed the semantic Stroop task prior to the rating tasks.

Experimental Items

The new 14 items were created by the author in the following way. There were two points to consider. First, the object should have typical and atypical colors, which must be known to people regardless of their L1 and L2. For example, people know that the typical color of a *tomato* is red and the atypical color is green because we generally agree that a red tomato was green when it was not ripe. However, there is much less agreement about the typicality of colors of some objects, such as a bicycle. Such objects cannot be included in the experiment.

The second constraint is the number of colors used in the experiment. Connell and Lynott (2009) performed the semantic Stroop task with oral production. They did not have to restrict the colors because it was a free production task. However, in this study, the semantic Stroop task was performed with a *QWERTY* keyboard; participants were asked to memorize the color-key correspondence. The author also had to pay attention to the ease of keystroke as reaction times were recorded. Considering these, the number of keys that recorded their reaction times was set to four: “S,” “D,” “K,” and “L,” which were mapped to red, green, white, and brown. The different color-key correspondence was tested in the pilot study.

Consequently, the 14 new items were *apple, ball, cake, cloud, ice cream, kiwi, lipstick, mountain, onion, popcorn, traffic light, plum, vegetable, and watermelon*. In the experimental sentences, the past tense was used following Connell and Lynott (2009). The Japanese version of the experimental materials was created by translating the English materials.

Procedure

The same procedure and rating tasks were used as in pilot study 1. Although the L1 translation was not provided in the sentence typicality rating task in Pilot Study 1, it was provided in Pilot Study 2 to test whether the newly constructed items had the same meaning in L1 and L2, since Experiment 1 was conducted with native speakers of English and Japanese.

Results

In evaluating the rating tasks, the author followed the same criteria as in Pilot Study 1. Items that received less than 50 percent on one of the word typicality tests were eliminated from the test items (for example, *lipstick*, *mountain*, *traffic light*, *vegetable*, and *watermelon* showed less than 50 percent of their atypical colors).

The results of the sentence typicality rating task showed that rating scores were more converged than the scores in Pilot study 1, except for *ice cream*. Only 16.7 percent of participants rated “*Nick liked to eat ice cream in the park*” as implying white ice cream, which is lower than the chance rate (25 percent). The author eliminated six items (*lipstick*, *mountain*, *traffic light*, *vegetable*, *watermelon*, and *ice cream*) based on the results of both the word and sentence typicality rating tasks.

Based on pilot studies 1 and 2, the total number of critical words was 15. For a key word, there were six conditions (2 types of sentence typicality and three types of combinations between ink and color typicality); thus, the number of experimental items was 90. The detailed results of pilot study 2 can be found in Appendix B.

Determining the Sample Size

Experiment 1

A power analysis was conducted to determine the sample size using the *samplesize_mixed* function of the *sjstats* package version 0.18.1 (Ludecke, 2021) for R version 4.1.1 (R Core Team, 2021). This function computes the sample size for two-level designs of linear mixed models. Experiment 1 investigated whether the results replicated Connell and Lynott (2009). They found the significant main effect of word typicality only. Thus, the targeted power, the degrees of freedom for the numerator (the number of predictors in the model), effect size, alpha level, and expected intraclass correlation coefficient were set to 80 percent, 2 (3 levels (typical, atypical, unrelated) - 1), $R^2 = .02$ (small), .05, and .05, respectively. The number of observations per cluster group was set at 60, as each participant completed 60 items for each word type. The results showed that to achieve the targeted power and effect size, a total of 1,915 observations were required. Therefore, the number of participants in Experiment 1 should be at least 32. For more details on the analysis, see Appendix C.

Experiment 2

The number of participants for Experiment 2 was also determined in the same way. Experiment 2 was designed to investigate whether the results of Connell and Lynott (2009) were replicated in L2. In addition, the interaction of word typicality and L2 proficiency, operationalized as a test score for vocabulary range, was to be investigated. Thus, the targeted power was increased by a moderate amount ($R^2 = .005$) as the predictors in the model were more than in Experiment 1. The degrees of freedom for the numerator, effect size, alpha level, and expected intraclass correlation coefficient were set to 80 percent, 5 (word typicality: 3 levels - 1; vocabulary size: 1, the interaction: (3 levels - 1) \times 1)), R^2

= .025, .05, and .05 respectively. The number of observations per cluster group was set at 60, as each participant completed 60 items for each word type. The results showed that to achieve the targeted power and effect size, a total of 2,049 observations were required. Therefore, the number of participants in Experiment 2 should be at least 35. For more details on the analysis, see Appendix C.

Chapter 4: Experiment 1

Aim of Experiment 1

Experiment 1 was conducted to obtain baseline data from L1 speakers (English and Japanese). The semantic Stroop task was performed with the two groups of native speakers.

Method

Participants

Native English Speakers. 37 native English speakers were recruited (see Chapter 3). None of the participants were involved in the pilot studies. However, the author excluded two participants because (1) one experienced a technical problem with the computer and (2) the other did not speak English as an L1 but as a primary language. Thus, data from 35 participants were analyzed (15 females, 19 males, and 1 other). Participants were provided with a questionnaire asking about their background information, such as nationalities and their L2s. The results of the questionnaire showed that 22 participants were American, followed by British ($n = 6$), Australian ($n = 4$), and Canadian ($n = 3$). All participants spoke English as their L1, and 29 participants reported speaking more than one language. Table 1 reports native English speakers' age.

Table 1

Native English Speakers' Age

	<i>N</i>	<i>M</i>	<i>SD</i>	<i>Median</i>	<i>Minimum</i>	<i>Maximum</i>
Age	35	30.09	11.28	28	20	74

Native Japanese Speakers. A total of 36 participants were recruited for the study (15 females, 20 males, and 1 other) (see Chapter 3). None of them participated in the pilot studies. Thirty-one were graduate or undergraduate students at Japanese universities. Their majors are German, music, engineering, education, economics, humanities, agriculture, and literature, among others. They learned English mainly in Japan. A background questionnaire indicated that 4 participants had experience studying abroad in English-speaking countries. The average duration of study abroad for the four learners was 11.25 months. The results of the V_YesNo v1.1 test (Meara & Miralpeix, 2016) showed that the participants' L2 proficiency was between beginner and advanced levels. The descriptive statistics of participants are shown in Table 2.

Table 2

Native Japanese Speakers' Descriptive Statistics

	<i>N</i>	<i>M</i>	<i>SD</i>	<i>Median</i>	<i>Minimum</i>	<i>Maximum</i>
Vocabulary size scores	36	3,820.33	1,531.52	4,125.0	850	7,574
Age	36	22.33	5.59	20.5	19	50
Years learning English	36	10.97	5.24	10.0	7	35
Self-reported English proficiency scores						
Reading	36	3.83	1.59	4.0	1	6
Writing	36	2.94	1.60	3.0	1	6
Listening	36	3.17	1.40	3.0	1	7
Speaking	36	2.97	1.70	3.0	1	6

Note. Vocabulary size scores were V_YesNo v1.0 test scores (Meara & Miralpeix, 2016). Self-reported English proficiency scores were calculated from rating scores on a 7-point Likert scale (1 = very poor, 7 = very good).

Tasks

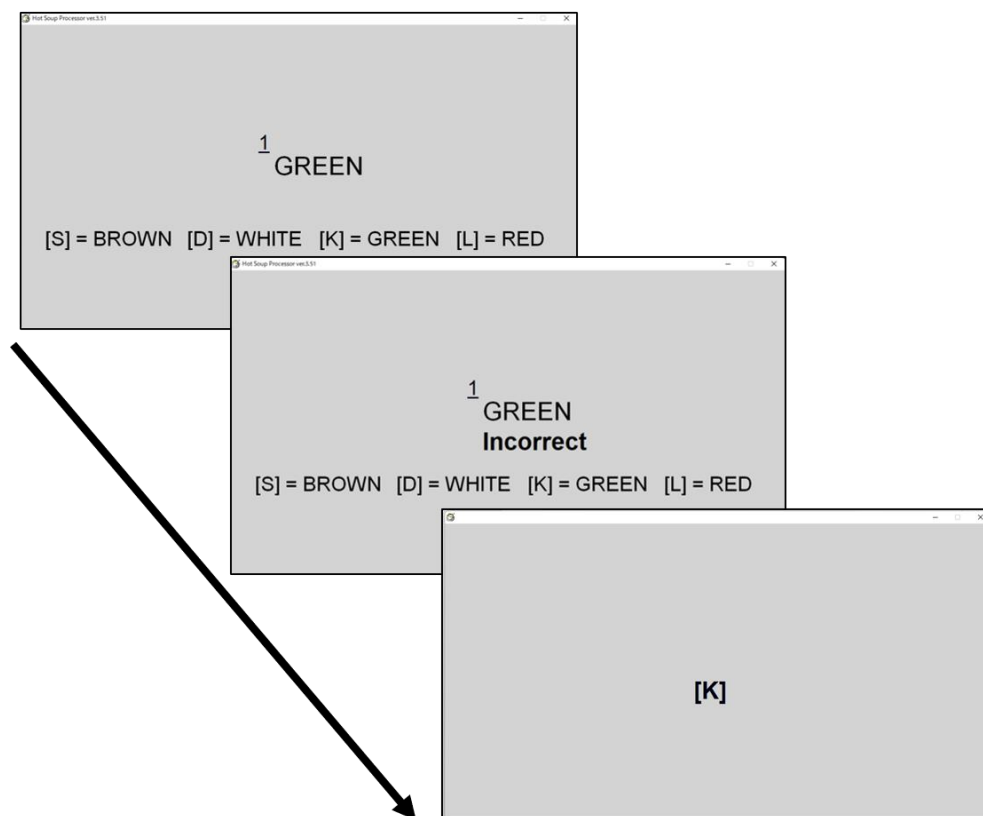
Semantic Stroop Task. Connell and Lynott (2009) performed the semantic Stroop task. In their task, participants verbally named the colors. However, in the present study, participants responded with a keyboard. To reiterate, the aim of Experiment 1 is to obtain the baseline to be compared to the L2 performance to avoid possible effects of slow L2 speaking—non-verbal response is desirable for L2 learners. Therefore, a keyboard task has also been used for L1 speakers. Studies of the Stroop effect have shown that the Stroop effect is observed in both the naming task and the manual task, although the magnitude of the Stroop effect was greater in the naming task (e.g., Augustinova et al., 2019; see for Parris et al., 2022 review of response mode).

The author coded the program with Hot Soup Processor version 3.5.1. (<http://hsp.tv/>). Two programs were created: the practice program and the main program. In some studies of Stroop tasks with manual responses, colored stickers are used, and the keys are covered to show which color the keys correspond to (e.g., Augustinova et al., 2019). However, in the study in which the Stroop task was performed both face-to-face and online, it was virtually impossible for participants to use the colored stickers in the online task because they were not using the author's keyboard. As an alternative, the study used a practice program so that participants could memorize the key for each color. In the practice program, the instructions were presented first. Then, the participants practiced pressing the keys. In the practice program (Figure 3), either of the words “BROWN,” “WHITE,” “GREEN,” and “RED” were provided with black ink in the middle of the screen in Arial (MS Gothic for Japanese items) 50-point on a light gray background (211, 211, 211 in RGB). The RGB rates were identical in the main phase.

In addition, the table with the corresponding color keys was presented at the same time. The participant pressed the keys corresponding to the presented word. *QWERTY* keyboards were used in this task. In some previous studies in which a Stroop task was performed using *QWERTY* keyboards, the S, D, K, and L keys were assigned to colors (e.g., Abrahamse et al., 2013). In this study, the S, D, K, and L keys were also used and assigned to the colors brown, white, green, and red, respectively. The words were presented in either English or Japanese (e.g., GREEN or 緑), with feedback immediately following the responses: If the answer was correct, the message “Correct” was presented, and if the answer was incorrect, the message “Incorrect” and the correct color were presented. Participants repeated the task 40 times (10 times for each color).

Figure 3

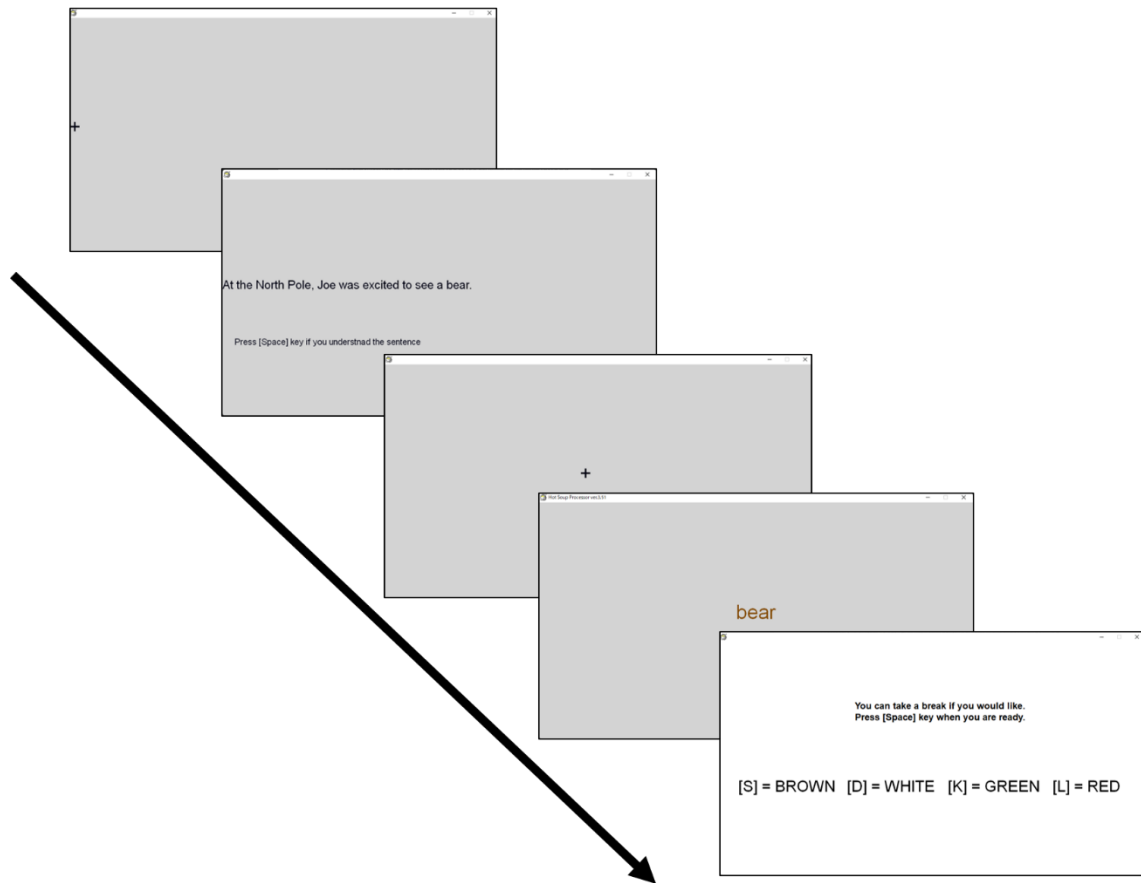
Diagram of the Practice Program



The second exercise was the semantic Stroop task (Figure 4). First, a plus sign was displayed on the far left of the screen for one second to help participants focus on the screen. Second, a sentence in Arial font (MS Gothic for Japanese items) was displayed in 30-point black on a light gray background. Participants pressed the space bar to move to the next sentence after understanding the meaning of the sentence. Third, a plus sign was displayed in the center of the screen for 500 milliseconds. Fourth, a colored word was displayed in either brown (132, 75, 0 in RGB), white (255, 255, 255 in RGB), green (0, 128, 0 in RGB), or red (255, 0, 0 in RGB) in Arial (MS Gothic for Japanese items) in 50-point font on a light gray background. However, Connell and Lynott (2009) changed the RGB rate depending on the item (e.g., leaf in green: 0, 130, 0; bananas in green: 181, 228, 36), the RGB between items were all identical within colors in the study. This is to avoid interaction between color difference and response. For example, the darker green might be easier to respond to than the lighter green. Participants were asked to respond to the color of the ink by pressing the keys. They were instructed to respond as quickly and accurately as possible, as the response latency was recorded.

Figure 4

Diagram of the Semantic Stroop Task (Critical Items)

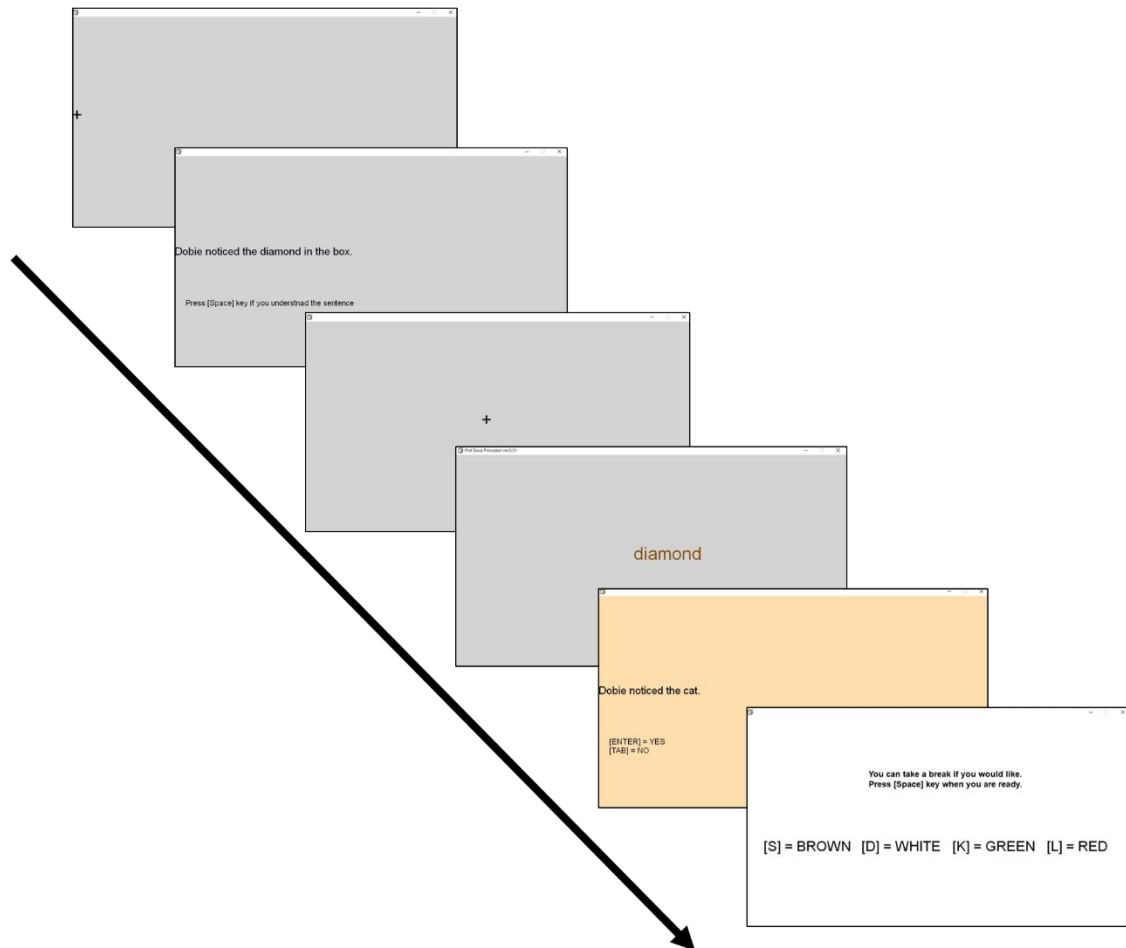


Note. The white screen with the table of color-key combinations was displayed between the trials. Participants could take a break if needed.

In the filler items, a comprehension question was asked after the color assessment (Figure 5). The background was changed to bright orange (255, 222, 173 in RGB) so that participants could see that it was the follow-up question. Participants pressed the enter key if the meaning of the sentence presented matched the sentence they had seen before the color judgment. If the sentences did not match, they pressed the tab key. Between trials, a screen with the table of color key combinations was displayed, and participants were told on a white background (without RGB indication) that they could pause if necessary.

Figure 5

Diagram of the Semantic Stroop Task (Filler Items)



Note. The white screen with the table of color-key combinations was displayed between the trials. Participants could take a break if needed.

Experimental Sentences. There were two types of sentences in the semantic Stroop task: critical and filler.

Critical Sentences. Based on the results of the two pilot studies (see Chapter 3), a total of 180 sentences were created. All 180 sentences were divided into six conditions. In a typical-typical condition, the sentence implied the typical color of the object, and the color of the ink also represented the typical color of the object. In a typical-atypical condition, the sentence implied the typical color of the object, but the color of the ink was

an atypical color of the object. A typical-unrelated condition implied the typical color of the object, but the color of the ink was not associated with the object. The atypical-atypical, atypical-typical, and atypical-unrelated conditions were the same, except that the sentence always implied the atypical color of the object (Table 3).

Table 3

An Example of Each Condition (bear)

Conditions: Sentence-Word (Ink color)	Sentence	Word (Ink color)
Typical-typical	<i>Joe was excited to see a bear in the woods.</i>	Brown
Typical-atypical		White
Typical-unrelated		Green
Atypical-typical	<i>Joe was excited to see a bear at the North Pole.</i>	Brown
Atypical-atypical		White
Atypical-unrelated		Green

The sentences were further divided into pre-context conditions and post-context conditions. In pre-context sentences, the critical word is preceded by a phrase that decides the typicality of the object (e.g., “*At the North Pole, Joe was excited to see a bear*”). In contrast, in post-context sentences, a phrase that decides the typicality comes after the critical words (e.g., “*Joe was excited to see the bear at the North Pole*”). All English and Japanese critical sentences were reviewed by one native English speaker and two native Japanese speakers, respectively. The critical sentences used in Experiment 1 can be found in Appendix D and Appendix E.

Filler Sentences. An equal number of filler sentences was created. There were only two conditions in the filler sentences: pre-context and post-context. A comprehension question followed the color judgment. The number of yes/no answers was the same. To equalize the color key responses, the number of ink colors was based on the critical sentences (Table 4). All filler sentences and comprehension questions were checked by a native English speaker and a native Chinese speaker whose L2 is English. All Japanese filler sentences were reviewed by two native Japanese speakers. The filler sentences used in Experiment 1 can be found in Appendix D and Appendix E.

Table 4

The Number of Trials in Each Condition

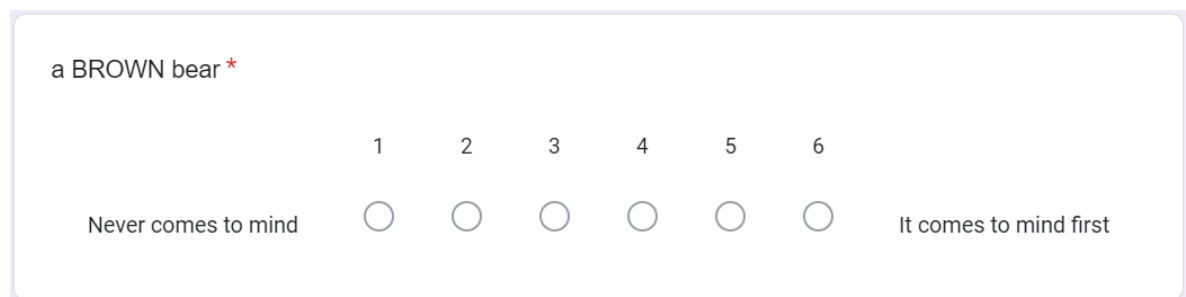
Colors	Critical/Filler	Comprehension Questions		Total
		Yes Response	No Response	
Brown	Critical	–	–	52
	Filler	19	19	38
White	Critical	–	–	44
	Filler	23	23	46
Green	Critical	–	–	44
	Filler	23	23	46
Red	Critical	–	–	40
	Filler	25	25	50
		90	90	360

Word Typicality Rating Task. The task was conducted to confirm the typicality of the critical words (determined in Pilot Studies 1 and 2) with the Experiment 1 participants. Participants were asked to rate the typicality of the color-word association on a 6-point Likert scale, where one means *never comes to mind*, and six means *comes to*

mind first. In the task, all keywords that were used in the semantic Stroop task were presented with each of the typicality (e.g., a *brown* bear: typical, a *white* bear: atypical, a *green* bear: unrelated). The task was performed using Google Forms and there were no time constraints. The order of presentation was randomized using the Google Forms function. Figure 6 shows an example of the word rating task in English.

Figure 6

Word Typicality Rating Task Example (English Version)



a BROWN bear *

1 2 3 4 5 6

Never comes to mind It comes to mind first

Sentence Typicality Rating Task. Immediately after the word typicality rating task, the participants performed a sentence-level rating task. This task checked whether the typicality of the image in a sentence matched what the author had determined. This task was identical to the task performed in the pilot test. There were two versions of the task (Set A and Set B), and the order of the tasks was counterbalanced. Thus, half of the participants answered Set A first and then Set B. The other half of the participants answered Set B first and then Set A. There were no time constraints, and participants answered the questions using Google Forms. The order of presentation was randomized using the Google Forms function.

Vocabulary Size Test. Japanese participants completed the online vocabulary size test: V_YesNo v1.1 test (Meara & Miralpeix, 2016). This was done to measure the

participants' L2 proficiency for comparison with participants in Experiment 2. In the test, 200 words or non-words were presented. Participants were instructed to choose *YES* only if they knew what the word meant. If they did not know the word's meaning, although they were familiar with the spelling, or if they were unsure, they must choose *NEXT*. The test calculates scores of 10,000. According to the criteria from Meara & Miralpeix (2016), scores from 2,000 to 3,500 were beginner levels (scores below 2,500 were considered probably unreliable), scores from 3,500 to 6,000 were intermediate levels, and scores from 6,000 to 10,000 were proficient learners: in their terms, "*good for non-native speakers.*"

Procedure

The experiment was conducted either by zoom or face-to-face. Those who participated via Zoom were asked to continue sharing their screen during the experiment. This allowed the author to observe how the participants worked on the tasks. The experiment began with practice on the semantic Stroop task. In practice, there were two phases: First, they practiced pressing color keys. Second, they practiced the same format as in the main session. In the first phase, the following instructions were given (the Japanese versions of the instructions are provided in Appendix F):

"First, we'll start the practice session. In this task, you will use the [Tab], [Enter], [Space], [L], [K], [D], and [S] keys. The speed you press [L], [K], [D], and [S] is recorded. Always keep your hands positioned over these four keys (Picture) (On the right corner, a picture [Figure 7] was presented to show the positions of the fingers on the keys). Reaction times for pressing the [Tab], [Enter], and [Space] keys are not recorded. [L], [K], [D], and [S] correspond to the following: [L] is for RED, [K] is for GREEN, [D] is for WHITE, [S]

is for BROWN. You are going to practice memorizing the color-key correspondence. You do not have to memorize what [Tab], [Enter], and [Space] keys correspond to.”

Figure 7

The Positions of the Fingers on the Keys



After confirming the instructions, another instruction asked the participants to press each of the keys from [S] to [L] so that they could confirm the position of the keys. Then they practiced pressing the key 40 times. Participants pressed one of the keys to respond to the colors. Immediately after pressing a key, feedback was given. When the answer was correct, the message “Correct” was shown on the screen. In contrast, when the answer was incorrect, the message “Incorrect” and the letter of the correct key were shown (e.g., “[L]”).

Immediately after the first stage, they moved on to the second stage of the practice. In the second stage, initially, the following instructions were provided (the Japanese versions of the instructions were provided in Appendix F):

“First, a plus sign (+) will be presented on the far-left side of the screen for 1 second. Second, a sentence will be presented. Press [Space] if you understand the meaning of this sentence. The reaction time is not recorded for this portion, so you can take your time when answering. Third, a plus sign (+) is presented centrally on-screen for 0.5 seconds. Fourth, a colored word is presented. L = RED, K = GREEN, D = WHITE, and S = BROWN. Respond to the color by pressing the keys. The reaction time will be RECORDED, so please try to answer as quickly and accurately as possible. Sometimes, a comprehension question will be asked after the presentation of a colored word. Answer if the information in the sentence is correct or not by pressing [Enter (YES)] or [TAB (NO)]. This task is untimed.”

In each trial, they received feedback (true/false) immediately after answering the color and comprehension questions. In total, five trials were conducted in this exercise: three trials were conducted under critical conditions, and two trials were conducted under filler conditions. The expected responses for the color questions were “brown” ($n = 1$), “white” ($n = 2$), “green” ($n = 1$), and “red” ($n = 1$); for the comprehension questions, the expected responses were “yes” ($n = 0$) and “no” ($n = 2$). The main program was the same except for the feedback; no feedback was provided.

After they finished the practice session, they moved on to the main session (the semantic Stroop task). The main session was divided into two sessions. The pre- and post-context conditions were separated because learners might notice the differences between the conditions. Therefore, the author created a separate set of tasks. The first set contained the critical items in the pre-context condition and the filler items in the post-context condition. The second set contained the critical items in the post-context condition and the

filler items in the pre-context condition. The order of the set was counterbalanced between participants to avoid the order effect. Between the two sets, participants took part in another experiment unrelated to the semantic Stroop task. The task was not to make participants aware of the differences between the pre- and post-context materials. The native Japanese speakers took an online vocabulary size test (i.e., the V_YesNo v1.1 test [Meara & Miralpeix, 2016]) to measure their L2 proficiency. The native English speakers took an article task, wherein a sentence was presented (e.g., *President of the United States lives in White House*); the participants were asked to insert the article “the” wherever they believed it necessary. There were 91 sentences in total.

After the semantic Stroop task, participants answered the word typicality rating task. Immediately after, they completed two sentence typicality rating tasks. The order of the sentence typicality rating tasks was counterbalanced to avoid practice and order effects.

At the end of the experiment, participants answered the background questionnaire. They were asked about their nationality, native language, history of learning a foreign language, self-reported proficiency in a foreign language, and so on.

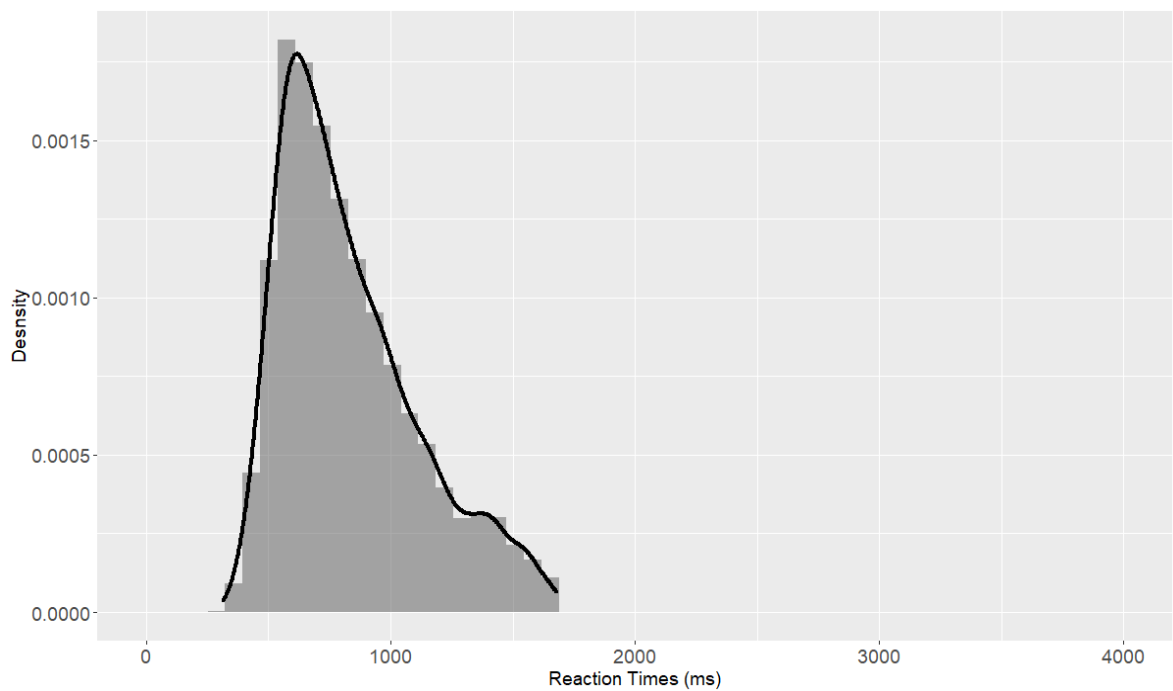
Analysis

All statistical analyses were performed using R version 4.1.1 (R Core Team, 2021). The same analysis procedure was applied to the data of native English and Japanese speakers. Before analysis, incorrect responses, filler words, and all data from participants with an overall accuracy of less than 80% on the color decision and with an overall accuracy of less than 50% on the comprehension question were excluded. In addition, reaction times that deviated more than \pm three median absolute deviations (MAD) (Leys et al., 2013) from the median were excluded. Reaction times were measured from the presentation of the colored word to the moment the participant pressed one of the S, D, K,

or L buttons. Figure 8 and Figure 9 show the distribution of reaction time data for English and Japanese after data treatment. 9.21 percent of the data for the critical words were deleted from the data for native English speakers. 9.24 percent of the data for the critical items were deleted from the data for the native Japanese speakers.

Figure 8

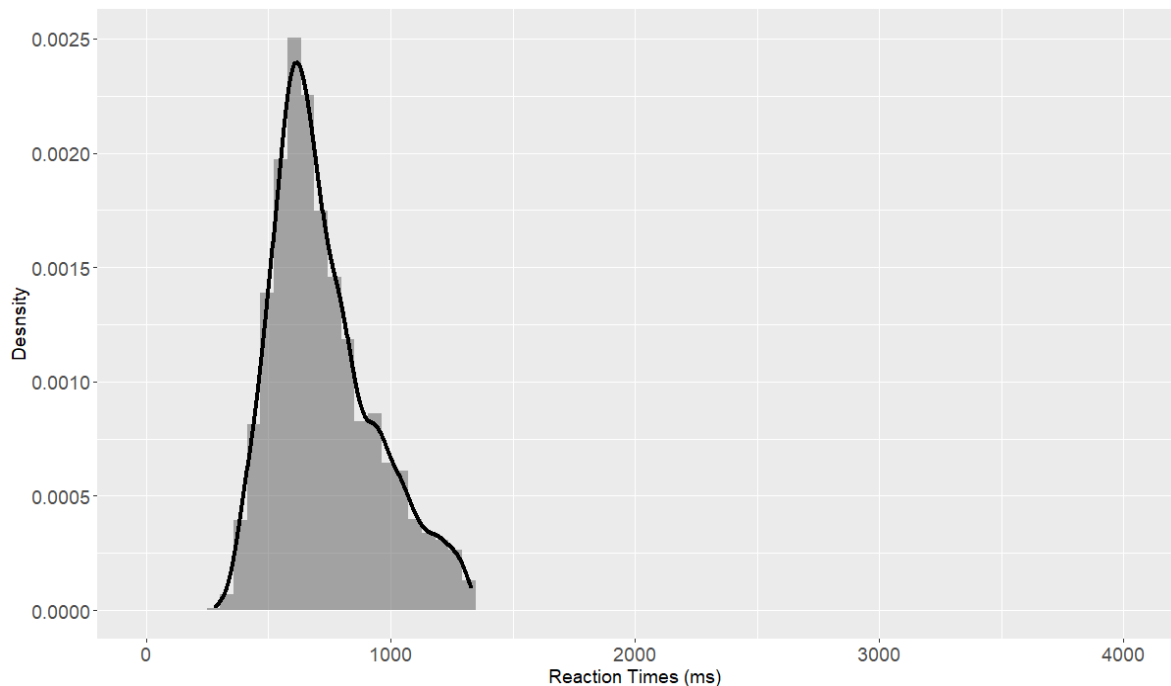
The Distribution of the Reaction Times of the Semantic Stroop Task (Native English Speakers)



Note. Density was calculated using kernel density estimation.

Figure 9

The Distribution of the Reaction Times of the Semantic Stroop Task (Native Japanese Speakers)



Note. Density was calculated using kernel density estimation.

After preliminary data processing, the probabilistic distributions were selected in the following procedures (Kusanagi, 2017). First, the possible probabilistic distributions were selected based on the characteristics of each distribution. Weibull, gamma, lognormal, and normal distributions were selected as possible probabilistic distributions because they are commonly used to analyze reaction time data. Next, *fitdistrplus* package 1.1-6 (Delignette-Muller & Dutang, 2015) was used to choose the best probabilistic distribution for the data. Considering the following goodness-of-fit statistics for parametric distributions that were computed with the *fitdist* function of the package (Table 5 and Table 6), the log-normal distribution was chosen as the probabilistic distribution for the present study.

Table 5*The Goodness-of-Fit Statistics and Information Criterion (Native English Speakers)*

	Weibull	Gamma	Log-normal	Normal
Kolmogorov-Smirnov	0.08	0.06	0.05	0.10
Cramér-von Mises	15.80	6.97	3.60	18.56
Anderson-Darling	101.74	43.57	22.99	114.68
AIC	79,620.30	78,844.60	78,634.99	79,729.57
BIC	79,633.58	78,857.87	78,648.27	79,742.84

Note. AIC refers to Akaike Information Criterion. BIC refers to Bayesian Information Criterion.

Table 6*The Goodness-of-Fit Statistics and Information Criterion (Native Japanese Speakers)*

	Weibull	Gamma	Log-normal	Normal
Kolmogorov-Smirnov	0.09	0.06	0.04	0.09
Cramér-von Mises	16.19	5.10	2.18	14.96
Anderson-Darling	97.89	29.10	13.02	87.31
AIC	79,674.56	78,880.16	78,752.49	79,520.64
BIC	79,687.92	78,893.52	78,765.85	79,534.00

Note. AIC refers to Akaike Information Criterion. BIC refers to Bayesian Information Criterion.

After the data treatment, linear mixed-effects modeling was performed using the *lme4* package 1.1-27.1 (Bates et al., 2021). The dependent variable was reaction time in the semantic Stroop task. Reaction time was log-transformed. The independent variables were

sentence typicality, word typicality, and context position. The number of trials (order of presentation) and the reading time of each sentence were also included as possible covariates. The reading time of a sentence was measured from the presentation of the sentence to the time when participants pressed the space bar. Reading time was scaled to avoid convergence problems with *scale* function of the *base* package 4.1.1. All categorical variables were contrasted (repeated)-coded to compare neighboring factor levels with the function of *contr.sdif* the *MASS* package 7.3-54 (Venables & Ripley, 2002). Table 7 shows how each condition was coded.

Table 7

Dependent Variables and Their Assigned Codes

Levels	2-1	3-2
Sentence Typicality		
1. Typical	-0.5	–
2. Atypical	0.5	–
Word Typicality		
1. Unrelated	-0.667	-0.333
2. Typical	0.333	-0.333
3. Atypical	0.333	0.667
Position		
1. Pre	-0.5	–
2. Post	0.5	–

Note. The numbers in the column of *Levels* refer to the levels of the variables. 2-1 refers to the comparison of level 2 - level 1, and 3-2 refers to the comparison of level 3 - level 2.

The best model structure was determined by the following procedures. First, the possible covariates were considered to decide which covariate to include in the final model. The possible covariates were reading time and order of presentation for the native English and Japanese data. The null model was compared to the model that included the possible covariates. For the English data, the results showed that the model with the scaled sentence reading time had the lowest AIC among the three models. Next, the model with scaled sentence reading time was compared with the model containing both presentation order and scaled sentence reading time. The models with both covariates showed the lowest AIC. Thus, the final model contains sentence typicality, word typicality, the interaction of the two, context position, presentation order, and scaled sentence reading time as independent variables.

For the Japanese data, the model including sentence reading time showed the lowest AIC among the three models. Then, the model was compared with the model that included both order of presentation and sentence reading time. The models with both covariates had the lowest AIC. Thus, the final model included sentence typicality, word typicality, the interaction of the two, context position, presentation order, and scaled sentence reading time as independent variables.

Next, the random structure was considered using the *rePCA* function of *lme4* package 1.1-27.1 (Bates et al., 2021). Bates et al. (2015) proposed a way to specify the random structure of mixed models using the *rePCA* function. The function displays the variance-covariance parameters and allows us to identify which parameters should continue to be included in the model. Following Bates et al. (2015), the maximum random effects model was built to include all independent variables as slopes. Then, the *rePCA* function was applied to the maximum random effects model with random structure

correlation and no correlation parameters. After checking the number of dimensions, one model term was taken out at a time to investigate whether it significantly increased the goodness-of-fit. The procedure was repeated as long as the goodness of fit increased significantly (i.e., until the lowest AIC was obtained). Finally, the goodness of fit was compared between the model with correlation parameters and the model without correlation parameters. The correlation parameters were included if they significantly increased the goodness-of-fit.

The variance inflation factors (VIF) were reviewed with the *check_collinearity* function of the performance package 0.9.0 (Lüdtke et al., 2021) to determine if there were any multicollinearity issues once the final model was established. The VIF threshold was set at 5. The analysis confirmed that the final models had no multicollinearity problems.

Results and Discussion

Word Typicality Rating Task

Native English Speakers. The results of the word typicality rating task were reviewed prior to the modeling procedures. The author checked whether the results of the typical colors of each word were higher than those of the atypical and unrelated colors. There was one item (*onion*) that needed a change in item coding. The scores of the unrelated color of *onion* (red) ($M = 3.77, SD = 1.82$) were higher than the scores of the typical color of *onion* (brown) ($M = 3.49, SD = 1.88$) (i.e., red was more typical than brown for the participants). This is due to the presence of red onions. An American participant mentioned that red onions are widely used in the United States, while they are less common in Japan. The data obtained show that for native English speakers, onions with a red surface color are more typical than onions with a brown surface color. Therefore, the typical color was changed to red for onions and the unrelated color was

changed to brown for native English speakers. Further details of this analysis can be found in Appendix G.

Native Japanese Speakers. The results of the native Japanese speakers were also reviewed before modeling. The results of the native Japanese speakers showed that all typical items were classified as more typical than atypical and unrelated items. The results showed that the typicality of the critical items decided in the pilot studies was consistent with the typicality of the native Japanese speakers. The details of the results can be found in Appendix H.

Sentence Typicality Rating Task

Native English Speakers. The task results revealed that the intended typicality was chosen for all sentences, and they were above the chance rate (25 percent). The details of the agreement rates will be found in Appendix G.

Native Japanese Speakers. The results of the sentence typicality rating task showed that the sentences reflected the intended typicality. The details of the agreement rates can be found in Appendix H.

As shown above, the word and sentence tasks confirmed that the intended typicality of the critical words and sentences corresponded to what the participants had in mind. Therefore, the analysis moved to the results of the semantic Stroop task.

Semantic Stroop Task

Native English Speakers.

Descriptive Statistics. The data treatment eliminated 9.21 percent of the experimental data (trials excluding filler items), and 5,638 observations were analyzed. The descriptive statistics showed that the reaction times to the typical color in atypical sentences were the fastest ($M = 812.46$, $SD = 287.70$), followed by typical colors in typical

sentences ($M = 823.37$, $SD = 297.94$). The slowest condition was unrelated color words in atypical sentences ($M = 846.62$, $SD = 287.67$). Table 8 shows the descriptive statistics of reaction times. The tendency for typical color words to respond faster than atypical and unrelated words was consistent regardless of context position (Figure 10). Table 9 and Table 10 show the descriptive statistics for the pre- and post-context conditions. The pre-context conditions showed slower reaction times than the post-context conditions. Figure 11 and Figure 12 illustrate the average response times for each condition.

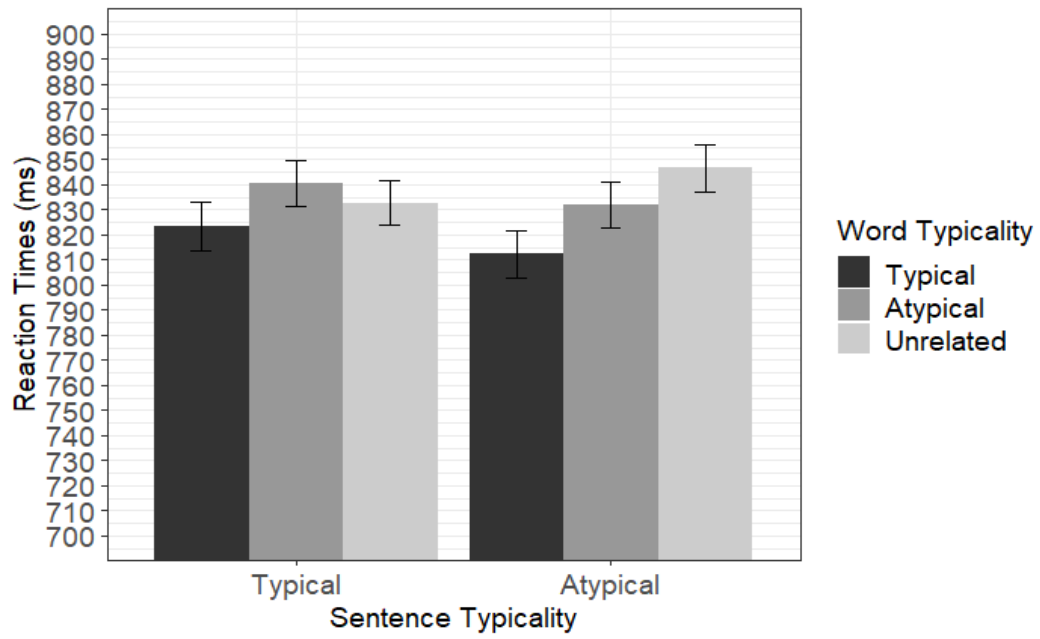
Table 8

Descriptive Statistics of the Reaction Times of the Semantic Stroop Task (Native English Speakers)

Word Typicality	<i>n</i>	<i>M</i>	<i>SD</i>	<i>Median</i>	<i>Minimum</i>	<i>Maximum</i>
Sentence: Typical						
Typical	929	823.37	297.94	747.0	334	1,679
Atypical	956	840.55	282.43	766.0	346	1,679
Unrelated	938	832.56	272.10	768.5	360	1,676
Sentence: Atypical						
Typical	939	812.46	287.70	738.0	327	1,677
Atypical	932	832.00	279.41	769.5	341	1,663
Unrelated	944	846.62	287.67	779.0	313	1,678

Figure 10

Mean Reaction Times of the Semantic Stroop Task (Native English Speakers)



Note. Error bars represent standard error.

Table 9

Descriptive Statistics of the Reaction Times of the Semantic Stroop Task (Native English Speakers: Pre-Context Condition)

Word Typicality	<i>n</i>	<i>M</i>	<i>SD</i>	<i>Median</i>	<i>Minimum</i>	<i>Maximum</i>
Sentence: Typical						
Typical	456	830.42	303.81	752.5	336	1,675
Atypical	477	856.28	286.71	793.0	376	1,679
Unrelated	458	843.71	281.93	770.5	360	1,673
Sentence: Atypical						
Typical	459	820.11	287.53	762.0	367	1,670
Atypical	463	845.29	286.14	792.0	372	1,663
Unrelated	467	862.70	285.94	796.0	402	1,678

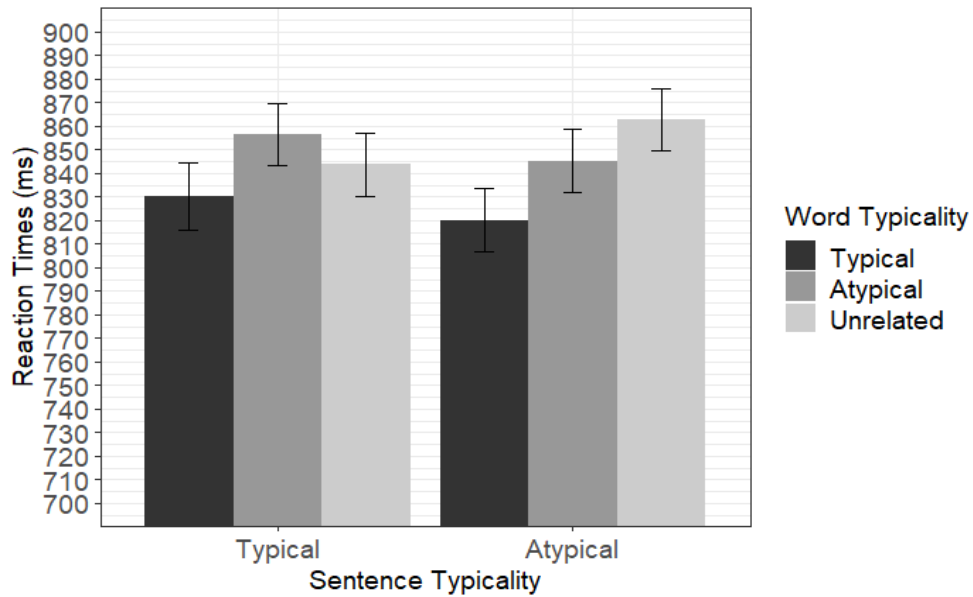
Table 10

Descriptive Statistics of the Reaction Times of the Semantic Stroop Task (Native English Speakers: Post-Context Condition)

Word Typicality	<i>n</i>	<i>M</i>	<i>SD</i>	<i>Median</i>	<i>Minimum</i>	<i>Maximum</i>
Sentence: Typical						
Typical	473	816.58	292.33	741.0	334	1,679
Atypical	479	824.88	277.53	748.0	346	1,664
Unrelated	480	821.93	262.23	766.5	388	1,676
Sentence: Atypical						
Typical	480	805.15	287.96	722.5	327	1,677
Atypical	469	818.88	272.26	761.0	341	1,636
Unrelated	477	830.88	288.80	753.0	313	1,675

Figure 11

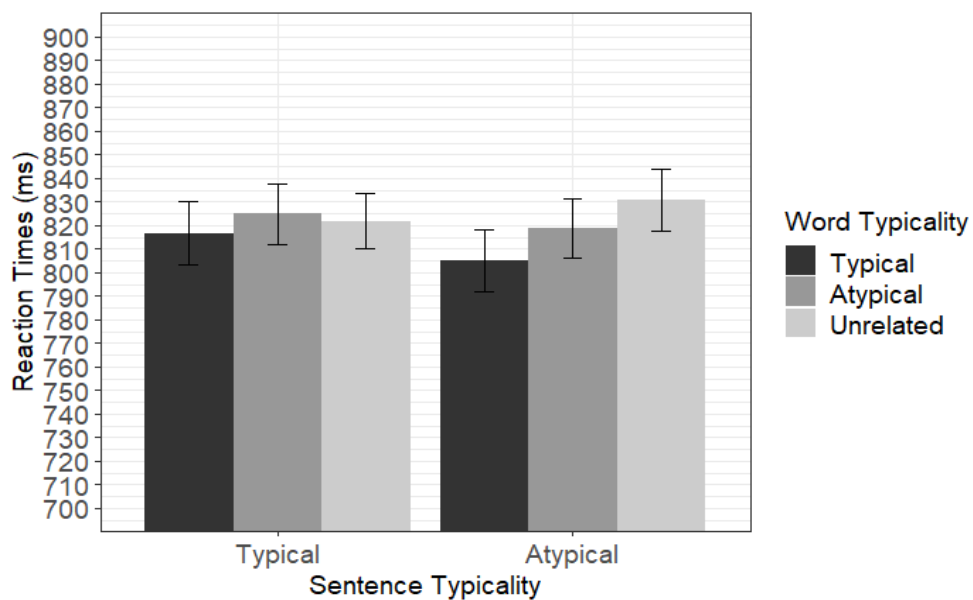
Mean Reaction Times of the Semantic Stroop Task (Native English Speakers: Pre-Context Condition)



Note. Error bars represent standard error.

Figure 12

Mean Reaction Times of the Semantic Stroop Task (Native English Speakers: Post-Context Condition)



Note. Error bars represent standard error.

Modeling Results. Table 11 summarizes the results of the mixed-effects regression modeling. The final model showed that the main effects of word typicality 2-1 (typical-unrelated: $Estimate = -0.03$, $SE = 0.01$, $t = -2.60$, $p = .010$, 95% CI [-0.06, -0.01]) and word typicality 3-2: (atypical-typical: $Estimate = 0.03$, $SE = 0.01$, $t = 2.01$, $p = .046$, 95% CI [0.00, 0.05]). There was no significant interaction between sentence typicality and word typicality. Position in context also showed no significant main effect. To examine whether there was a significant difference between atypical color words and unrelated color words, the values of the variables were changed. Word typicality 2-1 contrasted atypical and unrelated, and word typicality 3-2 contrasted typical and atypical (levels 2 - level 1: unrelated = $-\frac{2}{3}$, atypical = $\frac{1}{3}$, typical = $\frac{1}{3}$). The results showed that there was no significant difference between word typicality 2-1 (atypical-unrelated: $Estimate = -0.01$, $SE = 0.01$, $t = -0.59$, $p = .557$, 95% CI [-0.03, 0.02]) nor significant interaction between word typicality 2-1 and sentence typicality ($Estimate = -0.01$, $SE = 0.03$, $t = -0.54$, $p = .590$, 95% CI [-0.07, 0.04]). Figure 13 illustrates the relation between sentence typicality and word typicality on the reaction times of the semantic Stroop task.

Table 11*Results of Mixed-Effects of the Native English Speakers*

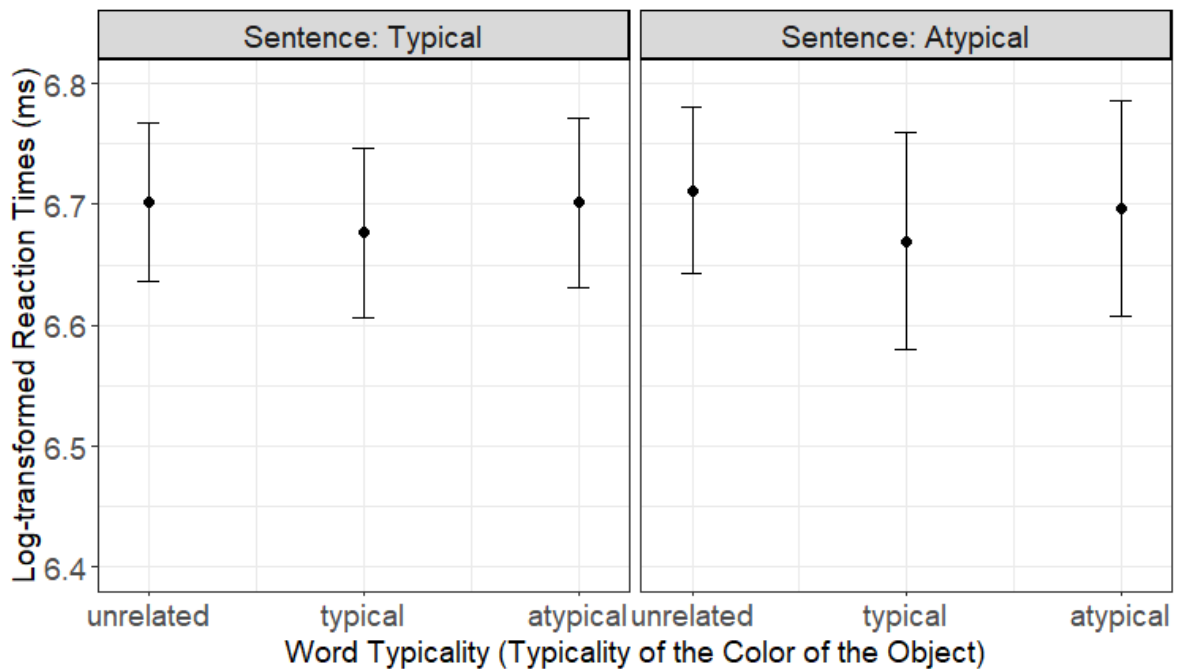
Predictors	Random Effects					
	Fixed Effects				By Subject	By Item
	Estimates	<i>SE</i>	<i>t</i>	<i>p</i>	<i>SD</i>	<i>SD</i>
Intercept	6.72	0.03	197.60	< .001	0.19	0.06
z RT Sentence	0.06	0.01	6.08	< .001	0.04	0.02
Pres Order	-0.00	0.00	-4.94	< .001	–	–
Sentence.Typicality 2-1	-0.00	0.01	-0.08	.940	–	–
Word.Typicality 2-1 (a)	-0.03	0.01	-2.60	.010	–	–
Word.Typicality 2-1 (b)	-0.01	0.01	-0.59	.557	–	–
Word.Typicaliy 3-2	0.03	0.01	2.01	.046	–	–
Position 2-1	-0.02	0.01	-1.48	.140	–	–
Sentence.Typicality 2-1*	-0.02	0.03	-0.63	.532	–	–
Word.Typicality 2-1 (a)						
Sentence.Typicality 2-1*	-0.01	0.03	-0.54	.590	–	–
Word.Typicality 2-1 (b)						
Sentence.Typicality 2-1*	0.00	0.03	0.09	.931	–	–
Word.Typicality 3-2						

Note. z RT Sentence: scaled reading time of each sentence; Pres Order: the order of presentation; Sentence.Typicality 2-1: typicality of sentences (atypical - typical); Word.Typicality 2-1 (a): typicality of word colors (typical - unrelated); Word.Typicality 2-1 (b): typicality of word colors (atypical - unrelated); Word.Typicality 3-2: typicality of word colors (atypical - typical); Position 2-1: position of context (post - pre). Model

formula: $\log(\text{RT.Stroop}) \sim z \text{ RT Sentence} + \text{Pres Order} + \text{Position} + \text{Sentence.Typicality} * \text{Word.Typicality} + (1 + z \text{ RT Sentence} \parallel \text{SubjectID}) + (1 + z \text{ RT Sentence} \parallel \text{ItemID})$

Figure 13

Effects of Sentence Typicality and Word Typicality on the Reaction Times of the Semantic Stroop Task (Native English Speakers)



Note. The y-axis represents the log-transformed reaction times of the semantic Stroop task, while the x-axis represents the three levels in word typicality. Unrelated, typical, and atypical represent respectively unrelated colors, typical colors, and atypical colors of an object color.

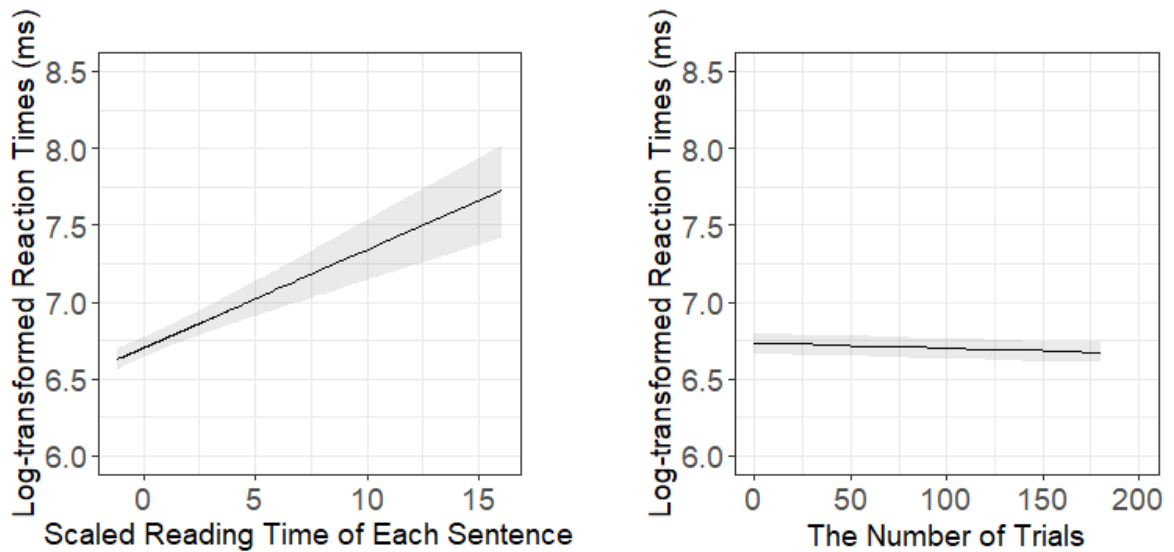
In answering research questions 1 through 3, the current results support only research hypothesis 1 (responding to keys is facilitated when the color of the words matches the color implied in the sentences). Participants responded significantly faster to typical colors than to atypical and unrelated colors, regardless of how typical the sentence was. The results suggest that readers simulate the typical color of objects when they read the L1 sentences. However, reaction time for atypical colors was not significantly faster

than for typical and unrelated colors after reading sentences that implied an atypical color of objects. The results did not support research hypothesis 2 (color simulation depends on the color implied in the L1 sentence) because no significant interaction was found between sentence typicality and word typicality. In addition, there was no significant main effect of context position. The study could not confirm research hypothesis 3 (reaction times for typical and atypical colors do not differ after reading atypical sentences when context is introduced after the keywords).

The two covariates showed the significant main effects (the scaled sentence reading time: $Estimate = 0.06$, $SE = 0.01$, $t = 6.80$, $p < .001$, 95% CI [0.04, 0.08]; presentation order: $Estimate = -0.00$, $SE = 0.00$, $t = -4.94$, $p < .001$, 95% CI [-0.00, -0.00]). Figure 14 visualizes the results of the two variables. The main effect of scaled sentence reading time suggests that the longer readers take to comprehend the sentences, the slower they respond to the color words. This could be because readers do not simulate the color if they do not understand the meaning of the sentence. The main effect of presentation order showed that reaction time decreased as the experiment progressed. This tendency is due to the fact that as the experiment progressed, the participants became more accustomed to the task.

Figure 14

The Scaled Reading Time of Each Sentence and the Presentation Order Variable Included in the Final Model (Native English Speakers)



Note. The y-axis represents the log-transformed reaction times of the semantic Stroop task, while the x-axis represents the scaled reading time of each sentence and the number of trials that were up to 180. For both plots, the grey areas represent 95% confidence intervals.

The model that included only the significant independent variables was also built (Table 12) to test whether the significant variables remained significant even in the absence of nonsignificant main effects and interactions. The model included word typicality 2-1, word typicality 3-2, scaled sentence reading time, and order of presentation. The levels of word typicality were set as follows: unrelated as level 1, typical as level 2, and atypical as level 3. For word typicality 2-1, reaction times to typical colors and reaction times to unrelated colors were compared, and for word typicality 3-2, reaction times to atypical colors and reaction times to typical colors were compared. The main effects of word typicality 2-1 and word typicality 3-2 remained significant in the model (word typicality 2-1: *Estimate* = -0.03, *SE* = 0.01, *t* = -2.58, *p* = .011, 95% CI [-0.06, -0.01]; word typicality 3-2: *Estimate* = 0.03, *SE* = 0.01, *t* = 2.00, *p* = .047, 95% CI [0.00, 0.05]). The coding of

word typicality was changed to compare the response times of atypical and unrelated colors. The levels were set as follows: unrelated as level 1, atypical as level 2, and typical as level 3. The results did not show a significant main effect, which is consistent with the model with all independent variables (word typicality 2-1 (atypical-unrelated): *Estimate* = -0.01, *SE* = 0.01, *t* = -0.58, *p* = .560, 95% CI [-0.03, 0.02]). Thus, the model without the non-significant variables showed similar results to the models with the significant independent variables. For more details on the procedures in R and their results, see Appendix I.

Table 12

Results of Mixed-Effects of the Native English Speakers (Only Significant Variables)

Predictors	Fixed Effects				Random Effects	
	Estimates	<i>SE</i>	<i>t</i>	<i>p</i>	By Subject <i>SD</i>	By Item <i>SD</i>
Intercept	6.72	0.03	197.46	< .001	0.20	0.06
z RT Sentence	0.06	0.01	6.86	< .001	0.04	0.02
Pres Order	-0.00	0.00	-4.91	< .001	–	–
Word.Typicality 2-1 (a)	-0.03	0.01	-2.58	.011	–	–
Word.Typicality 2-1 (b)	-0.01	0.01	-0.58	.560	–	–
Word.Typicaliy 3-2	0.03	0.01	2.00	.047	–	–

Note. z RT Sentence: scaled reading time of each sentence; Pres Order: the order of presentation; Sentence.Typicality 2-1: typicality of sentences (atypical - typical); Word.Typicality 2-1 (a): typicality of word colors (typical - unrelated); Word.Typicality 2-1 (b): typicality of word colors (atypical - unrelated); Word.Typicality 3-2: typicality of word colors (atypical - typical). Model formula: $\log(\text{RT.Stroop}) \sim \text{z RT Sentence} + \text{Pres Order} + \text{Word.Typicality} + (1 + \text{z RT Sentence} \parallel \text{SubjectID}) + (1 + \text{z RT Sentence} \parallel \text{ItemID})$

Native Japanese Speakers.

Descriptive Statistics. A rate of 9.24 percent of the data was excluded from the experimental trials, yielding a total of 5,881 observations. Table 13 shows the descriptive statistics on reaction times. Interestingly, the results were similar to native English speakers. Regardless of sentence type, typical word colors responded faster than atypical and unrelated color conditions. The differences were much more pronounced than the results for native English speakers (Figure 15). The reaction times were fastest in typical word colors after reading typical sentences ($M = 695.77$, $SD = 216.26$), and typical word colors after reading atypical sentences came second ($M = 701.78$, $SD = 213.85$). The slowest condition was atypical word colors in atypical sentences ($M = 745.78$, $SD = 216.60$), followed by atypical word colors in typical sentences ($M = 742.75$, $SD = 198.80$). Table 14 and 15 reports the descriptive statistics of pre- and post-context conditions. Both conditions showed that reaction times were faster for typical word colors than for all other conditions. Remarkably, although the difference was small, participants responded faster to atypical word colors only in the pre-context condition than to unrelated colors in atypical sentence conditions. In the post-context condition, the opposite trend was observed: Unrelated colors were much faster than atypical color words in atypical sentences. Figure 16 and Figure 17 illustrate the average reaction times of the individual context conditions.

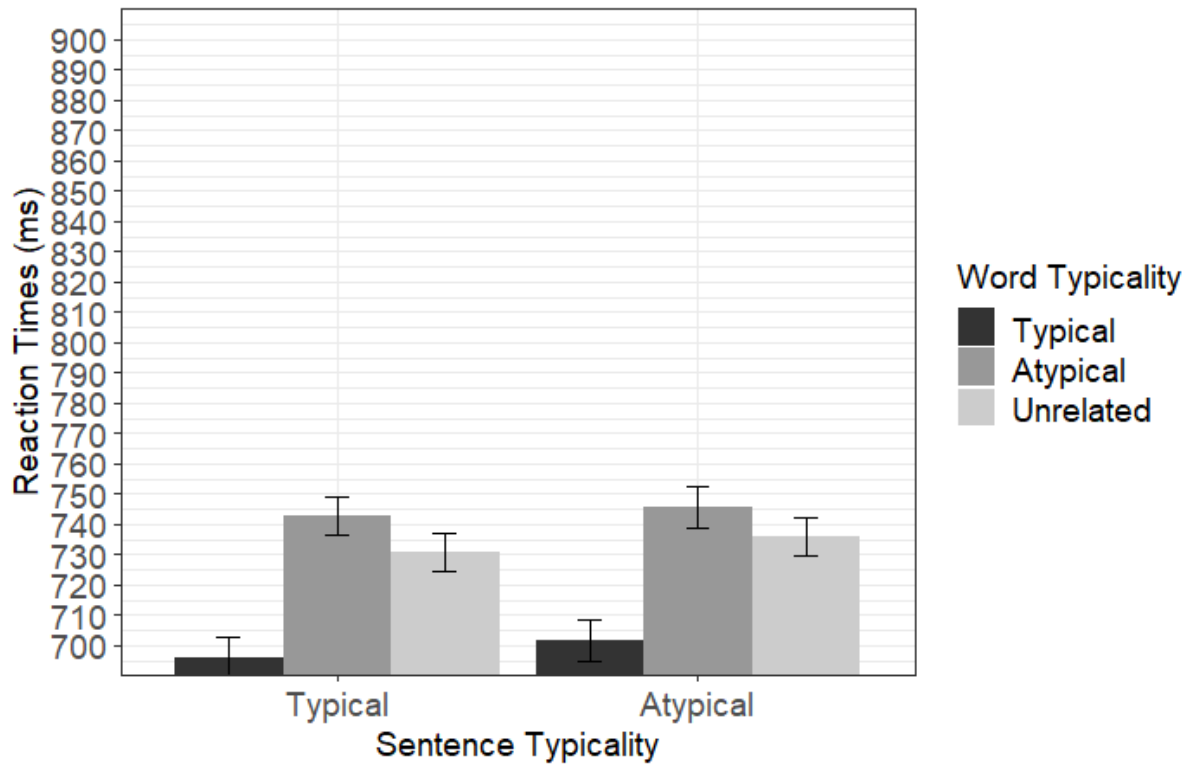
Table 13

Descriptive Statistics of the Reaction Times of the Semantic Stroop Task (Native Japanese Speakers)

Word: Typicality	<i>n</i>	<i>M</i>	<i>SD</i>	<i>Median</i>	<i>Minimum</i>	<i>Maximum</i>
Sentence: Typical						
Typical	975	695.77	216.26	653.0	283	1,326
Atypical	989	742.75	198.80	700.0	352	1,329
Unrelated	988	730.76	198.85	683.0	347	1,328
Sentence: Atypical						
Typical	976	701.78	213.85	657.5	283	1,320
Atypical	970	745.78	216.60	696.5	302	1,329
Unrelated	983	735.94	203.23	688.0	351	1,328

Figure 15

Mean Reaction Times of the Semantic Stroop Task (Native Japanese Speakers)



Note. Error bars represent standard error.

Table 14

Descriptive Statistics of the Reaction Times of the Semantic Stroop Task (Native Japanese Speakers: Pre-Context Condition)

Word: Typicality	<i>n</i>	<i>M</i>	<i>SD</i>	<i>Median</i>	<i>Minimum</i>	<i>Maximum</i>
Sentence: Typical						
Typical	492	692.86	214.94	651	283	1,326
Atypical	492	742.43	196.67	707	352	1,294
Unrelated	491	727.65	199.05	681	384	1,328
Sentence: Atypical						
Typical	493	705.03	216.77	657	283	1,304
Atypical	486	735.55	207.38	693	302	1,329
Unrelated	494	744.04	206.04	695	388	1,290

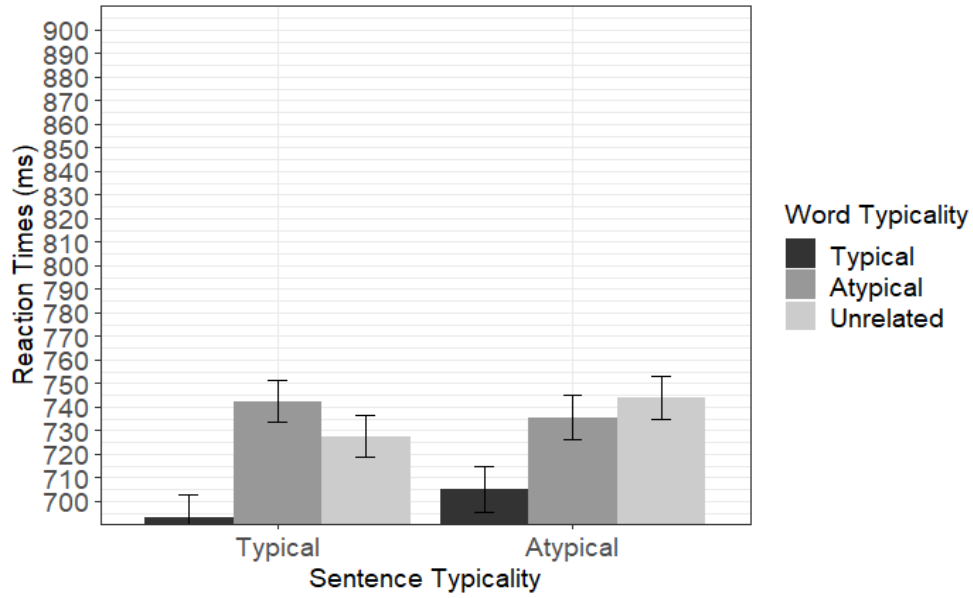
Table 15

Descriptive Statistics of the Reaction Times of the Semantic Stroop Task (Native Japanese Speakers: Post-Context Condition)

Word: Typicality	<i>n</i>	<i>M</i>	<i>SD</i>	<i>Median</i>	<i>Minimum</i>	<i>Maximum</i>
Sentence: Typical						
Typical	483	698.74	217.79	657	333	1,313
Atypical	497	743.07	201.09	695	398	1,329
Unrelated	497	733.83	198.81	687	347	1,327
Sentence: Atypical						
Typical	483	698.46	211.01	659	331	1,320
Atypical	484	756.06	225.22	699	372	1,318
Unrelated	489	727.76	200.22	677	351	1,328

Figure 16

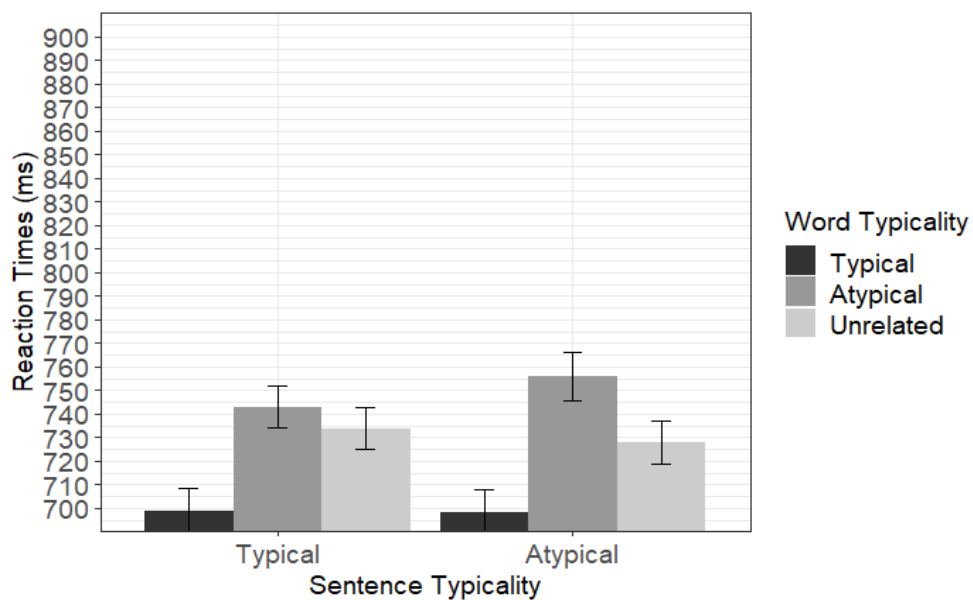
Mean Reaction Times of the Semantic Stroop Task (Native Japanese Speakers: Pre-Context Condition)



Note. Error bars represent standard error.

Figure 17

Mean Reaction Times of the Semantic Stroop Task (Native Japanese Speakers: Post-Context Condition)

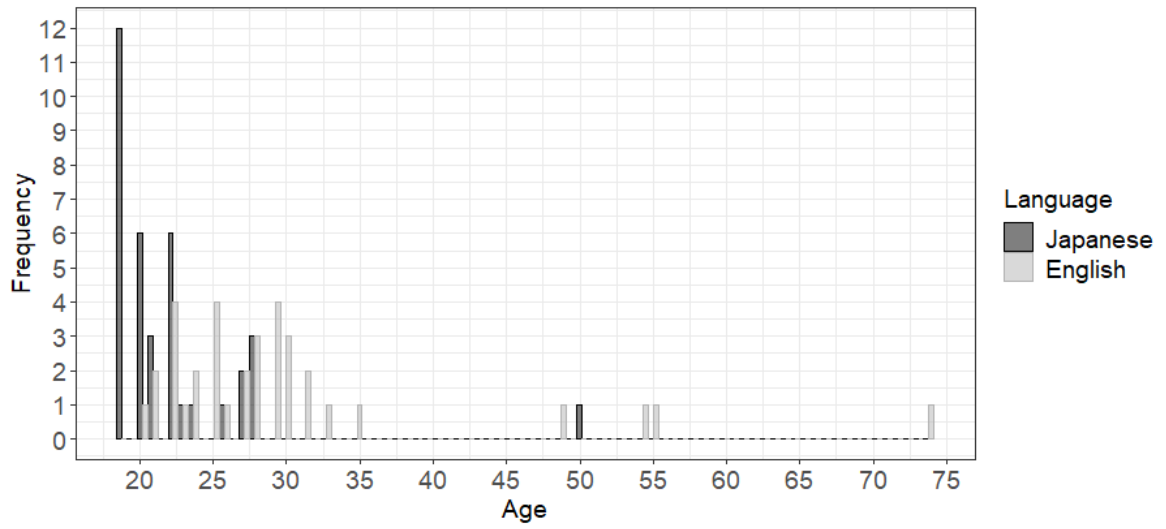


Note. Error bars represent standard error.

The mean reaction times of native Japanese speakers were much faster than native English speakers (Table 8 and Table 13). This is likely due to the different age distributions of the participants. The age of native Japanese speakers ranges from 19 to 50, and the age of native English speakers ranges from 20 to 74 (Figure 18). Most of the Japanese participants were under 25. However, the ages of the native English-speaking participants varied widely. Previous studies examining the influence of age on reaction times have found that reaction times increase as participants age (e.g., Hardwick et al., 2021; Woods et al., 2015). In this study, similar trends were observed in reaction times in the semantic Stroop task, but only in native English speakers' data (Figure 19). The slope of the line is much less steep for native Japanese speakers than for native English speakers because the age range is smaller. This difference in mean reaction times and the influence of participants' age is not important for the interpretation of the following results, including modeling. All independent variables in the study were examined within subjects, and the main focus of the study was on the differences in reaction times within each subject. Furthermore, the difference in average reaction times between languages is not the research interest.

Figure 18

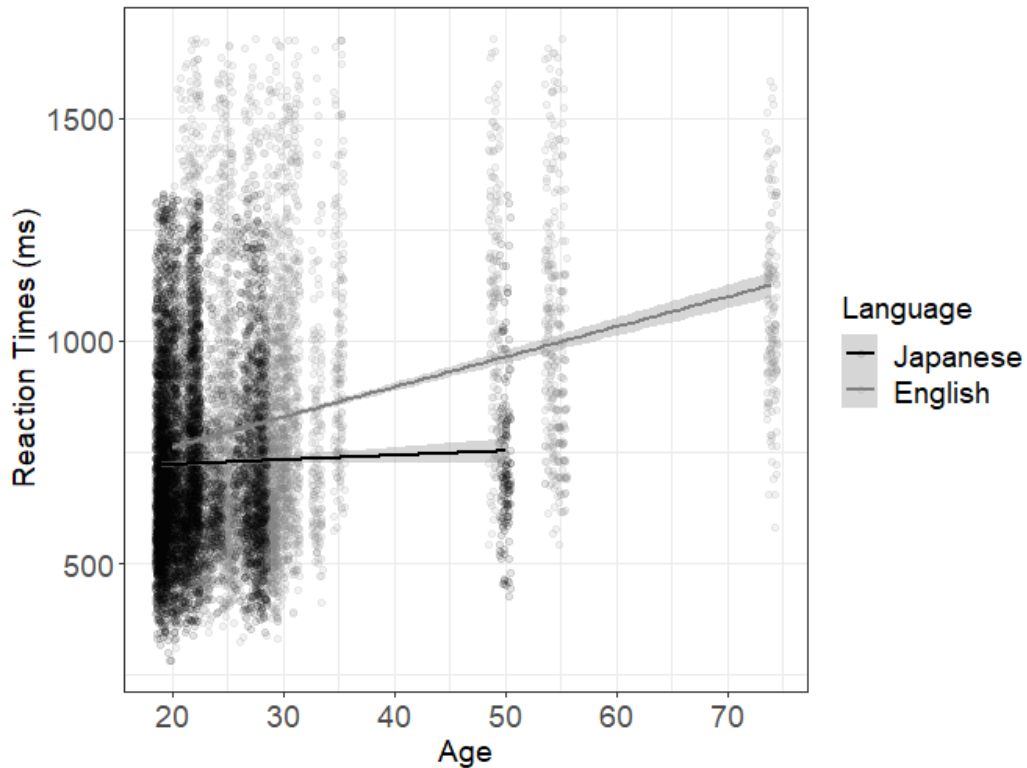
The Distributions of Native English and Japanese Participants' Age



Note. The y-axis represents the frequency of counts.

Figure 19

The Relation of Native English and Japanese Participants' Age and Reaction Times of the Semantic Stroop Task



Note. The y-axis represents the reaction times of the semantic Stroop task, while the x-axis represents the participants' age. The black line and the grey lines are regression lines. The grey areas represent 95% confidence intervals.

Modeling Results. As Table 16 shows, the final model showed that the main effects of word typicality 2-1 (typical-unrelated: $Estimate = -0.06$, $SE = 0.02$, $t = -2.83$, $p = .005$, 95% CI [-0.09, -0.02]) and word typicality 3-2: (atypical-unrelated: $Estimate = 0.07$, $SE = 0.02$, $t = 3.60$, $p < .001$, 95% CI [0.03, 0.11]) were significant. There was no significant interaction between sentence typicality and word typicality. The comparisons between atypical colors and unrelated colors did not show a significant main effect (atypical-unrelated: $Estimate = 0.02$, $SE = 0.02$, $t = 0.77$, $p = .445$, 95% CI [-0.02, 0.05]). Furthermore, there was no significant interaction between word typicality 2-1 and sentence

typicality ($Estimate = -0.00$, $SE = 0.04$, $t = -0.12$, $p = .906$, 95% CI [-0.08, 0.07]). There was no significant main effect of context position ($Estimate = 0.00$, $SE = 0.02$, $t = 0.22$, $p = .828$, 95% CI [-0.03, 0.04]), either. The lack of significant main effect of context position indicates that the position of the context did not affect the simulation in the L1 reading. Figure 20 illustrates the relationship between sentence typicality and word typicality in the reaction times.

Table 16*Results of Mixed-Effects of the Native Japanese Speakers*

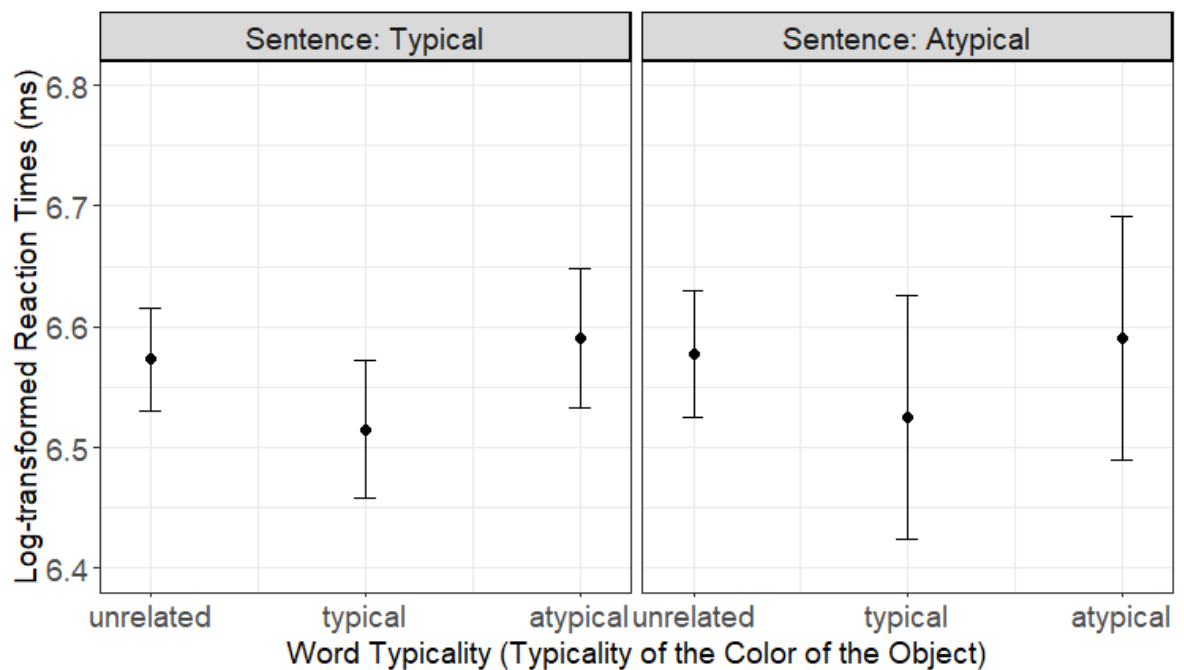
Predictors	Random Effects					
	Fixed Effects				By Subject	By Item
	Estimates	<i>SE</i>	<i>t</i>	<i>p</i>	<i>SD</i>	<i>SD</i>
Intercept	6.59	0.02	295.25	< .001	0.12	0.10
z RT Sentence	0.06	0.01	9.55	< .001	0.03	0.02
Pres Order	-0.00	0.00	-4.72	< .001	—	—
Sentence.Typicality 2-1	0.00	0.02	0.27	.787	—	—
Word.Typicaliy 2-1 (a)	-0.06	0.02	-2.83	.005	—	—
Word.Typicaliy 2-1 (b)	0.02	0.02	0.77	.445	—	—
Word.Typicaliy 3-2	0.07	0.02	3.60	< .001	—	—
Position 2-1	0.00	0.02	0.22	.828	—	—
Sentence.Typicality 2-1*	0.01	0.04	0.14	.887	—	—
Word.Typicality 2-1 (a)						
Sentence.Typicality 2-1*	-0.00	0.04	-0.12	.906	—	—
Word.Typicality 2-1 (b)						
Sentence.Typicality 2-1*	-0.01	0.04	-0.26	.795	—	—
Word.Typicality 3-2						

Note. z RT Sentence: scaled reading time of each sentence; Pres Order: the order of presentation; Sentence.Typicality 2-1: typicality of sentences (atypical - typical); Word.Typicality 2-1 (a): typicality of word colors (typical - unrelated); Word.Typicality 2-1 (b): typicality of word colors (atypical - unrelated); Word.Typicality 3-2: typicality of word colors (atypical - typical); Position 2-1: position of context (post - pre). Model

formula: $\log(\text{RT.Stroop}) \sim z \text{ RT Sentence} + \text{Pres Order} + \text{Position} + \text{Sentence.Typicality} * \text{Word.Typicality} + (1 + z \text{ RT Sentence} | \text{SubjectID}) + (1 + z \text{ RT Sentence} | \text{ItemID})$

Figure 20

Effects of Sentence Typicality and Word Typicality on the Reaction Times of the Semantic Stroop Task (Native Japanese Speakers)



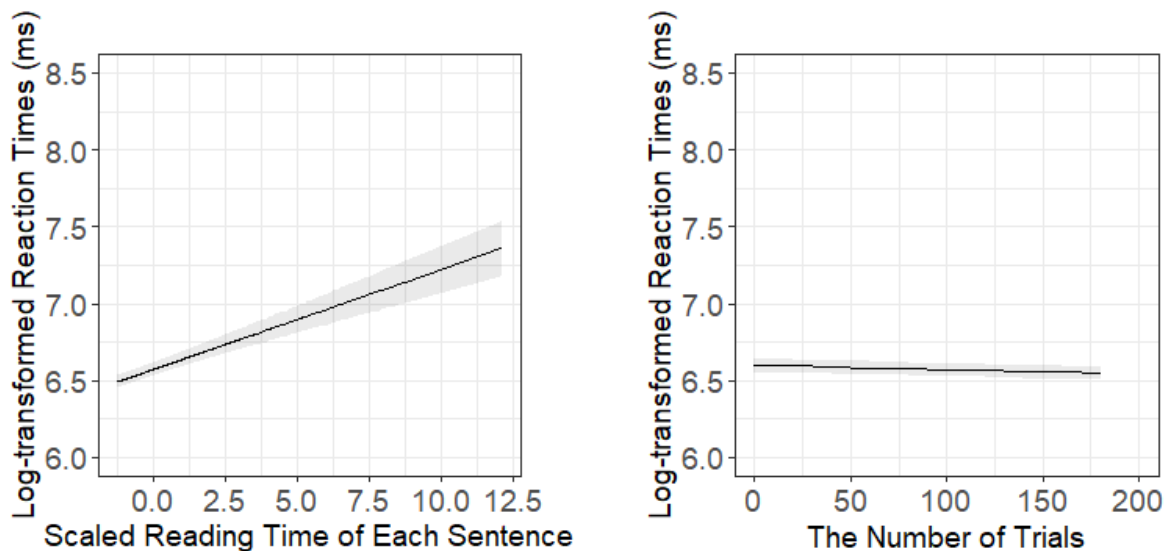
Note. The y-axis represents the log-transformed reaction times of the semantic Stroop task, while the x-axis represents the three levels in word typicality. Unrelated, typical, and atypical represent unrelated colors, typical colors, and atypical colors of object color.

The current results of native Japanese speakers are consistent with those of native English speakers, and only research hypothesis 1 was supported. Participants responded significantly faster to typical colors than to atypical and unrelated colors, regardless of the typicality of the sentence. Research hypothesis 2 was not supported because there was no significant interaction between sentence typicality and word typicality. The results contradict research hypothesis 3 (i.e., the rejection of context position effect).

The scaled sentence reading time and presentation order showed the significant main effects (the scaled sentence reading time: $Estimate = 0.06$, $SE = 0.01$, $t = 9.55$, $p < .001$, 95% CI [0.05, 0.08]; presentation order: $Estimate = -0.00$, $SE = 0.00$, $t = -4.72$, $p < .001$, 95% CI [-0.00, -0.00]) (Figure 21). As discussed for native English speakers' data, the main effect of scaled sentence reading time indicated that readers had difficulty simulating the color information when they had difficulty understanding the sentence. The main effect of presentation order indicated that participants became relatively more accustomed to the Stroop task.

Figure 21

The Scaled Reading Time of Each Sentence and the Presentation Order Variable Included in the Final Model (Native Japanese Speakers)



Note. The y-axis represents the log-transformed reaction times of the semantic Stroop task, while the x-axis represents the scaled reading time of each sentence and the number of trials that were up to 180. For both plots, the grey areas represent 95% confidence intervals.

The model that only included the significant independent variables was built (Table 17). The model included word typicality 2-1, word typicality 3-2, scaled sentence reading

time, and presentation order. The levels of word typicality were set as follows: unrelated as level 1, typical as level 2, and atypical as level 3. As with the results of native English speakers, the main effects of word typicality 2-1 and word typicality 3-2 remained significant in the model (word typicality 2-1: *Estimate* = -0.06, *SE* = 0.02, *t* = -2.83, *p* = .005, 95% CI [-0.09, -0.02]; word typicality 3-2: *Estimate* = 0.07, *SE* = 0.02, *t* = 3.60, *p* < .001, 95% CI [0.03, 0.11]). The reaction times of atypical and unrelated colors were not significantly different (word typicality 2-1 (atypical-unrelated): *Estimate* = 0.02, *SE* = 0.02, *t* = 0.77, *p* < .445, 95% CI [-0.02, 0.05]). Thus, the model without nonsignificant variables showed very similar results to the models with all independent variables. In addition, the results of the native English and Japanese speakers were consistent. The details of the results can be found in Appendix J.

Table 17*Results of Mixed-Effects of the Native Japanese Speakers (Only Significant Variables)*

Predictors	Fixed Effects					Random Effects	
	Estimates	<i>SE</i>	<i>t</i>	<i>p</i>	By Subject <i>SD</i>	By Item <i>SD</i>	
Intercept	6.59	0.02	295.24	< .001	0.12	0.10	
z RT Sentence	0.06	0.01	9.55	< .001	0.03	0.02	
Pres Order	-0.00	0.00	-4.72	< .001	—	—	
Word.Typicality 2-1 (a)	-0.06	0.02	-2.83	.005	—	—	
Word.Typicality 2-1 (b)	0.02	0.02	0.77	.445	—	—	
Word.Typicality 3-2	0.07	0.02	3.60	< .001	—	—	

Note. z RT Sentence: scaled reading time of each sentence; Pres Order: the order of presentation; Sentence.Typicality 2-1: typicality of sentences (atypical - typical); Word.Typicality 2-1 (a): typicality of word colors (typical - unrelated); Word.Typicality 2-1 (b): typicality of word colors (atypical - unrelated); Word.Typicality 3-2: typicality of word colors (atypical - typical). Model formula: $\log(\text{RT.Stroop}) \sim \text{z RT Sentence} + \text{Pres Order} + \text{Word.Typicality} + (1 + \text{z RT Sentence} \mid \text{SubjectID}) + (1 + \text{z RT Sentence} \mid \text{ItemID})$

Summary of Experiment 1

Research hypothesis 1 (reaction time is reduced when the color of the words presented matches the color implied by the sentences) was supported. Research hypothesis 2 (the simulation of the color depends on the color implied by the L1 sentence) was not supported. Research hypothesis 3 (the position of the context phrases would change the simulation in L1) was not supported. These results were the same for native English and Japanese speakers.

Chapter 5: Experiment 2

Aim of Experiment 2

Experiment 2 was conducted to investigate whether L2 learners simulate the colors of the object and to what extent their L2 proficiency affects the simulation. The results of Experiment 1 served as the baseline L1 data.

Method

Participants

A total of 36 participants were recruited for the study (20 females and 16 males). None of them took part in the pilot studies. The number of participants was determined based on power analysis (see Chapter 3). They were native Japanese speakers who learned English mainly in Japan. Twenty-nine of them were graduate or undergraduate students at Japanese universities. Their fields of study varied, including science, literature, agriculture, physics, biology, law, computer science, international development, engineering, government, economics, foreign languages, education, English, and global and regional. A background questionnaire indicated that 15 participants had experience studying abroad in English-speaking countries. Although two participants indicated that they had lived in an English-speaking country (3.5 years and five years), the rest of the participants learned English primarily in a foreign language context. The study addressed the learners' L2 proficiencies. The V_YesNo v1.1 test (Meara & Miralpeix, 2016) revealed that participants' L2 proficiency ranged from beginner to advanced. Table 18 provides the descriptive statistics of the learners' backgrounds. L2 learners show higher mean scores of the vocabulary size test compared to native Japanese speakers (cf., Table 2, Chapter 4) (3,820.33 vs. 5,325.14).

Table 18*Descriptive Statistics of the Japanese Learners of English*

	<i>N</i>	<i>M</i>	<i>SD</i>	<i>Median</i>	<i>Minimum</i>	<i>Maximum</i>
Vocabulary Size	36	5,325.14	1,065.10	5,118	3,033	7,721
Age	35	23.17	2.84	23	19	29
Years Learning English	36	12.08	4.05	12	7	24
Self-reported proficiency scores						
Reading	36	4.64	1.20	5	1	6
Listening	36	4.22	1.38	4	2	7
Speaking	36	3.69	1.33	4	1	7
Writing	36	3.81	1.41	4	1	6
Grammar	36	4.06	1.43	4	1	7

Note. Vocabulary size scores were the of V_YesNo v1.0 test scores (Meara & Miralpeix, 2016). One participant declined to provide their age. Self-reported proficiency scores were calculated from rating scores on a 7-point Likert scale (1 = very poor, 7 = very good).

Tasks and Materials

The semantic Stroop task was performed with manual responses in the same way as in Experiment 1 for native English speakers.

All other experimental tasks and materials were the same as in Experiment 1 for native English speakers, except for the following three points. First, all instructions were given in Japanese. Second, participants were given the V_YesNo v1.1 test (Meara & Miralpeix, 2016) instead of the article task. Third, some items were corrected because some participants in Experiment 1 pointed out spelling and grammatical errors after completing the task. These were mostly local grammatical errors, and none of the

participants who reported these errors indicated that the errors affected their understanding. Details of the corrections can be found in Appendix K.

Procedure

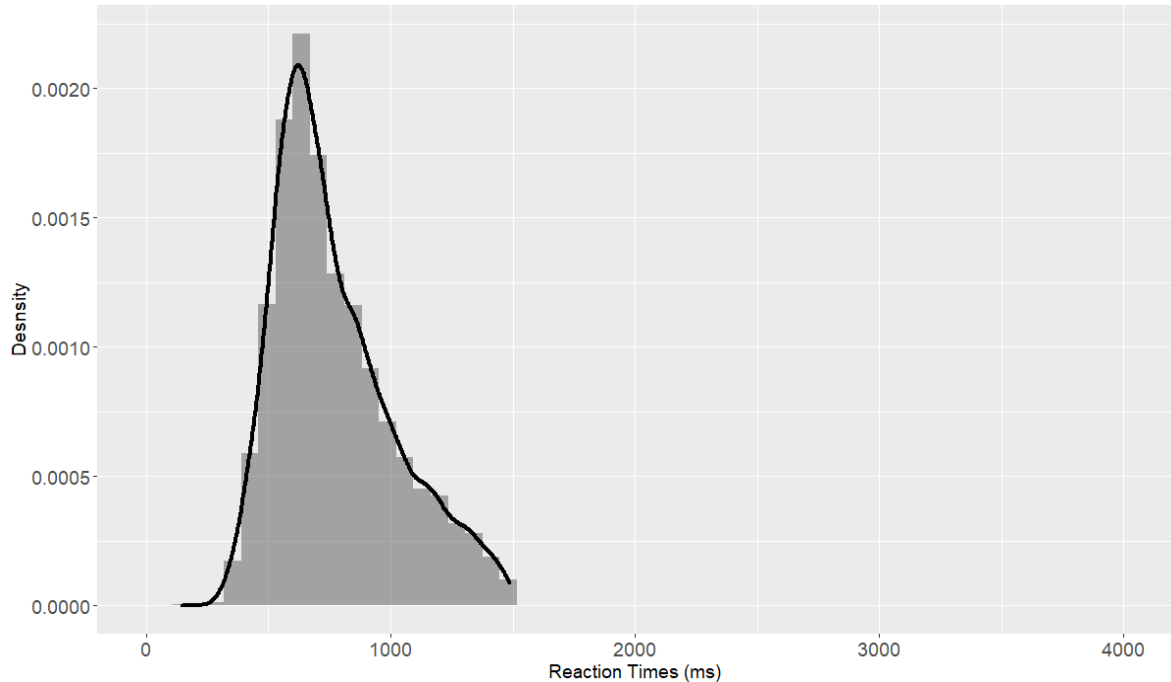
The procedure was identical to Experiment 1 for native Japanese speakers, except that the tasks were presented in English. The instructions can be found in Appendix F.

Analysis

All statistical analyses were performed using R version 4.1.1 (R Core Team, 2021). Incorrect responses, filler items, and all data from participants with an overall accuracy of less than 80% on the color decision and with an overall accuracy of less than 50% on the comprehension question were excluded before analysis. In addition, reaction times that had more than \pm three median absolute deviations from the median were excluded (MAD) (Leys et al., 2013). Reaction times were measured from the presentation of the colored word to the moment the participant pressed one of the S, D, K, or L keys. Figure 22 shows the distribution of reaction times. With the exception of the filler words, 9.03 percent of the data were deleted, and 5,895 observations were used as dependent variables.

Figure 22

The Distribution of the Reaction Times of the Semantic Stroop Task (Japanese Learners of English)



Note. Density was calculated using kernel density estimation.

After data processing, the probabilistic distributions were selected using the same procedure as in Experiment 1 (see *Analysis* in Chapter 4). Considering the following goodness-of-fit statistics for parametric distributions computed with the *fitdist* function of the package (Table 19), the log-normal distribution was chosen as the probabilistic distribution for the current investigation.

Table 19*Goodness-of-Fit Statistics and Information Criterion*

	Weibull	Gamma	Log-normal	Normal
Kolmogorov-Smirnov	0.09	0.06	0.04	0.10
Cramér-von Mises	16.67	6.42	2.94	17.90
Anderson-Darling	101.99	37.03	17.45	105.65
AIC	81,646.62	80,873.82	80,722.09	81,643.33
BIC	81,659.99	80,887.18	80,735.46	81,656.69

Note. AIC refers to Akaike Information Criterion. BIC refers to Bayesian Information Criterion.

After the data treatment, a series of linear mixed-effects modeling was performed, and the *lme4* package 1.1-27.1 was used (Bates et al., 2021). The dependent variable was log-transformed reaction time in the semantic Stroop task. The independent variables were sentence typicality, word typicality, L2 proficiency (the results of the vocabulary size test), the interaction of the three variables, and context position. The number of trials and the reading time of each sentence was also included as possible covariates. The reading time of a sentence was measured from the presentation of the sentence to the time when the participants pressed the space bar. All categorical variables were contrasted (repeated)-coded to compare neighboring factor levels with *contr.sdif* function of the *MASS* package 7.3-54 (Venables & Ripley, 2002). Table 20 shows how each condition was coded. The reading time and the scores of the vocabulary size were scaled to avoid convergence problems.

Table 20*Dependent Variables and Their Assigned Codes*

Levels	2-1	3-2
Sentence Typicality		
1. Typical	-0.5	
2. Atypical	0.5	
Word Typicality		
1. Unrelated	-0.667	-0.333
2. Typical	0.333	-0.333
3. Atypical	0.333	0.667
Position		
1. Pre	-0.5	
2. Post	0.5	

Note. The numbers in the column of *Levels* refer to the levels of the variables. 2-1 refers to the comparison of level 2 - level 1, and 3-2 refers to the comparison of level 3 - level 2.

The best LME model was determined by the following procedures. First, the possible covariates, reading time, and order of presentation were considered to decide which covariate to include in the final model. The null model was compared with the model that included the possible covariates. The results showed that the model with the presentation order had the lowest AIC among the three models. Then, the model with presentation order was compared with the model that included both presentation order and scaled reading time. The models with both covariates showed the lowest AIC. Thus, the model included sentence typicality, word typicality, L2 proficiency, the interaction of the three, context position, presentation order, and scaled reading time as independent variables.

Next, the best random effect structure was considered using the same procedure as in Experiment 1 (see *Analysis* in Chapter 4).

Finally, the variance inflation factors (VIF) were checked with the *check_collinearity* function of the *performance* package 0.9.0 (Lüdtke et al., 2021). The VIF threshold was set at 5. The analysis confirmed that the final models did not have multicollinearity problems.

Results and Discussion

Rating Tasks

Word Typicality Rating Task. Before reaction times were analyzed, the results of the word- and sentence-level assessment tasks were considered. In all items, typical colors were considered more typical than atypical and unrelated colors. It can be said that the typical color of the words in the experimental tasks reflected the typicality of what the participants had. The details of the rating scores can be found in Appendix L, including the descriptive statistics and plots.

Sentence Typicality Rating Task. In all sentences, the intended typicality was chosen. The agreement rates were above the chance levels (25 percent). Following Connell and Lynott, the author determined that each experimental sentence of the current study implied the intended colors. The details of the agreements and analysis will be found in Appendix L.

In summary, the results of the word and sentence rating task confirmed that the typicality of the experimental material implied the intended colors at both the word and sentence levels. Therefore, the analysis moved to the results of the semantic Stroop task.

Semantic Stroop Task

Descriptive Statistics. After the data treatment, reaction times of the semantic

Stroop task were analyzed. The descriptive statistics of reaction times are presented in Table 21. The typical color words in typical sentences were the fastest condition ($M = 752.30$, $SD = 254.22$), which was followed by the typical color words in atypical sentences ($M = 753.33$, $SD = 252.72$). Regardless of sentence types, reaction times were faster for typical colors than for other color types, atypical and unrelated (Figure 23). This trend is consistent with results from native English and Japanese speakers. Reaction times for atypical colors were slower than for unrelated colors, both pre-context and post-context (Table 22 and Table 23). Descriptive statistics showed that readers always simulated the typical color of the object, regardless of sentence typicality and context position. Figure 24 and Figure 25 show the average response times for each condition.

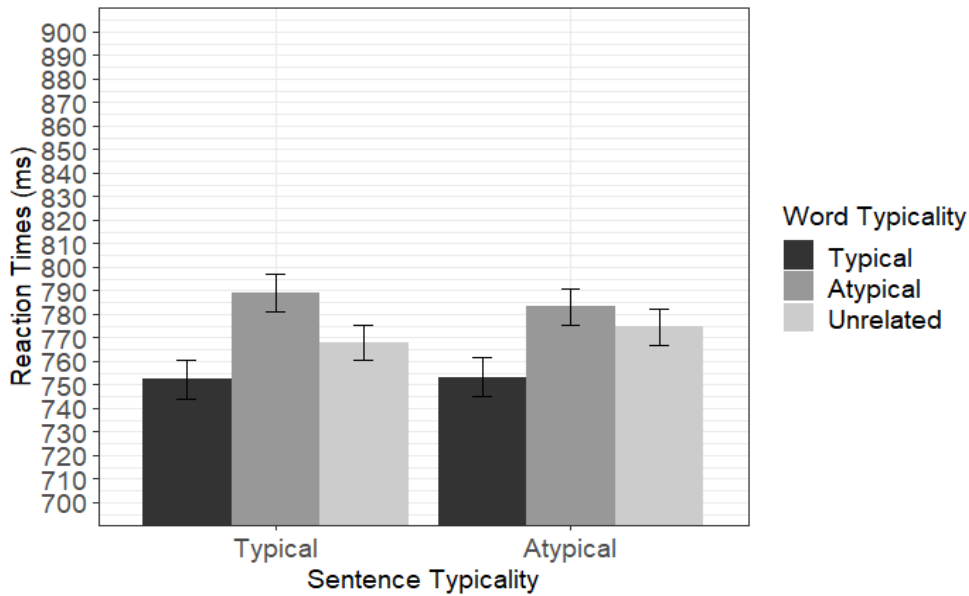
Table 21

Descriptive Statistics of the Reaction Times of the Semantic Stroop Task (Japanese Learners of English)

Word: Typicality	<i>n</i>	<i>M</i>	<i>SD</i>	<i>Median</i>	<i>Minimum</i>	<i>Maximum</i>
Sentence: Typical						
Typical	970	752.30	254.22	700.0	319	1,484
Atypical	976	789.11	243.11	739.5	300	1,482
Unrelated	994	767.98	238.48	709.5	335	1,483
Sentence: Atypical						
Typical	974	753.33	252.72	699.0	144	1,481
Atypical	974	783.22	244.42	723.0	319	1,463
Unrelated	1007	774.64	241.67	714.0	315	1,485

Figure 23

Mean Reaction Times of the Semantic Stroop Task (Japanese Learners of English)



Note. Error bars represent standard error.

Table 22

Descriptive Statistics of the Reaction Times of the Semantic Stroop Task (Japanese Learners of English: Pre-Context Condition)

Word: Typicality	<i>n</i>	<i>M</i>	<i>SD</i>	<i>Median</i>	<i>Minimum</i>	<i>Maximum</i>
Sentence: Typical						
Typical	495	747.66	259.68	686.0	319	1,474
Atypical	485	792.96	254.69	729.0	300	1,482
Unrelated	494	761.29	242.39	698.5	362	1,483
Sentence: Atypical						
Typical	488	750.87	266.13	690.0	144	1,475
Atypical	487	779.42	253.75	712.0	319	1,463
Unrelated	511	766.90	242.63	706.0	358	1,473

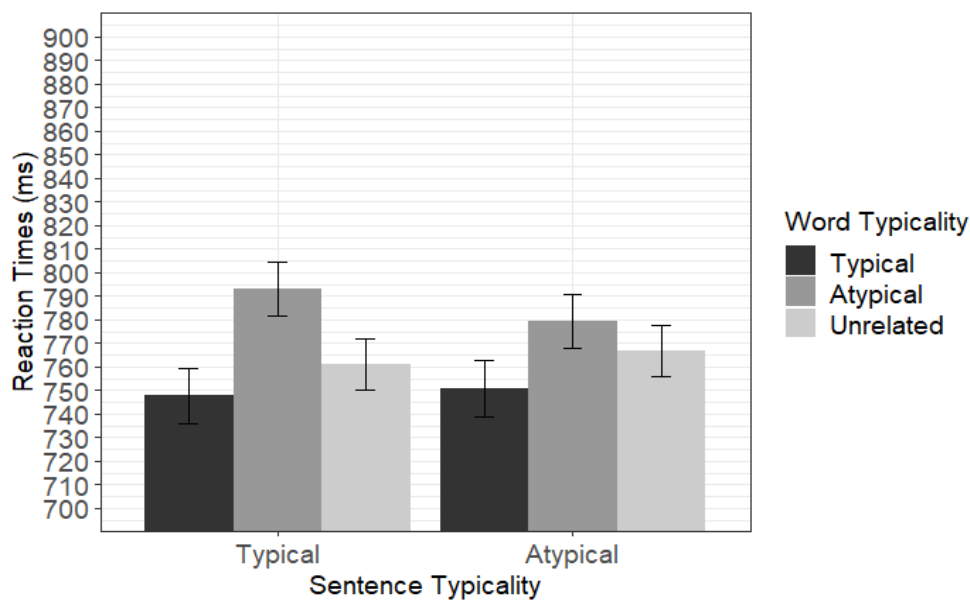
Table 23

Descriptive Statistics of the Reaction Times of the Semantic Stroop Task (Japanese Learners of English: Post-Context Condition)

Word: Typicality	<i>n</i>	<i>M</i>	<i>SD</i>	<i>Median</i>	<i>Minimum</i>	<i>Maximum</i>
Sentence: Typical						
Typical	475	757.13	248.59	708	333	1,484
Atypical	491	785.30	231.30	746	382	1,459
Unrelated	500	774.59	234.61	719	335	1,463
Sentence: Atypical						
Typical	486	755.81	238.74	703	338	1,481
Atypical	487	787.01	234.91	739	354	1,459
Unrelated	496	782.60	240.67	722	315	1,485

Figure 24

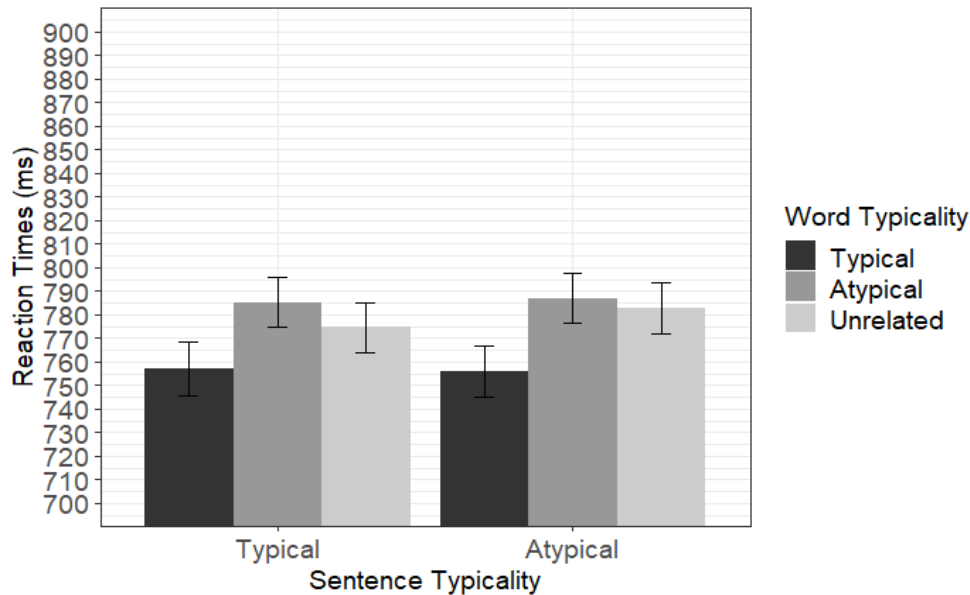
Mean Reaction Times of the Semantic Stroop Task (Japanese Learners of English: Pre-Context Condition)



Note. Error bars represent standard error.

Figure 25

Mean Reaction Times of the Semantic Stroop Task (Japanese Learners of English: Post-Context Condition)

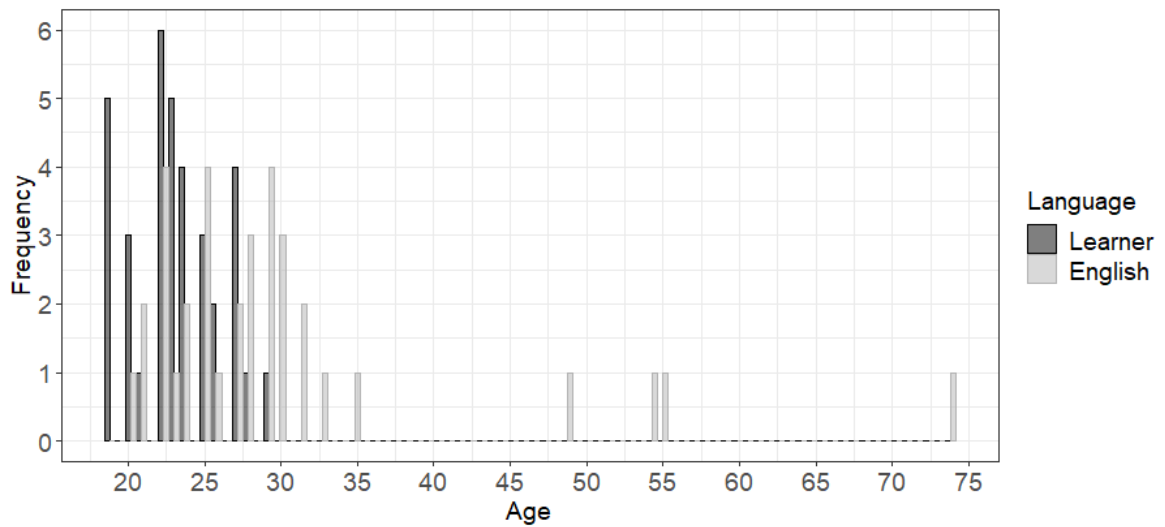


Note. Error bars represent standard error.

The average reaction times of L2 learners were faster than those of native English speakers. Intuitively, native English speakers should have shown faster reaction times than L2 learners because native English speakers solve the task in their L1. Experiment 1 showed that the average reaction times of native Japanese speakers were faster than those of native English speakers, probably due to the age difference between the participants. This was probably also true for native English speakers and L2 learners. Figure 26 illustrates the frequency of age of the participants in each group. All L2 learners, with the exception of the participant who refused to provide age, were under 30 years old, and the age of native English speakers ranged from 20 to 74.

Figure 26

The Distributions of Native English Speakers and English Learner' Age



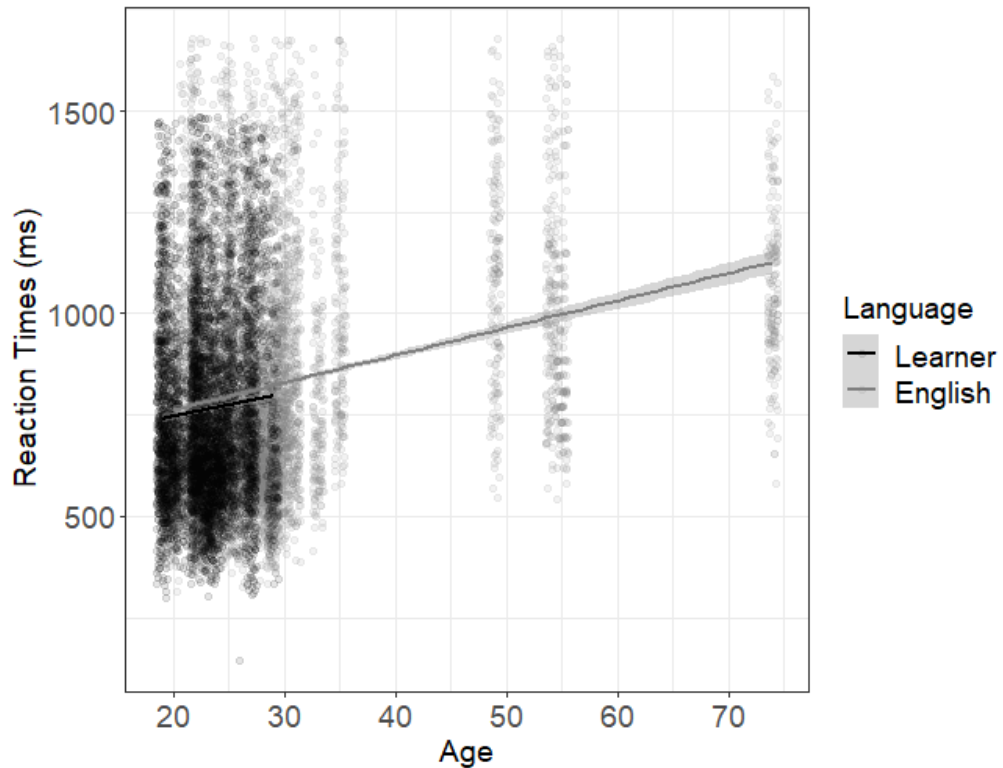
Note. The y-axis represents the frequency of counts. One L2 learner did not report its age.

Figure 27 depicts the relation between the reaction times and participants' age.

Although each of the slopes showed an upper trend, the slope of the native English speakers is steeper than that of the L2 learners. Thus, the difference in mean reaction times was likely due to the different distribution of the participants' ages.

Figure 27

The Relation of Native English Speakers and L2 Learners' Age and Reaction Times of the Semantic Stroop Task



Note. The y-axis represents the reaction times of the semantic Stroop task, while the x-axis represents the participants' age. The black line and the grey lines are regression lines. The grey areas represent 95% confidence intervals. One L2 learner did not report their age.

Modeling Results. Response times were analyzed with a series of linear mixed-effects models. The final model included log-transformed reaction time as a dependent variable; sentence typicality, word typicality, scaled vocabulary size, the interaction of the three, and position were the independent variables with covariates; the presentation order and the scaled reading time of each sentence were also included. Random effects included item intercept and subject intercept, as well as scaled reading time of each sentence for subjects without correlation parameters. The results of the model are shown in Table 24.

Table 24*Results of Mixed-Effects of the Japanese Learners of English*

Predictors	Estimates	Fixed Effects			Random Effects	
		<i>SE</i>	<i>t</i>	<i>p</i>	By Subject <i>SD</i>	By Item <i>SD</i>
Intercept	6.69	0.32	207.74	< .001	0.18	0.11
z RT Sentence	0.06	0.01	7.28	< .001	0.05	—
Pres Order	-0.00	0.00	-10.58	< .001	—	—
Sentence.Typicality 2-1	0.00	0.02	0.13	.901	—	—
Word. Typicality 2-1 (a)	-0.03	0.02	-1.27	.205	—	—
Word.Typicality 2-1 (b)	0.03	0.02	1.30	.194	—	—
Word.Typicaliy 3-2	0.05	0.02	2.57	.011	—	—
Position 2-1	0.02	0.02	1.18	.240	—	—
z VocabSize	0.01	0.03	0.26	.794	—	—
Sentence.Typicality 2-1*	-0.01	0.04	-0.29	.776	—	—
Word.Typicality 2-1 (a)						
Sentence.Typicality 2-1*	-0.03	0.04	-0.61	.545	—	—
Word.Typicality 2-1 (b)						
Sentence.Typicality 2-1*	-0.01	0.04	-0.32	.749	—	—
Word.Typicality 3-2						

Predictors	Estimates	Fixed Effects			Random Effects	
		<i>SE</i>	<i>t</i>	<i>p</i>	By Subject <i>SD</i>	By Item <i>SD</i>
Sentence.Typicality 2-1* z VocabSize	0.01	0.01	0.93	.351	—	—
Word.Typicality 2-1 (a)* z VocabSize	-0.02	0.01	-2.18	.029	—	—
Word.Typicality 2-1 (b)* z VocabSize	-0.00	0.01	-0.35	.724	—	—
Word.Typicality 3-2* z VocabSize	0.01	0.01	1.82	.069	—	—
Sentence.Typicality 2-1* Word.Typicality 2-1 (a)* z VocabSize	0.02	0.01	1.10	.274	—	—
Sentence.Typicality 2-1* Word.Typicality 2-1 (b)* z VocabSize	0.01	0.01	0.90	.370	—	—
Sentence.Typicality 2-1* Word.Typicality 3-2* z VocabSize	-0.00	0.01	-0.20	.844	—	—

Note. z RT Sentence: scaled reading time of each sentence; Pres Order: the order of presentation; Sentence.Typicality 2-1: typicality of sentences (atypical - typical); Word.Typicality 2-1 (a): typicality of word colors (typical - unrelated); Word.Typicality 2-1 (b): typicality of word colors (atypical - unrelated); Word.Typicality 3-2: typicality of word colors (atypical - typical); Position 2-1: position of context (post - pre); z VocabSize: scaled scores of the vocabulary size test. Model formula: $\log(\text{RT.Stroop}) \sim z \text{ RT Sentence} + \text{Pres Order} + \text{Position} + \text{Sentence.Typicality} * \text{Word.Typicality} * z \text{ VocabSize} + (1 + z \text{ RT Sentence} \parallel \text{SubjectID}) + (1 \parallel \text{ItemID})$

All variables had VIF scores below 5. The model revealed that L2 learners responded to atypical colors significantly slower than typical colors regardless of their L2 proficiency (word typicality 3-2: *Estimate* = 0.05, *SE* = 0.02, *t* = 2.57, *p* = .011, 95% CI [0.01, 0.09]). There was no significant difference in reaction time between typical color and unrelated color (word typicality 2-1: *Estimate* = -0.03, *SE* = 0.02, *t* = -1.27, *p* = .205, 95% CI [-0.07, 0.01]). Interestingly, however, there was a significant interaction between the typicality of word colors and vocabulary size. Specifically, the difference in reaction times between typical color and unrelated color interacted with L2 vocabulary size (word typicality 2-1 * scaled vocabulary size: *Estimate* = -0.02, *SE* = 0.01, *t* = -2.18, *p* = .029, 95% CI [-0.03, -0.00]). Figure 28 illustrates the interaction. As the figure shows, reaction time for typical colors decreased as the size of the vocabulary increased. In contrast, the reaction time for unrelated colors increased with increasing vocabulary size.

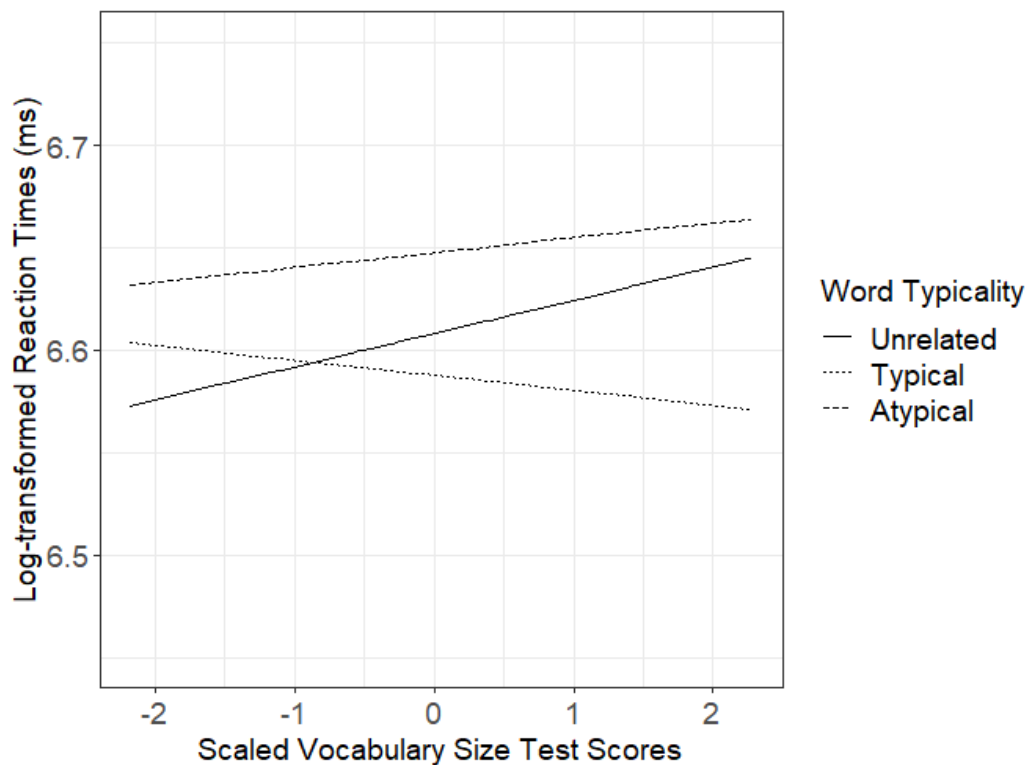
To compare the difference in reaction times between atypical color and unrelated color, the contrast coding of typicality of colors was changed (levels 2 - level 1: unrelated = - $\frac{2}{3}$, atypical = $\frac{1}{3}$, typical = $\frac{1}{3}$). The results revealed no significant difference in reaction times between atypical color and unrelated color (*Estimate* = 0.03, *SE* = 0.02, *t* = 1.30, *p* = .194, 95% CI [-0.01, 0.07]). Furthermore, there was no significant interaction between the typicality of word colors and the scaled vocabulary size test (*Estimate* = -0.00, *SE* = 0.01, *t* = -0.35, *p* = .724, 95% CI [-0.02, 0.01]). The results of the two models showed that readers tended to respond much faster to the typical color than to atypical or unrelated colors as their L2 proficiency increased. Moreover, it is not affected by the implicit colors of the sentence and whether the context was presented before or after the keywords.

The two covariates, scaled reading times of each sentence and presentation order showed significant main effects (scaled reading time: *Estimate* = 0.06, *SE* = 0.01, *t* = 7.28,

$p < .001$, 95% CI [0.05, 0.08]; presentation order: $Estimate = -0.00$, $SE = 0.00$, $t = -10.58$, $p < .001$, 95% CI [-0.00, -0.00]). The main effect of scaled reading time showed that reaction times became slower when they took time to read the sentences. In this task, there were no time constraints on reading the sentences. The significant effect of presentation order means that participants responded faster as the number of trials increased (Figure 29). The results of the two main effects indicate that participants took more time to respond to color when a sentence was difficult to understand or relatively unfamiliar to the task.

Figure 28

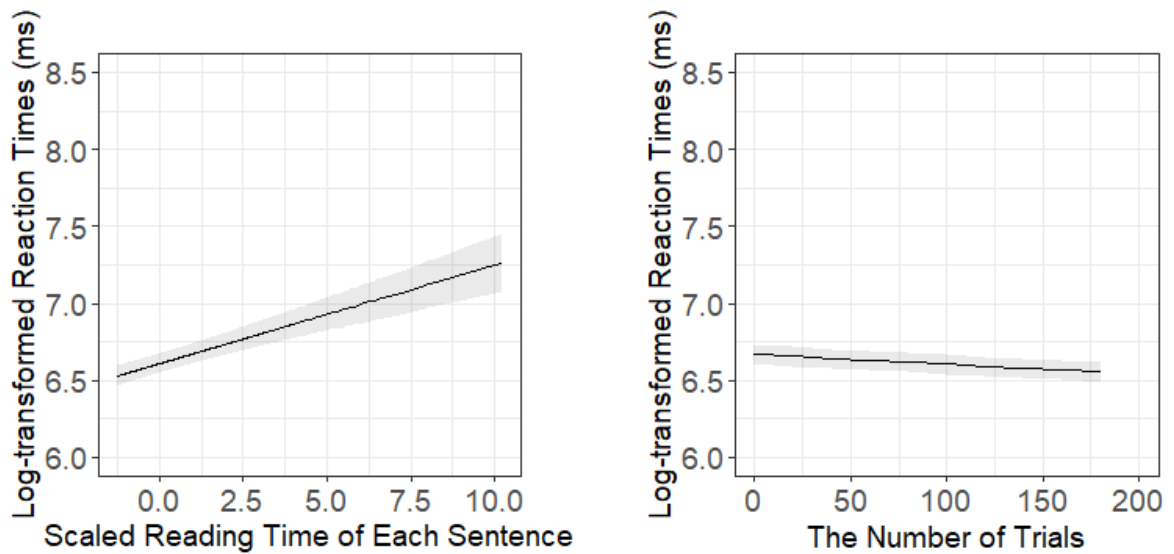
Effects of Word Typicality and Scaled Vocabulary Size Test Scores on the Reaction Times of the Semantic Stroop Task



Note. The y-axis represents the log-transformed reaction times of the semantic Stroop task, while the x-axis represents the scaled vocabulary size test scores.

Figure 29

The Scaled Reading Time of Each Sentence and Presentation Order Variable Included in the Final Model (Japanese Learners of English)



Note. The y-axis represents the log-transformed reaction times of the semantic Stroop task, while the x-axis represents the scaled reading time of each sentence and the number of trials that were up to 180. For both plots, the grey areas represent 95% confidence intervals.

The follow-up model that excluded nonsignificant independent variables was constructed (Table 25). The model included word typicality, scaled scores on the vocabulary size test, the interaction of word typicality and the scaled scores on the vocabulary size test, scaled sentence reading time, and presentation order as independent variables. The levels of word typicality were set as follows: unrelated as level 1, typical as level 2, and atypical as level 3.

Table 25*Results of Mixed-Effects of the Japanese Learners of English (Only Significant Variables)*

Predictors	Fixed Effects					Random Effects	
	Estimates	SE	<i>t</i>	<i>p</i>	By Subject <i>SD</i>	By Item <i>SD</i>	
Intercept	6.69	0.03	207.60	< .001	0.18	0.11	
z RT Sentence	0.06	0.01	7.30	< .001	0.05	–	
Pres Order	-0.00	0.00	-10.58	< .001	–	–	
Word.Typicality 2-1 (a)	-0.03	0.02	-1.27	.206	–	–	
Word.Typicality 2-1 (b)	0.03	0.02	1.30	.196	–	–	
Word.Typicaliy 3-2	0.05	0.02	2.57	.011	–	–	
z VocabSize	0.01	0.03	0.26	.796	–	–	
Word.Typicality 2-1 (a)* z VocabSize	-0.02	0.01	-2.18	.029	–	–	
Word.Typicality 2-1 (b)* z VocabSize	-0.00	0.01	-0.35	.726	–	–	
Word.Typicality 3-2* z VocabSize	0.01	0.01	1.82	.068	–	–	

Note. z RT Sentence: scaled reading time of each sentence; Pres Order: the order of presentation; Word.Typicality 2-1 (a): typicality of word colors (typical - unrelated); Word.Typicality 2-1 (b): typicality of word colors (atypical - unrelated); Word.Typicality 3-2: typicality of word colors (atypical - typical); z VocabSize: scaled scores of the vocabulary size test. Model formula: $\log(\text{RT.Stroop}) \sim z \text{ RT Sentence} + \text{Pres Order} + \text{Word.Typicality} * z \text{ VocabSize} + (1 + z.\text{RT.Sentence} \parallel \text{SubjectID}) + (1 \mid \text{ItemID})$

The results found that the main effects of word typicality 3-2 remained significant in the model (word typicality 3-2: *Estimate* = 0.05, *SE* = 0.02, *t* = 2.57, *p* = .011, 95% CI

[0.01, 0.09]). Furthermore, the interaction of word typicality 2-1 and the scaled scores of the vocabulary size test was significant ($Estimate = -0.02$, $SE = 0.01$, $t = -2.18$, $p = .029$, 95% CI [-0.03, -0.00]). Thus, the model without non-significant variables showed very similar results to the models with all independent variables. Further details of the analysis can be found in Appendix M.

Overall, a series of linear mixed-effects modeling showed that participants responded much faster to the typical color than to the atypical color, regardless of the size of the learners' L2 vocabulary. There was no significant difference between the atypical color and the unrelated color. However, as the learners' vocabulary size increased, the difference in reaction time between the unrelated color and the typical color increased significantly. The position of a context phrase did not affect these results, as we did not find a significant main effect of position.

On the surface, the results suggest that L2 learners simulate colors. However, this could be due to the specific influence of the color red (see below for details). With the exception of the color red, the data showed the following trend. When the color of an object was typical, participants tended to respond faster compared to other conditions. The study found a significant main effect of word typicality. That is, reaction times to typical color words were significantly faster than to atypical color words. However, there was no significant difference between typical and unrelated and atypical and unrelated colors. These results suggest that L2 learners do not simulate atypical colors of objects.

The color red showed some distinct patterns. Table 26 summarizes the items for which the reaction times of the typical color words were faster than those of the atypical color words by more than 100 milliseconds.

Table 26*Reaction Times of the Semantic Stroop Task for Individual Words*

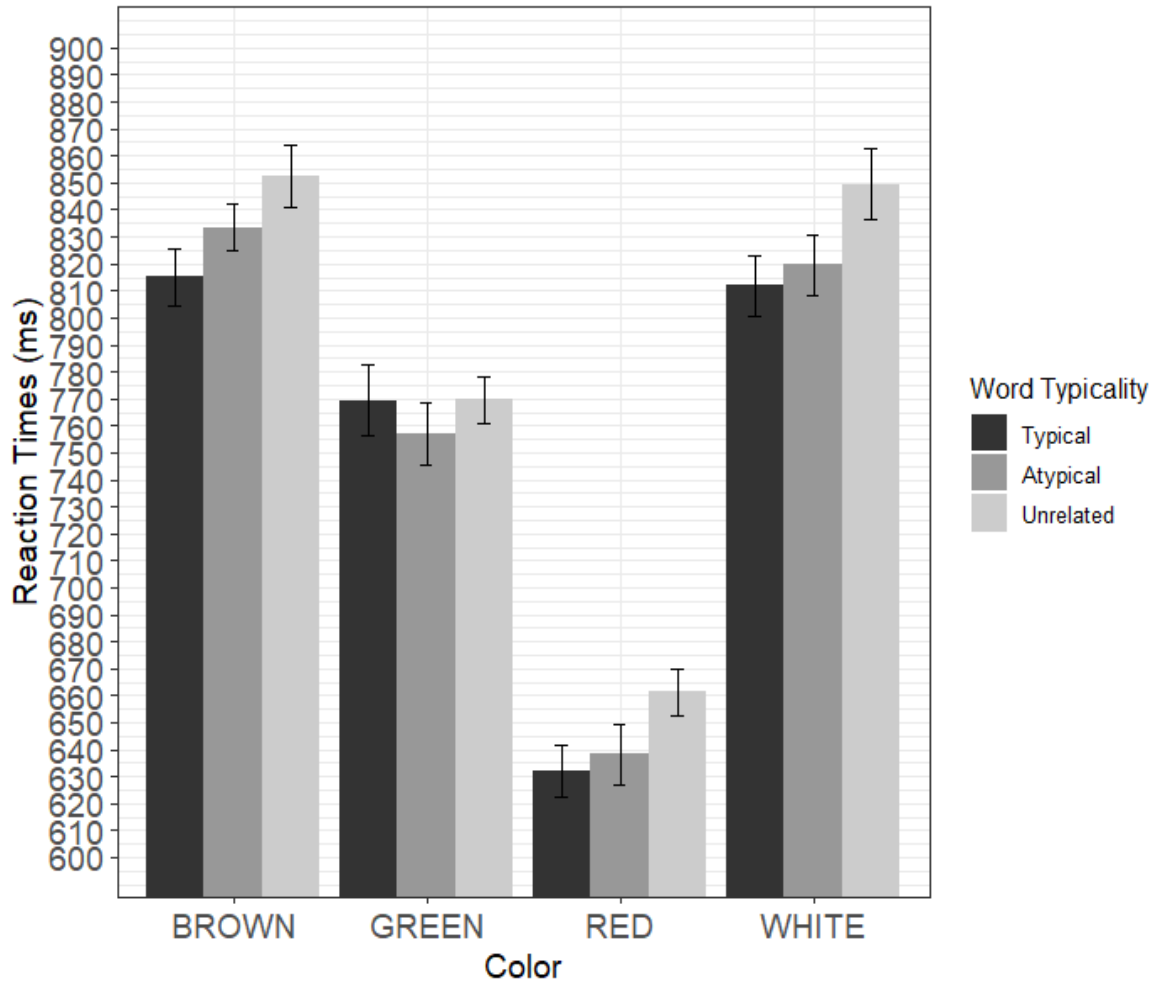
	<i>M</i> (Typical)	<i>M</i> (Atypical)	Differences
apple	617.15	840.71	-223.56
tomato	606.64	732.19	-125.55
strawberry	650.36	773.09	-122.73
plum	654.27	765.92	-111.65
kiwi	744.15	843.34	-99.19

Note. *M* (Typical) represents the mean reaction times of the typical color of the objects, and *M* (Atypical) represents the mean reaction times of the atypical color of the objects. *Differences* were calculated from *M* (Typical) - *M* (Atypical). The order of words was arranged in ascending order of *Differences*.

Except for *kiwi*, the typical colors of these words are all red. Figure 30 shows the reaction times according to colors, showing that participants responded to red much faster than any other colors, regardless of typicality. This trend was similar in the data on filler words as well (Figure 31).

Figure 30

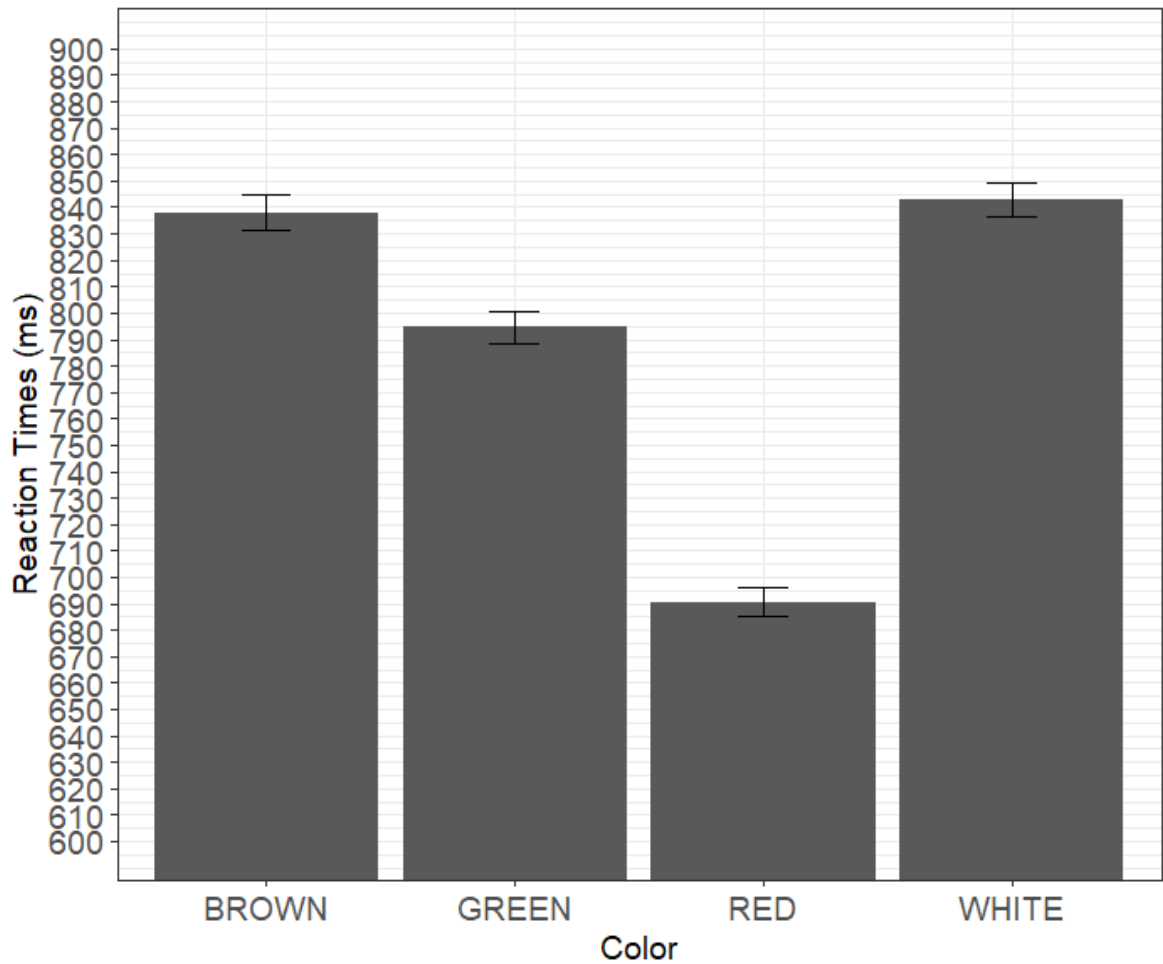
Reaction Times of the Semantic Stroop Task in Each Color (Critical Items)



Note. The data was after the data treatment (see *Analysis* in this chapter). Error bars represent standard error.

Figure 31

Reaction Times in Each Color (Filler Items)



Note. The data was after the data treatment (see Analysis in this chapter). Error bars represent standard error.

The items with a difference in the reaction times between typical and atypical conditions by more than 100 ms (*apple, tomato, strawberry, and plum*) showed another interesting trend: they received higher typicality rating scores in their typical colors (Table 27).

Table 27

Mean and Standard Deviations of Scores of Word Typicality Rating Task of Each Word

	Typical Color	<i>M</i>	<i>SD</i>
apple	Red	6.00	0.00
strawberry	Red	6.00	0.00
tomato	Red	5.94	0.23
plum	Red	4.69	1.56

Apple, *strawberry*, and *tomato* showed the highest rating scores with almost near-zero standard deviations (i.e., regardless of L2 proficiency). Taken together, it can be inferred that L2 learners simulate typical colors even if their L2 proficiency is lower when an object represents typical colors and the color is red.

The large differences between typical and atypical color words in these items may be why there was a significant main effect of word typicality (typical - atypical). Thus, with the exception of the color red, we cannot assume that learners made a direct connection from L2 forms to their concepts of typical colors.

The results show that participants do not simulate the atypical color of the object even when they read atypical sentences. Rather, they always simulate the typical color of the object for typical and atypical sentences. No significant main effect of context position was found in the study. Similar to the results for native Japanese and native English speakers, simulation may not be affected by context position.

Finally, the results of a significant interaction of word typicality and L2 vocabulary size demonstrate that response tendencies with higher L2 proficiency become similar to the results of the L1 task. As Figure 28 shows, the difference between the typical color and the unrelated color becomes more accentuated as the learners' L2 vocabulary size increases.

This trend suggests that higher L2 proficiency leads to richer mental simulation of colors during processing of L2 sentences in general.

Summary of Experiment 2

The results supported research hypothesis 4, which posited that when learners' L2 proficiency is not considered, there are not differences in reaction times between match and mismatch conditions, except for items whose typical color is red. Research hypothesis 5 predicted that when learners' L2 proficiency is not considered, colors implied by sentences do not affect reaction times. This research hypothesis was supported. Research hypothesis 6 posited that when learners' L2 proficiency is not considered, the position of the context phrases does not influence reaction times. This research hypothesis was also supported. Research hypothesis 7 posited that L2 proficiency affects the degree of simulation of objects' color and that research hypotheses 1–3 will hold true for higher proficiency L2 learners. The results indicated that only research hypothesis 1 (the color is simulated) was supported for higher-proficiency L2 learners. However, Experiment 1 did not support research hypotheses 2 and 3. Thus, research hypothesis 7 was supported in the sense that higher L2 proficiency leads to a pattern similar to the L1 results.

Chapter 6: General Discussion

Summary of the Results

Experiment 1

Experiment 1 was conducted to examine whether readers simulate the colors of objects while reading an L1 sentence. Native English and Japanese speakers performed the semantic Stroop task in their L1. In both groups, they responded significantly faster to typical color words (e.g., *bear* in brown) than atypical (e.g., *bear* in white) and unrelated color words (e.g., *bear* in green). The results are partially consistent with the findings of Connell and Lynott (2009). They found that native English speakers showed the fastest reaction time to typical color words among atypical and unrelated color words. Further, they argued that readers responded faster to atypical words (e.g., *bear* in white) after reading atypical sentences (e.g., *Joe was excited to see a bear at the North Pole*) than reading typical sentences (e.g., *Joe was excited to see a bear in the woods*). However, this study did not confirm this result. Moreover, the tendency did not depend on the position of the context.

Experiment 2

Experiment 2 aimed to investigate whether readers simulate the colors of objects when reading L2 sentences. In addition, Experiment 2 investigated whether L2 proficiency influenced the simulation process. Japanese learners of English performed the semantic Stroop task in English. They completed the L2 vocabulary size test (Meara & Miralpeix, 2016) to measure their L2 proficiency. The results showed a significant interaction between word typicality and vocabulary size: the difference between typical color words and unrelated color words in reaction time increases as learners' L2 proficiency increases. This means that higher proficiency learners respond faster to typical color words than to

unrelated words (i.e., they simulate colors). Interestingly, L2 learners responded significantly faster to the color red, regardless of their L2 proficiency. Jiang (2000) suggested that the development of lexical items varies from word to word. Thus, even for learners with lower language proficiency, words whose referents were red might have developed further than other words.

Simulation of Object Colors

Color Simulation in L1 Processing

The study found that L1 readers simulate the typical color of objects when processing vocabulary. The results are consistent with a previous study that found simulation of color using a semantic Stroop task (Connell & Lynott, 2009) and an SPV task (e.g., de Koning et al., 2017; Hoeben Mannaert et al., 2017; Zwaan & Pecher, 2012).

The study does not support the atypical color simulation results of Connell and Lynott (2009). The interaction of sentence typicality and word typicality did not reach significance. As mentioned earlier, Connell and Lynott (2009) argued that mean reaction times of atypical color became faster under atypical color conditions than under typical color conditions. However, there was actually no significant interaction ($p = .057$) between sentence typicality and word color typicality in their study. With a normal interpretation of the p -value, Connell and Lynott's (2009) results can be considered an insignificant interaction between sentence typicality and word color typicality. If this is the case, this study has the same result as their study.

The study considered the influence of context position to consider the multiple-color simulation that was found in Connell and Lynott (2009). According to previous studies (Sato et al., 2013), readers simulate typical and atypical sentences when the keyword is presented before the context. However, readers may simulate only what the

sentence implies if a context is presented before the keyword. However, the results of this study suggest that the context position does not affect the simulation. That is, L1 readers did not simulate the atypical color of the objects (*bear in white*) neither when the following information indicated the atypical color (post-context) nor when it was presented before the keywords (pre-context).

The results contrast with Connell and Lynott (2009). They also disagree with the studies that reported that L1 readers simulate visual aspects of the objects and update the image with the following information (e.g., Sato et al., 2013; Horchak and Garrido, 2021; Kang et al., 2020). The different results might be due to the difference in the change implied in a sentence between this study and previous studies. The object's state changed substantially in the experimental sentences in the previous studies, for example, "*The woman dropped the ice cream.*" In contrast, it did not significantly change in the current study (e.g., *Joe was excited to see a bear at the North Pole*). This difference in the degree of change in the target object could have led to different results. It is possible that the state changes of the object in this study material are not strong enough to activate atypical information. In addition, the lack of evidence that readers simulate atypical colors, even when the pre-context conditions are atypical, suggests the robustness of simulating typical colors.

Color Simulation in L2 Processing

Experiment 2 showed that L2 learners could simulate typical color words with increasing L2 proficiency. The results suggest that L2 learners, in this case those who learned the L2 in the context of English as a foreign language, can simulate colors of objects using verbal information as they improve their L2 proficiency. The results suggest that the more the lexical items in the L2 mental lexicon develop, the more the mental status

of the L2 vocabulary approaches the L1 pattern. The lack of significant effects of context is consistent with the results for the L1.

The typical color simulation in L2 is robust in higher-level learners. The result is in line with the L2 studies that found simulation of visual aspects (e.g., Ahn & Jiang, 2018; Vukovic & Williams, 2014) and other aspects of embodied knowledge such as a motor (e.g., Buccino et al., 2017; Dudschig et al., 2014). However, studies that explored L2 simulation of sensorimotor have suggested that not only high-proficiency learners but also lower-proficiency learners simulate non-linguistic information (e.g., Awazu & Suzuki, 2020; Kogan et al., 2020). This result may seem to contradict the present finding, but the discrepancy can be explained as follows. The results of the present study suggest that simulation may vary from word to word. As previously reported, the color red showed a different result than other colors, suggesting that even learners with lower L2 proficiency simulate objects whose color is red. Thus, different aspects of embodied knowledge could lead to different levels of simulation. This view could be supported by studies that found that even learners with higher levels of knowledge did not simulate some visual aspects of knowledge (e.g., Chen et al., 2020; Norman & Peleg, 2021). Since color typicality varies from word to word, the object's color could contribute to the degree of embodiment. However, this question needs further research. The influence of color is discussed further in the Limitations and Directions for Future Research section.

It has been controversial whether embodied knowledge is present in the processing of lately-acquired languages (e.g., Kogan et al., 2020; Monaco et al., 2019). This is because L2 is usually learned explicitly in a classroom context without the inclusion of embodied knowledge (e.g., Chen et al., 2020; Monaco et al., 2019), especially in English as a foreign language. In some studies, simulation has not been observed in higher-level

learners (e.g., Norman & Peleg, 2021). However, the present study argues that the foreign language context does not necessarily lead to disembodiment of language processing. Most previous studies have not included L2 proficiency in statistical analyses to examine their effects on the degree of L2 embodiment (e.g., Ahlberg et al., 2018; Buccino et al., 2017; Vukovic and Williams, 2014). In the present study, this analysis was conducted using objective scores for vocabulary size and found that an increase in L2 proficiency enables EFL learners to simulate anchored embodied knowledge in the L2. This suggests that learning context may not be a significant predictor of embodiment. In addition, the EFL participants in the study started learning English at the mean age of 11.31 ($SD = 2.56$). This also implies that infant language exposure might not be necessary for embodiment.

Overall, the study suggests that L2 learners can simulate the typical color of objects as their L2 proficiency increases. For some objects, even learners with lower L2 proficiency simulate the typical color of the objects during vocabulary processing. In the next section, the author presents the implications of the results for L2 vocabulary research.

The Representation of L2 Mental Lexicon

The present study suggests that L2 processing involves both linguistic and non-linguistic processing. As L2 proficiency increases, the relationship changes not only between forms and conceptual representation but also in what L2 learners understand during vocabulary processing. Moreover, the non-linguistic features of the word itself influence the development of L2 vocabulary representation and processing (e.g., the color red). These findings have implications for models of L2 vocabulary processing. As reviewed earlier (see Chapter 2), existing models of L2 vocabulary processing, such as the Revised Hierarchical Model and the Three-Stage Model, cannot fully explain the results. Although these models assume the existence of a conceptual representation, the specific

nature of the concept has not been considered. Furthermore, these models assume that as L2 proficiency increases, the relationship between lexical stores and concepts changes; however, the extent to which this change affects L2 learners' comprehension (or mental representation) of linguistic stimuli is not elaborated. The current study has shown that establishing a direct link between L2 form and conceptual representation involves the activation of embodied knowledge (color in this case). As in the case of L1 processing (e.g., Barsalou et al., 2008; Connell & Lynott, 2009; Hoeben Mannaert et al., 2017; Zwaan et al., 2002; Zwaan & Pecher, 2012; as well as Experiment 1 in this study), improving L2 proficiency allows L2 learners to develop a comprehensive understanding of words.

The results of the current study suggest that studies of L2 vocabulary need to account for the use of non-linguistic information in order to fully understand the mechanisms of L2 vocabulary processing and representation. Research on embodied cognition is one of the approaches that complement L2 vocabulary studies. The mixture of the two research paradigms has already appeared. Studies on embodied cognition in the L2 use the L2 vocabulary model as a theoretical background (e.g., Chen et al., 2020). Pavlenko (2009) proposed a modified version of the Revised Hierarchical Model (Kroll & Stewart, 1994). Her Modified Revised Hierarchical Model retains the assumptions of the Revised Hierarchical Model, such as the separate lexical stores of L1 and L2, and the developmental aspect, but assumes a different conceptual representation. It has three concepts: L1-specific, L2-specific, and shared concepts. This more specific conceptual assumption is consistent with the findings of bilingual research reporting on culturally specific conceptual representation (e.g., Jared et al., 2013). Pavlenko (2009) highlighted the importance of the embodied cognition paradigm for L2 vocabulary processing.

Limitations and Directions for Future Research

This study has three major limitations: (1) the impact of word frequency and word length was not considered in the selection of experimental items, (2) the influence of specific colors was not considered, and (3) the difference between an object's surface color and its inside color was not taken into account. First, the selection of items did not take into account a characteristic of each word, such as word frequency and word length. This is because (1) most of the experimental words were high-frequency words, such as *apple*, *strawberry*, *bear*, and *ice cream* (Table 28), and (2) Connell and Lynott (2009) did not control the word length of words. In addition, item construction would be more difficult if word length were controlled, as there are several criteria to consider when constructing items (see Pilot Study 2, Experimental Items section). Nevertheless, to address these limitations, a post-hoc analysis was conducted.

To account for frequency's influence on simulation, the relationship between frequency and reaction times was examined in both L1 and L2. Thus, correlation analyses were performed. The frequencies of English words were extracted from the Corpus of Contemporary American English (COCA) (Davies, 2008-), which contains more than one billion words from spoken sources, fiction, popular magazines, newspapers, academic texts, TV, movie subtitles, blogs, and other web pages. The frequencies of Japanese words were extracted from the Balanced Corpus of Contemporary Written Japanese (short-unit) (BCCWJ) (National Institute for Japanese Language and Linguistics, Center for Corpus Development, 2021), which contains more than 100 million words from books, magazines, newspapers, government white papers, an Internet bulletin board, blogs, school textbooks, national state legislature minutes, local government promotional letters, laws, and poetry (Maekawa et al., 2014) (Table 28).

Table 28*Frequencies of the Experimental Items*

	Frequency in English (COCA)		Frequency in Japanese (BCCWJ)
ball	90,520	ボール	7,084
apple	61,123	リンゴ	2,408
bear	53,333	クマ	1,824
horse	44,525	馬	6,354
cloud	28,089	雲	3,944
cake	27,942	ケーキ	2,926
onion	13,694	たまねぎ	2,087
leaf	12,897	葉っぱ	663
tomato	10,932	トマト	2,570
steak	9,196	ステーキ	550
popcorn	5,484	ポップコーン	89
strawberry	4,740	イチゴ	1,221
plum	3,619	梅	1,352
chameleon	1,043	カメレオン	61
kiwi	940	キウイ	175

Note. The order of items was arranged in descending order of frequencies in English.

Because the frequencies of the individual items did not follow the normal distribution, Spearman's rank correlation coefficient was calculated. The analyzes revealed no significant correlation between the variables (Table 29). Therefore, the word frequency does not seem to influence the result.

Table 29

Spearman's Rank Correlation Coefficients and 95% Confidence Interval Between Frequency and the Reaction Times of the Semantic Stroop Task

	<i>r_s</i>	95% Confidence Interval	
		Lower	Upper
Native English	-.00	-.03	.03
Native Japanese	.01	-.02	.03
English Learner	.01	-.01	.04

Next, the correlation between the length of each word (i.e., the number of letters) and reaction times were computed. The number of letters varies from four (e.g., *ball*, *kiwi*) to ten (*strawberry*) in English (Table 30). Spearman's rank correlation coefficient was computed because the word length did not follow a normal distribution. Although native English speakers' reaction times showed a significant correlation with word length ($r_s = -.03$, 95% CI [-.05, .00], $p = .03$), the correlation was small and could be considered negligible. Since the correlation was rather low, word length did not affect the result. Although the post-hoc analysis revealed only a negligible effect on word frequency and length, it would be interesting to examine the influence of these two factors in future studies.

Table 30*The Number of Letters in English and Japanese Experimental Items*

English	Letters	Japanese	Letters
strawberry	10	イチゴ	3
chameleon	9	カメレオン	5
popcorn	7	ポップコーン	6
tomato	6	トマト	3
apple	5	リンゴ	3
cloud	5	雲	1
horse	5	馬	1
onion	5	たまねぎ	4
steak	5	ステーキ	4
ball	4	ボール	3
bear	4	クマ	2
cake	4	ケーキ	3
kiwi	4	キウイ	3
leaf	4	葉っぱ	3
plum	4	梅	1

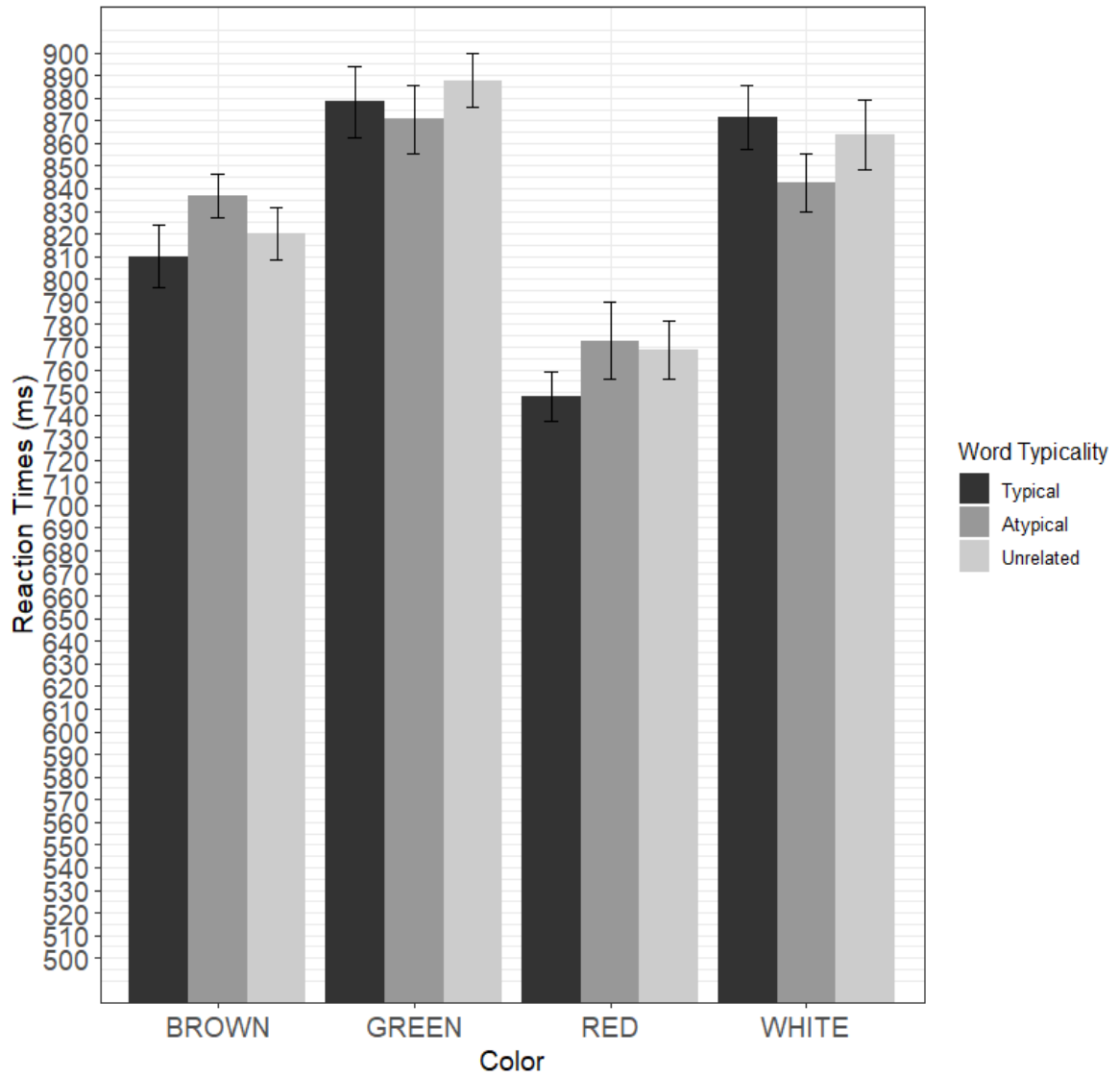
Note: The order of items was arranged in descending order of word length in English. *English* and *Japanese* columns are the translation pairs.

Another limitation is that the study did not consider the influence of specific colors. Brown, green, white, and red were used in the study. The reason for choosing these four colors was that it was easier to create experimental materials with typical, atypical, and unrelated colors for an object. The possible effects of other colors were not considered, as this was not the main objective of the study. As found in Experiment 2 with L2 learners,

the color red could have some distinct influence in cognitive processing. To investigate this issue, a post-hoc analysis was conducted for native English (Figure 32) and native Japanese speakers (Figure 33). Both figures describe the mean reaction times of the semantic Stroop task across colors and the typicality of the colors of objects for critical items. As with L2 learners, native speakers responded faster to the color red than to other colors, regardless of the typicality of the object color. Combining these results with those of L2 learners (Figure 30 in Chapter 5), all results show that the color red elicits significantly faster reaction times, regardless of language (L1 and L2) and object color typicality.

Figure 32

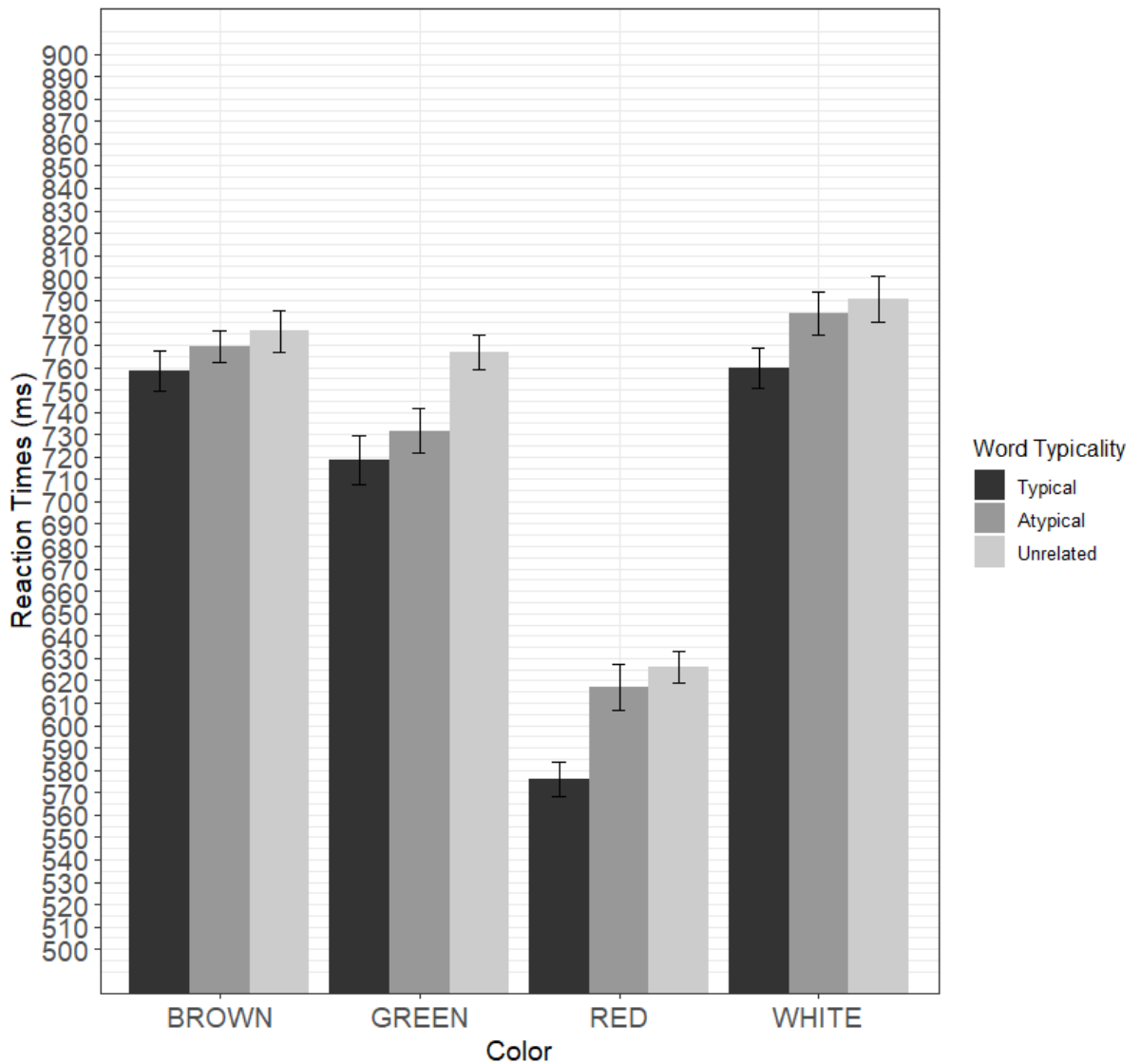
Reaction Times of the Semantic Stroop Task in Each Color (Native English Speakers: Critical Items)



Note. The data was after the data treatment (see *Analysis* in Chapter 4). Error bars represent standard error.

Figure 33

Reaction Times of the Semantic Stroop Task in Each Color (Native Japanese Speakers: Critical Items)



Note. The data was after the data treatment (see *Analysis* in Chapter 4). Error bars represent standard error.

Additionally, reaction times for filler items were analyzed for native English speakers and native Japanese speakers. The combination of colors and words was irrelevant for filler items (e.g., *banana-RED*, *butter-BROWN*); therefore, the author was

able to examine the influence of color, excluding the influence of word typicality. The analysis of filler items also showed that L1 and L2 readers responded much faster to red than to other colors (Table 31, Table 32, Table 33, and Figure 31). One might think that reaction times for red were fastest because the color red was assigned to keys typed with the right hand (note that the majority of participants were right-handed: native English speakers = 32 participants, native Japanese speakers = 33 participants, English learners = 34 participants). This is unlikely, however, because reaction times for the color green, which was also assigned to the keys typed with the right hand, were much slower than those for the color red.

Table 31

Descriptive Statistics of the Reaction Times of the Semantic Stroop Task (Native English Speakers: Filler Items)

	<i>n</i>	<i>M</i>	<i>SD</i>	<i>Median</i>	<i>Minimum</i>	<i>Maximum</i>
Red	1,629	789.72	269.82	718	350	1,705
Brown	1,205	836.90	281.12	769	383	1,706
White	1,416	868.83	300.11	803	348	1,708
Green	1,403	890.09	285.50	829	365	1,711

Note: *n* = the number of observations. The items whose reaction times that exceed \pm three median absolute deviations from median were excluded. The order of the colors was arranged in ascending order of mean reaction times.

Table 32

Descriptive Statistics of the Reaction Times of the Semantic Stroop Task (Native Japanese Speakers: Filler Items)

	<i>n</i>	<i>M</i>	<i>SD</i>	<i>Median</i>	<i>Minimum</i>	<i>Maximum</i>
Red	1,747	634.71	175.36	598.0	343	1,358
Green	1,464	777.26	216.59	743.5	340	1,361
Brown	1,266	779.58	203.47	731.5	365	1,367
White	1,458	790.59	220.68	744.5	349	1,368

Note: *n* = the number of observations. The items whose reaction times that exceed \pm three median absolute deviations from median were excluded. The order of the colors was arranged in ascending order of mean reaction times.

Table 33

Descriptive Statistics of the Reaction Times of the Semantic Stroop Task (Japanese Learners of English: Filler Items)

	<i>n</i>	<i>M</i>	<i>SD</i>	<i>Median</i>	<i>Minimum</i>	<i>Maximum</i>
Red	1,735	690.51	223.47	633	315	1,506
Green	1,523	794.49	245.58	742	338	1,509
Brown	1,247	837.95	236.80	796	374	1,506
White	1,431	842.87	253.02	786	342	1,503

Note: *n* = the number of observations. The items whose reaction times that exceed \pm three median absolute deviations from median were excluded. The order of the colors was arranged in ascending order of mean reaction times.

Thus, both L1 and L2 speakers demonstrated the effect of the color red. However, it is beyond the scope of this study to investigate this issue further. The effects of the different colors remain a question for future studies.

The third limitation is that the study did not consider the difference in color typicality between objects' surface and inside colors. Three of the 15 items were compared vis-à-vis the typicality of objects' surface and inside colors. For example, the typical color of *tomato* was implied with the sentence "*Jane ate the tomato because it was ready to eat*" (a ripe tomato), and atypical color was implied with the sentence "*Jane ate the tomato before it was ready to eat*" (an unripe tomato). Both were the surface colors of the tomato (red and green). On the other hand, the typical *kiwi* color was implied with the sentence "*Roy found a kiwi at the bottom of the parfait*" (without peels), and atypical color was implied with the sentence "*Roy found a kiwi at the bottom of the basket*" (with peels). The typical color was the interior color of a kiwi (green), but the atypical color was the surface color of a *kiwi* (brown). Future research could take this problem into account when creating experimental materials.

Chapter 7: Conclusions

The study investigated whether readers mentally represent the color of an object in L1 and L2. Models of the L2 mental lexicon showed how L1 and L2 words relate to their concepts and how an increase in L2 proficiency affects the representation. However, these models cannot fully explain what happens after concepts are accessed. Some models, such as the Modified Revised Hierarchical Model (Pavlenko, 2009), take into account the more specific nature of the concept. This model provides a better explanation of the cross-linguistic differences between concepts, but does not provide accurate information about the nature of the concept within a language. This limitation can be complemented by the embodied cognition account (e.g., Barsalou, 1999; Barsalou et al., 2008), which has been used to study mental representations during language processing. In the present study, the author used the representations and methods of embodied cognition to uncover non-linguistic information during L2 vocabulary processing.

In the study, a psycholinguistic experiment was conducted by administering a semantic Stroop task to 35 native English speakers, 36 native Japanese speakers, and 36 Japanese learners of English. L1 readers simulated the typical color of an object (e.g., a brown *bear*) but not the atypical color. Even when a sentence implied an atypical color for an object (e.g., *bear at the North Pole*), the readers always simulated the typical color of an object. This was also true when a context (e.g., *in the woods/at the North Pole*) was placed before the keywords (e.g., *bear*). L2 readers also simulated a typical color of an object at higher proficiency levels. The simulation pattern resembled that of L1 readers in that they always simulated a typical color of an object, regardless of the color implied by the sentence and position in context. Interestingly, even learners with lower language proficiency simulated an object color when the typical color was red. This suggests that

non-linguistic information (e.g., color) influences lexical representation in the L2 and that representation may be different for different words.

The results support the embodied cognition approach. Moreover, they offer new insights into research in this area. Although the visual aspects of the object are simulated in L1, readers may update their simulated mental representation only when significant state changes have been implied in the sentence (e.g., *drop an ice cream*). The simple change in color of an object (e.g., brown bear - white bear) may not be sufficient to trigger the simulation update. In L2, the current results have shown that L2 simulation is possible if readers improve their L2 proficiency to the point in which they can make more direct mapping between an L2 form and its concept (stages 2 and 3 in Jiang [2000]).

The results also have implications for L2 vocabulary research. The color simulation results suggest that L2 vocabulary processing involves not only linguistic but also non-linguistic information. Models of L2 word processing that incorporate developmental aspects, such as the Three-Stage Model (Jiang, 2000) and the Revised Hierarchical Model (Kroll & Stewart, 1994), assume that the relationship between L1 word, L2 word, and their concept changes as the learner's L2 proficiency increases. However, these models did not fully explain the difference between what learners understand before and after the increase in L2 proficiency. The results suggest that an increase in L2 proficiency enables learners to make a direct connection between an L2 form and its embodied concept; therefore, learners can represent a richer mental image than they do in the L1 at higher proficiency levels. Moreover, words whose most typical color is red might develop earlier than others. All these results imply that models of L2 word processing must incorporate the use of non-linguistic information. Therefore, the embodied cognition paradigm will complement studies of L2 word representation and processing.

Future studies on color simulation need to consider the frequency of words, the effects of colors, and the difference in color typicality between objects' surface and inside colors. In the study, these factors were not considered because the influence of these factors was not the main goal of the research. Although strong influences of these factors were not suggested in the post-hoc analyses, future studies should consider these factors.

Finally, simulating the visual aspects of an object has not been an important research topic in studies of L2 vocabulary representation and processing. However, L2 learners are not machines programmed with artificial symbols. L2 learners (and, of course, L1 language users) use not only the linguistic information they have learned, but also their experiences with their physical senses, such as seeing, touching, tasting, hearing, and smelling. For example, reading about *a fresh lemon* activates the representation of a yellow object in the readers' minds and the sourness of the lemon. This information is crucial to comprehend what a word means to humans. One of the most important features that distinguish humans from machines is the use of non-linguistic information. The author hopes the study will encourage more L2 linguistic studies to consider these "non-linguistic" aspects of language processing to reveal L2 word representation and processing.

References

- Abrahamse, E. L., Duthoo, W., Notebaert, W., & Risko, E. F. (2013). Attention modulation by proportion congruency: The asymmetrical list shifting effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(5), 1552–1562.
- Ahlberg, D. K., Bischoff, H., Kaup, B., Bryant, D., & Strozyk, J. V. (2018). Grounded cognition: Comparing language× space interactions in first language and second language. *Applied Psycholinguistics*, *39*(2), 437–459.
- Ahn, S., & Jiang, N. (2018). Automatic semantic integration during L2 sentential reading. *Bilingualism: Language and Cognition*, *21*(2), 375–383.
- Augustinova, M., Parris, B. A., & Ferrand, L. (2019). The loci of Stroop interference and facilitation effects with manual and vocal responses. *Frontiers in Psychology*, *10*, 1786.
- Awazu, S., & Suzuki, A. (2020). Daini gengo teijukutatsusha ni yoru daini gengo bun rikai no sintaisei [Embodiment of L2 sentence comprehension by L2 beginners]. *Cognitive Studies: Bulletin of the Japanese Cognitive Science Society*. Advance online publication.
- Bai, B., Yang, C., & Fan, J. (2022). Semantic integration of multidimensional perceptual information in L1 sentence comprehension. *Language and Cognition*, *14*(1), 109–130.
- Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577–660.
- Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617–645.
- Barsalou, L. W., Dutriaux, L., & Scheepers, C. (2018). Moving beyond the distinction between concrete and abstract concepts. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1752), 20170144.

- Barsalou, L. W., Santos, A., Simmons, W. K., & Wilson, C. D. (2008). Language and simulation in conceptual processing. In M. De Vega, A. M. Glenberg, & A. C. A. Graesser (Eds.), *Symbols, Embodiment, and Meaning* (pp. 245–283). Oxford University Press.
- Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. <https://arxiv.org/abs/1506.04967v2>
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2021). lme4: Linear mixed-effects models using Eigen and S4 (Version 1.1-27.1) [Computer software]. <https://cran.r-project.org/web/packages/lme4/index.html>
- Bergen, B. K., Lindsay, S., Matlock, T., & Narayanan, S. (2007). Spatial and linguistic aspects of visual imagery in sentence comprehension. *Cognitive Science*, *31*(5), 733–764.
- Brysbaert, M., & Duyck, W. (2010). Is it time to leave behind the Revised Hierarchical Model of bilingual language processing after fifteen years of service? *Bilingualism: Language and cognition*, *13*(3), 359–371.
- Brysbaert, M., Verreyt, N., & Duyck, W. (2010). Models as hypothesis generators and models as roadmaps. *Bilingualism: Language and Cognition*, *13*(3), 383–384.
- Buccino, G., Marino, B. F., Bulgarelli, C., & Mezzadri, M. (2017). Fluent speakers of a second language process graspable nouns expressed in L2 like in their native language. *Frontiers in Psychology*, *8*, 1306.
- Chen, D., Wang, R., Zhang, J., & Liu, C. (2020). Perceptual Representations in L1, L2 and L3 Comprehension: Delayed Sentence–Picture Verification. *Journal of Psycholinguistic Research*, *49*(1), 41–57.

- Chuang, Y. Y., Bell, M. J., Banke, I., & Baayen, R. H. (2021). Bilingual and multilingual mental lexicon: a modeling study with linear discriminative learning. *Language Learning, 71*(S1), 219–292.
- Connell, L. (2007). Representing object colour in language comprehension. *Cognition, 102*(3), 476–485.
- Connell, L., & Lynott, D. (2009). Is a bear white in the woods? Parallel representation of implied object color during language comprehension. *Psychonomic Bulletin & Review, 16*(3), 573–577.
- Connell, L., & Lynott, D. (2011). Modality switching costs emerge in concept creation as well as retrieval. *Cognitive Science, 35*(4), 763–778.
- Connell, L., & Lynott, D. (2012). Strength of perceptual experience predicts word processing performance better than concreteness or imageability. *Cognition, 125*(3), 452–465.
- Davies, M. (2008-). *The Corpus of Contemporary American English (COCA): 520 Million Words, 1990–Present*. Retrieved June 17, 2022, from <https://www.english-corpora.org/coca/>
- de Groot, A. M. (1992). Determinants of word translation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(5), 1001–1018.
- de Koning, B. B., Wassenburg, S. I., Bos, L. T., & van der Schoot, M. (2017). Mental simulation of four visual object properties: similarities and differences as assessed by the sentence–picture verification task. *Journal of Cognitive Psychology, 29*(4), 420–432.
- Delignette-Muller, ML, Dutang C (2015). fitdistrplus: An R Package for Fitting Distributions. *Journal of Statistical Software, 64*(4), 1–34.
<https://www.jstatsoft.org/article/view/v064i04>

- Dijkstra, T., & Van Heuven, W. J. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition*, 5(3), 175–197.
- Dijkstra, T., Wahl, A., Buytenhuijs, F., Van Halem, N., Al-Jibouri, Z., De Korte, M., & Rekké, S. (2019). Multilink: A computational model for bilingual word recognition and word translation. *Bilingualism: Language and Cognition*, 22(4), 657–679.
- Dudschig, C., de la Vega, I., & Kaup, B. (2014). Embodiment and second-language: Automatic activation of motor responses during processing spatially associated L2 words and emotion L2 words in a vertical Stroop paradigm. *Brain and Language*, 132, 14–21.
- Fodor, J. A. (1979). *The language of thought*. Harvard university press. (Original work published 1975).
- Garofalo, G., & Riggio, L. (2022). Influence of colour on object motor representation. *Neuropsychologia*, 164, 108103.
- Glenberg, A. M., & Gallese, V. (2012). Action-based language: A theory of language acquisition, comprehension, and production. *Cortex*, 48(7), 905–922.
- Glenberg, A. M., & Robertson, D. A. (1999). Indexical understanding of instructions. *Discourse Processes*, 28(1), 1–26.
- Hardwick, R. M., Forrence, A. D., Costello, M. G., Zachowski, K., & Haith, A. M. (2021). Age-related increases in reaction time result from slower preparation, not delayed initiation. *Journal of Neurophysiology*, 128, 582–592.
- Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1–3), 335–346.

- Hayakawa, S., & Keysar, B. (2018). Using a foreign language reduces mental imagery. *Cognition*, *173*, 8–15.
- Hoeben Mannaert, L. N., Dijkstra, K., & Zwaan, R. A. (2017). Is color an integral part of a rich mental simulation? *Memory & Cognition*, *45*(6), 974–982.
- Horchak, O. V., Giger, J. C., Cabral, M., & Pochwatko, G. (2014). From demonstration to theory in embodied language comprehension: A review. *Cognitive Systems Research*, *29*, 66–85.
- Jared, D., Poh, R. P. Y., & Paivio, A. (2013). L1 and L2 picture naming in Mandarin–English bilinguals: A test of bilingual dual coding theory. *Bilingualism: Language and Cognition*, *16*(2), 383–396.
- Jiang, N. (2000). Lexical representation and development in a second language. *Applied Linguistics*, *21*(1), 47–77.
- Jiang, N. (2002). Form–meaning mapping in vocabulary acquisition in a second language. *Studies in Second Language Acquisition*, *24*(4), 617–637.
- Kang, X., Eerland, A., Joergensen, G. H., Zwaan, R. A., & Altmann, G. (2020). The influence of state change on object representations in language comprehension. *Memory & Cognition*, *48*(3), 390–399.
- Kogan, B., Muñoz, E., Ibáñez, A., & García, A. M. (2020). Too late to be grounded? Motor resonance for action words acquired after middle childhood. *Brain and Cognition*, *138*, 105509.
- Kroll, J. F., & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language*, *33*, 149–174.

- Kroll, J. F., & Sunderman, G. (2003). Cognitive processes in second language learners and bilinguals: The development of lexical and conceptual representations. In C. J. Doughty & M. H. Long (Eds.), *The Handbook of Second Language Acquisition* (pp. 104–129). Blackwell.
- Kroll, J. F., Michael, E., Tokowicz, N., & Dufour, R. (2002). The development of lexical fluency in a second language. *Second Language Research*, 18(2), 137–171.
- Kroll, J. F., Van Hell, J. G., Tokowicz, N., & Green, D. W. (2010). The Revised Hierarchical Model: A critical review and assessment. *Bilingualism: Language and Cognition*, 13(3), 373–381.
- Kühne, K., & Gianelli, C. (2019). Is embodied cognition bilingual? Current evidence and perspectives of the embodied cognition approach to bilingual language processing. *Frontiers in Psychology*, 10, 108.
- Kusanagi, K. (2017). Kakuritsubunpu karamiru gaikokugokyouikukenyudeta [Analyzing data of the studies in foreign language education with probability distributions]. *Reports Vol. 10 of 2017 Studies in Japan Association for Language Education and Technology, Kansai Chapter, Methodology Special Interest Group (SIG)*, 10, 1–40.
- Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4), 764–766.
- Li, P., & Xu, Q. (2022). Computational modeling of bilingual language learning: current models and future directions. *Language Learning*, 1–48.
- Louwerse, M., & Connell, L. (2011). A taste of words: Linguistic context and perceptual simulation predict the modality of words. *Cognitive Science*, 35(2), 381–398.

- Ludecke, D. (2021). *sjstats: Statistical Functions for Regression Models* (Version 0.18.1) [Computer software]. <https://doi.org/10.5281/zenodo.1284472>
- Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). *performance: An R package for assessment, comparison and testing of statistical models* (Version 0.9.0). *Journal of Open Source Software*, 6(60), 3139. doi.org/10.21105/joss.03139
- Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M., & Den, Y. (2014). Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*, 48(2), 345–371.
- Meara, P., & Miralpeix, I. (2016). *Tools for Researching Vocabulary*. Multilingual Matters.
- Mochizuki, M. (2015). Sintaika sa re ta ninchi wa gengo rikai ni dono teido juuyou na no ka [How important is embodied cognition to language comprehension?]. *Japanese Psychological Review*, 58(4), 485–505.
- Mochizuki, M. (2021). Gainen wa nani ni kibanka sarete iru no ka: Sintaika sa re ta ninchi to kibanka sa re ta ninchi ni moto zuku gainenshori to tango ninchi [What are concepts grounded in? A brief review of concept processing and word recognition on the framework of embodied and grounded cognition]. *Cognitive Studies: Bulletin of the Japanese Cognitive Science Society*, 28(4), 629–641.
- Monaco, E., Jost, L. B., Lancheros, M., Harquel, S., Schmidlin, E., & Annoni, J. M. (2021). First and Second Language at Hand: A Chronometric Transcranial-Magnetic Stimulation Study on Semantic and Motor Resonance. *Journal of Cognitive Neuroscience*, 33(8), 1563–1580.
- National Institute for Japanese Language and Linguistics, Center for Corpus Development. (2021). *Balanced Corpus of Contemporary Written Japanese* (short-unit) (Version

- 1.0) [Data set]. National Institute for Japanese Language and Linguistics. Retrieved August 14, 2022.
- Norman, T., & Peleg, O. (2021). The reduced embodiment of a second language. *Bilingualism: Language and Cognition*, 25(3), 406–416.
<https://doi.org/10.1017/s1366728921001115>
- Paivio, A., Clark, J. M., & Lambert, W. E. (1988). Bilingual dual-coding theory and semantic repetition effects on recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(1), 163–172.
- Parris, B. A., Hasshim, N., Wadsley, M., Augustinova, M., & Ferrand, L. (2022). The loci of Stroop effects: a critical review of methods and evidence for levels of processing contributing to color-word Stroop effects and the implications for the loci of attentional selection. *Psychological Research*, 86, 1029–1053.
- Patterson, A. (2021). Predicting second language listening functor comprehension probability with usage-based and embodiment approaches. *International Journal of Bilingualism*, 25(3), 772–788.
- Pavlenko, A. (2009). Conceptual representation in the bilingual lexicon and second language vocabulary learning. In Pavlenko, A. (ed.), *The Bilingual Mental Lexicon: Interdisciplinary Approaches*, (pp. 125–160). Multilingual Matters.
- Pavlenko, A. (2017). Do you wish to waive your rights? Affect and decision-making in multilingual speakers. *Current opinion in Psychology*, 17, 74–78.
- Pecher, D., van Dantzig, S., Zwaan, R. A., & Zeelenberg, R. (2009). Short article: Language comprehenders retain implied shape and orientation of objects. *Quarterly Journal of Experimental Psychology*, 62(6), 1108–1114.

- Poarch, G. J., Van Hell, J. G., & Kroll, J. F. (2015). Accessing word meaning in beginning second language learners: Lexical or conceptual mediation? *Bilingualism: Language and Cognition*, *18*(3), 357–371.
- Potter, M. C., So, K.-F., Von Eckardt, B., & Feldman, L. B. (1984). Lexical and conceptual representation in beginning and more proficient bilinguals. *Journal of Verbal Learning and Verbal Behavior*, *23*, 23–38.
- R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. [Computer software]. <http://www.R-Project.org>
- Richter, T., & Zwaan, R. A. (2009). Processing of color words activates color representations. *Cognition*, *111*(3), 383–389.
- Rommers, J., Meyer, A. S., & Huettig, F. (2013). Object shape and orientation do not routinely influence performance during language processing. *Psychological Science*, *24*(11), 2218–2225.
- Sato, M., Schafer, A. J., & Bergen, B. K. (2013). One word at a time: mental representations of object shape change incrementally during sentence processing. *Language and Cognition*, *5*(4), 345–373.
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, *3*(3), 417–424.
- Stanfield, R. A., & Zwaan, R. A. (2001). The effect of implied orientation derived from verbal context on picture recognition. *Psychological Science*, *12*(2), 153–156.
- Talamas, A., Kroll, J. F., & Dufour, R. (1999). From form to meaning: Stages in the acquisition of second-language vocabulary. *Bilingualism: Language and Cognition*, *2*(1), 45–58.

- Terai, M. (2019). Does the difference in learning directions affect the amount of knowledge gained from paired-association learning? *LET journal of Central Japan*, 30, 11–20.
- Terai, M., Fukuta, J., & Tamura, Y. (under review). Learnability and L1 Influence on L2 Collocational Representations of Japanese Learners of English.
- Terai, M., Yamashita, J., & Pasich, K. E. (2021). Effects of learning direction in retrieval practice on EFL vocabulary learning. *Studies in Second Language Acquisition*, 43(5), 1116–1137.
- Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Version 7.3-54). Fourth Edition. Springer.
- Vukovic, N., & Williams, J. N. (2014). Automatic perceptual simulation of first language meanings during second language sentence processing in bilinguals. *Acta Psychologica*, 145, 98–103.
- Wolter, B., & Gyllstad, H. (2011). Collocational links in the L2 mental lexicon and the influence of L1 intralexical knowledge. *Applied Linguistics*, 32(4), 430–449.
- Wolter, B., & Gyllstad, H. (2013). Frequency of input and L2 collocational processing: A comparison of congruent and incongruent collocations *Studies in Second Language Acquisition*, 35(3), 451–482.
- Wolter, B., & Yamashita, J. (2015). Processing collocations in a second language: A case of first language activation? *Applied Psycholinguistics*, 36(5), 1193–1221.
- Wolter, B., & Yamashita, J. (2018). Word frequency, collocational frequency, L1 congruency, and proficiency in L2 collocational processing: What accounts for L2 performance? *Studies in Second Language Acquisition*, 40(2), 395–416.

- Woods, D. L., Wyma, J. M., Yund, E. W., Herron, T. J., & Reed, B. (2015). Age-related slowing of response selection and production in a visual choice reaction time task. *frontiers in Human Neuroscience*, *9*, 193.
- Wu, Z., & Juffs, A. (2019). Revisiting the Revised Hierarchical Model: Evidence for concept mediation in backward translation. *Bilingualism: Language and Cognition*, *22*(2), 285–299.
- Yamashita, J., & Jiang, N. (2010). L1 influence on the acquisition of L2 collocations: Japanese ESL users and EFL learners acquiring English collocations. *TESOL Quarterly*, *44*(4), 647–668.
- Yaxley, R. H., & Zwaan, R. A. (2007). Simulating visibility during language comprehension. *Cognition*, *105*(1), 229–236.
- Zwaan, R. A. (2004). The immersed experiencer: Toward an embodied theory of language comprehension. In B. H. Ross (Ed.), *The Psychology of Learning and Motivation* (Vol. 44, pp. 35–62). Academic Press.
- Zwaan, R. A., & Pecher, D. (2012). Revisiting mental simulation in language comprehension: Six replication attempts. *PloS one*, *7*(12), e51382.
- Zwaan, R. A., Stanfield, R. A., & Yaxley, R. H. (2002). Language comprehenders mentally represent the shapes of objects. *Psychological Science*, *13*(2), 168–171.

Appendix A: The Results of the Rating Tasks (Pilot Study 1)

Word

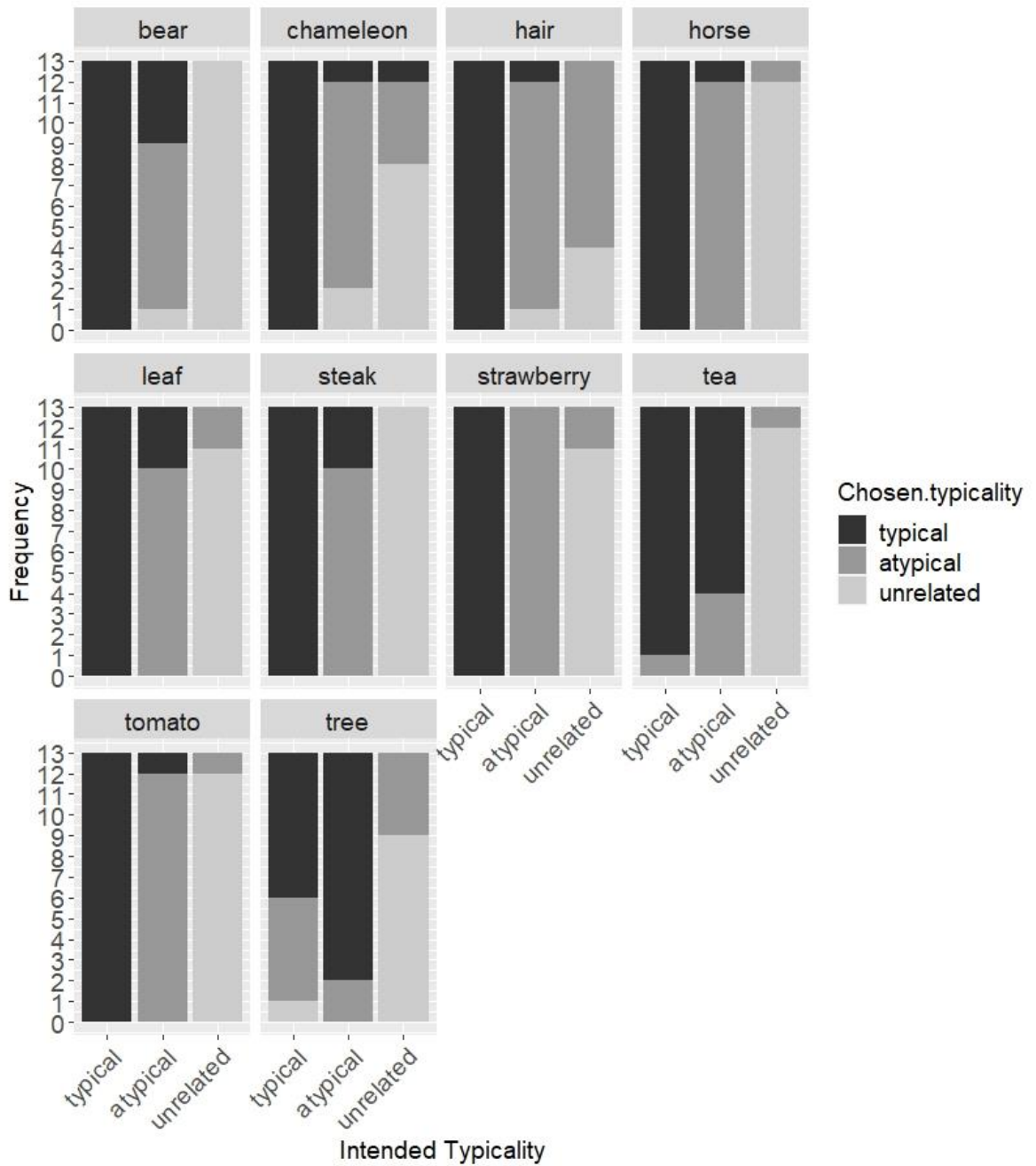
Rating Scores

The following table summarizes the agreement rates (%) of each typicality of objects' colors.

<i>Word</i>	<i>Typical</i>	<i>Atypical</i>	<i>Unrelated</i>
strawberry	100.00	100.00	84.62
bear	100.00	61.54	100.00
chameleon	100.00	76.92	61.54
hair	100.00	84.62	30.77
horse	100.00	92.31	92.31
leaf	100.00	76.92	84.62
tea	92.31	30.77	92.31
steak	100.00	76.92	100.00
tomato	100.00	92.31	92.31
tree	53.85	15.38	69.23

Stacked Bar Chart

In the following figure, the y-axis represents the frequency of the counts for each typicality. The x-axis (intended typicality) represents the typicality of colors that the author determined. *Chosen typicality* means that the typicality of colors that the participants chose. For example, the author determined that the typical color of bear is brown. The chart shows that all of the participants ($N = 13$) also considered the brown as typical color of bear.



Sentence

Agreement Rates

The following tables summarize the agreement rates (%) of each typicality that was implied by the sentences.

Entirely.

<i>Typicality</i>	<i>Agreement Rates</i>
typical	76.00
atypical	76.80

Each Sentence.

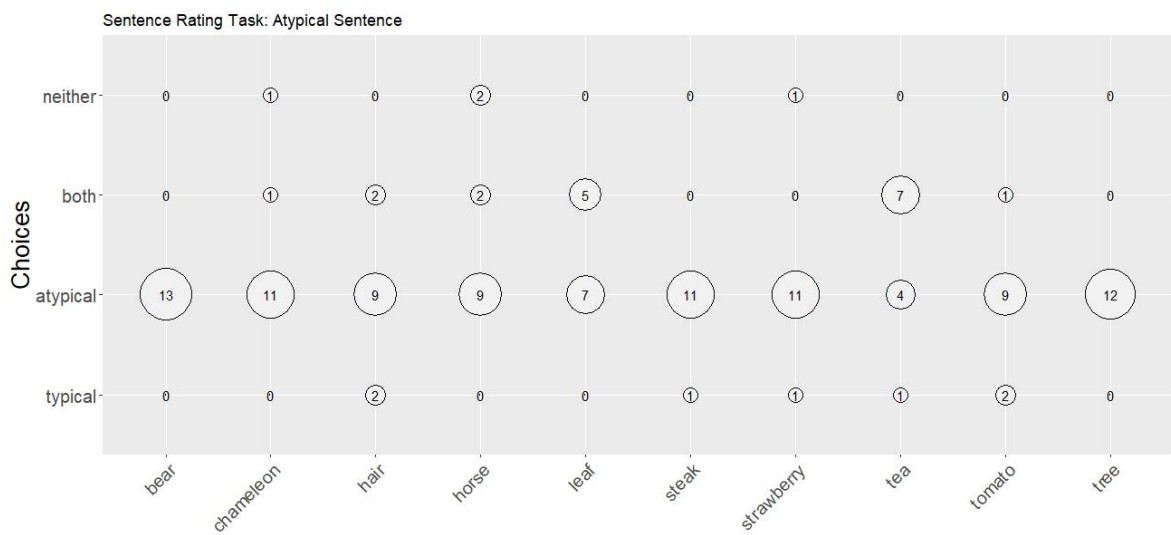
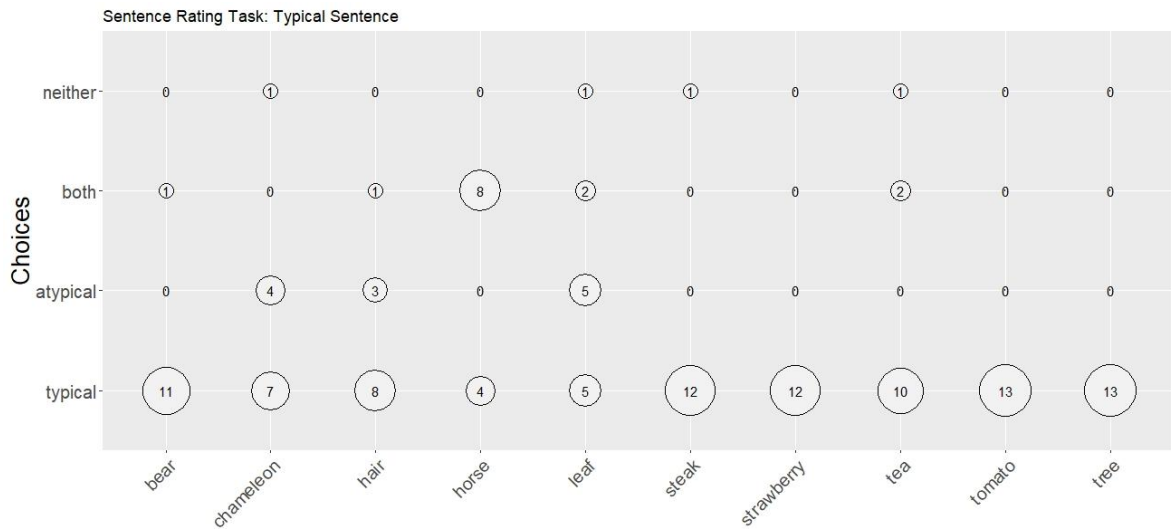
<i>Word</i>	<i>Typical</i>	<i>Atypical</i>
bear	91.67	100.00
chameleon	58.33	84.62
hair	66.67	69.23
horse	33.33	69.23
leaf	38.46	58.33
steak	92.31	91.67
strawberry	100.00	84.62
tea	76.92	33.33
tomato	100.00	75.00
tree	100.00	100.00

Balloon Plot

In the following Figure, the numbers in each balloon refers to the number of the participants who chose the choice. The test sentences were presented with two pictures and four forced choice alternatives:

- typical: best matched by the first picture (the first pictures were always typical objects)

- atypical: best matched by the second picture (the first pictures were always atypical objects)
- both: matched by both pictures equally
- neither: matched by neither picture



Appendix B: The Results of the Rating Tasks (Pilot Study 2)

Word

Rating Scores

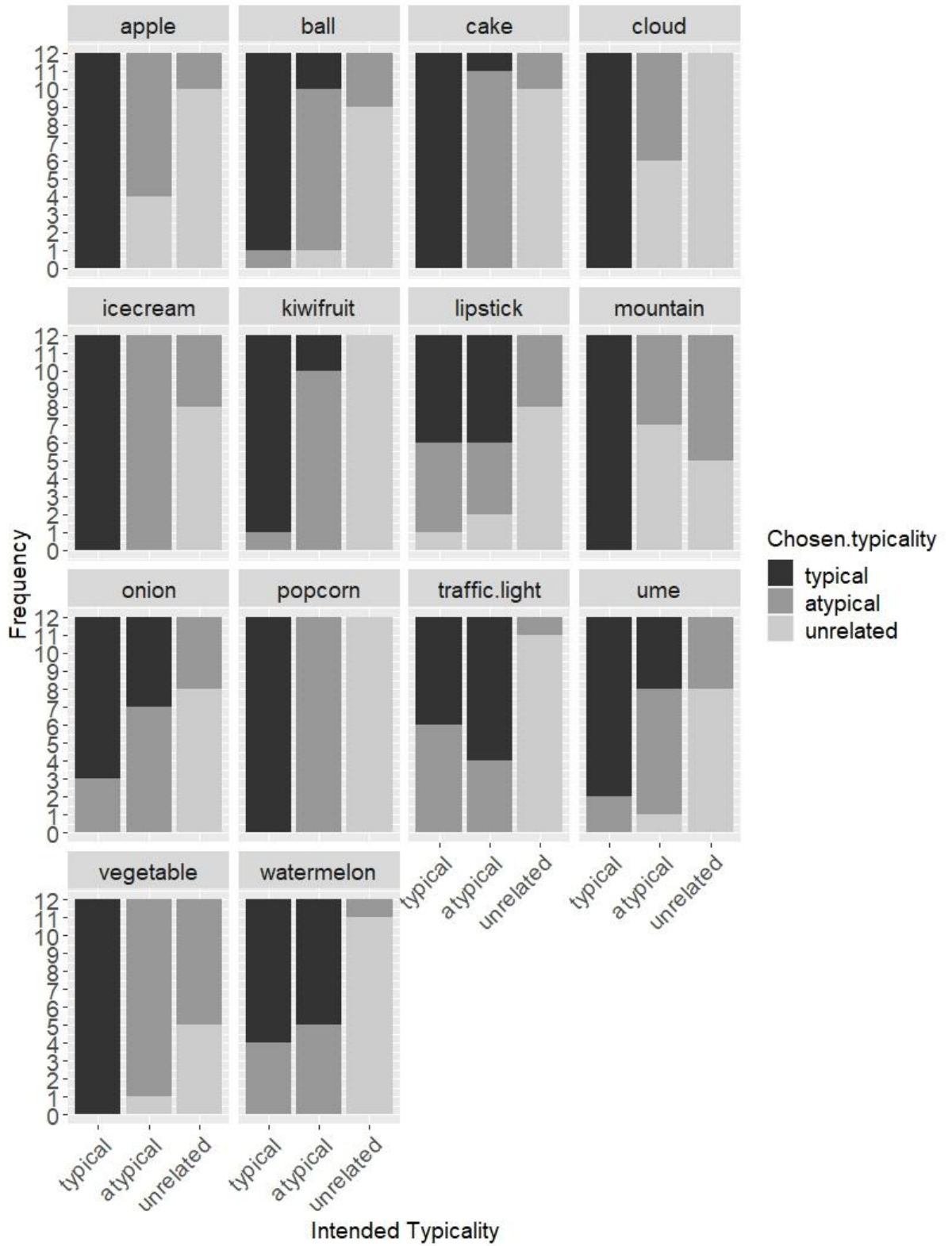
The following table summarizes the agreement rates (%) of each typicality of objects' colors.

<i>Word</i>	<i>Typical</i>	<i>Atypical</i>	<i>Unrelated</i>
icecream	100.00	100.00	66.67
vegetable	100.00	91.67	41.67
watermelon	66.67	41.67	91.67
mountain	100.00	41.67	41.67
traffic.light	50.00	33.33	91.67
ume	83.33	58.33	66.67
onion	75.00	58.33	66.67
ball	91.67	75.00	75.00
cake	100.00	91.67	83.33
apple	100.00	66.67	83.33
lipstick	50.00	33.33	66.67
popcorn	100.00	100.00	100.00
kiwifruit	91.67	83.33	100.00
cloud	100.00	50.00	100.00

Stacked Bar Chart

In the figure, the y-axis represents the frequency of the counts for each typicality. The x-axis (intended typicality) represents the typicality of colors that the author determined. *Chosen typicality* means that the typicality of colors that the participants

chose. For example, the author determined that the typical color of *apple* is red. The chart shows that all of the participants ($N = 12$) also considered the red as typical color of *apple*.



Sentence

Agreement Rates

The following tables summarize the agreement rates (%) of each typicality that was implied by the sentences.

Entirely.

<i>Typicality</i>	<i>Agreement Rates</i>
typical	82.14
atypical	86.90

Each Sentence.

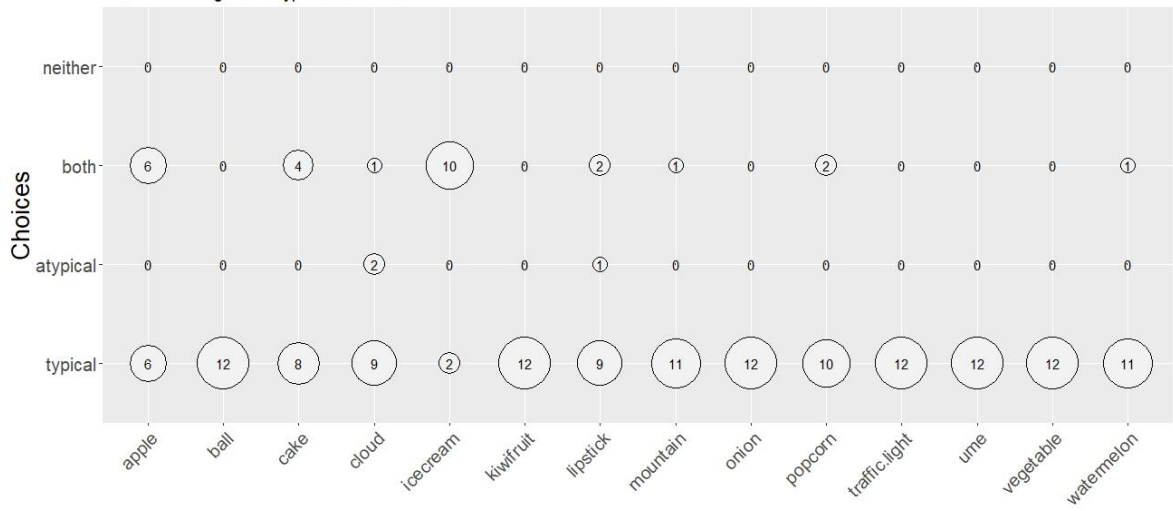
<i>Word</i>	<i>Typical</i>	<i>Atypical</i>
apple	50.00	75.00
ball	100.00	100.00
cake	66.67	41.67
cloud	75.00	100.00
icecream	16.67	66.67
kiwifruit	100.00	100.00
lipstick	75.00	91.67
mountain	91.67	100.00
onion	100.00	83.33
popcorn	83.33	100.00
traffic.light	100.00	100.00
ume	100.00	75.00
vegetable	100.00	91.67
watermelon	91.67	91.67

Balloon Plot

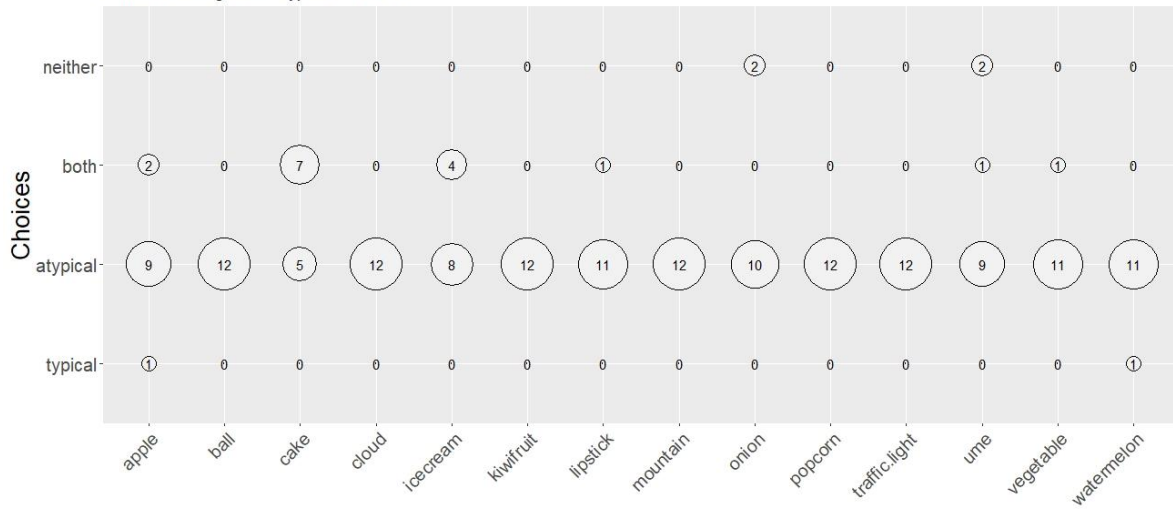
In the Figure, the numbers in each balloon refers to the number of the participants who chose the choice. The test sentences were presented with two pictures and four forced choice alternatives:

- typical: best matched by the first picture (the first pictures were always typical objects)
- atypical: best matched by the second picture (the first pictures were always atypical objects)
- both: matched by both pictures equally
- neither: matched by neither picture

Sentence Rating Task: Typical Sentence



Sentence Rating Task: Atypical Sentence



Appendix C: Power Analysis

Experiment 1

Native English and Native Japanese Speakers

- Targeted power
 - 80 percent
- Degrees of freedom for the numerator (the number of predictors in the model)
 - 2 (3 levels (typical, atypical, unrelated) -1)
- Effect size
 - $R^2 = .02$
- Alpha level
 - .05
- Expected intraclass correlation coefficient
 - .05
- The number of observations per cluster group
 - 60 (each participant processed 60 items for each word typicality)

Results.

```
power.sjstats <- sjstats::samplesize_mixed(  
  eff.size = 0.02,  
  df.n = 2,  
  power = 0.8,  
  sig.level = 0.05,  
  k = NULL,  
  n = 60,  
  icc = 0.05  
)  
  
power.sjstats  
  
## $`Subjects per Cluster`  
## numeric(0)  
##  
## $`Total Sample Size`  
## [1] 1915  
  
power.sjstats$`Total Sample Size`  
## [1] 1915
```

The results showed that to achieve the targeted power and the effect size, a total of 1915 observations were needed.

The Required Sample Sizes.

```
power.sjstats$`Total Sample Size` / 60  
## [1] 31.91667
```

The results showed that at least 32 participants were needed for the study.

Experiment 2

Native Japanese Speakers Learning English

- Targeted power
 - 80 percent
- Degrees of freedom for the numerator (the number of predictors in the model)
 - 5 (Word Typicality (3 levels -1), Vocabulary Size, the Interaction of the two (3 levels -1 × 1))
- Effect size
 - $R^2 = .025$
- Alpha level
 - .05
- Expected intraclass correlation coefficient
 - .05
- The number of observations per cluster group
 - 60 (each participant processed 60 items for each word typicality)

Results.

```
power.sjstats <- sjstats::samplesize_mixed(  
  eff.size = 0.025,  
  df.n = 5,  
  power = 0.8,  
  sig.level = 0.05,  
  k = NULL,  
  n = 60,  
  icc = 0.05  
)  
  
power.sjstats$`Total Sample Size`  
## [1] 2049
```

The results showed that to achieve the targeted power and the effect size, a total of 2049 observations were needed.

The Required Sample Sizes.

```
power.sjstats$`Total Sample Size` / 60
```

```
## [1] 34.15
```

The results showed that at least 35 participants were needed for the study.

Appendix D: The Experimental Items Used in Experiment 1 (Native English Speakers)

Critical Sentences

- Sentence. (Typicality of the Sentence: Typical)
- Sentence. (Typicality of the Sentence: Atypical)
 - Keyword: Font Color (Typical), Font Color (Atypical), Font Color (Unrelated)

Example

- The bananas that Mark bought looked ready to eat. (Typical)
- The bananas that Mark bought didn't look ready to eat. (Atypical)
 - bananas: yellow (Typical) / green (Atypical) / red (Unrelated)

Filler Sentences

- Sentence.
 - Keyword: Font Color
 - (Correct Answer: True/False) Comprehension Question.

Example

- The bird couldn't fly because it had a broken wing.
 - bird: white
 - (T) The bird couldn't fly because it broke its leg.

Practice Session

Critical Sentences (Three Sentences)

- She didn't like to wear a mask.
 - mask: red
- Ken always used his favorite cup when he had tea.
 - cup: green
- The kids looked happy when they saw a new computer.
 - computer: white

Filler Sentences (Two Sentences)

- The bird couldn't fly because it had a broken wing.
 - bird: white
 - (F) The bird couldn't fly because it broke its leg.
- Matt had drunk five beers before his friend had finished two.
 - beer: brown
 - (F) Matt's friend had drunk five beers before Matt had finished two.

Main Session

Critical Sentences (Before) (Ninety Sentences)

- It looked ready to eat when Mark bought the strawberry. (Typical)

- It didn't look ready to eat when Mark bought the strawberry. (Atypical)
 - strawberry: red (Typical), green (Atypical), brown (Unrelated)
- In the woods, Joe was excited to see a bear. (Typical)
- At the North Pole, Joe was excited to see a bear. (Atypical)
 - bear: brown (Typical), white (Atypical), green (Unrelated)
- The teacher pointed to the grass when he found a chameleon lying camouflaged. (Typical)
- The teacher pointed to the sand when he found a chameleon lying camouflaged. (Atypical)
 - chameleon: green (Typical), brown (Atypical), white (Unrelated)
- Sam liked to ride on his horse. (Typical)
- Sam liked to ride on the prince's horse. (Atypical)
 - horse: brown (Typical), white (Atypical), red (Unrelated)
- Sarah stopped in front of a tree and pick a leaf off. (Typical)
- Sarah sat on the ground and pick a leaf up. (Atypical)
 - leaf: green (Typical), brown (Atypical), white (Unrelated)
- At the restaurant, John looked at the steak. (Typical)
- At the meat-shop, John looked at the steak. (Atypical)
 - steak: brown (Typical), red (Atypical), green (Unrelated)
- Because it was ready to eat, Jane ate the tomato. (Typical)
- Before it was ready to eat, Jane ate the tomato. (Atypical)
 - tomato: red (Typical), green (Atypical), white (Unrelated)
- Inside the rice ball, Marie noticed the plum. (Typical)
- In the tree, Marie noticed the plum. (Atypical)
 - plum: red (Typical), green (Atypical), brown (Unrelated)
- From out of the field, Liz took an onion. (Typical)
- From out of the pot, Liz took an onion. (Atypical)
 - onion: brown (Typical), white (Atypical), red (Unrelated)
- Ray went out to the baseball field with the ball. (Typical)
- Ray went out to the the basketball court with the ball. (Atypical)
 - ball: white (Typical), brown (Atypical), green (Unrelated)
- For her wedding, Lynn ordered a cake. (Typical)
- For Valentine's Day, Lynn ordered a cake. (Atypical)
 - cake: white (Typical), brown (Atypical), green (Unrelated)
- When we went to the orchard, Amy ate an apple. (Typical)
- When she had a cold, Amy ate an apple. (Atypical)
 - apple: red (Typical), white (Atypical), brown (Unrelated)
- It tasted sour when Ben popped a piece of popcorn in his mouth. (Typical)
- It tasted sweet when Ben popped a piece of popcorn in his mouth. (Atypical)
 - popcorn: white (Typical), brown (Atypical), red (Unrelated)
- At the bottom of the parfait, Roy found a kiwi. (Typical)
- At the bottom of the basket, Roy found a kiwi. (Atypical)

- kiwi: green (Typical), brown (Atypical), red (Unrelated)
- Claire felt it was a beautiful summer sky when she saw the cloud. (Typical)
- Claire felt it was a beautiful sunset when she saw the cloud. (Atypical)
 - cloud: white (Typical), red (Atypical), green (Unrelated)

Critical Sentences (After) (Ninety Sentences)

- The strawberry that Mark bought looked ready to eat. (Typical)
- The strawberry that Mark bought didn't look ready to eat. (Atypical)
 - strawberry: red (Typical), green (Atypical), brown (Unrelated)
- Joe was excited to see a bear in the woods. (Typical)
- Joe was excited to see a bear at the North Pole. (Atypical)
 - bear: brown (Typical), white (Atypical), green (Unrelated)
- The teacher pointed to the chameleon lying camouflaged in the grass. (Typical)
- The teacher pointed to the chameleon lying camouflaged in the sand. (Atypical)
 - chameleon: green (Typical), brown (Atypical), white (Unrelated)
- Sam liked the horse which he was riding. (Typical)
- Sam liked the horse which the prince was riding. (Atypical)
 - horse: brown (Typical), white (Atypical), red (Unrelated)
- Sarah stopped in the woods to pick a leaf off a tree. (Typical)
- Sarah stopped in the woods to pick a leaf off the ground. (Atypical)
 - leaf: green (Typical), brown (Atypical), white (Unrelated)
- John looked at the steak on his plate. (Typical)
- John looked at the steak in the meat-shop. (Atypical)
 - steak: brown (Typical), red (Atypical), green (Unrelated)
- Jane ate the tomato because it was ready to eat. (Typical)
- Jane ate the tomato before it was ready to eat. (Atypical)
 - tomato: red (Typical), green (Atypical), white (Unrelated)
- Marie noticed the plum inside the rice ball. (Typical)
- Marie noticed the plum in the tree. (Atypical)
 - plum: red (Typical), green (Atypical), brown (Unrelated)
- Liz took an onion from out of the field. (Typical)
- Liz took an onion from out of the pot. (Atypical)
 - onion: brown (Typical), white (Atypical), red (Unrelated)
- Ray took the ball and went out to the baseball field. (Typical)
- Ray took the ball and went out to the the basketball court. (Atypical)
 - ball: white (Typical), brown (Atypical), green (Unrelated)
- Lynn ordered a cake for her wedding. (Typical)
- Lynn ordered a cake for Valentine's Day. (Atypical)
 - cake: white (Typical), brown (Atypical), green (Unrelated)
- Amy ate an apple when we went to the orchard. (Typical)
- Amy ate an apple when she had a cold. (Atypical)

- apple: red (Typical), white (Atypical), brown (Unrelated)
- Ben popped a piece of popcorn in his mouth and it tasted sour. (Typical)
- Ben popped a piece of popcorn in his mouth and it tasted sweet. (Atypical)
 - popcorn: white (Typical), brown (Atypical), red (Unrelated)
- Roy found a kiwi at the bottom of the parfait. (Typical)
- Roy found a kiwi at the bottom of the basket. (Atypical)
 - kiwi: green (Typical), brown (Atypical), red (Unrelated)
- Claire saw the cloud and felt it was a beautiful summer sky. (Typical)
- Claire saw the cloud and felt it was a beautiful sunset. (Atypical)
 - cloud: white (Typical), red (Atypical), green (Unrelated)

Filler Sentences (Before) (Ninety Sentences)

- To go to Kyoto, his parents took an airplane.
 - airplane: brown
 - (F) They went to Europe.
- At the market, John was eating a banana that he bought.
 - banana: red
 - (F) John ate a cookie.
- To see the beautiful moon, his daughter used to go to the beach.
 - moon: green
 - (F) His daughter used to go to the mountain.
- Before Aaron slept, he smoked in bed.
 - bed: white
 - (F) Aaron smoked in the park.
- On bread, Dan liked to spread butter.
 - butter: brown
 - (F) Dan liked to spread chocolate on bread.
- On the eighth of August, the girl had to return the book to the library.
 - book: red
 - (F) The girl did not have to return the book.
- Because he had to stay alone in the house, the boy was worried.
 - house: green
 - (T) The boy had to stay alone.
- It looked very expensive when George bought the chocolate.
 - chocolate: red
 - (F) George bought some books.
- Noah was interested in the Japanese history of coffee.
 - coffee: red
 - (T) Noah was interested in the history of coffee.
- On the branch, the kids found a praying mantis.
 - praying mantis: red

- (T) The kids found a praying mantis.
- At the zoo, the children watched the cat.
 - cat: red
 - (T) The children watched the cat.
- When Erika was a child, she always had stew for dinner.
 - stew: white
 - (F) Erika always ate pasta.
- In the morning, Logan stopped at the bar to pick up his salad.
 - salad: red
 - (F) Logan got his alchohole.
- Although Emma cooked the chicken for too long, she wanted to make a good dinner.
 - chicken: red
 - (T) Emma cooked chicken for dinner.
- Before May left the kitchen, she put the avocado in the pot.
 - avocado: green
 - (F) May put the beef in the pot.
- For the first time, Milo tasted the rice but he didn't like it with vinegar.
 - rice: red
 - (T) Milo ate the rice.
- In the cold weather, Ted thought the pear outside his house looked delicious.
 - pear: red
 - (F) Ted thought about a friend.
- In the park, Nick liked to eat ice cream.
 - ice cream: white
 - (F) Nick did not like ice cream.
- From out of the fridge, Davis took the vegetable.
 - vegetable: green
 - (F) Davis took the chicken.
- In a field, Simon saw a watermelon.
 - watermelon: green
 - (F) Simon saw a cat.
- In the summer, Paula thought the mountain outside her window looked beautiful.
 - mountain: green
 - (F) Paula thought the mountain looked ugly.
- Ben kept going after he checked the traffic light.
 - traffic light: green
 - (F) Ben stopped at the traffic light.
- For makeup, Beth put on lipstick.
 - lipstick: red
 - (F) Beth did not have a lipstick.
- In the winter, Posy often went to see her favorite tree.

- tree: white
 - (F) Posy did not like a tree.
- When her granddaughter wore hair up, Susan liked it better.
 - hair: brown
 - (F) Susan liked her mother to wear updos.
- When Mike went to Japan, he bought tea.
 - tea: white
 - (F) Mike bought coffee in Japan.
- With his friends, the kid decided to stay home.
 - home: red
 - (F) The kid decided to go out.
- At the restaurant, Robert asked the waiter to bring him the check.
 - check: red
 - (F) Robert was at the university.
- At the station, his girlfriend lost her wallet.
 - wallet: red
 - (F) The girlfriend found her wallet.
- Outside, the children were playing Cowboys.
 - cowboy: red
 - (F) The adults were playing outside.
- On the weekends, his father always enjoyed driving the car.
 - car: red
 - (F) The father enjoyed driving on the weekdays.
- At the station, Alyce bought a newspaper.
 - newspaper: red
 - (F) Alyce bought a sandwich.
- In the basket, she found an eggplant.
 - eggplant: red
 - (F) The egg was in the basket.
- At the park, Amy strained a muscle in her leg.
 - muscle: red
 - (T) Amy strained a muscle in her leg.
- On the table, Anika saw a dog sitting.
 - dog: red
 - (T) Anika saw a dog.
- After dinner, Bella washed her plate.
 - plate: red
 - (T) Bella washed a plate after dinner.
- At the office, he was using a computer.
 - computer: red
 - (T) He was using a computer.
- At the dentist, Barrett bought a toothbrush.

- toothbrush: red
 - (T) Barrett bought a toothbrush at the dentist.
- At the store, Becky sold the pyjamas that she liked.
 - pyjamas: red
 - (T) Becky sold the pyjamas that she liked.
- In 1927, Benny first visited the restaurant with his friends.
 - restaurant: red
 - (T) Benny first visited the restaurant in 1927.
- In late summer, Benny took a lot of pictures of a beautiful flower.
 - flower: brown
 - (F) Benny took pictures of a car.
- Last week, Bret bought a chair that looked expensive.
 - chair: brown
 - (F) Bret sold a chair last week.
- For her family, she baked a tart that was covered with chocolate.
 - tart: brown
 - (F) She baked a cake.
- Before the end of the year, they completed the new road.
 - road: brown
 - (F) The road was not completed before the end of the year.
- At the pub, Bryan ordered his favorite beer.
 - beer: brown
 - (F) Bryan ordered a coffee.
- Last year, the bike that Cale wanted was sold out.
 - bike: brown
 - (F) Cale did not want the bike.
- At the port, Cary was surprised when he found a battleship.
 - battleship: brown
 - (T) Cary saw a battleship at the port.
- After eating breakfast, the man rushed to the parking lot.
 - parking lot: brown
 - (T) The man went to the parking lot after eating breakfast.
- In 2005, Chad visited a famous office.
 - office: brown
 - (T) Chad visited a famous office in 2005.
- In the morning, Chuck stopped at a cafe to get milk.
 - milk: brown
 - (T) Chuck got milk in the morning.
- In the town, Clint opened a map to find a place.
 - map: brown
 - (T) Clint was looking for a place.
- Dean tasted the French wine that he imported.

- wine: brown
 - (T) Dean tasted the wine.
- In the box, Dobie noticed the diamond.
 - diamond: brown
 - (T) Dobie noticed the diamond.
- From the kitchen, Eddie took a knife.
 - knife: brown
 - (T) Eddie took a knife from the kitchen.
- After a busy week, Frank bought a magazine.
 - magazine: brown
 - (T) Frank bought a magazine.
- For Christmas party, Gabe ordered a pizza.
 - pizza: brown
 - (T) Gabe ordered a pizza.
- When Hank was a student, he liked to collect sneakers.
 - sneaker: green
 - (T) Hank liked to collect sneakers.
- With a knife, Hal opened the can.
 - can: green
 - (T) Hal opened the can with a knife.
- At the park, Hilary found an orange.
 - orange: green
 - (T) Hilary found an orange.
- At the shopping mall, India saw an actor.
 - actor: green
 - (T) India saw an actor at the shopping mall.
- In the pasture, there were a lot of sheep.
 - sheep: green
 - (T) There were a lot of sheep in the pasture.
- When Nana studied, she always listened to the radio.
 - radio: green
 - (T) Nana listened to the radio.
- To make the pudding, it was very important to choose good quality sugar.
 - sugar: green
 - (T) Sugar was important for the pudding.
- At the door, Ryan showed his ticket for the movie.
 - ticket: green
 - (T) Ryan had a ticket for the movie.
- From space, Maggie wanted to see the earth.
 - earth: green
 - (T) Maggie wanted to see the earth.
- In the band, a ten-year-old girl played the keyboard.

- keyboard: green
 - (T) The girl played the keyboard in the band.
- Due to heavy snow, Pat ended up staying all day on the train.
 - train: green
 - (T) Pat ended up staying on the train.
- Every morning, they drew water from the well.
 - water: green
 - (F) They bought water from the store.
- At the small store, Jake bought a pen.
 - pen: green
 - (F) Jake bought a ruler.
- In the 1950s, most of children owned a doll and played with it.
 - doll: green
 - (F) Dolls were not popular in the 1950s.
- For her birthday, the kid received a small box as a present.
 - box: green
 - (F) The kid gave his friend a present.
- During the 1960s, some people considered TV bad for kids.
 - TV: green
 - (F) All the people used to consider TV good for kids.
- When Jacki was in college, she spent a lot of money on piano lessons.
 - piano: white
 - (T) Jacki spent a lot of money on piano lessons.
- Yesterday, Jeff was asked to close the door.
 - door: white
 - (T) Jeff was asked to close the door yesterday.
- During the day, the window was kept open.
 - window: white
 - (T) The door was being opened all day.
- Last week, the crow that attacked Jed was finally caught.
 - crow: white
 - (T) The crow attacked Jed.
- Today was the day that the student was supposed to hand in his paper.
 - paper: white
 - (T) The student had to submit his paper.
- At night, the children were always scared when they saw the statue.
 - statue: white
 - (T) The statue scared the children.
- Because it allowed Kasey to buy coffee, she was happy to receive a coin.
 - coin: white
 - (T) Kasey was happy when she received the coin.
- Because Lizzy had long hair, she liked her comb.

- comb: white
 - (T) Lizzy had long hair.
- Because the actor mentioned the factory in the press conference, it became famous.
 - factory: white
 - (T) The factory became famous because of the actor.
- Because it is very relaxing, Neal liked to watch the flame.
 - flame: white
 - (T) Neal liked to watch the flame.
- In the past, a fur coat was a popular gift.
 - fur coat: white
 - (T) A fur coat was a popular gift.
- From Japan came a story about a ghost.
 - ghost: white
 - (F) A story came from Korea.
- For pasta, Mercy went to a mountain to find a certain mushroom.
 - mushroom: white
 - (F) Mercy looked for a smartphone.
- To spread the skin cream, her mother would always use cotton.
 - cotton: white
 - (F) Her mother did not have a skin cream.
- In the experiment, the professor used a pigeon.
 - pigeon: white
 - (F) A rat was used in the experiment.
- For hunting, Lyle was looking for a gun.
 - gun: white
 - (F) Lyle was looking for a knife.
- For his dog, his father bought a bar of soap.
 - soap: white
 - (F) His father bought a bar of chocolate.
- Off the coast of Japan, the whale was found.
 - whale: white
 - (F) A dolphine was found near Japan.

Filler Sentences (After) (Ninety Sentences)

- His parents took an airplane to go to Kyoto.
 - airplane: brown
 - (T) They went to Kyoto.
- John was eating a banana that he bought at the market.
 - banana: red
 - (T) John ate a banana.
- His daughter used to go to the beach to see the beautiful moon.

- moon: green
 - (T) His daughter used to go to the beach.
- Aaron smoked in bed before he slept.
 - bed: white
 - (T) Aaron smoked in bed.
- Dan liked to spread butter on bread.
 - butter: brown
 - (T) Dan liked to spread butter on bread.
- The girl had to return the book to the library on the eighth of August.
 - book: red
 - (T) The girl had to return the book.
- The boy was worried because he had to stay alone in the house.
 - house: green
 - (F) The boy had to go to the neighbor.
- The chocolate that George bought looked very expensive.
 - chocolate: red
 - (T) George bought some chocolate.
- Noah was interested in the history of coffee in Japan.
 - coffee: red
 - (F) Noah was interested in soccer.
- The kids found a praying mantis on the branch.
 - praying mantis: red
 - (F) The kids killed a praying mantis.
- The children watched the cat at the zoo.
 - cat: red
 - (F) The children watched an elephant.
- Erika always had stew for dinner when she was a child.
 - stew: white
 - (T) Erika always ate stew.
- Logan stopped at the bar to pick up his salad in the morning.
 - salad: red
 - (T) Logan got his salad.
- Emma wanted to make a good dinner but she cooked the chicken for too long.
 - chicken: red
 - (F) Emma cooked beef for dinner.
- May put the avocado in the pot and left the kitchen.
 - avocado: green
 - (T) May put the avocado in the pot.
- Milo tasted the rice for the first time but he didn't like it with vinegar.
 - rice: red
 - (F) Milo did not eat the rice.
- Ted thought the pear outside his house looked delicious in the cold weather.

- pear: red
 - (T) Ted thought about a pear.
- Nick liked to eat ice cream in the park.
 - ice cream: white
 - (T) Nick liked to eat ice cream.
- Davis took the vegetable from out of the fridge.
 - vegetable: green
 - (T) Davis took the vegetable.
- Simon saw a watermelon in a field.
 - watermelon: green
 - (T) Simon saw a watermelon.
- Paula thought the mountain outside her window looked beautiful in the summer.
 - mountain: green
 - (T) Paula thought the mountain looked beautiful.
- Ben checked the traffic light and kept going.
 - traffic light: green
 - (T) Ben kept going.
- Beth put on lipstick for makeup.
 - lipstick: red
 - (T) Beth put on lipstick.
- Posy often went to see her favorite tree in the winter.
 - tree: white
 - (T) Posy has a favorite tree.
- Susan liked it better when her granddaughter wore her hair up.
 - hair: brown
 - (T) Susan liked her granddaughter to wear updos.
- Mike bought tea when he went to Japan
 - tea: white
 - (T) Mike bought tea in Japan.
- The kid decided to stay home with his friends.
 - home: red
 - (T) The kid stayed home.
- Robert asked the waiter to bring him the check at the restaurant.
 - check: red
 - (T) Robert was at the restaurant.
- His girlfriend lost her wallet at the station.
 - wallet: red
 - (T) The girlfriend lost her wallet.
- The children were playing Cowboys outside.
 - cowboy: red
 - (T) The children were playing outside.
- His father always enjoyed driving the car on the weekends.

- car: red
 - (T) The father enjoyed driving on the weekends.
- Alyce bought a newspaper at the station.
 - newspaper: red
 - (T) Alyce bought a newspaper.
- She found an eggplant in the basket.
 - eggplant: red
 - (T) The eggplant was in the basket.
- Amy strained a muscle in her leg at the park.
 - muscle: red
 - (F) Amy strained a muscle in her arm.
- Anika saw a dog sitting on the table.
 - dog: red
 - (F) Anika saw a cat sitting on the table.
- Bella washed her plate after dinner.
 - plate: red
 - (F) Bella washed a plate after breakfast.
- He was using a computer at the office.
 - computer: red
 - (F) He could not use a computer.
- Barrett bought a toothbrush at the dentist.
 - toothbrush: red
 - (F) Barrett bought a toothbrush at the supermarket.
- Becky sold the pyjamas that she liked at the store.
 - pyjamas: red
 - (F) Becky bought the pyjamas that she liked.
- Benny first visited the restaurant with his friends in 1927.
 - restaurant: red
 - (F) Benny first opened the restaurant in 1927.
- Berny took a lot of pictures of a beautiful flower in late summer.
 - flower: brown
 - (T) Berny took pictures of a flower.
- Bret bought a chair that looked expensive last week.
 - chair: brown
 - (T) Bret bought a chair last week.
- She baked a tart that was covered with chocolate for her family.
 - tart: brown
 - (T) The tart was covered with chocolate.
- They completed the new road before the end of the year.
 - road: brown
 - (T) The road was completed before the end of the year.
- Bryan ordered his favorite beer at the pub.

- beer: brown
 - (T) Bryan ordered a beer.
- The bike that Cale wanted was sold out last year.
 - bike: brown
 - (T) The bike was sold out last year.
- Cary was surprised when he found a battleship at the port.
 - battleship: brown
 - (F) Cary saw a battleship at the museum.
- The man rushed to the parking lot after eating breakfast.
 - parking lot: brown
 - (F) The man went to the parking lot before eating breakfast.
- Chad visited a famous office in 2005.
 - office: brown
 - (F) Chad visited a famous office in 2020.
- Chuck stopped at a cafe to get milk in the morning.
 - milk: brown
 - (F) Chuck got milk in the evening.
- Clint opened a map to find a place in the town.
 - map: brown
 - (F) Clint asked a policeman to find a place.
- Dean tasted the wine that he imported from France.
 - wine: brown
 - (F) Dean tasted the chocolate bar.
- Dobie noticed the diamond in the box.
 - diamond: brown
 - (F) Dobie noticed the cat.
- Eddie took a knife from the kitchen.
 - knife: brown
 - (F) Eddie took a fork from the kitchen.
- Frank bought a magazine after a busy week.
 - magazine: brown
 - (F) Frank bought a newspaper.
- Gabe ordered a pizza for the Christmas party.
 - pizza: brown
 - (F) Gabe ordered a cake.
- Hank liked to collect sneakers when he was a student.
 - sneaker: green
 - (F) Hank liked to collect jackets.
- Hal opened the can with a knife.
 - can: green
 - (F) Hal opened the box with a knife.
- Hilary found an orange at the park.

- orange: green
 - (F) Hilary found a strawberry.
- India saw an actor at the shopping mall.
 - actor: green
 - (F) India saw her friend at the shopping mall.
- There were a lot of sheep in the pasture.
 - sheep: green
 - (F) There were a lot of cows in the pasture.
- Nana always listened to the radio when she studied.
 - radio: green
 - (F) Nana did not listened to the radio when she studied.
- It was very important to choose good quality sugar to make the pudding.
 - sugar: green
 - (F) Sugar was not important for the pudding.
- Ryan showed his ticket for the movie at the door.
 - ticket: green
 - (F) Ryan had a ticket for the zoo.
- Maggie wanted to see the earth from space.
 - earth: green
 - (F) Maggie had a picture of the earth.
- A ten-year-old girl played the keyboard in the band.
 - keyboard: green
 - (F) The girl played the guitar in the band.
- Pat ended up staying all day on the train due to heavy snow.
 - train: green
 - (F) Pat ended up staying at a hotel.
- They drew water from the well every morning.
 - water: green
 - (T) They drew water from the well.
- Jake bought a pen at the small store.
 - pen: green
 - (T) Jake bought a pen.
- Most of children owned a doll and played with it in the 1950s.
 - doll: green
 - (T) Most of children owned a doll in the 1950s.
- The kid received a small box as a present for her birthday.
 - box: green
 - (T) The kid received a present for her birthday.
- Some people cosidered TV bad for kids during the 1960s.
 - TV: green
 - (T) Some people used to consider TV bad for kids.
- Jacki spent a lot of money on piano lessons when she was in college.

- piano: white
 - (F) Jacki spent a lot of money on drum lessons.
- Jeff was asked to close the door yesterday.
 - door: white
 - (F) Jeff was asked to open the door yesterday.
- The window was kept open during the day.
 - window: white
 - (F) The door was closed all day.
- The crow that attacked Jed was finally caught last week.
 - crow: white
 - (F) Jed attacked the crow.
- The student was supposed to hand in his paper by today.
 - paper: white
 - (F) The student had nothing to submit.
- The children were always scared when they saw the statue at night.
 - statue: white
 - (F) The children liked the statue.
- Kasey was happy to receive a coin because she could buy coffee.
 - coin: white
 - (F) Kasey was sad when she received the coin.
- Lizzy liked her comb because she had long hair.
 - comb: white
 - (F) Lizzy had short hair.
- The factory became famous because an actor mentioned it at the press conference.
 - factory: white
 - (F) The factory was not famous.
- Neal liked to watch the flame because it is very relaxing.
 - flame: white
 - (F) Neal liked to watch the ocean.
- A fur coat was a popular gift in the past.
 - fur coat: white
 - (F) In the past, a fur coat was not a common gift.
- A story about a ghost came from Japan.
 - ghost: white
 - (T) A story came from Japan.
- Mercy went to a mountain to find a certain mushroom for pasta.
 - mushroom: white
 - (T) Mercy wanted a mushroom.
- Her mother would always use cotton to spread her skin cream.
 - cotton: white
 - (T) Her mother used skin cream.
- The professor used a pigeon in the experiment.

- pigeon: white
 - (T) A pigeon was used in the experiment.
- Lyle was looking for a gun to take hunting.
 - gun: white
 - (T) Lyle was looking for a gun.
- His father bought a bar of soap for his dog.
 - soap: white
 - (T) His father bought a bar of soap.
- The whale was found off the coast of Japan.
 - whale: white
 - (T) The whale was found near Japan.

Appendix E: The Experimental Items Used in Experiment 1 (Native Japanese Speakers)

Practice Session

Critical Sentences (Three Sentences)

- 彼女はマスクをするのが好きではなかった。
○ マスク: red
- ケンはスープを飲むときいつもお気に入りのカップを使っていた。
○ コップ: green
- 子どもたちは新しいコンピュータを見て嬉しそうだった。
○ コンピュータ: white

Filler Sentences (Two Sentences)

- その鳥は翼が折れていたのでは飛べなかった。
○ 鳥: white
■ (F) その鳥は足を折ってしまったので飛べなかった。
- マットは友人がビールを2本飲み終わる前に5本のビールを飲んでいて。
○ ビール: brown
■ (F) マットの友人は、マットがビールを2本飲み終わる前に5本のビールを飲んでいて。

Main Session

Critical Sentences (Before) (Ninety Sentences)

- マークはすぐに食べられそうなイチゴを買ってきた。
- マークはまだ食べられそうにはないイチゴを買ってきた。
○ イチゴ: red (Typical), green (Atypical), brown (Unrelated)
- ジョーは森の中でクマを見て興奮した。
- ジョーは北極でクマを見て興奮した。
○ クマ: brown (Typical), white (Atypical), green (Unrelated)
- 先生は草むらでカモフラージュしているカメレオンに気付いた。
- 先生は砂の中でカモフラージュしているカメレオンに気付いた。
○ カメレオン: green (Typical), brown (Atypical), white (Unrelated)
- サムは自分の馬に乗るのが好きだった。
- サムは王子様の馬に乗るのが好きだった。
○ 馬: brown (Typical), white (Atypical), red (Unrelated)
- サラは木の前で立ち止まって葉っぱを摘み取った。
- サラは地面に座って葉っぱを拾い上げた。
○ 葉っぱ: green (Typical), brown (Atypical), white (Unrelated)
- ジョンはレストランでステーキを見た。

- ジョンは精肉店でステーキを見た。
 - ステーキ: brown (Typical), red (Atypical), green (Unrelated)
- 食べごろだったのでジェーンはトマトを食べた。
- 食べごろになる前にジェーンはトマトを食べた。
 - トマト: red (Typical), green (Atypical), white (Unrelated)
- マリーはおにぎりの中に梅が入っているのに気付いた。
- マリーは木に実っている梅に気付いた。
 - 梅: red (Typical), green (Atypical), brown (Unrelated)
- リズは畑からたまねぎを取ってきた。
- リズは鍋からたまねぎを取り出した。
 - たまねぎ: brown (Typical), white (Atypical), red (Unrelated)
- レイは野球場へボールを持って出かけた。
- レイはバスケットコートへボールを持って出かけた。
 - ボール: white (Typical), brown (Atypical), green (Unrelated)
- リンは結婚式のためにケーキを頼んだ。
- リンはバレンタインデーのためにケーキを頼んだ。
 - ケーキ: white (Typical), brown (Atypical), green (Unrelated)
- 果樹園に行った時、エイミーはリンゴを食べた。
- 風邪をひいた時、エイミーはリンゴを食べた。
 - リンゴ: red (Typical), white (Atypical), brown (Unrelated)
- ベンはしょっぱいとポップコーンを口に入れた時に感じた。
- ベンは甘いとおポップコーンを口に入れた時に感じた。
 - ポップコーン: white (Typical), brown (Atypical), red (Unrelated)
- パフェの底で、ロイはキウイを見つけた。
- かごの底で、ロイはキウイを見つけた。
 - キウイ: green (Typical), brown (Atypical), red (Unrelated)
- クレアはきれいな夏の空だと雲を見て感じた。
- クレアはきれいな夕焼けだと雲を見て感じた。
 - 雲: white (Typical), red (Atypical), green (Unrelated)

Critical Sentences (After) (Ninety Sentences)

- マークが買ってきたイチゴはすぐに食べられそうだった。
- マークが買ってきたイチゴはまだ食べられそうになかった。
 - イチゴ: red (Typical), green (Atypical), brown (Unrelated)
- ジョーはクマを森の中で見て興奮した。
- ジョーはクマを北極で見て興奮した。
 - クマ: brown (Typical), white (Atypical), green (Unrelated)
- 先生はカメレオンが草むらでカモフラージュしているのに気付いた。
- 先生はカメレオンが砂の中でカモフラージュしているのに気付いた。
 - カメレオン: green (Typical), brown (Atypical), white (Unrelated)
- サムは馬が好きで特に自分が乗るのが好きだった。

- サムは馬が好きで特に王子様が乗るのが好きだった。
 - 馬: brown (Typical), white (Atypical), red (Unrelated)
- サラは森の中で立ち止まって葉っぱを木から摘み取った。
- サラは森の中で立ち止まって葉っぱを地面から拾い上げた。
 - 葉っぱ: green (Typical), brown (Atypical), white (Unrelated)
- ジョンはステーキを自分の皿の上で見た。
- ジョンはステーキを精肉店で見た。
 - ステーキ: brown (Typical), red (Atypical), green (Unrelated)
- ジェーンがトマトを食べたのはそれが食べごろだったからだ。
- ジェーンはトマトを食べたがそれは食べごろになる前だった。
 - トマト: red (Typical), green (Atypical), white (Unrelated)
- マリーは梅がおにぎりの中に入っているのに気付いた。
- マリーは梅が木に実っていることに気付いた。
 - 梅: red (Typical), green (Atypical), brown (Unrelated)
- リズはたまねぎを畑から取ってきた。
- リズはたまねぎを鍋から取り出した。
 - たまねぎ: brown (Typical), white (Atypical), red (Unrelated)
- レイはボールを持って野球場へ出かけた。
- レイはボールを持ってバスケットコートへ出かけた。
 - ボール: white (Typical), brown (Atypical), green (Unrelated)
- リンはケーキを結婚式のために頼んだ。
- リンはケーキをバレンタインデーのために頼んだ。
 - ケーキ: white (Typical), brown (Atypical), green (Unrelated)
- エイミーはリンゴを果樹園に行った時に食べた。
- エイミーはリンゴを風邪をひいた時に食べた。
 - リンゴ: red (Typical), white (Atypical), brown (Unrelated)
- ベンはポップコーンを口に入れた時にしょっぱさを感じた。
- ベンはポップコーンを口に入れた時に甘さを感じた。
 - ポップコーン: white (Typical), brown (Atypical), red (Unrelated)
- ロイはキウイをパフェの底で見つけた。
- ロイはキウイをかごの底で見つけた。
 - キウイ: green (Typical), brown (Atypical), red (Unrelated)
- クレアは雲を見てきれいな夏の空だと感じた。
- クレアは雲を見てきれいな夕焼けだと感じた。
 - 雲: white (Typical), red (Atypical), green (Unrelated)

Filler Sentences (Before) (Ninety Sentences)

- 彼の両親は京都に飛行機で行った。
 - 飛行機: brown
 - (F) 彼らはヨーロッパに行った。
- ジョンは市場で買ったバナナを食べていた。

- バナナ: red
 - (F) ジョンはクッキーを食べた。
- 彼の娘はよく浜辺に、美しい月を見るために行っていた。
 - 月: green
 - (F) 彼の娘は山によく行っていた。
- アーロンは眠る前にベッドでたばこを吸った。
 - ベッド: white
 - (F) アーロンは公園でたばこを吸った。
- ダンはパンにバターを塗るのが好きだった。
 - バター: brown
 - (F) ダンはチョコレートにパンに塗るのが好きだった。
- その少女は8月8日に本を図書館に返さなければならなかった。
 - 本: red
 - (F) その少女は本を返さなくてもよかった。
- 一人で家にいなければならなかったので少年は心配だった。
 - 家: green
 - (T) その少年は家にいなければならなかった。
- ジョージはとても高そうなチョコレートを買った。
 - チョコレート: red
 - (F) ジョージは本を何冊か買った。
- ノアは日本のコーヒーの歴史に興味があった。
 - コーヒー: red
 - (T) ノアはコーヒーの歴史に興味があった。
- 子どもたちは枝にカマキリがいるのを見つけた。
 - カマキリ: red
 - (T) 子どもたちはカマキリを見つけた。
- 子どもたちは動物園でネコを観察した。
 - ネコ: red
 - (T) 子どもたちはネコを観察した。
- エリカは子どもの頃夕食に必ずシチューを食べていた。
 - シチュー: white
 - (F) エリカはいつもパスタを食べていた。
- ローガンは朝、サラダを受け取るためにバーに立ち寄った。
 - サラダ: red
 - (F) ローガンは酒を手に入れた。
- エマはおいしい夕食を作りたかったが鶏肉を長く煮込みすぎてしまった。
 - 鶏肉: red
 - (T) エマは夕食に鶏肉を調理した。
- メイはキッチンを後にする前にアボカドを鍋に入れた。
 - アボカド: green
 - (F) メイは牛肉を鍋に入れた。

- マイロは初めて米を味わったが酢飯は苦手だった。
 - 米: red
 - (T) マイロは米を食べた。
- 寒空の下、家の外にある梨がおいしそうに見えるとテッドは思った。
 - 梨: red
 - (F) テッドは友達のことを考えた。
- ニックは公園でアイスクリームを食べるのが好きだった。
 - アイスクリーム: white
 - (F) ニックはアイスを食べるのが嫌いだった。
- デイビスは冷蔵庫から野菜を取り出した。
 - 野菜: green
 - (F) デイビスは鶏肉を取り出した。
- シモンは畑でスイカを見た。
 - スイカ: green
 - (F) シモンはネコを見た。
- 夏、窓の外に見える山が美しいなとポーラは思って見ていた。
 - 山: green
 - (F) ポーラはその山が不格好だと思った。
- ベンはそのまま歩き続けたが、それは信号機を確認した後だった。
 - 信号機: green
 - (F) ベンは信号機で止まった。
- ベスは化粧をするためにリップスティックを塗った。
 - リップスティック: red
 - (F) ベスはリップスティックを持っていなかった。
- 冬にポージーはお気に入りの木をよく見に行った。
 - 木: white
 - (F) ポージーは木が嫌いだった。
- スーザンは孫娘がアップスタイルの髪にしている時の方が好きだった。
 - 髪: brown
 - (F) スーザンは髪をアップスタイルにしている母が好きだった。
- マイクは日本に行った時お茶を買った。
 - お茶: white
 - (F) マイクは日本でコーヒーを買った。
- その子どもは友達と家にいることにした。
 - 家: red
 - (F) その子どもは外出することにした。
- ロバートはレストランでウェーターに領収書を持ってくるように頼んだ。
 - 領収書: red
 - (F) ロバートは大学にいた。
- 彼のガールフレンドは駅で財布をなくした。
 - 財布: red

- (F) 彼のガールフレンドは財布を見つけた。
- その子どもたちは外でカウボーイごっこをした。
 - カウボーイ: red
 - (F) 大人たちは外で遊んでいた。
- 毎週末、彼の父親は車を運転するのを楽しんでいた。
 - 車: red
 - (F) 父親は平日に運転を楽しんでいた。
- アリセは駅で新聞紙を買った。
 - 新聞紙: red
 - (F) アリセはサンドイッチを買った。
- 彼女はかごの中になすびがあるのを見つけた。
 - なすび: red
 - (F) その卵はかごの中にあった。
- エイミーは公園で足の筋肉を痛めた。
 - 筋肉: red
 - (T) エイミーは足の筋肉を痛めた。
- アニカはテーブルの上に犬が座っているのを見た。
 - 犬: red
 - (T) アニカは犬を見た。
- ベラは夕食後に皿を洗った。
 - 皿: red
 - (T) ベラは皿を夕食後に洗った。
- 彼は会社でコンピューターを使っていた。
 - コンピューター: red
 - (T) 彼はコンピューターを使っていた。
- バレットは歯医者で歯ブラシを買った。
 - 歯ブラシ: red
 - (T) バレットは歯ブラシを歯医者で買った。
- ベッキーはそのお店にお気に入りのパジャマを売った。
 - パジャマ: red
 - (T) ベッキーは気に入っていたパジャマを売った。
- ベニーは1927年に初めてそのレストランを友人たちと訪れた。
 - レストラン: red
 - (T) ベニーは1927年に初めてそのレストランを訪れた。
- バーニーは夏の終わりに美しい花の写真をたくさん撮った。
 - 花: brown
 - (F) バーニーは車の写真を撮った。
- 先週、ブレットは高そうなイスを買った。
 - イス: brown
 - (F) ブレットは先週イスを売った。
- 彼女は家族のためにチョコレートをかけたタルトを焼いた。

- タルト: brown
 - (F) 彼女はケーキを焼いた。
- 彼らはその年が終わる前に新しい道路を完成させた。
 - 道路: brown
 - (F) その道路は年が終わる前に完成していなかった。
- 居酒屋で、ブライアンはお気に入りのビールを注文した。
 - ビール: brown
 - (F) ブライアンはコーヒーを注文した。
- 昨年、ケイルが欲しがっていた自転車は完売していた。
 - 自転車: brown
 - (F) ケイルは自転車が欲しくなかった。
- ケイリーは港で戦艦を見つけたとき驚いた。
 - 戦艦: brown
 - (T) ケイリーは港で戦艦を見た。
- 朝食を食べた後、男は駐車場へ急いで向かった。
 - 駐車場: brown
 - (T) 男は朝食を食べた後で駐車場へ向かった。
- 2005年にチャドはある有名な事務所を訪れた。
 - 事務所: brown
 - (T) チャドはある有名な事務所を2005年に訪れた。
- 朝、チャックは牛乳を買うためにカフェに立ち寄った。
 - 牛乳: brown
 - (T) チャックは朝、牛乳を手に入れた。
- クリントは町の中で地図を開いて、とある場所を探した。
 - 地図: brown
 - (T) クリントはある場所を探していた。
- ディーンはフランスから取り寄せたワインを試飲した。
 - ワイン: brown
 - (T) ディーンはワインを試飲した。
- ドビーは箱の中にダイヤモンドがあるのに気づいた。
 - ダイヤモンド: brown
 - (T) ドビーはダイヤモンドに気づいた。
- エディはキッチンからナイフを持ってきた。
 - ナイフ: brown
 - (T) エディはナイフをキッチンから持ってきた。
- 忙しい一週間を終えてフランクは雑誌を買った。
 - 雑誌: brown
 - (T) フランクは雑誌を買った。
- クリスマスパーティーのためにゲイブはピザを注文した。
 - ピザ: brown
 - (T) ゲイブはピザを注文した。

- ハンクは学生時代スニーカーを集めるのが好きだった。
 - スニーカー: green
 - (T) ハンクはスニーカーを集めるのが好きだった。
- ハルはナイフで缶を開けた。
 - 缶: green
 - (T) ハルは缶をナイフで開けた。
- ヒラリーは公園でミカンを見つけた。
 - ミカン: green
 - (T) ヒラリーはミカンを見つけた。
- インディアはショッピングモールで俳優を見かけた。
 - 俳優: green
 - (T) インディアは俳優をショッピングモールで見た。
- 牧草地にたくさんの羊がいた。
 - 羊: green
 - (T) たくさんの羊が牧草地にいた。
- 勉強するときにナナはラジオをいつも聞いていた。
 - ラジオ: green
 - (T) ナナはラジオを聞いていた。
- プリンを作る際には質の良い砂糖を選ぶことがとても重要だった。
 - 砂糖: green
 - (T) 砂糖はプリンを作るのに重要だった。
- ライアンは入り口で映画のチケットを見せた。
 - チケット: green
 - (T) ライアンは映画のチケットを持っていた。
- マギーは宇宙から地球を見たかった。
 - 地球: green
 - (T) マギーは地球を見たかった。
- バンドで 10 歳の少女がキーボードを弾いていた。
 - キーボード: green
 - (T) 少女はバンドでキーボードを弾いていた。
- 大雪によりパットは電車の中で一日中過ごすはめになった。
 - 電車: green
 - (T) パットは電車の中で過ごすことになった。
- 彼らは毎朝水を井戸からくんでいた。
 - 水: green
 - (F) 彼らはお店で水を買った。
- ジェイクは小さなお店でペンを買った。
 - ペン: green
 - (F) ジェイクは定規を買った。
- 1950 年代にはほとんどの子どもたちが人形を持ちそして遊んでいた。
 - 人形: green

- (F) 人形は 1950 年代に人気ではなかった。
- その子どもは誕生日プレゼントとして小さな箱を渡された。
 - 箱: green
 - (F) その子どもは友達にプレゼントを渡した。
- 1960 年代にはテレビは子どもにとって良くないという意見があった。
 - テレビ: green
 - (F) 昔、すべての人はテレビが子どもにとって良いと考えていた。
- ジャッキーは大学時代ピアノを習うのに沢山のお金を払った。
 - ピアノ: white
 - (T) ジャッキーはピアノのレッスンに大金を払った。
- 昨日、ジェフはドアを閉めるように言われた。
 - ドア: white
 - (T) ジェフはドアを閉めるように昨日言われた。
- 一日中その窓は開いたままになっていた。
 - 窓: white
 - (T) その窓は一日中開いていた。
- 先週、ジェドを襲ったカラスがようやく捕まった。
 - カラス: white
 - (T) カラスがジェドを襲った。
- 今日、その学生はレポートを提出することになっていた。
 - レポート: white
 - (T) その学生はレポートを提出しなければならなかった。
- 夜にその像を見ると子どもたちはいつも怖がっていた。
 - 像: white
 - (T) その像は子どもたちを怖がらせていた。
- ケーシーはそれでコーヒーが買えるのでコインをもらって喜んだ。
 - コイン: white
 - (T) ケーシーはコインを貰ったとき嬉しかった。
- 髪が長いのでリージーはくしを愛用していた。
 - くし: white
 - (T) リージーは髪が長かった。
- ある俳優が記者会見で言及したためその工場は有名になった。
 - 工場: white
 - (T) その工場は俳優のおかげで有名になった。
- とてもリラックスできるのでニールは炎を見るのが好きだった。
 - 炎: white
 - (T) ニールは炎を見るのが好きだった。
- 昔は、毛皮のコートはプレゼントとして人気だった。
 - 毛皮のコート: white
 - (T) 昔、毛皮のコートは贈り物として人気だった。
- 日本からある幽霊の話が伝わった。

- 幽霊: white
 - (F) ある話が韓国から伝わった。
- マーシーはパスタに使うため、山にきのこを探しに行った。
 - きのこ: white
 - (F) マーシーはスマートフォンを探していた。
- スキンクリームを伸ばすために、彼女の母親はいつもコットンを使っていた。
 - コットン: white
 - (F) 彼女の母はスキンクリームを持っていなかった。
- その教授は実験でハトを使用した。
 - ハト: white
 - (F) ラットが実験で使われた。
- ライラは狩りに持っていくために銃を探していた。
 - 銃: white
 - (F) ライラはナイフを探していた。
- 彼の父親は愛犬のためにせっけんを買ってきた。
 - せっけん: white
 - (F) 彼の父親はチョコレートを買った。
- 日本の沖合でそのクジラは発見された。
 - クジラ: white
 - (F) 日本付近でイルカが見つかった。

Filler Sentences (After) (Ninety Sentences)

- 彼の両親は飛行機で京都に行った。
 - 飛行機: brown
 - (T) 彼らは京都に行った。
- ジョンはバナナを市場で買って食べていた。
 - バナナ: red
 - (T) ジョンはバナナを食べた。
- 彼の娘は美しい月を見るためによく浜辺に行っていた。
 - 月: green
 - (T) 彼の娘は昔よく浜辺に行っていた。
- アーロンはベッドで、眠る前にたばこを吸った。
 - ベッド: white
 - (T) アーロンはベッドでたばこを吸った。
- ダンはバターをパンに塗るのが好きだった。
 - バター: brown
 - (T) ダンはパンにバターを塗るのが好きだった。
- その少女は本を8月8日に図書館へ返さなければならなかった。
 - 本: red
 - (T) その少女は本を返却しなければならなかった。
- 少年は家に一人でいなければならなかったため心配だった。

- 家: green
 - (F) 男の子は近所に行かなければならなかった。
- ジョージが買ったチョコレートはとても高そうだった。
 - チョコレート: red
 - (T) ジョージはチョコレートを買った。
- ノアはコーヒーの日本における歴史に興味があった。
 - コーヒー: red
 - (F) ノアはサッカーに興味があった。
- 子どもたちはカマキリが枝にいるのを見つけた。
 - カマキリ: red
 - (F) 子どもたちはカマキリを殺した。
- 子どもたちはネコを動物園で観察した。
 - ネコ: red
 - (F) 子どもたちはゾウを観察した
- エリカはシチューを子どもの頃、必ず夕食で食べていた。
 - シチュー: white
 - (T) エリカは昔、よくシチューを食べていた。
- ローガンはサラダを受け取るためにバーへ朝、立ち寄った。
 - サラダ: red
 - (T) ローガンはサラダを受け取った。
- エマは鶏肉を長く煮込みすぎたが、本来はおいしい夕食を作りたかったのだ。
 - 鶏肉: red
 - (F) エマは夕食に牛肉を調理した。
- メイはアボカドを鍋に入れキッチンを後にした。
 - アボカド: green
 - (T) メイはアボカドを鍋に入れた。
- マイロは米を初めて味わったが酢飯は苦手だった。
 - 米: red
 - (F) マイロは米を食べなかった。
- テッドは家の外にある梨が寒空の下で美味しそうに見えると思った。
 - 梨: red
 - (T) テッドは梨について考えた。
- ニックはアイスクリームを公園で食べるのが好きだった。
 - アイスクリーム: white
 - (T) ニックはアイスクリームを食べるのが好きだった。
- デイビスは野菜を冷蔵庫から取り出した。
 - 野菜: green
 - (T) デイビスは野菜を取った。
- シモンはスイカを畑を見た。
 - スイカ: green
 - (T) シモンはスイカを見た。

- ポーラは窓の外に見える山が、夏には美しく見えると思っていた。
 - 山: green
 - (T) ポーラは山が美しいと思った。
- ベンは信号機を確認してそのまま歩き続けた。
 - 信号機: green
 - (T) ベンは歩き続けた。
- ベスはリップスティックを化粧をするために塗った。
 - リップスティック: red
 - (T) ベスはリップスティックを塗った。
- ポージーはお気に入りの木を冬によく見に行った。
 - 木: white
 - (T) ポージーにはお気に入りの木がある。
- スーザンは孫娘が髪を結んでいる方が好きだった。
 - 髪: brown
 - (T) スーザンは孫娘の髪を結んだ姿が好きだった。
- マイクはお茶を日本に行った時に買った。
 - お茶: white
 - (T) マイクは日本でお茶を買った。
- その子どもは家に友達といることにした。
 - 家: red
 - (T) その子どもは家にいた。
- ロバートはウェ이터に領収書を持ってくるようにレストランで頼んだ。
 - 領収書: red
 - (T) ロバートはレストランにいた。
- 彼のガールフレンドは財布を駅でなくした。
 - 財布: red
 - (T) 彼のガールフレンドは財布をなくした。
- 子どもたちはカウボーイごっこを外でやった。
 - カウボーイ: red
 - (T) 子どもたちは外で遊んでいた。
- 彼の父親は車を毎週末、運転するのを楽しんでいた。
 - 車: red
 - (T) 彼の父親は毎週末、運転を楽しんでいた。
- アリセは新聞紙を駅で買った。
 - 新聞紙: red
 - (T) アリセは新聞紙を買った。
- 彼女はなすびがかごの中にあるのを見つけた。
 - なすび: red
 - (T) なすびはかごの中にあった。
- エイミーは足の筋肉を公園で痛めた。
 - 筋肉: red

- アニカは犬がテーブルの上に座っているのを見た。
 - 犬: red
 - (F) エイミーは腕の筋肉を痛めた。
 - ベラは皿を夕食後に洗った。
 - 皿: red
 - (F) アニカは猫がテーブルの上に座っているのを見た。
 - 彼はコンピューターを会社で使っていた。
 - コンピューター: red
 - (F) ベラは朝食後に皿を洗った。
 - バレットは歯ブラシを歯医者で買った。
 - 歯ブラシ: red
 - (F) 彼はコンピューターを使うことができなかった。
 - ベッキーはお気に入りのパジャマをそのお店に売った。
 - パジャマ: red
 - (F) バレットはスーパーマーケットで歯ブラシを買った。
 - ベニーが友人たちと初めてこのレストランを訪れたのは 1927 年のことだった。
 - レストラン: red
 - (F) ベッキーは気に入ったパジャマを買った。
 - バーニーは美しい花の写真を夏の終わりにたくさん撮った。
 - 花: brown
 - (F) ベニーが初めてレストランを開いたのは 1927 年だった。
 - ブレットは高そうなイスを先週買った。
 - イス: brown
 - (T) バーニーは花の写真を撮った。
 - 彼女はチョコレートがかかったタルトを家族のために焼いた。
 - タルト: brown
 - (T) ブレットは先週イスを買った。
 - 彼らは新しい道路をその年が終わる前に完成させた。
 - 道路: brown
 - (T) そのタルトにはチョコレートがかけられていた。
 - ブライアンはお気に入りのビールを居酒屋で注文した。
 - ビール: brown
 - (T) 道路はその年が終わる前に完成した。
 - ケイルが欲しがっていた自転車は昨年完売した。
 - 自転車: brown
 - (T) ブライアンはビールを注文した。
 - ケイリーは戦艦を港で見つけたとき驚いた。
 - 戦艦: brown
 - (T) 自転車は昨年完売した。
 - 男は駐車場へ、朝食を食べた後急いで向かった。
 - 戦艦: brown
 - (F) ケイリーは博物館で戦艦を見た。

- 駐車場: brown
 - (F) 男は朝食を食べる前に駐車場へ行った。
- チャドはある有名な事務所を 2005 年に訪れた。
 - 事務所: brown
 - (F) チャドは 2020 年に有名な事務所を訪れた。
- チャックは牛乳を買うために朝、カフェに立ち寄った。
 - 牛乳: brown
 - (F) チャックは夕方に牛乳を買った。
- クリントは地図を町の中で開いて、とある場所を探した。
 - 地図: brown
 - (F) クリントは警察官に頼んでとある場所を探してもらった。
- ディーンはワインを試飲したがそれはフランスから取り寄せたものだった。
 - ワイン: brown
 - (F) ディーンはチョコレートバーを試食した。
- ドビーはダイヤモンドが箱の中にあるのに気づいた。
 - ダイヤモンド: brown
 - (F) ドビーは猫に気づいた。
- エディはナイフをキッチンから持ってきた。
 - ナイフ: brown
 - (F) エディはキッチンからフォークを取ってきた。
- フランクは雑誌を、忙しい一週を終えた後に買った。
 - 雑誌: brown
 - (F) フランクは新聞紙を買った。
- ゲイブはピザをクリスマスパーティーのために注文した。
 - ピザ: brown
 - (F) ゲイブはケーキを注文した。
- ハンクはスニーカーを集めるのが学生時代好きだった。
 - スニーカー: green
 - (F) ハンクは上着を集めるのが好きだった。
- ハルは缶をナイフで開けた。
 - 缶: green
 - (F) ハルはナイフで段ボールを開けた。
- ヒラリーはミカンを公園で見つけた。
 - ミカン: green
 - (F) ヒラリーはイチゴを見つけた。
- インディアは俳優をショッピングモールで見かけた。
 - 俳優: green
 - (F) インディアはショッピングモールで友人を見かけた。
- たくさんの羊が牧草地にいた。
 - 羊: green
 - (F) 牧草地にはたくさんの牛がいた。

- ナナはラジオを、勉強するときにもいつも聞いていた。
 - ラジオ: green
 - (F) ナナは勉強するときにはラジオを聞かなかった。
- 質の良い砂糖を選ぶことがプリンを作る際にとっても重要だった。
 - 砂糖: green
 - (F) そのプリンに砂糖は重要ではなかった。
- ライアンは映画のチケットを入り口で見せた。
 - チケット: green
 - (F) ライアンは動物園のチケットを持っていた。
- マギーは地球を宇宙から見たかった。
 - 地球: green
 - (F) マギーは地球の写真を持っていた。
- 10歳の少女がキーボードをバンドで弾いていた。
 - キーボード: green
 - (F) 女の子はバンドでギターを弾いていた。
- パットは電車の中で、大雪により一日中過ごすはめになった。
 - 電車: green
 - (F) パットはホテルに泊まるはめになった。
- 彼らは水を毎朝井戸からくんでいた。
 - 水: green
 - (T) 彼らは井戸から水をくんでいた。
- ジェイクはペンを小さなお店で買った。
 - ペン: green
 - (T) ジェイクはペンを買った。
- ほとんどの子どもたちが人形を持ちそれで遊んでいたのは1950年代だった。
 - 人形: green
 - (T) 1950年代にはほとんどの子どもが人形を所有していた。
- その子どもは小さな箱を誕生日プレゼントとして渡された。
 - 箱: green
 - (T) その子どもは誕生日プレゼントをもらった。
- テレビは子どもにとって良くないという意見が1960年代にあった。
 - テレビ: green
 - (T) テレビは子どもに悪影響だと考える人も過去にはいた。
- ジャッキーはピアノを習うのに大学時代沢山のお金を払った。
 - ピアノ: white
 - (F) ジャッキーはドラムのレッスンに大金を費やした。
- ジェフはドアを閉めるように昨日言われた。
 - ドア: white
 - (F) ジェフは昨日ドアを開けるように言われた。
- その窓は一日中開いたままになっていた。
 - 窓: white

- ジェドを襲ったカラスは先週ようやく捕まった。
 - カラス: white
 - (F) 窓は一日中閉まっていた。
 - その学生はレポートを今日提出することになっていた。
 - レポート: white
 - (F) ジェドはカラスを襲った。
 - その像を夜に見ると子どもたちはいつも怖がった。
 - 像: white
 - (F) その学生は何も提出するものがなかった。
 - ケーシーはコインをもらって、コーヒーが買えるからと喜んでいて。
 - コイン: white
 - (F) 子どもたちはその像を気に入っていた。
 - リージーはくしを、髪が長いので愛用していた。
 - くし: white
 - (F) ケーシーはコインを受け取った時悲しかった。
 - その工場が有名になったのはある俳優が記者会見でそのことを口にしたからだ。
 - 工場: white
 - (F) リージーの髪は短かった。
 - ニールが炎を見るのが好きだったのは、とてもリラックスできるからだった。
 - 炎: white
 - (F) その工場は以前から有名だった。
 - 毛皮のコートは、昔はプレゼントとして人気だった。
 - 毛皮のコート: white
 - (F) ニールは海を見るのが好きだった。
 - ある幽霊の話が日本から伝わった。
 - 幽霊: white
 - (F) 昔、毛皮のコートはプレゼントとして人気ではなかった。
 - マーシーが山にきのこを探しに行ったのはパスタに使うためだった。
 - きのこと: white
 - (T) ある話が日本から伝わった。
 - 彼女の母親はいつもコットンでスキนครリームを伸ばしていた。
 - コットン: white
 - (T) マーシーはきのこが欲しかった。
 - その教授はハトを実験で使用した。
 - ハト: white
 - (T) 母親はスキนครリームを使っていた。
 - ライラは銃を、狩りに持っていくために探していた。
 - 銃: white
 - (T) 実験にハトが使われた。
 - 父親はせっけんを愛犬のために買ってきた。
 - 銃: white
 - (T) ライラは銃を探していた。

- せっけん: white
 - (T) 父親はせっけんを買った。
- そのクジラは日本の沖合で発見された。
 - クジラ: white
 - (T) そのクジラは日本付近で見つかった。

Appendix F: The Japanese Version of the Instructions

The following instructions were provided during the semantic Stroop task to native Japanese (Experiment 1) and native Japanese speakers learning English (Experiment 2).

Practice Session

これから練習を始めます。この課題では、[Tab]、[Enter]、[Space]、[L]、[K]、[D]、[S]のキーを使います。[L]、[K]、[D]、[S]を押す速さを記録していますので、常にこの四つのキーの上に手を構えておいてください（写真）。これ以外の[Tab]、[Enter]、[Space]キーを押す時間は記録していません。[L]、[K]、[D]、[S]は以下に対応しています。[L] は 赤、[K] は 緑、[D] は 白、[S] は 茶。以上の色とキーの対応を、繰り返し練習して覚えていただきます。[Tab]、[Enter]、[Space]キーに関しては、対応は覚えなくても大丈夫です。

Main Session

初めに、画面左端に注視点（+）が1秒間提示される。次に、日本語（for Experiment 2: 英語）の文が一文提示される。文が理解できたら[Space]キーを押す。この時間は計測されていないので、自分のペースで文を読んでよい。再度、画面中央に注視点（+）が0.5秒提示される。その後、色付きの日本語（for Experiment 2: 英語）の単語が一語提示される。L = 赤、K = 緑、D = 白、S = 茶のいずれかを押し、その色が何色か答える。ボタンを押す速度を測定しているため、できるだけ速く、そして正確に答える。色のついた単語が提示された直後、文の内容理解を問う問題が出る場合がある。その場合、文が正しいかを、[Enter]キー（正しい）または[TAB]キー（正しくない）を押して答える（この時間は測定していない）。

Appendix G: Rating Task (Native English Speakers)

Word

Rating Scores

The following table summarizes the descriptive statistics of the word typicality rating task for each word. The second column represents the typicality of the combinations that were shown in the first column.

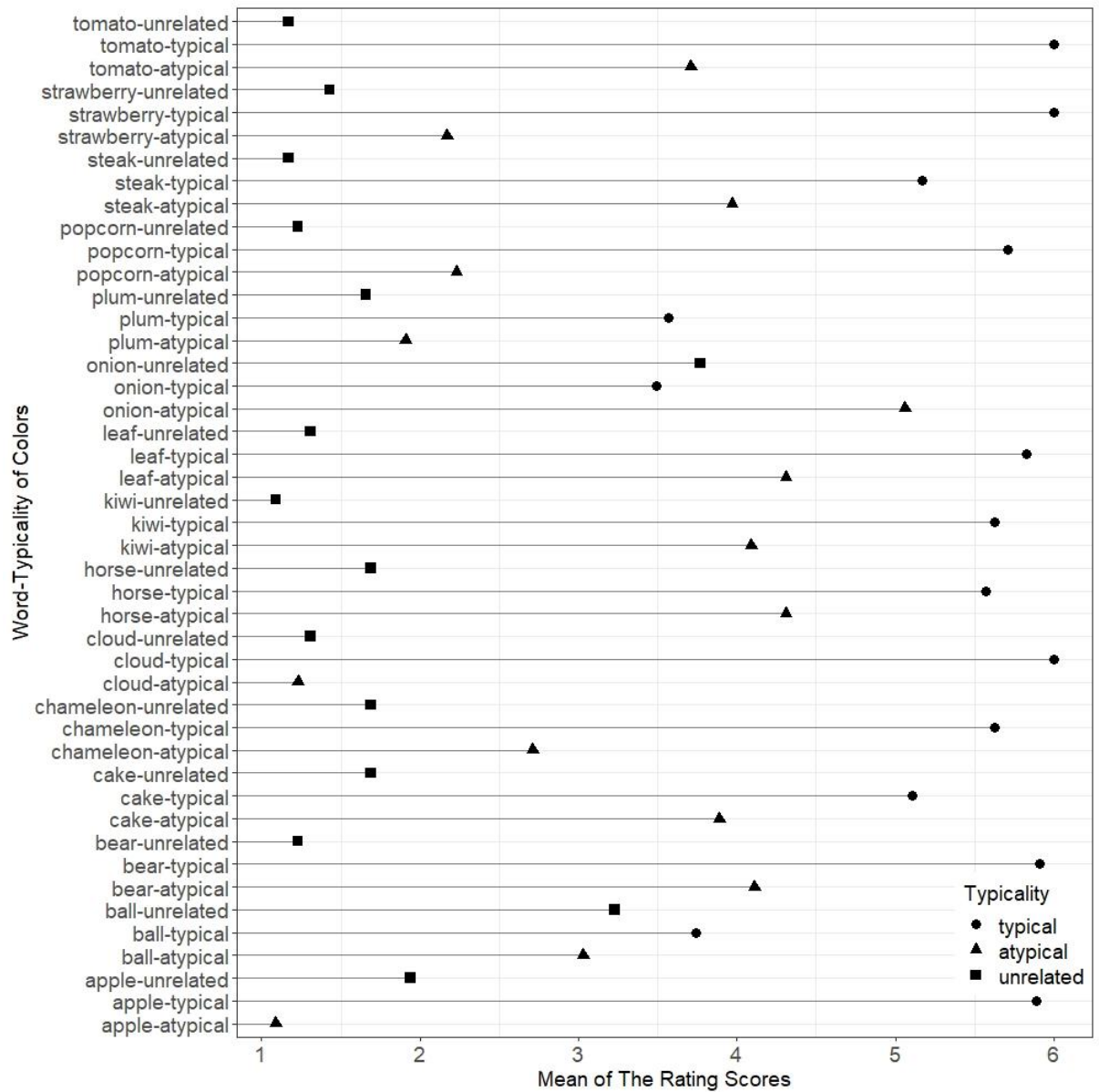
<i>Word-Color</i>	<i>Typicality</i>	<i>M</i>	<i>SD</i>	<i>Median</i>	<i>Min</i>	<i>Max</i>
apple-BROWN	unrelated	1.94	1.28	1.00	1.00	6.00
apple-RED	typical	5.89	0.40	6.00	4.00	6.00
apple-WHITE	atypical	1.09	0.28	1.00	1.00	2.00
ball-BROWN	atypical	3.03	1.48	3.00	1.00	6.00
ball-GREEN	unrelated	3.23	1.73	3.00	1.00	6.00
ball-WHITE	typical	3.74	1.72	4.00	1.00	6.00
bear-BROWN	typical	5.91	0.28	6.00	5.00	6.00
bear-GREEN	unrelated	1.23	0.84	1.00	1.00	5.00
bear-WHITE	atypical	4.11	1.32	4.00	1.00	6.00
cake-BROWN	atypical	3.89	1.69	4.00	1.00	6.00
cake-GREEN	unrelated	1.69	1.11	1.00	1.00	6.00
cake-WHITE	typical	5.11	0.96	5.00	2.00	6.00
chameleon-BROWN	atypical	2.71	1.49	2.00	1.00	6.00
chameleon-GREEN	typical	5.63	0.69	6.00	3.00	6.00
chameleon-WHITE	unrelated	1.69	1.08	1.00	1.00	5.00
cloud-GREEN	unrelated	1.31	0.76	1.00	1.00	4.00
cloud-RED	atypical	1.23	0.43	1.00	1.00	2.00
cloud-WHITE	typical	6.00	0.00	6.00	6.00	6.00
horse-BROWN	typical	5.57	0.61	6.00	4.00	6.00
horse-RED	unrelated	1.69	1.18	1.00	1.00	5.00
horse-WHITE	atypical	4.31	1.32	4.00	2.00	6.00
kiwi-BROWN	atypical	4.09	1.90	4.00	1.00	6.00
kiwi-GREEN	typical	5.63	0.65	6.00	4.00	6.00
kiwi-RED	unrelated	1.09	0.28	1.00	1.00	2.00
leaf-BROWN	atypical	4.31	1.11	4.00	1.00	6.00

<i>Word-Color</i>	<i>Typicality</i>	<i>M</i>	<i>SD</i>	<i>Median</i>	<i>Min</i>	<i>Max</i>
leaf-GREEN	typical	5.83	0.45	6.00	4.00	6.00
leaf-WHITE	unrelated	1.31	0.90	1.00	1.00	6.00
onion-BROWN	typical	3.49	1.88	4.00	1.00	6.00
onion-RED	unrelated	3.77	1.82	4.00	1.00	6.00
onion-WHITE	atypical	5.06	1.11	5.00	2.00	6.00
plum-BROWN	unrelated	1.66	1.28	1.00	1.00	6.00
plum-GREEN	atypical	1.91	1.12	2.00	1.00	4.00
plum-RED	typical	3.57	1.48	4.00	1.00	6.00
popcorn-BROWN	atypical	2.23	1.42	2.00	1.00	5.00
popcorn-RED	unrelated	1.23	0.77	1.00	1.00	5.00
popcorn-WHITE	typical	5.71	0.67	6.00	3.00	6.00
steak-BROWN	typical	5.17	1.32	6.00	1.00	6.00
steak-GREEN	unrelated	1.17	0.71	1.00	1.00	4.00
steak-RED	atypical	3.97	1.82	4.00	1.00	6.00
strawberry-BROWN	unrelated	1.43	1.01	1.00	1.00	5.00
strawberry-GREEN	atypical	2.17	1.48	2.00	1.00	6.00
strawberry-RED	typical	6.00	0.00	6.00	6.00	6.00
tomato-GREEN	atypical	3.71	1.43	4.00	1.00	6.00
tomato-RED	typical	6.00	0.00	6.00	6.00	6.00
tomato-WHITE	unrelated	1.17	0.57	1.00	1.00	4.00

Cleveland Dot Plot

In the following figure, the y-axis represents the word-typicality combinations. For example, “bear-typical” is equal to “a BROWN bear.” The x-axis represents the mean of

the word typicality rating scores. The legend represents the correspondence of the shapes of the plots and the typicality of colors.



Sentence

Agreement Rates

The following tables summarize the agreement rates (%) of each typicality that was implied by the sentences.

Entirely.

<i>Typicality</i>	<i>Agreement Rates</i>
typical	84.57
atypical	73.52

Each Sentence.

<i>Word</i>	<i>Typical</i>	<i>Atypical</i>
apple	97.14	28.57
ball	97.14	97.14
bear	97.14	94.29
cake	82.86	25.71
chameleon	80.00	88.57
cloud	77.14	82.86
horse	57.14	54.29
kiwi	94.29	82.86
leaf	85.71	60.00
onion	88.57	68.57
plum	88.57	71.43
popcorn	40.00	80.00
steak	94.29	94.29
strawberry	97.14	91.43
tomato	91.43	82.86

The atypical sentence of *cake* had the smallest agreement rate (25.7 percent), followed by the atypical sentence of *apple* (28.6 percent). These two agreement rates were much lower than the rates of the third lowest item (typical popcorn: 40.0). The reason for the relatively low rates for the two sentences is due to the number of participants who indicated that the atypical sentences could correspond to the typical or both the typical and atypical color of the objects (cake: typical = 29 percent, both = 46 percent; apple: typical = 43 percent, both = 29 percent). Thus, the reason for the relatively low rates was that the sentences implied not only the atypical image of the objects but also the typical image of the objects.

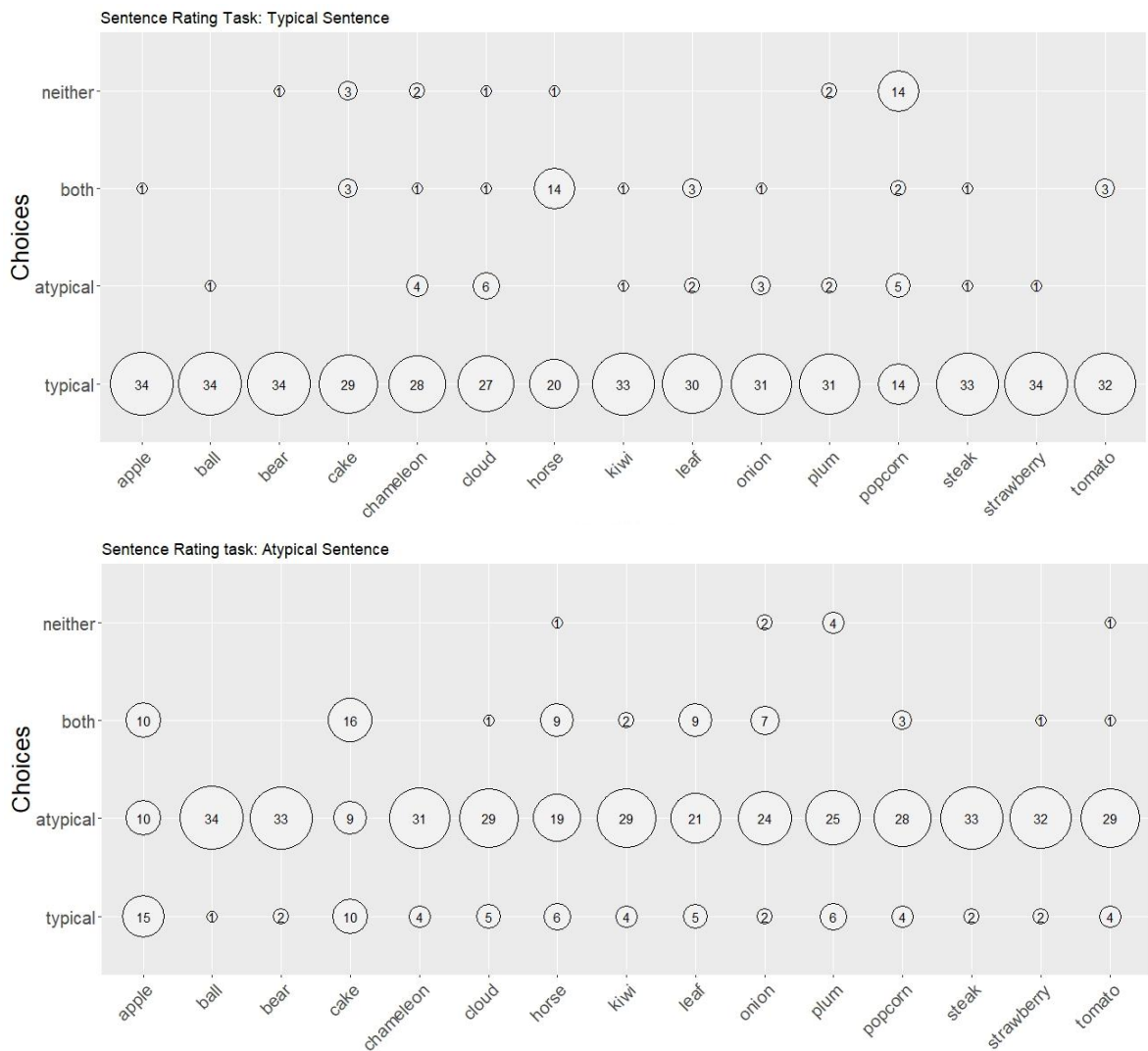
Balloon Plot

The test sentences were presented with two pictures and four forced choice

alternatives:

- typical: best matched by the first picture (the first pictures were always typical objects)
- atypical: best matched by the second picture (the second pictures were always atypical objects)
- both: matched by both pictures equally
- neither: matched by neither picture

The numbers in each balloon refers to the number of the participants who selected the choice.



Appendix H: Rating Task (Native Japanese Speakers)

Word

Rating Scores

The following table summarizes the descriptive statistics of the word typicality rating task for each word. The second column represents the typicality of the combinations that were shown in the first column.

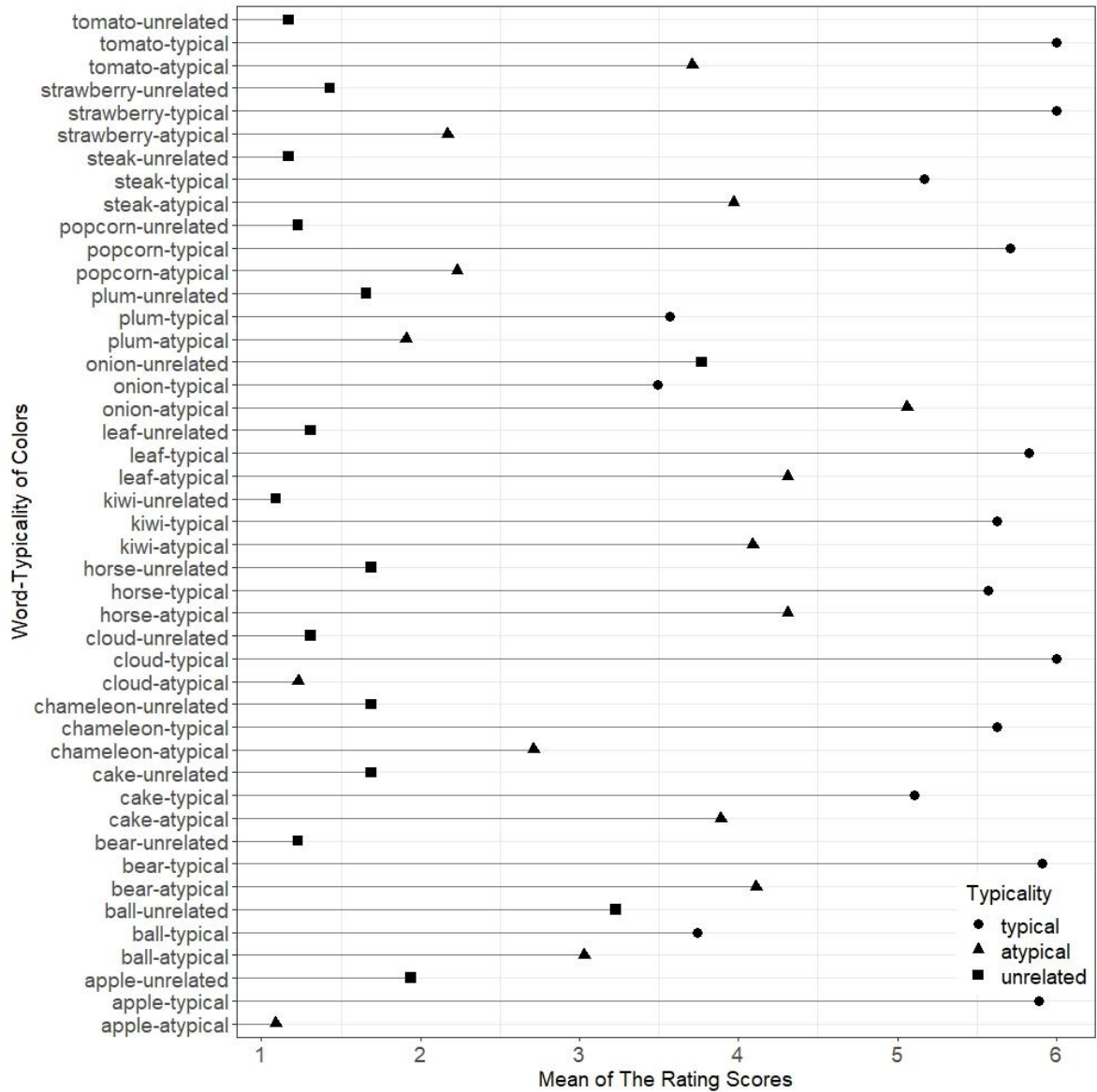
<i>Word-Color</i>	<i>Typicality</i>	<i>M</i>	<i>SD</i>	<i>Median</i>	<i>Min</i>	<i>Max</i>
apple-BROWN	unrelated	1.86	0.93	2.00	1.00	5.00
apple-RED	typical	5.97	0.17	6.00	5.00	6.00
apple-WHITE	atypical	1.72	1.14	1.00	1.00	5.00
ball-BROWN	atypical	3.22	1.53	3.00	1.00	6.00
ball-GREEN	unrelated	2.67	1.26	3.00	1.00	6.00
ball-WHITE	typical	5.17	1.30	6.00	2.00	6.00
bear-BROWN	typical	5.69	0.67	6.00	3.00	6.00
bear-GREEN	unrelated	1.31	0.79	1.00	1.00	5.00
bear-WHITE	atypical	4.50	1.30	5.00	1.00	6.00
cake-BROWN	atypical	4.28	1.32	4.50	1.00	6.00
cake-GREEN	unrelated	2.17	1.38	2.00	1.00	6.00
cake-WHITE	typical	5.53	0.88	6.00	2.00	6.00
chameleon-BROWN	atypical	3.22	1.48	3.00	1.00	6.00
chameleon-GREEN	typical	5.94	0.23	6.00	5.00	6.00
chameleon-WHITE	unrelated	2.56	1.18	2.50	1.00	5.00
cloud-GREEN	unrelated	1.17	0.45	1.00	1.00	3.00
cloud-RED	atypical	2.56	1.52	2.50	1.00	5.00
cloud-WHITE	typical	6.00	0.00	6.00	6.00	6.00
horse-BROWN	typical	5.92	0.28	6.00	5.00	6.00
horse-RED	unrelated	1.75	1.34	1.00	1.00	5.00
horse-WHITE	atypical	4.56	1.03	5.00	2.00	6.00
kiwi-BROWN	atypical	4.17	1.56	4.00	1.00	6.00
kiwi-GREEN	typical	5.64	0.68	6.00	4.00	6.00
kiwi-RED	unrelated	1.08	0.28	1.00	1.00	2.00
leaf-BROWN	atypical	4.78	0.87	5.00	3.00	6.00

<i>Word-Color</i>	<i>Typicality</i>	<i>M</i>	<i>SD</i>	<i>Median</i>	<i>Min</i>	<i>Max</i>
leaf-GREEN	typical	5.94	0.23	6.00	5.00	6.00
leaf-WHITE	unrelated	1.94	1.07	2.00	1.00	5.00
onion-BROWN	typical	5.00	1.39	5.50	1.00	6.00
onion-RED	unrelated	2.28	1.39	2.00	1.00	5.00
onion-WHITE	atypical	4.06	1.71	5.00	1.00	6.00
plum-BROWN	unrelated	2.17	1.13	2.00	1.00	5.00
plum-GREEN	atypical	3.53	1.65	4.00	1.00	6.00
plum-RED	typical	5.78	0.48	6.00	4.00	6.00
popcorn-BROWN	atypical	3.67	1.47	4.00	1.00	6.00
popcorn-RED	unrelated	1.67	0.96	1.00	1.00	4.00
popcorn-WHITE	typical	5.78	0.59	6.00	3.00	6.00
steak-BROWN	typical	5.72	0.74	6.00	3.00	6.00
steak-GREEN	unrelated	1.00	0.00	1.00	1.00	1.00
steak-RED	atypical	4.22	1.22	4.00	2.00	6.00
strawberry-BROWN	unrelated	1.50	1.00	1.00	1.00	6.00
strawberry-GREEN	atypical	3.08	1.48	3.00	1.00	6.00
strawberry-RED	typical	5.94	0.33	6.00	4.00	6.00
tomato-GREEN	atypical	3.64	1.40	4.00	1.00	6.00
tomato-RED	typical	6.00	0.00	6.00	6.00	6.00
tomato-WHITE	unrelated	1.56	0.88	1.00	1.00	4.00

Cleveland Dot Plot

In the following figure, the y-axis represents the word-typicality combinations. For example, “bear-typical” is equal to “a BROWN bear.” The x-axis represents the mean of

the word typicality rating scores. The legend represents the correspondence of the shapes of the plots and the typicality of colors.



Sentence

Agreement Rates

The following tables summarize the agreement rates (%) of each typicality that was implied by the sentences.

Entirely.

<i>Typicality</i>	<i>Agreement Rates</i>
typical	87.96
atypical	81.30

Each Sentence.

<i>Word</i>	<i>Typical</i>	<i>Atypical</i>
strawberry	97.22	86.11
chameleon	86.11	86.11
kiwi	97.22	91.67
bear	100.00	97.22
cake	80.56	66.67
steak	97.22	80.56
onion	97.22	80.56
tomato	100.00	88.89
ball	97.22	91.67
popcorn	86.11	83.33
apple	72.22	77.78
cloud	77.78	94.44
horse	58.33	75.00
plum	94.44	77.78
leaf	77.78	41.67

Each sentence received more than 25 percent agreement with its intended typicality, which is higher than the chance rate. The item that received the lowest agreement rate was *leaf* in the atypical sentence (41.7 percent). The reason for the result was similar to the result of

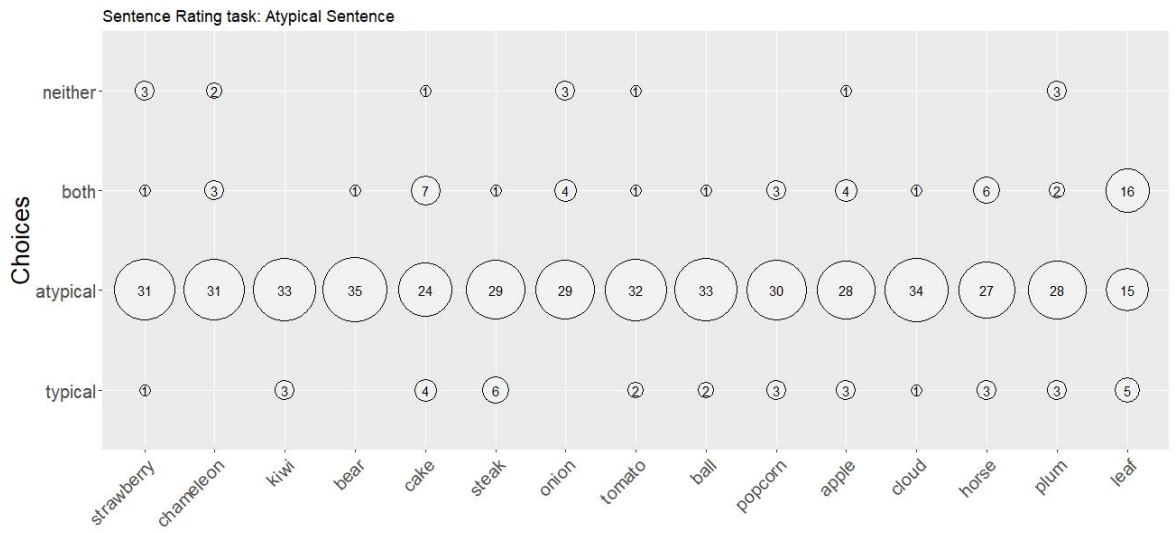
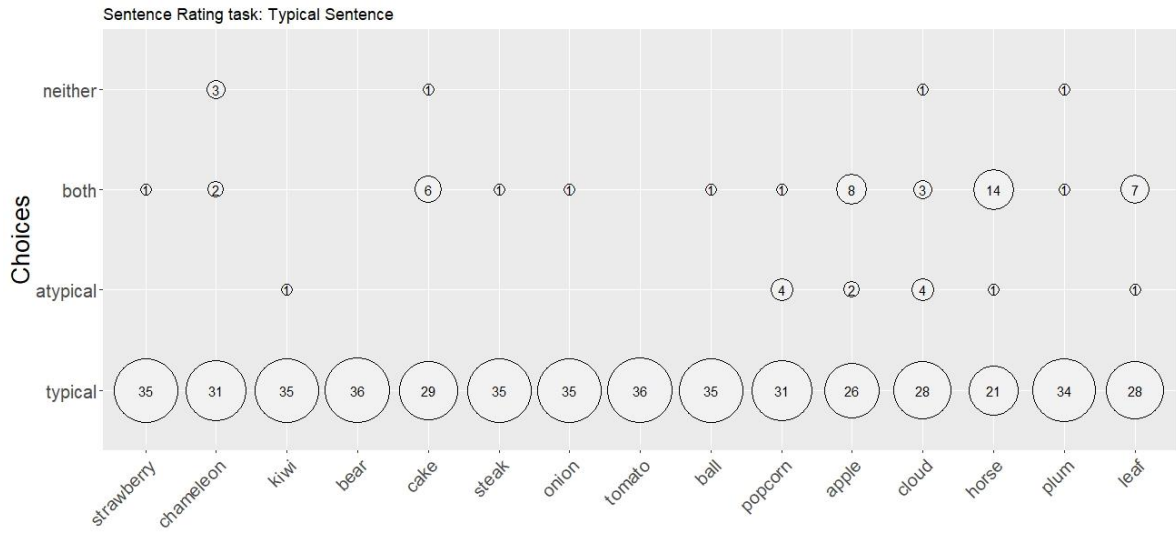
native English speakers. The sixteen participants (44 percent) rated “both,” and five rated (14 percent) “typical” for *leaf* in the atypical sentence.

Balloon Plot

The test sentences were presented with two pictures and four forced choice alternatives:

- typical: best matched by the first picture (the first pictures were always typical objects)
- atypical: best matched by the second picture (the second pictures were always atypical objects)
- both: matched by both pictures equally
- neither: matched by neither picture

The numbers in each balloon refers to the number of the participants who selected the choice.



Appendix I: Statistical Modeling (Native English Speakers)

List of Variables

- SubjectID: Subject ID
- ItemID: Item ID
- Set: Set number
- Position: Whether the phrases that determine the color are placed before or after the keywords
- Pres.Order: Presentation order
- Sentence.Typicality: Typicality of the colors that sentences implied (e.g., *bear in the woods* implies a brown bear [typical], and *bear at the North Pole* implies a white bear [atypical])
- Word: Stimuli (Word)
- Word.Typicality: Typicality of the colors of the fonts (e.g., a brown *bear* represents a typical bear, a white *bear* represents an atypical bear)
- RT.Stroop: Reaction times of the semantic Stroop task
- RT.Sentence: Reading times for each sentence
- z.RT.Sentence: Scaled reading times for each sentence

Change Coding of the Typicality

Based on the word typicality rating task, the typical color of *onion* was changed to red, and the atypical color was changed to brown.

```
pacman::p_load(forcats)
fct_list <- c(
  "typical" = "unrelated",
  "unrelated" = "typical")

EN.model1 <- EN.model %>%
  filter(Word == "onion")
EN.model1 <- EN.model1 %>%
  dplyr::mutate(Word.Typicality = fct_recode(Word.Typicality, !!!fct_list))

EN.model0 <- EN.model %>%
  filter(Word != "onion")
EN.model <- rbind(EN.model0, EN.model1)

EN.model <- EN.model %>%
```

```
mutate(Combination = paste(!!!rlang::syms(c("Sentence.Typicality", "Word.Typicality")), sep="-"))
```

Change Coding of the Categorical Variables

Sentence

```
EN.model$Sentence.Typicality <- factor(EN.model$Sentence.Typicality, levels = c("typical", "atypical"))
contrasts(EN.model$Sentence.Typicality) <- fractions(contr.sdif(2))
contrasts(EN.model$Sentence.Typicality)
```

```
##           2-1
## typical  -1/2
## atypical  1/2
```

Word

```
EN.model$Word.Typicality <- factor(EN.model$Word.Typicality, levels = c("unrelated", "typical", "atypical"))
contrasts(EN.model$Word.Typicality) <- fractions(contr.sdif(3))
contrasts(EN.model$Word.Typicality)
```

```
##           2-1  3-2
## unrelated -2/3 -1/3
## typical   1/3 -1/3
## atypical  1/3  2/3
```

Position

```
EN.model$Position <- factor(EN.model$Position, levels = c("Pre", "Post"))
contrasts(EN.model$Position) <- fractions(contr.sdif(2))
contrasts(EN.model$Position)
```

```
##           2-1
## Pre      -1/2
## Post     1/2
```

Scaling the Continuous Variables

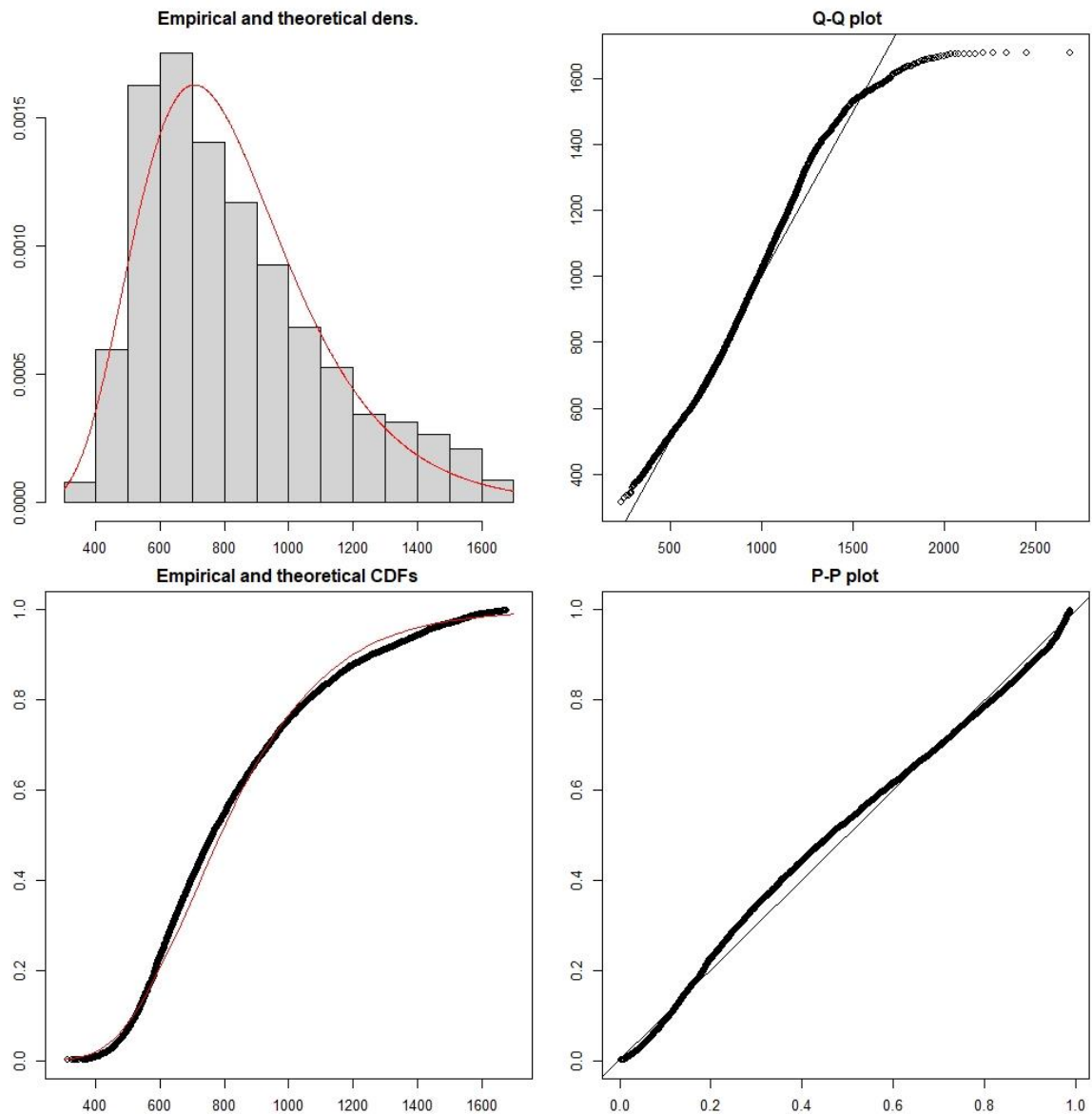
Sentence Reading Time

```
EN.model%>%
  mutate(across(RT.Sentence, ~scale(.x)[,1], .names = "z.{.col}")) -> EN.model
```

Choose Probabilistic Distributions for the Observed Data

According to the goodness-of-fit statistics and information criterion, log-normal distribution was chosen. The top-left panel:

- Histogram: The observed data (Reaction times of the semantic Stroop task)
- Red line: The density curve



Possible Covariates

The null model was compared with the model including the possible covariates.

```

model_EN_backward <- list()
model_EN_backward[[1]] <- lmer(log(RT.Stroop) ~ + (1|SubjectID)+(1|ItemID),
                               data = EN.model,
                               REML = FALSE, lmerControl(optimizer = "bobyqa"),

```

```

200000),
optCtrl=list(maxfun=
check.conv.singular
= .makeCC(action = "ignore", tol = 1e-4))

model_EN_backward[[2]] <- lmer(log(RT.Stroop) ~ + Pres.Order + (1|Subject
ID)+(1|ItemID),
data = EN.model,
REML = FALSE, lmerControl(optimizer = "bobyq
a",
optCtrl=list(maxfun=
200000),
check.conv.singular
= .makeCC(action = "ignore", tol = 1e-4))

model_EN_backward[[3]] <- lmer(log(RT.Stroop) ~ +z.RT.Sentence + (1|Subje
ctID)+(1|ItemID),
data = EN.model,
REML = FALSE, lmerControl(optimizer = "bobyq
a",
optCtrl=list(maxfun=
200000),
check.conv.singular =
.makeCC(action = "ignore", tol = 1e-4))

sapply(model_EN_backward, AIC) %>%
data.frame

##
## 1 1011.5617
## 2 935.0368
## 3 875.3540

sapply(model_EN_backward, AIC) %>%
which.min

## [1] 3

```

The model including the scaled sentence reading time showed the lowest AIC among the three models. Then the model with the scaled sentence reading time was compared with the model with the both presentation order and scaled sentence reading time.

```

model_EN_backward_2 <- list()
model_EN_backward_2[[1]] <- model_EN_backward[[3]]
model_EN_backward_2[[2]] <- stats::update(model_EN_backward_2[[1]], .~.+P
res.Order)

sapply(model_EN_backward_2, AIC)%>%
data.frame

```

```
##
## 1 875.3540
## 2 842.8381

sapply(model_EN_backward_2, AIC)%>%
  which.min

## [1] 2
```

The models with the both covariates showed the lowest AIC. The final model included presentation order and scaled sentence reading time as covariates.

```
model_EN <- model_EN_backward_2[[2]]
```

Specification of the Best Random-Effects Structure

Maximal Model

```
model_EN_1 <- list()
model_EN_1[[1]] <- model_EN
model_EN_1[[2]] <- update(model_EN_1[[1]], .~.-(1|SubjectID)-(1|ItemID)+
  Sentence.Typicality*Word.Typicality + Position +
  (1+Sentence.Typicality+Word.Typicality + Positio
n + Pres.Order +z.RT.Sentence|SubjectID) +
  (1+Sentence.Typicality+Word.Typicality + Positio
n + Pres.Order + z.RT.Sentence|ItemID))
```

Output.

Warning messages:

1: In optwrap(optimizer, devfun, getStart(start, rhopp), lower = rho_lower, : convergence code 1 from bobyqa: bobyqa – maximum number of function evaluations exceeded
2: Model failed to converge with 4 negative eigenvalues: -1.3e+01 -2.4e+01 -4.3e+01 -1.4e+02

Random-Effects Principal Components Analysis.

Subject: first 7 components capture 100% of the random variance.

Item: first 6 components capture 100% of the random variance.

Maximal Model (Zero-Correlation-Parameter)

```
model_EN_1_nocor <- list()
model_EN_1_nocor[[1]] <- model_EN
model_EN_1_nocor[[2]] <- update(model_EN_1_nocor[[1]], .~.-(1|SubjectID)-
(1|ItemID)+
  Sentence.Typicality*Word.Typicality + Position +
  (1+Sentence.Typicality+Word.Typicality + Positio
n + Pres.Order +z.RT.Sentence||SubjectID) +
```

```
(1+Sentence.Typicality+Word.Typicality + Position + Pres.Order + z.RT.Sentence||ItemID))
```

Output.

Warning messages:

- 1: In UseMethod("depth") : no applicable method for 'depth' applied to an object of class "NULL"
- 2: In optwrap(optimizer, devfun, getStart(start, rhopp), lower = rho.lower, : convergence code 1 from bobyqa: bobyqa – maximum number of function evaluations exceeded
- 3: Model failed to converge with 6 negative eigenvalues: -6.0e-02 -1.1e+00 -1.5e+00 -3.5e+00 -3.5e+00 -1.1e+01

Random-Effects Principal Components Analysis.

Subject: first 6 components capture 100% of the random variance.

Item: first 9 components capture 100% of the random variance.

Dropping Variance Components.

Pres.Order was eliminated from the both item and subject random effects.

```
model_EN_2_nocor <- list()
model_EN_2_nocor[[1]] <- model_EN
model_EN_2_nocor[[2]] <- update(model_EN_2_nocor[[1]], .~.-(1|SubjectID)-
(1|ItemID)+
Sentence.Typicality*Word.Typicality + Position +
(1+Sentence.Typicality+Word.Typicality + Position +z.RT.Sentence||SubjectID) +
(1+Sentence.Typicality+Word.Typicality + Position + z.RT.Sentence||ItemID))
```

Output.

Warning messages:

- 1: In UseMethod("depth") : no applicable method for 'depth' applied to an object of class "NULL"
- 2: In UseMethod("depth") : no applicable method for 'depth' applied to an object of class "NULL"
- 3: In UseMethod("depth") : no applicable method for 'depth' applied to an object of class "NULL"
- 4: In UseMethod("depth") : no applicable method for 'depth' applied to an object of class "NULL"
- 5: Model failed to converge with 1 negative eigenvalue: -5.2e-03

Random-Effects Principal Components Analysis.

Subject: first 5 components capture 100% of the random variance.

Item: first 9 components capture 100% of the random variance.

Dropping Variance Components.

Position was eliminated from the subject random effect.

```
model_EN_3_nocor <- list()
model_EN_3_nocor[[1]] <- model_EN
model_EN_3_nocor[[2]] <- update(model_EN_3_nocor[[1]], .~.-(1|SubjectID)-
(1|ItemID)+
                                Sentence.Typicality*Word.Typicality + Position +
                                (1+Sentence.Typicality+Word.Typicality +z.RT.Sen
tence||SubjectID) +
                                (1+Sentence.Typicality+Word.Typicality + Positio
n + z.RT.Sentence||ItemID))
```

Output.

Warning message:

Model failed to converge with 4 negative eigenvalues: -1.4e-04 -9.9e-04 -1.3e-03 -4.8e-02

Random-Effects Principal Components Analysis.

Subject: first 4 components capture 100% of the random variance.

Item: first 9 components capture 100% of the random variance.

Dropping Variance Components.

Sentence.Typicality was eliminated from the subject random effect.

```
model_EN_4_nocor <- list()
model_EN_4_nocor[[1]] <- model_EN
model_EN_4_nocor[[2]] <- update(model_EN_4_nocor[[1]], .~.-(1|SubjectID)-
(1|ItemID)+
                                Sentence.Typicality*Word.Typicality + Position +
                                (1+Word.Typicality +z.RT.Sentence||SubjectID) +
                                (1+Sentence.Typicality+Word.Typicality + Positio
n + z.RT.Sentence||ItemID))
```

Output.

Warning message:

Model failed to converge with 2 negative eigenvalues: -7.9e-06 -2.8e-02

Random-Effects Principal Components Analysis.

Subject: first 3 components capture 100% of the random variance.

Item: first 8 components capture 100% of the random variance.

Dropping Variance Components.

Sentence.Typicality was eliminated from the item random effect.

```
model_EN_5_nocor <- list()
model_EN_5_nocor[[1]] <- model_EN
model_EN_5_nocor[[2]] <- update(model_EN_5_nocor[[1]], .~.-(1|SubjectID)-
(1|ItemID)+
Sentence.Typicality*Word.Typicality + Position +
(1+Word.Typicality +z.RT.Sentence||SubjectID) +
(1+Word.Typicality + Position + z.RT.Sentence||
ItemID))
```

Output.

Warning message:

Model failed to converge with 4 negative eigenvalues: -1.1e-04 -3.4e-04 -6.9e-04 -3.0e-03

Random-Effects Principal Components Analysis.

Subject: first 3 components capture 100% of the random variance.

Item: first 7 components capture 100% of the random variance.

Dropping Variance Components.

Word.Typicality was eliminated from the subject random effect.

```
model_EN_6_nocor <- list()
model_EN_6_nocor[[1]] <- model_EN
model_EN_6_nocor[[2]] <- update(model_EN_6_nocor[[1]], .~.-(1|SubjectID)-
(1|ItemID)+
Sentence.Typicality*Word.Typicality + Position +
(1 +z.RT.Sentence||SubjectID) +
(1+Word.Typicality + Position + z.RT.Sentence||
ItemID))
```

Output.

Warning messages: 1: In checkConv(attr(opt, "derivs"), optpar, ctrl = controlcheckConv, : unable to evaluate scaled gradient

2: In checkConv(attr(opt, "derivs"), optpar, ctrl = controlcheckConv, : Model failed to converge: degenerate Hessian with 1 negative eigenvalues

Random-Effects Principal Components Analysis.

Subject: first 2 components capture 100% of the random variance.

Item: first 7 components capture 100% of the random variance.

Dropping Variance Components.

Position was eliminated from the item random effect.

```
model_EN_7_nocor <- list()
model_EN_7_nocor[[1]] <- model_EN
model_EN_7_nocor[[2]] <- update(model_EN_7_nocor[[1]], .~.-(1|SubjectID)-
(1|ItemID)+
                                Sentence.Typicality*Word.Typicality + Position +
                                (1 +z.RT.Sentence||SubjectID) +
                                (1+Word.Typicality + z.RT.Sentence||ItemID))
```

Output.

Warning messages:

- 1: In checkConv(attr(opt, "derivs"), optpar, ctrl = controlcheckConv, : unable to evaluate scaled gradient
- 2: In checkConv(attr(opt, "derivs"), optpar, ctrl = controlcheckConv, : Model failed to converge: degenerate Hessian with 1 negative eigenvalues
- 3: Model failed to converge with 2 negative eigenvalues: -2.7e-04 -5.8e-04

Random-Effects Principal Components Analysis.

Subject: first 2 components capture 100% of the random variance.

Item: first 4 components capture 100% of the random variance.

Dropping Variance Components.

Word.Typicality was eliminated from the item random effect.

```
model_EN_8_nocor <- list()
model_EN_8_nocor[[1]] <- model_EN
model_EN_8_nocor[[2]] <- update(model_EN_8_nocor[[1]], .~.-(1|SubjectID)-
(1|ItemID)+
                                Sentence.Typicality*Word.Typicality + Position +
                                (1 +z.RT.Sentence||SubjectID) +
                                (1 + z.RT.Sentence||ItemID))
```

Output.

```
summary(model_EN_8_nocor[[2]])
## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
## method [lmerModLmerTest]
## Formula: log(RT.Stroop) ~ z.RT.Sentence + Pres.Order + Sentence.Typicality +
## Word.Typicality + Position + (1 + z.RT.Sentence || SubjectID) +
## (1 + z.RT.Sentence || ItemID) + Sentence.Typicality:Word.Typicalit
```

```

y
## Data: EN.model
## Control: lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+
05),
## check.conv.singular = .makeCC(action = "ignore", tol = 1e-04))
##
## AIC BIC logLik deviance df.resid
## 782.1 875.0 -377.1 754.1 5624
##
## Scaled residuals:
## Min 1Q Median 3Q Max
## -2.8227 -0.6648 -0.1058 0.5847 4.1045
##
## Random effects:
## Groups Name Variance Std.Dev.
## ItemID z.RT.Sentence 0.0003673 0.01916
## ItemID.1 (Intercept) 0.0030813 0.05551
## SubjectID z.RT.Sentence 0.0018939 0.04352
## SubjectID.1 (Intercept) 0.0380003 0.19494
## Residual 0.0623724 0.24974
## Number of obs: 5638, groups: ItemID, 180; SubjectID, 35
##
## Fixed effects:
## Estimate Std. Error df
## (Intercept) 6.716e+00 3.399e-02 3.864e+0
1
## z.RT.Sentence 6.370e-02 9.364e-03 3.098e+
01
## Pres.Order -3.436e-04 6.953e-05 5.538e+0
3
## Sentence.Typicality2-1 -8.083e-04 1.065e-02 1.773e
+02
## Word.Typicality2-1 -3.389e-02 1.305e-02 1.779e+
02
## Word.Typicality3-2 2.622e-02 1.304e-02 1.775e+
02
## Position2-1 -1.608e-02 1.086e-02 1.914e+
02
## Sentence.Typicality2-1:Word.Typicality2-1 -1.633e-02 2.610e-02 1.778
e+02
## Sentence.Typicality2-1:Word.Typicality3-2 2.267e-03 2.608e-02 1.774
e+02
## t value Pr(>|t|)
## (Intercept) 197.597 < 2e-16 ***
## z.RT.Sentence 6.803 1.28e-07 ***
## Pres.Order -4.942 7.96e-07 ***
## Sentence.Typicality2-1 -0.076 0.9396
## Word.Typicality2-1 -2.597 0.0102 *
## Word.Typicality3-2 2.010 0.0459 *
## Position2-1 -1.481 0.1403
## Sentence.Typicality2-1:Word.Typicality2-1 -0.626 0.5323
## Sentence.Typicality2-1:Word.Typicality3-2 0.087 0.9308

```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##          (Intr) z.RT.S Prs.Or Sn.T2-1 W.T2-1 W.T3-2 Pst2-1 S.T2-1:W.
T2
## z.RT.Sentnc -0.019
## Pres.Order  -0.182  0.181
## Sntnc.Ty2-1  0.000 -0.001 -0.003
## Wrđ.Typc2-1 -0.002 -0.004  0.009 -0.001
## Wrđ.Typc3-2  0.001 -0.003 -0.008  0.003  -0.501

## Position2-1 -0.007  0.024  0.020  0.000  0.001  0.000

## S.T2-1:W.T2 -0.001  0.006  0.007  0.001  -0.002  0.001 -0.002

## S.T2-1:W.T3  0.002 -0.005 -0.008 -0.002  0.001  0.002 -0.001 -0.501
```

Random-Effects Principal Components Analysis.

Subject: first 2 components capture 100% of the random variance.

Item: first 2 components capture 100% of the random variance.

Dropping Variance Components.

z.RT.Sentence was eliminated from the item random effect.

```
model_EN_9_nocor <- list()
model_EN_9_nocor[[1]] <- model_EN
model_EN_9_nocor[[2]] <- update(model_EN_9_nocor[[1]], .~. - (1|SubjectID) -
(1|ItemID) +
          Sentence.Typicality*Word.Typicality + Position +
          (1+z.RT.Sentence||SubjectID) +
          (1|ItemID))
```

Output.

```
summary(model_EN_9_nocor[[2]])
## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
## method [lmerModLmerTest]
## Formula: log(RT.Stroop) ~ z.RT.Sentence + Pres.Order + Sentence.Typicality +
## Word.Typicality + Position + (1 + z.RT.Sentence || SubjectID) +
## (1 | ItemID) + Sentence.Typicality:Word.Typicality
## Data: EN.model
## Control: lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+05),
```

```

##      check.conv.singular = .makeCC(action = "ignore", tol = 1e-04))
##
##      AIC      BIC    logLik deviance df.resid
##      783.1    869.4   -378.6   757.1    5625
##
## Scaled residuals:
##      Min      1Q  Median      3Q      Max
## -3.0304 -0.6656 -0.1027  0.5911  4.0951
##
## Random effects:
##      Groups      Name      Variance Std.Dev.
##      ItemID      (Intercept)  0.003070 0.05541
##      SubjectID    z.RT.Sentence 0.001839 0.04288
##      SubjectID.1 (Intercept)  0.037883 0.19464
##      Residual                0.062702 0.25040
## Number of obs: 5638, groups:  ItemID, 180; SubjectID, 35
##
## Fixed effects:
##
##              Estimate Std. Error      df
## (Intercept)      6.717e+00  3.394e-02  3.864e+0
## 1
## z.RT.Sentence      6.154e-02  9.103e-03  2.967e+
## 01
## Pres.Order        -3.529e-04  6.946e-05  5.533e+0
## 3
## Sentence.Typicality2-1
## +02                -1.056e-03  1.063e-02  1.772e
## +02
## Word.Typicality2-1
## 02                 -3.378e-02  1.303e-02  1.778e+
## 02
## Word.Typicality3-2
## 02                 2.605e-02  1.302e-02  1.774e+
## 02
## Position2-1
## 02                 -1.631e-02  1.084e-02  1.910e+
## 02
## Sentence.Typicality2-1:Word.Typicality2-1 -1.652e-02  2.605e-02  1.777
## e+02
## Sentence.Typicality2-1:Word.Typicality3-2  2.896e-03  2.604e-02  1.773
## e+02
##
##              t value Pr(>|t|)
## (Intercept)      197.929 < 2e-16 ***
## z.RT.Sentence      6.760 1.80e-07 ***
## Pres.Order        -5.080 3.91e-07 ***
## Sentence.Typicality2-1
## -0.099  0.9210
## Word.Typicality2-1
## -2.593  0.0103 *
## Word.Typicality3-2
## 2.001  0.0469 *
## Position2-1
## -1.505  0.1340
## Sentence.Typicality2-1:Word.Typicality2-1 -0.634  0.5268
## Sentence.Typicality2-1:Word.Typicality3-2  0.111  0.9116
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##      (Intr) z.RT.S Prs.Or Sn.T2-1 W.T2-1 W.T3-2 Pst2-1 S.T2-1:W.

```

```

T2
## z.RT.Sentnc -0.019
## Pres.Order -0.182 0.180
## Sntnc.Ty2-1 0.000 -0.001 -0.003
## Wrđ.Typc2-1 -0.002 -0.003 0.009 -0.001
## Wrđ.Typc3-2 0.001 -0.003 -0.008 0.003 -0.501

## Position2-1 -0.006 0.023 0.018 0.000 0.001 0.000

## S.T2-1:W.T2 -0.001 0.007 0.007 0.001 -0.001 0.001 -0.002

## S.T2-1:W.T3 0.002 -0.006 -0.009 -0.002 0.001 0.002 -0.001 -0.501

```

Random-Effects Principal Components Analysis.

Subject: first 2 components capture 100% of the random variance.

Item: first 1 component capture 100% of the random variance.

Dropping Variance Components.

z.RT.Sentence was eliminated from the item random effect.

```

model_EN_10_nocor <- list()
model_EN_10_nocor[[1]] <- model_EN
model_EN_10_nocor[[2]] <- update(model_EN_10_nocor[[1]], .~.-(1|SubjectID)
-(1|ItemID)+
                                Sentence.Typicality*Word.Typicality + Position +
                                (1|SubjectID) +
                                (1+z.RT.Sentence || ItemID))

```

Model Comparisons.

A log likelihood ratio test showed that the model that included the scaled sentence reading time for the both item and subject random effects (model_EN_8_nocor) showed significantly lower AIC than the model that included the scaled sentence reading time for the item random effect (model_EN_10_nocor). Thus, the model_EN_8_nocor was chosen.

```

anova(model_EN_8_nocor[[2]], model_EN_10_nocor[[2]])

## Data: EN.model
## Models:
## model_EN_10_nocor[[2]]: log(RT.Stroop) ~ z.RT.Sentence + Pres.Order +
Sentence.Typicality + Word.Typicality + Position + (1 | SubjectID) + (1 +
z.RT.Sentence || ItemID) + Sentence.Typicality:Word.Typicality
## model_EN_8_nocor[[2]]: log(RT.Stroop) ~ z.RT.Sentence + Pres.Order + S
entence.Typicality + Word.Typicality + Position + (1 + z.RT.Sentence || S
ubjectID) + (1 + z.RT.Sentence || ItemID) + Sentence.Typicality:Word.Typi

```

```

cality
##              npar    AIC    BIC  logLik deviance  Chisq Df Pr(>
Chisq)
## model_EN_10_nocor[[2]]    13 835.99 922.28 -405.00    809.99

## model_EN_8_nocor[[2]]    14 782.12 875.04 -377.06    754.12 55.874  1
7.728e-14
##
## model_EN_10_nocor[[2]]
## model_EN_8_nocor[[2]] ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

anova(model_EN_9_nocor[[2]],model_EN_10_nocor[[2]])

## Data: EN.model
## Models:
## model_EN_9_nocor[[2]]: log(RT.Stroop) ~ z.RT.Sentence + Pres.Order + S
entence.Typicality + Word.Typicality + Position + (1 + z.RT.Sentence || S
ubjectID) + (1 | ItemID) + Sentence.Typicality:Word.Typicality
## model_EN_10_nocor[[2]]: log(RT.Stroop) ~ z.RT.Sentence + Pres.Order +
Sentence.Typicality + Word.Typicality + Position + (1 | SubjectID) + (1 +
z.RT.Sentence || ItemID) + Sentence.Typicality:Word.Typicality
##              npar    AIC    BIC  logLik deviance  Chisq Df Pr(>C
hisq)
## model_EN_9_nocor[[2]]    13 783.13 869.41 -378.56    757.13

## model_EN_10_nocor[[2]]    13 835.99 922.28 -405.00    809.99    0  0

```

Checking If Including Correlation Parameter Increases the Goodness-of-Fit

The correlation parameter was added to the model_EN_8_nocor and compare the AIC with the zero-correlation-parameter model.

```

model_EN_8_cor <- list()
model_EN_8_cor[[1]] <- model_EN
model_EN_8_cor[[2]] <-update(model_EN_8_cor[[1]],.~.(1|SubjectID)-(1|It
emID)+
                                Sentence.Typicality*Word.Typicality + Position +
                                (1 +z.RT.Sentence|SubjectID) +
                                (1 + z.RT.Sentence|ItemID))

```

Model Comparisons.

A log likelihood ratio test showed that including the correlation parameter did not significantly reduce the AIC score. Therefore, the model without the correlation parameter was chosen as the final model.

```

anova(model_EN_8_nocor[[2]],model_EN_8_cor[[2]])

```



```

## Data: EN.model
## Models:
## model_EN_8_nocor[[2]]: log(RT.Stroop) ~ z.RT.Sentence + Pres.Order + S
entence.Typicality + Word.Typicality + Position + (1 + z.RT.Sentence || S
ubjectID) + (1 + z.RT.Sentence || ItemID) + Sentence.Typicality:Word.Typi
cality
## model_EN_8_cor[[2]]: log(RT.Stroop) ~ z.RT.Sentence + Pres.Order + Sen
tence.Typicality + Word.Typicality + Position + (1 + z.RT.Sentence | Subj
ectID) + (1 + z.RT.Sentence | ItemID) + Sentence.Typicality:Word.Typicali
ty
##
##          npar    AIC    BIC  logLik deviance  Chisq Df Pr(>C
hisq)
## model_EN_8_nocor[[2]]    14 782.12 875.04 -377.06   754.12
## model_EN_8_cor[[2]]     16 785.23 891.42 -376.61   753.23 0.8941  2
0.6395

```

Results of the Final Model

Summary of the Final Model

```

## Linear mixed model fit by maximum likelihood . t-tests use Satterthwai
te's
## method [lmerModLmerTest]
## Formula: log(RT.Stroop) ~ z.RT.Sentence + Pres.Order + Sentence.Typica
lity +
##   Word.Typicality + Position + (1 + z.RT.Sentence || SubjectID) +
##   (1 + z.RT.Sentence || ItemID) + Sentence.Typicality:Word.Typicalit
y
## Data: EN.model
## Control: lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+
05),
##   check.conv.singular = .makeCC(action = "ignore", tol = 1e-04))
##
##      AIC      BIC  logLik deviance df.resid
##  782.1    875.0  -377.1   754.1    5624
##
## Scaled residuals:
##   Min      1Q  Median      3Q      Max
## -2.8227 -0.6648 -0.1058  0.5847  4.1045
##
## Random effects:
## Groups      Name                Variance Std.Dev.
## ItemID      z.RT.Sentence 0.0003673 0.01916
## ItemID.1    (Intercept)  0.0030813 0.05551
## SubjectID   z.RT.Sentence 0.0018939 0.04352
## SubjectID.1 (Intercept)  0.0380003 0.19494
## Residual                    0.0623724 0.24974
## Number of obs: 5638, groups: ItemID, 180; SubjectID, 35
##
## Fixed effects:
##
##                                     Estimate Std. Error      df

```

```

## (Intercept) 6.716e+00 3.399e-02 3.864e+0
1
## z.RT.Sentence 6.370e-02 9.364e-03 3.098e+
01
## Pres.Order -3.436e-04 6.953e-05 5.538e+0
3
## Sentence.Typicality2-1 -8.083e-04 1.065e-02 1.773e
+02
## Word.Typicality2-1 -3.389e-02 1.305e-02 1.779e+
02
## Word.Typicality3-2 2.622e-02 1.304e-02 1.775e+
02
## Position2-1 -1.608e-02 1.086e-02 1.914e+
02
## Sentence.Typicality2-1:Word.Typicality2-1 -1.633e-02 2.610e-02 1.778
e+02
## Sentence.Typicality2-1:Word.Typicality3-2 2.267e-03 2.608e-02 1.774
e+02
## t value Pr(>|t|)
## (Intercept) 197.597 < 2e-16 ***
## z.RT.Sentence 6.803 1.28e-07 ***
## Pres.Order -4.942 7.96e-07 ***
## Sentence.Typicality2-1 -0.076 0.9396
## Word.Typicality2-1 -2.597 0.0102 *
## Word.Typicality3-2 2.010 0.0459 *
## Position2-1 -1.481 0.1403
## Sentence.Typicality2-1:Word.Typicality2-1 -0.626 0.5323
## Sentence.Typicality2-1:Word.Typicality3-2 0.087 0.9308
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
## (Intr) z.RT.S Prs.Or Sn.T2-1 W.T2-1 W.T3-2 Pst2-1 S.T2-1:W.
T2
## z.RT.Sentnc -0.019
## Pres.Order -0.182 0.181
## Sntnc.Ty2-1 0.000 -0.001 -0.003
## Wrđ.Typc2-1 -0.002 -0.004 0.009 -0.001
## Wrđ.Typc3-2 0.001 -0.003 -0.008 0.003 -0.501
## Position2-1 -0.007 0.024 0.020 0.000 0.001 0.000
## S.T2-1:W.T2 -0.001 0.006 0.007 0.001 -0.002 0.001 -0.002
## S.T2-1:W.T3 0.002 -0.005 -0.008 -0.002 0.001 0.002 -0.001 -0.501

```

Variance Inflation Factors (VIF)

```

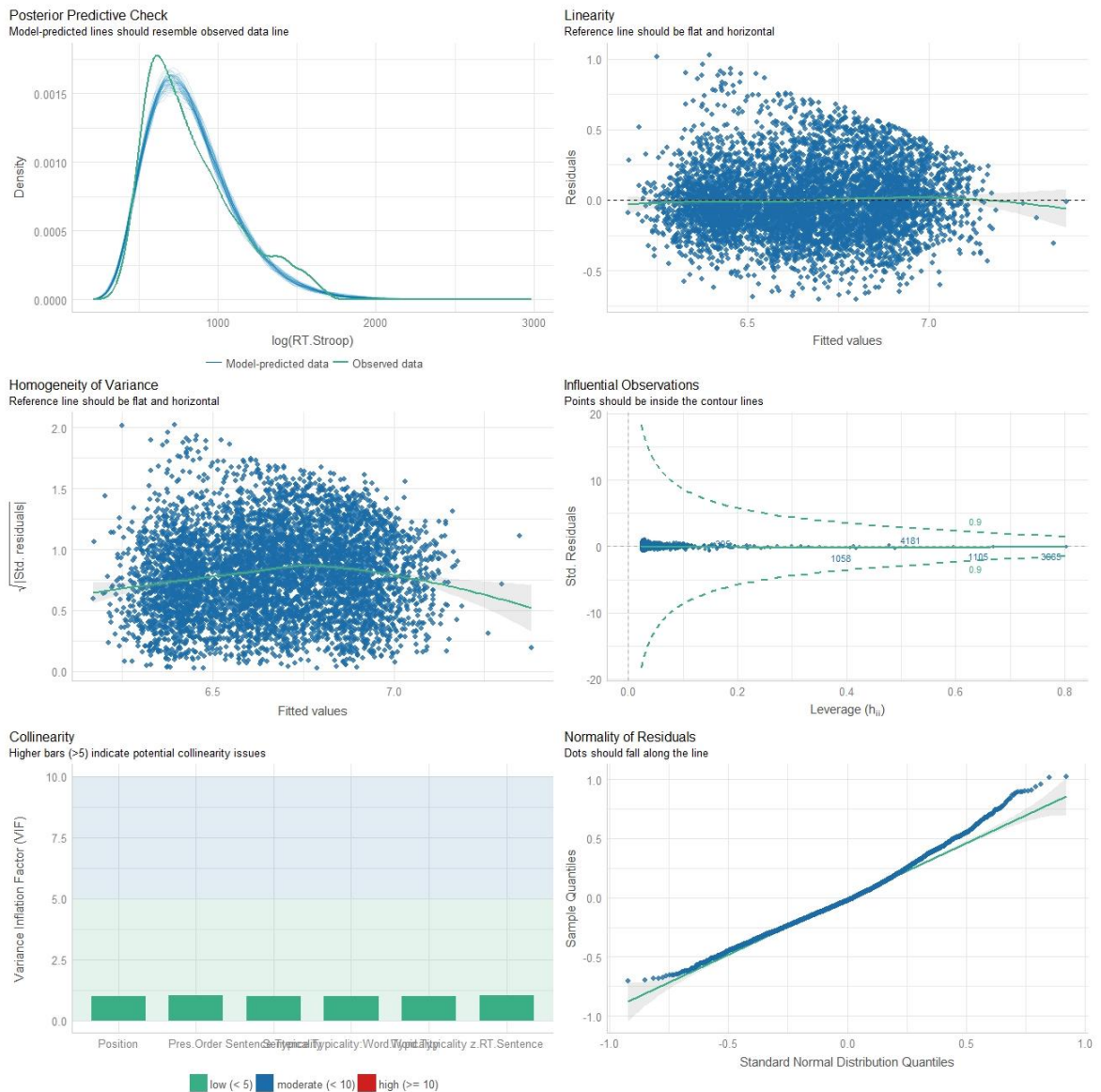
## # Check for Multicollinearity
##
## Low Correlation
##

```

##	Term	VIF	Increased SE	Tolerance
##	z.RT.Sentence	1.03	1.02	0.97
##	Pres.Order	1.03	1.02	0.97
##	Sentence.Typicality	1.00	1.00	1.00
##	Word.Typicality	1.00	1.00	1.00
##	Position	1.00	1.00	1.00
##	Sentence.Typicality:Word.Typicality	1.00	1.00	1.00

Model Diagnosis

Could not compute standard errors from random effects for diagnostic plot.



The Model That Only Includes Significant Predictors

```
final_model_EN_3_nocor <- list()
final_model_EN_3_nocor[[1]] <- model_EN
```

```

final_model_EN_3_nocor[[2]] <-update(final_model_EN_3_nocor[[1]],.~.- (1|
SubjectID)-(1|ItemID)+
                                Word.Typicality +
                                (1 +z.RT.Sentence||SubjectID) +
                                (1 + z.RT.Sentence||ItemID))

## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
## method [lmerModLmerTest]
## Formula: log(RT.Stroop) ~ z.RT.Sentence + Pres.Order + Word.Typicality
+
## (1 + z.RT.Sentence || SubjectID) + (1 + z.RT.Sentence || ItemID)
## Data: EN.model
## Control: lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+
05),
## check.conv.singular = .makeCC(action = "ignore", tol = 1e-04))
##
##      AIC      BIC   logLik deviance df.resid
##  776.8    843.1  -378.4   756.8    5628
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.8164 -0.6640 -0.1050  0.5851  4.1087
##
## Random effects:
## Groups      Name                Variance Std.Dev.
## ItemID      z.RT.Sentence 0.0003697 0.01923
## ItemID.1    (Intercept)  0.0031520 0.05614
## SubjectID   z.RT.Sentence 0.0018822 0.04338
## SubjectID.1 (Intercept)  0.0380389 0.19504
## Residual                    0.0623741 0.24975
## Number of obs: 5638, groups:  ItemID, 180; SubjectID, 35
##
## Fixed effects:
##              Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)    6.716e+00  3.401e-02  3.866e+01 197.463 < 2e-16 **
## z.RT.Sentence  6.406e-02  9.344e-03  3.106e+01   6.856 1.09e-07 **
## Pres.Order     -3.414e-04  6.952e-05  5.539e+03  -4.910 9.37e-07 **
## Word.Typicality2-1 -3.389e-02  1.314e-02  1.775e+02  -2.579  0.0107 *
## Word.Typicality3-2  2.623e-02  1.313e-02  1.772e+02   1.997  0.0473 *
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##              (Intr) z.RT.S Prs.Or W.T2-1
## z.RT.Sentnc -0.019

```

```
## Pres.Order -0.182 0.181
## Wrd.Typc2-1 -0.002 -0.004 0.009
## Wrd.Typc3-2 0.001 -0.003 -0.008 -0.501
```

Appendix J: Statistical Modeling (Native Japanese Speakers)

List of Variables

- SubjectID: Subject ID
- ItemID: Item ID
- Set: Set number
- Position: Whether the phrases that determine the color are placed before or after the keywords
- Pres.Order: Presentation order
- Sentence.Typicality: Typicality of the colors that sentences implied (e.g., *bear in the woods* implies a brown bear [typical], and *bear at the North Pole* implies a white bear [atypical])
- Word: Stimuli (Word)
- Word.Typicality: Typicality of the colors of the fonts (e.g., a brown *bear* represents a typical bear, a white *bear* represents an atypical bear)
- RT.Stroop: Reaction times of the semantic Stroop task
- RT.Sentence: Reading times for each sentence
- z.RT.Sentence: Scaled reading times for each sentence

Change Coding of the Categorical Variables

Sentence

```
JPN.model$Sentence.Typicality <- factor(JPN.model$Sentence.Typicality, 1
levels = c("typical", "atypical"))
contrasts(JPN.model$Sentence.Typicality) <- fractions(contr.sdif(2))
contrasts(JPN.model$Sentence.Typicality)

##          2-1
## typical -1/2
## atypical 1/2
```

Word

```
JPN.model$Word.Typicality <- factor(JPN.model$Word.Typicality, levels = c
("unrelated", "typical", "atypical"))
contrasts(JPN.model$Word.Typicality) <- fractions(contr.sdif(3))
contrasts(JPN.model$Word.Typicality)

##          2-1  3-2
## unrelated -2/3 -1/3
## typical   1/3 -1/3
## atypical  1/3  2/3
```

Position

```
JPN.model$Position <- factor(JPN.model$Position, levels = c("Pre", "Post"))
contrasts(JPN.model$Position) <- fractions(contr.sdif(2))
contrasts(JPN.model$Position)

##      2-1
## Pre -1/2
## Post 1/2
```

Scaling the Continuous Variables

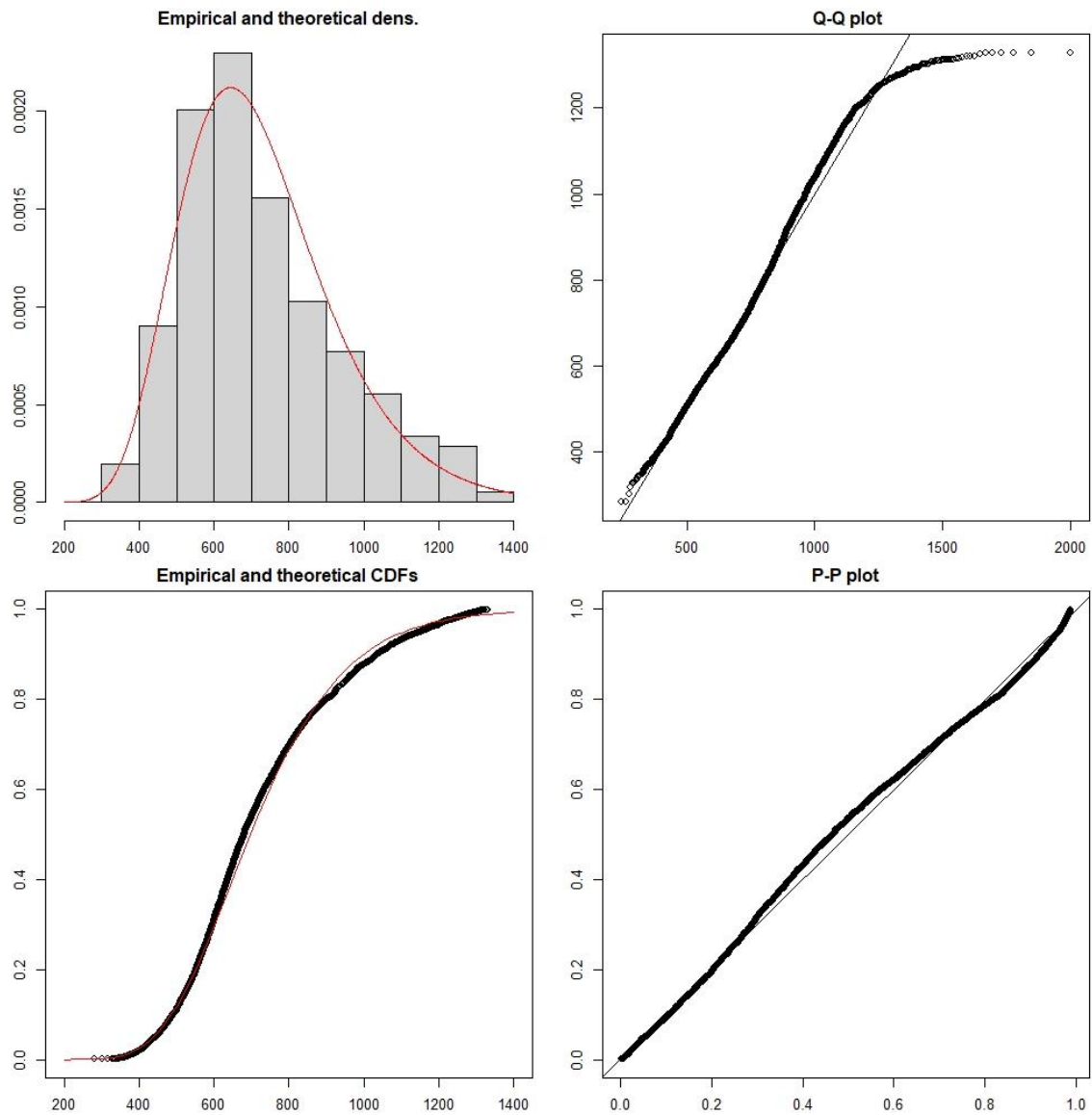
Sentence Reading Time

```
JPN.model%>%
  mutate(across(RT.Sentence, ~scale(.x)[,1], .names = "z.{.col}")) -> JPN.model
```

Choose Probabilistic Distributions for the Observed Data

According to the goodness-of-fit statistics and information criterion, log-normal distribution was chosen. The top-left panel:

- Histogram: The observed data (Reaction times of the semantic Stroop task)
- Red line: The density curve



Possible Covariates

The null model was compared with the model including the possible covariates.

```

model_JPN_backward <- list()
model_JPN_backward[[1]] <- lmer(log(RT.Stroop) ~ + (1|SubjectID)+(1|ItemID),
                                data = JPN.model,
                                REML = FALSE, lmerControl(optimizer = "bobyqa",
                                                                optCtrl=list(maxfun
                                                                =200000),
                                                                check.conv.singular
                                                                = .makeCC(action = "ignore", tol = 1e-4)))

```



```

model_JPN_backward[[2]] <- lmer(log(RT.Stroop) ~ + Pres.Order + (1|SubjectID)+(1|ItemID),
                                data = JPN.model,
                                REML = FALSE, lmerControl(optimizer = "bobyqa",
                                optCtrl=list(maxfun=200000),
                                check.conv.singular = .makeCC(action = "ignore", tol = 1e-4)))

model_JPN_backward[[3]] <- lmer(log(RT.Stroop) ~ +z.RT.Sentence + (1|SubjectID)+(1|ItemID),
                                data = JPN.model,
                                REML = FALSE, lmerControl(optimizer = "bobyqa",
                                optCtrl=list(maxfun=200000),
                                check.conv.singular = .makeCC(action = "ignore", tol = 1e-4)))

sapply(model_JPN_backward, AIC) %>%
  data.frame

##           .
## 1      8.117597
## 2     -69.179343
## 3    -213.262019

sapply(model_JPN_backward, AIC) %>%
  which.min

## [1] 3

```

The model including the scaled sentence reading time showed the lowest AIC among the three models. Then the model with the scaled sentence reading time was compared with the model with the both presentation order and scaled sentence reading time.

```

model_JPN_backward_2 <- list()
model_JPN_backward_2[[1]] <- model_JPN_backward[[3]]
model_JPN_backward_2[[2]] <- stats::update(model_JPN_backward_2[[1]], .~.
+Pres.Order)

sapply(model_JPN_backward_2, AIC)%>%
  data.frame

##           .
## 1    -213.2620
## 2    -243.2904

```

```
sapply(model_JPN_backward_2, AIC)%>%
  which.min
```

```
## [1] 2
```

The models with the both covariates showed the lowest AIC. The final model included presentation order and scaled sentence reading time as covariates.

```
model_JPN <- model_JPN_backward_2[[2]]
```

Specification of the Best Random-Effects Structure

Maximal model

```
model_JPN_1 <- list()
model_JPN_1[[1]] <- model_JPN
model_JPN_1[[2]] <- update(model_JPN_1[[1]], .~.-(1|SubjectID)-(1|ItemID)+
  Sentence.Typicality*Word.Typicality + Position
+
  (1+Sentence.Typicality+Word.Typicality + Position
+ z.RT.Sentence + Pres.Order|SubjectID) +
  (1+Sentence.Typicality+Word.Typicality + Position
+ z.RT.Sentence + Pres.Order|ItemID))
```

Output.

Warning messages:

1: In UseMethod("depth") :

no applicable method for 'depth' applied to an object of class "NULL"

2: In optwrap(optimizer, devfun, getStart(start, rho\$pp), lower = rho\$lower, :

convergence code 1 from bobyqa: bobyqa -- maximum number of function evaluations exceeded

3: Model failed to converge with 2 negative eigenvalues: -3.6e+01 -8.3e+01

Random-Effects Principal Components Analysis.

Subject: first 6 components capture 100% of the random variance.

Item: first 5 components capture 100% of the random variance.

Maximal Model (Zero-Correlation-Parameter)

```
model_JPN_1_nocor <- list()
model_JPN_1_nocor[[1]] <- model_JPN
model_JPN_1_nocor[[2]] <- update(model_JPN_1_nocor[[1]], .~.-(1|SubjectID)
-(1|ItemID)+
  Sentence.Typicality*Word.Typicality + Position
```

```

+
      (1+Sentence.Typicality+Word.Typicality + Position + z.RT.Sentence + Pres.Order||SubjectID) +
      (1+Sentence.Typicality+Word.Typicality + Position + z.RT.Sentence + Pres.Order||ItemID))

```

Output.

Warning in optwrap(optimizer, devfun, getStart(start, rhopp), lower = rholower, : convergence code 1 from bobyqa: bobyqa – maximum number of function evaluations exceeded

Warning: Model failed to converge with 2 negative eigenvalues: -4.7e-01 -1.4e+01

Random-Effects Principal Components Analysis.

Subject: first 9 components capture 100% of the random variance.

Item: first 8 components capture 100% of the random variance.

Dropping Variance Components.

Pres.order was removed from the item and subject random effects.

```

model_JPN_2_nocor <- list()
model_JPN_2_nocor[[1]] <- model_JPN
model_JPN_2_nocor[[2]] <- update(model_JPN_2_nocor[[1]], .~. -(1|SubjectID)
-(1|ItemID)+
      Sentence.Typicality*Word.Typicality + Position
+
      (1+Sentence.Typicality+Word.Typicality + Position + z.RT.Sentence ||SubjectID) +
      (1+Sentence.Typicality+Word.Typicality + Position + z.RT.Sentence||ItemID))

```

Output.

Warning: Model failed to converge with 6 negative eigenvalues: -1.5e-04 -1.9e-04 -4.1e-04 -1.2e-03 -6.6e-01 -8.8e+00

Random-effects Principal Components Analysis.

Subject: first 6 components capture 100% of the random variance.

Item: first 9 components capture 100% of the random variance.

Dropping Variance Components.

Position was removed from the subject random effect.

```

model_JPN_3_nocor <- list()
model_JPN_3_nocor[[1]] <- model_JPN
model_JPN_3_nocor[[2]] <- update(model_JPN_3_nocor[[1]], .~.-(1|SubjectID)
-(1|ItemID)+
                                Sentence.Typicality*Word.Typicality + Position
+
                                (1+Sentence.Typicality+Word.Typicality + z.RT.S
entence ||SubjectID) +
                                (1+Sentence.Typicality+Word.Typicality + Positi
on + z.RT.Sentence ||ItemID))

```

Output.

Warning: Model failed to converge with 2 negative eigenvalues: -5.2e-04 -6.6e-01

Random-Effects Principal Components Analysis.

Subject: first 4 components capture 100% of the random variance.

Item: first 9 components capture 100% of the random variance.

Dropping Variance Components.

Sentence.Typicality was removed from the subject random effect.

```

model_JPN_4_nocor <- list()
model_JPN_4_nocor[[1]] <- model_JPN
model_JPN_4_nocor[[2]] <- update(model_JPN_4_nocor[[1]], .~.-(1|SubjectID)
-(1|ItemID)+
                                Sentence.Typicality*Word.Typicality + Position
+
                                (1+Word.Typicality + z.RT.Sentence ||SubjectID)
+
                                (1+Sentence.Typicality+Word.Typicality + Positi
on + z.RT.Sentence ||ItemID))

```

Output.

Warning: Model failed to converge with 2 negative eigenvalues: -1.4e-04 -2.5e-02

Random-Effects Principal Components Analysis.

Subject: first 3 components capture 100% of the random variance.

Item: first 8 components capture 100% of the random variance.

Dropping Variance Components.

Word.Typicality was removed from the subject random effect.

```

model_JPN_5_nocor <- list()
model_JPN_5_nocor[[1]] <- model_JPN
model_JPN_5_nocor[[2]] <- update(model_JPN_5_nocor[[1]], .~.-(1|SubjectID)
-(1|ItemID)+
                                Sentence.Typicality*Word.Typicality + Position
+
                                (1 + z.RT.Sentence ||SubjectID) +
                                (1+Sentence.Typicality+Word.Typicality + Positi
on + z.RT.Sentence ||ItemID))

```

Output.

Warning in checkConv(attr(opt, "derivs"), optpar, ctrl = controlcheckConv, : unable to evaluate scaled gradient
Warning in checkConv(attr(opt, "derivs"), optpar, ctrl = controlcheckConv, : Model failed to converge: degenerate Hessian with 1 negative eigenvalues
Warning: Model failed to converge with 5 negative eigenvalues: -3.8e-04 -7.2e-04 -1.2e-03 -1.3e-03 -1.8e-03

Random-Effects Principal Components Analysis.

Subject: first 2 components capture 100% of the random variance.

Item: first 9 components capture 100% of the random variance.

Dropping Variance Components.

Position was removed from the item random effect.

```

model_JPN_6_nocor <- list()
model_JPN_6_nocor[[1]] <- model_JPN
model_JPN_6_nocor[[2]] <- update(model_JPN_6_nocor[[1]], .~.-(1|SubjectID)
-(1|ItemID)+
                                Sentence.Typicality*Word.Typicality + Position
+
                                (1 + z.RT.Sentence ||SubjectID) +
                                (1+Sentence.Typicality+Word.Typicality + z.RT.S
entence ||ItemID))

```

Output.

Warning: Model failed to converge with 1 negative eigenvalue: -1.1e-04

Random-Effects Principal Components Analysis.

Subject: first 2 components capture 100% of the random variance.

Item: first 7 components capture 100% of the random variance.

Dropping Variance Components.

Word.Typicality was removed from the item random effect.

```
model_JPN_7_nocor <- list()
model_JPN_7_nocor[[1]] <- model_JPN
model_JPN_7_nocor[[2]] <- update(model_JPN_7_nocor[[1]], .~.-(1|SubjectID)
-(1|ItemID)+
                                Sentence.Typicality*Word.Typicality + Position
+
                                (1 + z.RT.Sentence ||SubjectID) +
                                (1+Sentence.Typicality + z.RT.Sentence||ItemI
D))
```

Output.

Warning in checkConv(attr(opt, "derivs"), optpar, ctrl = controlcheckConv, : unable to evaluate scaled gradient

Warning in checkConv(attr(opt, "derivs"), optpar, ctrl = controlcheckConv, : Model failed to converge: degenerate Hessian with 1 negative eigenvalues

Warning: Model failed to converge with 2 negative eigenvalues: -2.1e-04 -3.7e-04

Random-Effects Principal Components Analysis.

Subject: first 2 components capture 100% of the random variance.

Item: first 4 components capture 100% of the random variance.

Dropping Variance Components.

z.RT.Sentence was removed from the item random effect.

```
model_JPN_8_nocor <- list()
model_JPN_8_nocor[[1]] <- model_JPN
model_JPN_8_nocor[[2]] <- update(model_JPN_8_nocor[[1]], .~.-(1|SubjectID)
-(1|ItemID)+
                                Sentence.Typicality*Word.Typicality + Position
+
                                (1 + z.RT.Sentence ||SubjectID) +
                                (1+ Sentence.Typicality||ItemID))
```

Output.

Warning: Model failed to converge with 2 negative eigenvalues: -3.7e-05 -3.3e-02

Random-Effects Principal Components Analysis.

Subject: first 2 components capture 100% of the random variance.

Item: first 2 components capture 100% of the random variance.

Dropping Variance Components.

z.RT.Sentence was added to the item random effect. Sentence.Typicality was removed from the item random effect.

```
model_JPN_9_nocor <- list()
model_JPN_9_nocor[[1]] <- model_JPN
model_JPN_9_nocor[[2]] <- update(model_JPN_9_nocor[[1]], .~. -(1|SubjectID)
-(1|ItemID)+
                                Sentence.Typicality*Word.Typicality + Position
+
                                (1 + z.RT.Sentence ||SubjectID) +
                                (1+ z.RT.Sentence||ItemID))
```

Output.

```
summary(model_JPN_9_nocor[[2]])

## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
## method [lmerModLmerTest]
## Formula: log(RT.Stroop) ~ z.RT.Sentence + Pres.Order + Sentence.Typicality +
## Word.Typicality + Position + (1 + z.RT.Sentence || SubjectID) +
## (1 + z.RT.Sentence || ItemID) + Sentence.Typicality:Word.Typicality
## Data: JPN.model
## Control: lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+05),
## check.conv.singular = .makeCC(action = "ignore", tol = 1e-04))
##
##      AIC      BIC   logLik deviance df.resid
## -278.4   -184.9   153.2  -306.4    5867
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.8249 -0.6656 -0.0827  0.6005  3.6808
##
## Random effects:
## Groups      Name                Variance Std.Dev.
## ItemID      z.RT.Sentence 0.0004504 0.02122
## ItemID.1    (Intercept)    0.0102160 0.10107
## SubjectID   z.RT.Sentence 0.0008753 0.02959
## SubjectID.1 (Intercept)    0.0144085 0.12004
## Residual                    0.0503459 0.22438
## Number of obs: 5881, groups: ItemID, 180; SubjectID, 36
##
## Fixed effects:
##
##              Estimate Std. Error      df
## (Intercept)  6.589e+00  2.228e-02  5.235e+0
## 1
## z.RT.Sentence  6.345e-02  6.614e-03  2.955e+01
```

```

## Pres.Order -2.879e-04 6.016e-05 5.721e+0
3
## Sentence.Typicality2-1 3.728e-03 1.618e-02 1.809e+
02
## Word.Typicality2-1 -5.383e-02 1.982e-02 1.808e+
02
## Word.Typicality3-2 6.837e-02 1.982e-02 1.809e+
02
## Position2-1 -1.128e-03 1.627e-02 1.850e+
02
## Sentence.Typicality2-1:Word.Typicality2-1 6.970e-03 3.963e-02 1.808
e+02
## Sentence.Typicality2-1:Word.Typicality3-2 -9.188e-03 3.964e-02 1.809
e+02
## t value Pr(>|t|)
## (Intercept) 295.751 < 2e-16 ***
## z.RT.Sentence 9.594 1.39e-10 ***
## Pres.Order -4.786 1.74e-06 ***
## Sentence.Typicality2-1 0.230 0.818080
## Word.Typicality2-1 -2.716 0.007239 **
## Word.Typicality3-2 3.450 0.000699 ***
## Position2-1 -0.069 0.944822
## Sentence.Typicality2-1:Word.Typicality2-1 0.176 0.860599
## Sentence.Typicality2-1:Word.Typicality3-2 -0.232 0.816967
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
## (Intr) z.RT.S Prs.Or Sn.T2-1 W.T2-1 W.T3-2 Pst2-1 S.T2-1:W.
T2
## z.RT.Sentnc -0.023
## Pres.Order -0.242 0.154
## Sntnc.Ty2-1 0.000 0.002 -0.001
## WrD.Typc2-1 0.001 0.004 -0.002 0.000
## WrD.Typc3-2 0.000 -0.001 -0.001 0.001 -0.500
## Position2-1 0.000 -0.010 -0.004 0.000 0.001 -0.001
## S.T2-1:W.T2 0.000 0.002 0.000 0.000 0.000 0.000 -0.001
## S.T2-1:W.T3 0.001 -0.001 -0.001 0.000 0.000 0.001 0.001 -0.500

```

Random-Effects Principal Components Analysis.

Subject: first 2 components capture 100% of the random variance.

Item: first 2 components capture 100% of the random variance.

Dropping Variance Components.

z.RT.Sentence was removed from the item random effect.


```

model_JPN_10_nocor <- list()
model_JPN_10_nocor[[1]] <- model_JPN
model_JPN_10_nocor[[2]] <- update(model_JPN_10_nocor[[1]], .~.-(1|SubjectID)-(1|ItemID)+
                                Sentence.Typicality*Word.Typicality + Position
+
                                (1 + z.RT.Sentence ||SubjectID) +
                                (1|ItemID))

```

Output.

Warning: Model failed to converge with 2 negative eigenvalues: -3.7e-05 -3.3e-02

Random-Effects Principal Components Analysis.

Subject: first 2 components capture 100% of the random variance.

Item: first 1 component capture 100% of the random variance.

Dropping Variance Components.

z.RT.Sentence was removed from the subject random effect.

```

model_JPN_11_nocor <- list()
model_JPN_11_nocor[[1]] <- model_JPN
model_JPN_11_nocor[[2]] <- update(model_JPN_11_nocor[[1]], .~.-(1|SubjectID)-(1|ItemID)+
                                Sentence.Typicality*Word.Typicality + Position
+
                                (1|SubjectID) +
                                (1+ z.RT.Sentence ||ItemID))

```

Model Comparisons.

A log likelihood ratio test showed that the model that included the scaled sentence reading time for the both item and subject random effects (model_JPN_9_nocor) showed significantly lower AIC than the model that included the scaled sentence reading time for item random effect (model_JPN_11_nocor). Thus, the model_JPN_9_nocor was chosen.

```

anova(model_JPN_9_nocor[[2]], model_JPN_11_nocor[[2]])

## Data: JPN.model
## Models:
## model_JPN_11_nocor[[2]]: log(RT.Stroop) ~ z.RT.Sentence + Pres.Order +
  Sentence.Typicality + Word.Typicality + Position + (1 | SubjectID) + (1
+ z.RT.Sentence || ItemID) + Sentence.Typicality:Word.Typicality
## model_JPN_9_nocor[[2]]: log(RT.Stroop) ~ z.RT.Sentence + Pres.Order +
  Sentence.Typicality + Word.Typicality + Position + (1 + z.RT.Sentence ||

```

```

SubjectID) + (1 + z.RT.Sentence || ItemID) + Sentence.Typicality:Word.Typicality
##              npar      AIC      BIC logLik deviance Chisq Df
## model_JPN_11_nocor[[2]]  13 -251.63 -164.80 138.81 -277.63
## model_JPN_9_nocor[[2]]   14 -278.43 -184.92 153.22 -306.43 28.8 1
##              Pr(>Chisq)
## model_JPN_11_nocor[[2]]
## model_JPN_9_nocor[[2]]   8.025e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Checking If Including Correlation Parameter Increases the Goodness-of-Fit

The correlation parameter was added to the model_EN_8_nocor and compare the AIC with the zero-correlation-parameter model.

```

model_JPN_9_withcor <- list()
model_JPN_9_withcor[[1]] <- model_JPN
model_JPN_9_withcor[[2]] <- update(model_JPN_9_withcor[[1]], .~. - (1|SubjectID) - (1|ItemID) +
                                Sentence.Typicality*Word.Typicality + Position
+
                                (1 + z.RT.Sentence | SubjectID) +
                                (1 + z.RT.Sentence | ItemID))

```

Model Comparisons.

A log likelihood ratio test showed that including the correlation parameter significantly reduce the AIC score. Therefore, the model without the correlation parameter was chosen as the final model.

```

anova(model_JPN_9_nocor[[2]], model_JPN_9_withcor[[2]])
## Data: JPN.model
## Models:
## model_JPN_9_nocor[[2]]: log(RT.Stroop) ~ z.RT.Sentence + Pres.Order +
Sentence.Typicality + Word.Typicality + Position + (1 + z.RT.Sentence ||
SubjectID) + (1 + z.RT.Sentence || ItemID) + Sentence.Typicality:Word.Typicality
## model_JPN_9_withcor[[2]]: log(RT.Stroop) ~ z.RT.Sentence + Pres.Order
+ Sentence.Typicality + Word.Typicality + Position + (1 + z.RT.Sentence |
SubjectID) + (1 + z.RT.Sentence | ItemID) + Sentence.Typicality:Word.Typicality
##              npar      AIC      BIC logLik deviance  Chisq Df
## model_JPN_9_nocor[[2]]   14 -278.43 -184.92 153.22 -306.43
## model_JPN_9_withcor[[2]]  16 -280.84 -173.97 156.42 -312.84 6.4119
2
##              Pr(>Chisq)

```

```
## model_JPN_9_nocor[[2]]
## model_JPN_9_withcor[[2]] 0.04052 *
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Results of the Final Model

Summary of the Final Model.

```
## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
## method [lmerModLmerTest]
## Formula: log(RT.Stroop) ~ z.RT.Sentence + Pres.Order + Sentence.Typicality +
## Word.Typicality + Position + (1 + z.RT.Sentence | SubjectID) +
## (1 + z.RT.Sentence | ItemID) + Sentence.Typicality:Word.Typicality
## Data: JPN.model
## Control: lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+
## 05),
## check.conv.singular = .makeCC(action = "ignore", tol = 1e-04))
##
##      AIC      BIC   logLik deviance df.resid
## -280.8   -174.0   156.4   -312.8    5865
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.7892 -0.6617 -0.0844  0.6032  3.6784
##
## Random effects:
## Groups      Name                Variance Std.Dev. Corr
## ItemID      (Intercept)          0.0102908 0.10144
##              z.RT.Sentence      0.0004432 0.02105  0.36
## SubjectID   (Intercept)          0.0144560 0.12023
##              z.RT.Sentence      0.0009546 0.03090  0.34
## Residual                    0.0503324 0.22435
## Number of obs: 5881, groups: ItemID, 180; SubjectID, 36
##
## Fixed effects:
##
##              Estimate Std. Error      df
## (Intercept)  6.589e+00  2.232e-02  5.230e+0
## 1
## z.RT.Sentence  6.487e-02  6.793e-03  2.748e+
## 01
## Pres.Order    -2.838e-04  6.015e-05  5.719e+0
## 3
## Sentence.Typicality2-1  4.336e-03  1.603e-02  1.817e+
## 02
## Word.Typicality2-1    -5.560e-02  1.963e-02  1.818e+
## 02
## Word.Typicality3-2     7.064e-02  1.963e-02  1.816e+
## 02
## Position2-1         3.499e-03  1.610e-02  1.850e+0
```

```

2
## Sentence.Typicality2-1:Word.Typicality2-1 5.592e-03 3.926e-02 1.818
e+02
## Sentence.Typicality2-1:Word.Typicality3-2 -1.021e-02 3.925e-02 1.816
e+02
## t value Pr(>|t|)
## (Intercept) 295.250 < 2e-16 ***
## z.RT.Sentence 9.549 3.18e-10 ***
## Pres.Order -4.719 2.43e-06 ***
## Sentence.Typicality2-1 0.271 0.787060
## Word.Typicality2-1 -2.832 0.005146 **
## Word.Typicality3-2 3.599 0.000411 ***
## Position2-1 0.217 0.828224
## Sentence.Typicality2-1:Word.Typicality2-1 0.142 0.886900
## Sentence.Typicality2-1:Word.Typicality3-2 -0.260 0.794989
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
## (Intr) z.RT.S Prs.Or Sn.T2-1 W.T2-1 W.T3-2 Pst2-1 S.T2-1:W.
T2
## z.RT.Sentnc 0.238
## Pres.Order -0.242 0.152
## Sntnc.Ty2-1 0.000 -0.001 -0.002
## Wrđ.Typc2-1 0.001 -0.003 -0.006 -0.001
## Wrđ.Typc3-2 0.000 0.001 0.001 0.001 -0.500
## Position2-1 0.000 -0.009 -0.006 0.000 0.000 0.001

## S.T2-1:W.T2 0.000 -0.001 -0.001 0.000 0.000 0.001 0.001
## S.T2-1:W.T3 0.001 -0.001 -0.001 0.000 0.001 0.000 0.000 -0.500

```

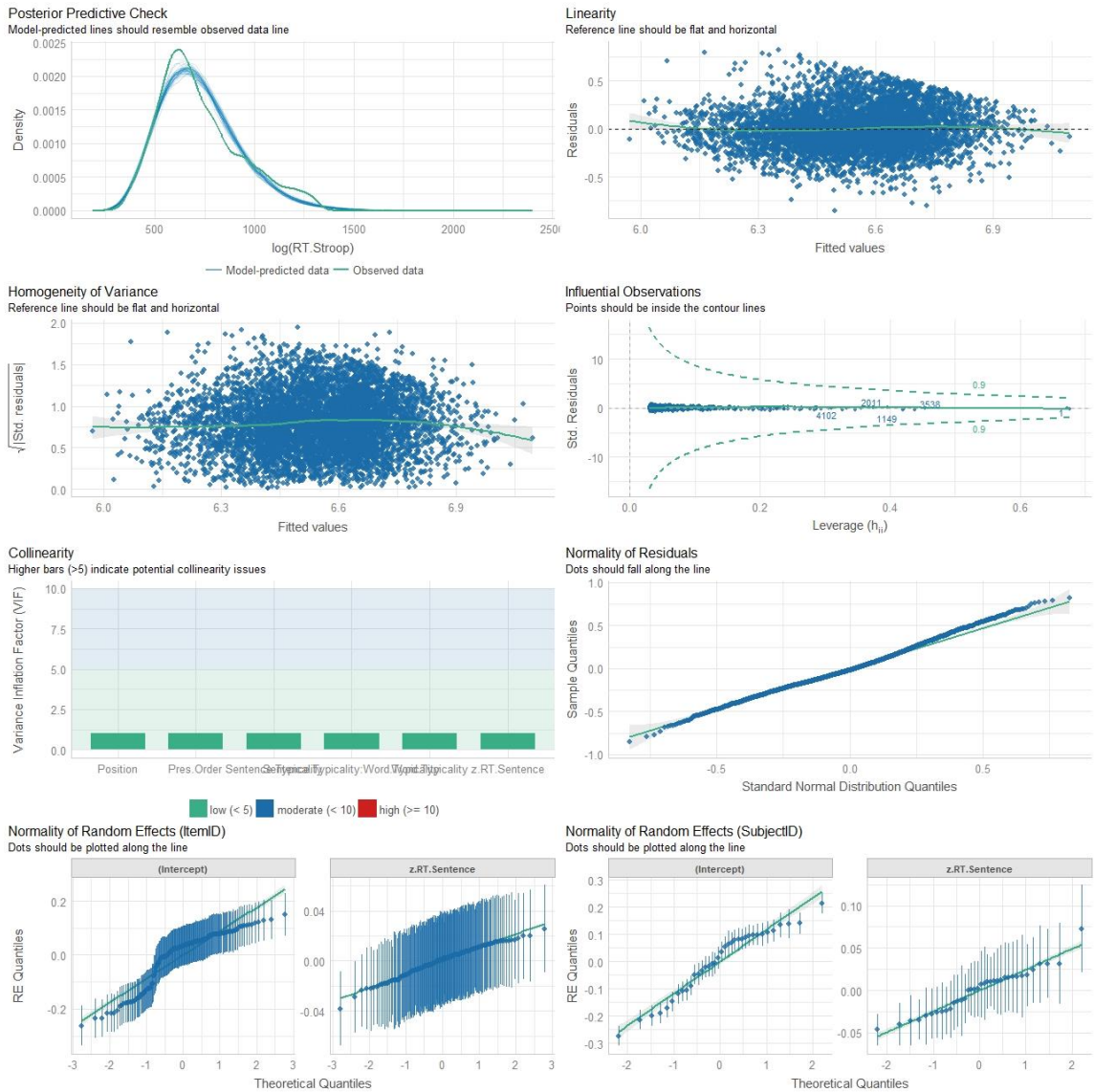
Variance Inflation Factors (VIF).

```

## # Check for Multicollinearity
##
## Low Correlation
##
## Term VIF Increased SE Tolerance
## z.RT.Sentence 1.02 1.01 0.98
## Pres.Order 1.02 1.01 0.98
## Sentence.Typicality 1.00 1.00 1.00
## Word.Typicality 1.00 1.00 1.00
## Position 1.00 1.00 1.00
## Sentence.Typicality:Word.Typicality 1.00 1.00 1.00

```

Model Diagnosis.



The Model That Only Includes Significant Predictors

```
final_model_JPN_3_withcor <- list()
final_model_JPN_3_withcor[[1]] <- model_JPN
final_model_JPN_3_withcor[[2]] <- update(final_model_JPN_3_withcor[[1]], .
~.-(1|SubjectID)-(1|ItemID)+
      Word.Typicality +
      (1 + z.RT.Sentence |SubjectID) +
      (1+ z.RT.Sentence|ItemID))

## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
## method [lmerModLmerTest]
## Formula: log(RT.Stroop) ~ z.RT.Sentence + Pres.Order + Word.Typicality
+
```

```

##      (1 + z.RT.Sentence | SubjectID) + (1 + z.RT.Sentence | ItemID)
##      Data: JPN.model
## Control: lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+
05),
##      check.conv.singular = .makeCC(action = "ignore", tol = 1e-04))
##
##      AIC      BIC    logLik deviance df.resid
## -288.7    -208.5    156.3   -312.7     5869
##
## Scaled residuals:
##      Min      1Q  Median      3Q      Max
## -3.7951 -0.6627 -0.0845  0.6032  3.6746
##
## Random effects:
## Groups      Name                Variance Std.Dev. Corr
## ItemID      (Intercept)          0.0102922 0.10145
##              z.RT.Sentence      0.0004433 0.02105  0.36
## SubjectID   (Intercept)          0.0144589 0.12024
##              z.RT.Sentence      0.0009573 0.03094  0.34
## Residual                    0.0503318 0.22435
## Number of obs: 5881, groups: ItemID, 180; SubjectID, 36
##
## Fixed effects:
##              Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)    6.589e+00  2.232e-02  5.230e+01 295.224 < 2e-16 **
## *
## z.RT.Sentence    6.489e-02  6.799e-03  2.744e+01   9.545 3.25e-10 **
## *
## Pres.Order      -2.838e-04  6.015e-05  5.719e+03  -4.717 2.45e-06 **
## *
## Word.Typicality2-1 -5.555e-02  1.964e-02  1.818e+02  -2.828 0.005202 *
## *
## Word.Typicality3-2  7.058e-02  1.964e-02  1.816e+02   3.595 0.000418 *
## **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##              (Intr) z.RT.S Prs.Or W.T2-1
## z.RT.Sentnc  0.238
## Pres.Order  -0.242  0.152
## Wrđ.Typc2-1  0.001 -0.002 -0.006
## Wrđ.Typc3-2  0.000  0.001  0.001 -0.500

```

Appendix K: List of Correction for the English Sentences

The italicized words in *Before* column were the words before correction and the words in *After* column were the words after correction.

Item ID	Before	After
Sentences		
3	It didn't <i>looked</i> ready to eat when Mark bought the strawberry.	look
4	It didn't <i>looked</i> ready to eat when Mark bought the strawberry.	look
6	It didn't <i>looked</i> ready to eat when Mark bought the strawberry.	look
25	Sarah stopped in front of a tree and <i>pick</i> a leaf off.	picked
26	Sarah sat on the ground and <i>pick</i> a leaf up.	picked
27	Sarah stopped in front of a tree and <i>pick</i> a leaf off.	picked
28	Sarah sat on the ground and <i>pick</i> a leaf up.	picked
29	Sarah stopped in front of a tree and <i>pick</i> a leaf off.	picked
30	Sarah sat on the ground and <i>pick</i> a leaf up.	picked
265	For pasta, Mercy went to a <i>moutain</i> to find a certain mushroom.	mountain
355	Mercy went to a <i>moutain</i> to find a certain mushroom for pasta.	mountain
252	During the 1960s, some people <i>cosidered</i> TV bad for kids.	considered
342	Some people <i>cosidered</i> TV bad for kids during the 1960s.	considered

Comprehension Questions

193	Logan got his <i>alchohole</i> .	alcohol
-----	----------------------------------	---------

(Sentence: In the morning, Logan stopped at the bar to pick up his salad)

270	<p>A <i>dolphine</i> was found near Japan. (Sentence: Off the coast of Japan, the whale was found)</p>	dolphin
255	<p>The <i>door</i> was being opened all day. (Sentence: During the day, the window was kept open)</p>	window
345	<p>The <i>door</i> was closed all day. (Sentence: The window was kept open during the day)</p>	window

Appendix L: Rating Task (Native Japanese Speakers Learning English)

Word

Rating Scores

The following table summarizes the descriptive statistics of the word typicality rating task for each word. The second column represents the typicality of the combinations that were shown in the first column.

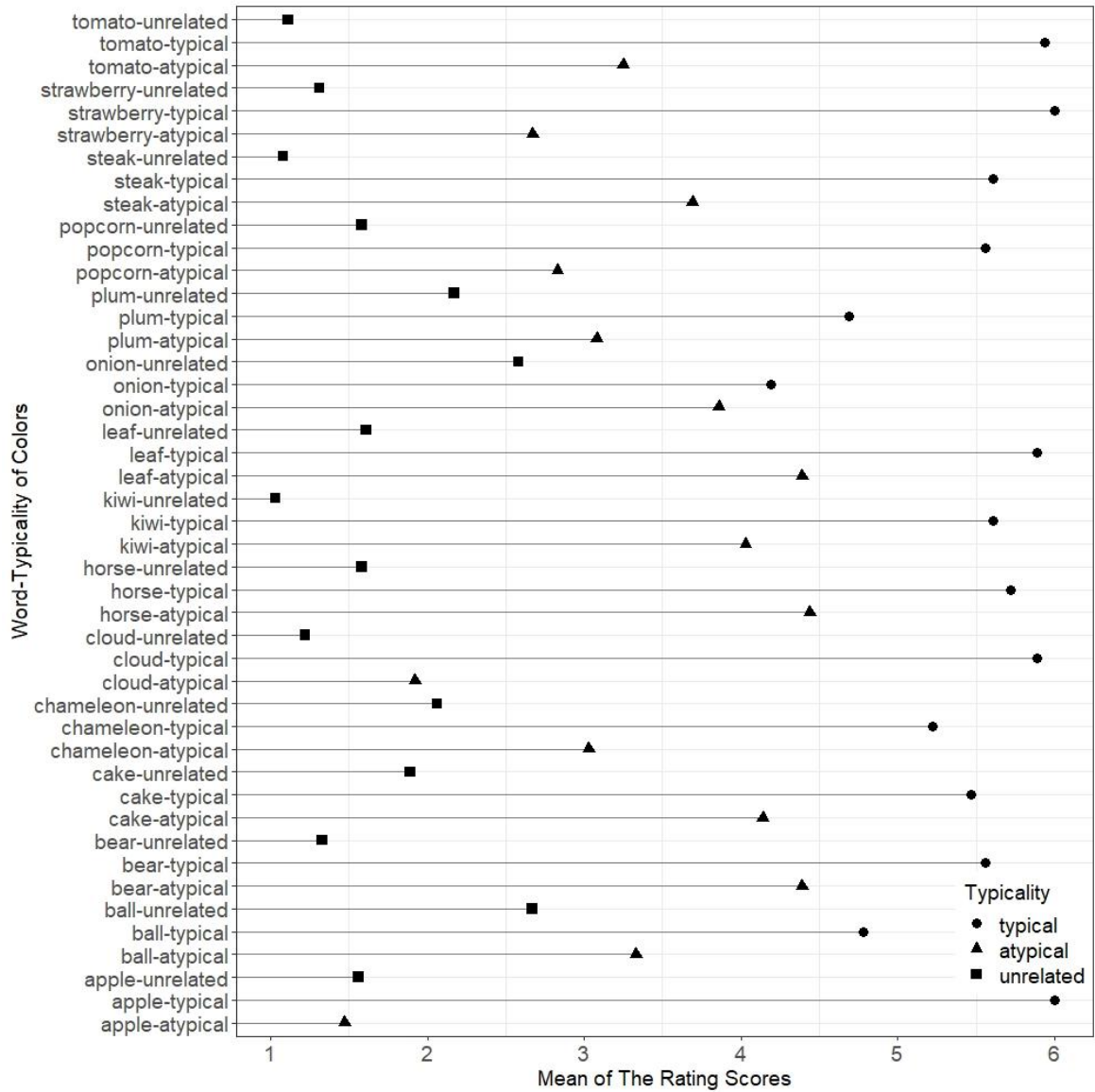
<i>Word-Color</i>	<i>Typicality</i>	<i>M</i>	<i>SD</i>	<i>Median</i>	<i>Min</i>	<i>Max</i>
apple-BROWN	unrelated	1.56	0.73	1.00	1.00	4.00
apple-RED	typical	6.00	0.00	6.00	6.00	6.00
apple-WHITE	atypical	1.47	0.77	1.00	1.00	3.00
ball-BROWN	atypical	3.33	1.35	3.00	1.00	6.00
ball-GREEN	unrelated	2.67	1.31	3.00	1.00	6.00
ball-WHITE	typical	4.78	1.38	5.00	1.00	6.00
bear-BROWN	typical	5.56	0.91	6.00	3.00	6.00
bear-GREEN	unrelated	1.33	0.63	1.00	1.00	3.00
bear-WHITE	atypical	4.39	1.13	5.00	2.00	6.00
cake-BROWN	atypical	4.14	1.27	5.00	1.00	6.00
cake-GREEN	unrelated	1.89	1.12	1.50	1.00	5.00
cake-WHITE	typical	5.47	0.70	6.00	4.00	6.00
chameleon-BROWN	atypical	3.03	1.56	3.00	1.00	6.00
chameleon-GREEN	typical	5.22	1.48	6.00	1.00	6.00
chameleon-WHITE	unrelated	2.06	1.24	2.00	1.00	6.00
cloud-GREEN	unrelated	1.22	0.48	1.00	1.00	3.00
cloud-RED	atypical	1.92	1.11	1.50	1.00	5.00
cloud-WHITE	typical	5.89	0.32	6.00	5.00	6.00
horse-BROWN	typical	5.72	0.45	6.00	5.00	6.00
horse-RED	unrelated	1.58	0.87	1.00	1.00	5.00
horse-WHITE	atypical	4.44	1.25	5.00	1.00	6.00
kiwi-BROWN	atypical	4.03	1.42	4.00	1.00	6.00
kiwi-GREEN	typical	5.61	0.69	6.00	3.00	6.00
kiwi-RED	unrelated	1.03	0.17	1.00	1.00	2.00
leaf-BROWN	atypical	4.39	1.10	5.00	2.00	6.00

<i>Word-Color</i>	<i>Typicality</i>	<i>M</i>	<i>SD</i>	<i>Median</i>	<i>Min</i>	<i>Max</i>
leaf-GREEN	typical	5.89	0.32	6.00	5.00	6.00
leaf-WHITE	unrelated	1.61	0.77	1.00	1.00	4.00
onion-BROWN	typical	4.19	1.55	4.50	1.00	6.00
onion-RED	unrelated	2.58	1.50	2.50	1.00	6.00
onion-WHITE	atypical	3.86	1.55	4.00	1.00	6.00
plum-BROWN	unrelated	2.17	0.97	2.00	1.00	4.00
plum-GREEN	atypical	3.08	1.54	3.00	1.00	6.00
plum-RED	typical	4.69	1.56	5.00	1.00	6.00
popcorn-BROWN	atypical	2.83	1.18	3.00	1.00	5.00
popcorn-RED	unrelated	1.58	0.84	1.00	1.00	4.00
popcorn-WHITE	typical	5.56	0.91	6.00	3.00	6.00
steak-BROWN	typical	5.61	0.77	6.00	2.00	6.00
steak-GREEN	unrelated	1.08	0.28	1.00	1.00	2.00
steak-RED	atypical	3.69	1.62	4.00	1.00	6.00
strawberry-BROWN	unrelated	1.31	0.67	1.00	1.00	4.00
strawberry-GREEN	atypical	2.67	1.47	2.50	1.00	6.00
strawberry-RED	typical	6.00	0.00	6.00	6.00	6.00
tomato-GREEN	atypical	3.25	1.32	3.00	1.00	6.00
tomato-RED	typical	5.94	0.23	6.00	5.00	6.00
tomato-WHITE	unrelated	1.11	0.32	1.00	1.00	2.00

Cleveland Dot Plot

In the following figure, the y-axis represents the word-typicality combinations. For example, “bear-typical” is equal to “a BROWN bear.” The x-axis represents the mean of

the word typicality rating scores. The legend represents the correspondence of the shapes of the plots and the typicality of colors.



Sentence

Agreement Rates

The following tables summarize the agreement rates (%) of each typicality that was implied by the sentences.

Entirely.

<i>Typicality</i>	<i>Agreement Rates</i>
typical	80.19
atypical	73.70

Each Sentence.

<i>Word</i>	<i>Typical</i>	<i>Atypical</i>
apple	61.11	47.22
ball	86.11	86.11
bear	91.67	91.67
cake	83.33	61.11
chameleon	83.33	80.56
cloud	72.22	86.11
horse	63.89	77.78
kiwi	69.44	83.33
leaf	36.11	58.33
onion	94.44	27.78
plum	91.67	86.11
popcorn	77.78	83.33
steak	97.22	75.00
strawberry	97.22	83.33
tomato	97.22	77.78

The atypical sentence of *onion* had the smallest agreement rate (27.8 percent) followed by the typical sentence of *leaf* (36.1 percent) and the atypical sentence of *apple* (47.2 percent). This is because 44.4 percent of the participants judged the atypical sentence of *onion* (*Liz took an onion from out of the pot*) that the matched with the image of the typical color of

onion (brown). Moreover, 22.2 percent of the participants judged the atypical sentence to match both the typical and atypical color of *onion*.

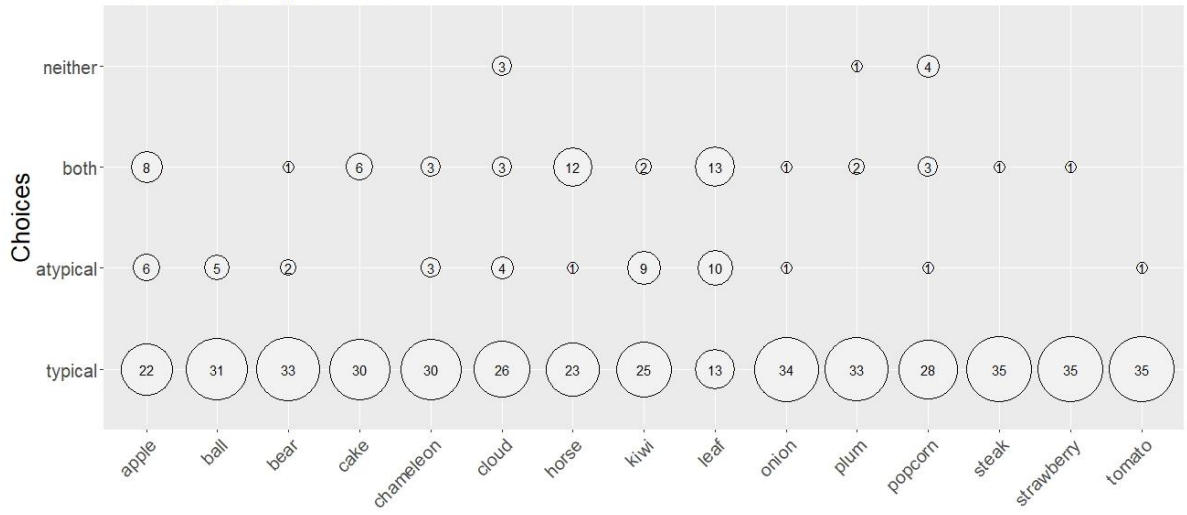
Balloon Plot

The test sentences were presented with two pictures and four forced choice alternatives:

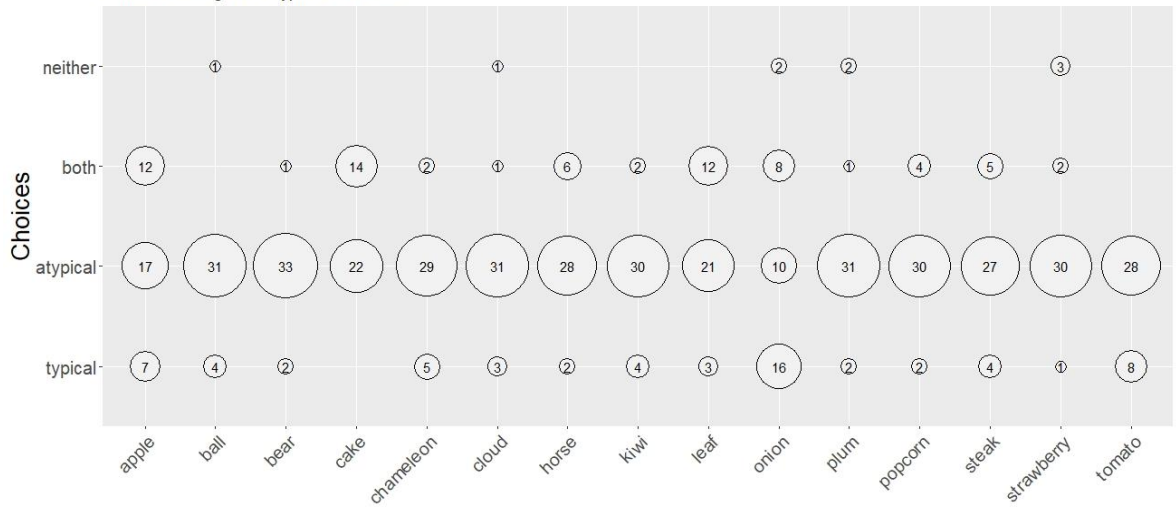
- typical: best matched by the first picture (the first pictures were always typical objects)
- atypical: best matched by the second picture (the second pictures were always atypical objects)
- both: matched by both pictures equally
- neither: matched by neither picture

The numbers in each balloon refers to the number of the participants who selected the choice.

Sentence Rating task: Typical Sentence



Sentence Rating task: Atypical Sentence



Appendix M: Statistical Modeling (Japanese Learners of English)

List of Variables

- SubjectID: Subject ID
- ItemID: Item ID
- Set: Set number
- Position: Whether the phrase that determine the color are placed before or after the keywords
- Pres.Order: Presentation order
- Sentence.Typicality: Typicality of the colors that sentences implied (e.g., *bear in the woods* implies a brown bear [typical], and *bear at the North Pole* implies a white bear [atypical])
- Word: Stimuli (Word)
- Word.Typicality: Typicality of the colors of the fonts (e.g., a brown *bear* represents a typical bear, a white *bear* represents an atypical bear)
- RT.Stroop: Reaction times of the semantic Stroop task
- RT.Sentence: Reading times for each sentence
- z.RT.Sentence: Scaled reading times for each sentence
- VocabSize: Scores of the vocabulary size test
- z.VocabSize: Scaled scores of the vocabulary size test

Change Coding of the Categorical Variables

Sentence

```
L2.model$Sentence.Typicality <- factor(L2.model$Sentence.Typicality, levels = c("typical", "atypical"))
contrasts(L2.model$Sentence.Typicality) <- fractions(contr.sdif(2))
contrasts(L2.model$Sentence.Typicality)
```

```
##           2-1
## typical  -1/2
## atypical  1/2
```

```
L2.model$Word.Typicality <- factor(L2.model$Word.Typicality, levels = c("unrelated", "typical", "atypical"))
contrasts(L2.model$Word.Typicality) <- fractions(contr.sdif(3))
contrasts(L2.model$Word.Typicality)
```

```
##           2-1  3-2
## unrelated -2/3 -1/3
```



```

## typical    1/3 -1/3
## atypical   1/3  2/3

L2.model$Position <- factor(L2.model$Position, levels = c("Pre", "Post"))
contrasts(L2.model$Position) <- fractions(contr.sdif(2))
contrasts(L2.model$Position)

##      2-1
## Pre -1/2
## Post 1/2

```

Scaling the Continuous Variables

Sentence Reading Time

```

L2.model%>%
  mutate(across(RT.Sentence, ~scale(.x)[,1], .names = "z.{.col}")) -> L2.model

```

Scores of the Vocabulary Size Test

```

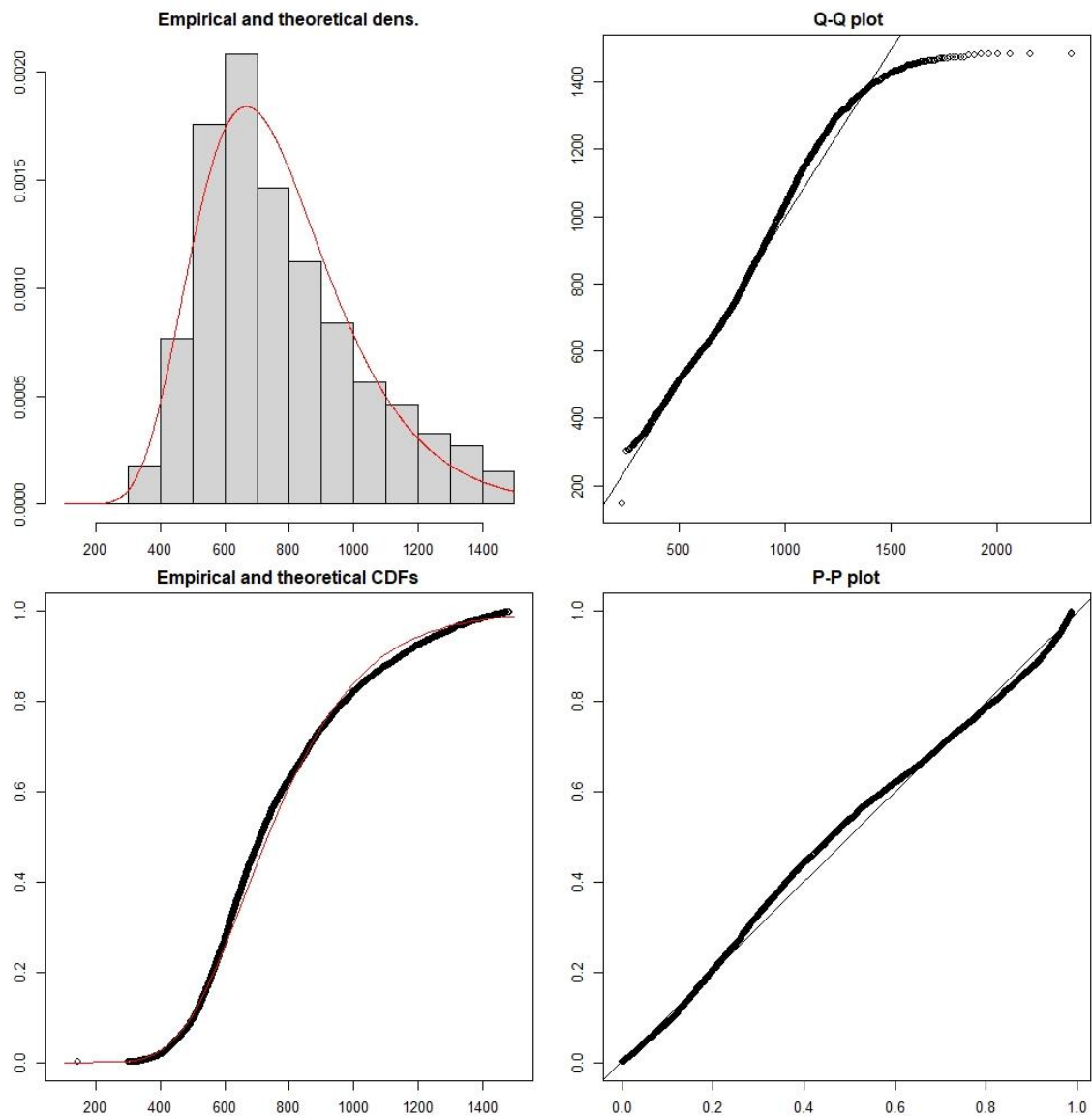
L2.model%>%
  mutate(across(VocabSize, ~scale(.x)[,1], .names = "z.{.col}")) -> L2.model

```

Choose Probabilistic Distributions for the Observed Data

According to the goodness-of-fit statistics and information criterion, Log-normal distribution was chosen for the probabilistic distribution. The top-left panel:

- Histogram: The observed data (Reaction times of the semantic Stroop task)
- Red line: The density curve



Possible Covariates

The null model was compared with the model including the possible covariates.

```

model_L2_backward <- list()
model_L2_backward[[1]] <- lmer(log(RT.Stroop) ~ + (1|SubjectID)+(1|ItemI
D),
                                data = L2.model,
                                REML = FALSE, lmerControl(optimizer = "bobyq
a",
                                                                optCtrl=list(maxfun=
200000),
                                                                check.conv.singular
= .makeCC(action = "ignore", tol = 1e-4)))

```

```

model_L2_backward[[2]] <- lmer(log(RT.Stroop) ~ + Pres.Order + (1|Subject
ID)+(1|ItemID),
                                data = L2.model,
                                REML = FALSE, lmerControl(optimizer = "bobyq
a",
                                                                optCtrl=list(maxfun=
200000),
                                                                check.conv.singular
= .makeCC(action = "ignore", tol = 1e-4)))

model_L2_backward[[3]] <- lmer(log(RT.Stroop) ~ +z.RT.Sentence + (1|Subje
ctID)+(1|ItemID),
                                data = L2.model,
                                REML = FALSE, lmerControl(optimizer = "bobyqa
",
                                                                optCtrl=list(maxfun=2
00000),
                                                                check.conv.singular =
.makeCC(action = "ignore", tol = 1e-4)))

sapply(model_L2_backward, AIC) %>%
  data.frame

##           .
## 1 135.74670
## 2  -83.79249
## 3  -70.51260

sapply(model_L2_backward, AIC) %>%
  which.min

## [1] 2

```

The model including the presentation order showed the lowest AIC among the three models. Then the model with presentation order was compared with the model with the both presentation order and scaled sentence reading time.

```

model_L2_backward_2 <- list()
model_L2_backward_2[[1]] <- model_L2_backward[[2]]
model_L2_backward_2[[2]] <- stats::update(model_L2_backward_2[[1]], .~.+
z.RT.Sentence)

sapply(model_L2_backward_2, AIC) %>%
  data.frame

##           .
## 1  -83.79249
## 2 -203.48339

sapply(model_L2_backward_2, AIC) %>%
  which.min

```

```
## [1] 2
```

Specification of the Best Random-Effects Structure

Maximal Model

```
model_L2_1 <- list()
model_L2_1[[1]] <- model_L2
model_L2_1[[2]] <- update(model_L2_1[[1]], .~.-(1|SubjectID)-(1|ItemID)+
                          Sentence.Typicality*Word.Typicality*z.VocabSize
                          + Position +
                          (1+Sentence.Typicality+Word.Typicality+z.VocabSi
                          ze + Position + Pres.Order + z.RT.Sentence|SubjectID) +
                          (1+Sentence.Typicality+Word.Typicality+z.VocabSi
                          ze + Position + Pres.Order + z.RT.Sentence|ItemID))
```

Output.

Warning messages: 1: In `optwrap(optimizer, devfun, getStart(start, rhopp), lower = rhollower, :`

convergence code 1 from bobyqa: bobyqa – maximum number of function evaluations exceeded 2: Model failed to converge with 5 negative eigenvalues: -4.8e+00 -3.9e+01 -5.5e+01 -5.9e+01 -1.1e+02

Random-Effects Principal Components Analysis.

Subject: first 7 components capture 100% of the random variance.

Item: first 7 components capture 100% of the random variance.

Maximal Model (Zero-Correlation-Parameter)

```
model_L2_1_nocor <- list()
model_L2_1_nocor[[1]] <- model_L2
model_L2_1_nocor[[2]] <- update(model_L2_1_nocor[[1]], .~.-(1|SubjectID)-
(1|ItemID)+
                          Sentence.Typicality*Word.Typicality*z.Voca
                          bSize + Position +
                          (1+Sentence.Typicality+Word.Typicality+z.V
                          ocabSize + Position + Pres.Order + z.RT.Sentence||SubjectID) +
                          (1+Sentence.Typicality+Word.Typicality+z.V
                          ocabSize + Position + Pres.Order + z.RT.Sentence||ItemID))
```

Output.

Warning messages:

1: In `optwrap(optimizer, devfun, getStart(start, rhopp), lower = rhollower, :` convergence code 1 from bobyqa: bobyqa – maximum number of function evaluations exceeded

2: In `checkConv(attr(opt, "derivs"), optpar, ctrl = controlcheckConv, :` unable to evaluate scaled gradient

3: In checkConv(attr(opt, "derivs"), optpar, ctrl = controlcheckConv, : Model failed to converge: degenerate Hessian with 7 negative eigenvalues

4: Model failed to converge with 7 negative eigenvalues: -1.1e-01 -1.5e-01 -1.1e+00 -7.8e+00 -9.6e+00 -2.7e+01 -3.4e+01

Random-Effects Principal Components Analysis.

Subject: first 10 components capture 100% of the random variance.

Item: first 9 components capture 100% of the random variance.

Dropping Variance Components.

Pres.Order is removed from the both subject and item random effects.

```
model_L2_2_nocor <- list()
model_L2_2_nocor[[1]] <- model_L2
model_L2_2_nocor[[2]] <- update(model_L2_2_nocor[[1]], .~.-(1|SubjectID)-
(1|ItemID)+ Sentence.Typicality*Word.Typicality*z.VocabSize + Position +
(1+Sentence.Typicality+Word.Typicality+z.VocabSize + Position + z.RT.Sentence|SubjectID) +
(1+Sentence.Typicality+Word.Typicality+z.VocabSize + Position + z.RT.Sentence|ItemID))
```

Output.

Warning message:

Model failed to converge with 3 negative eigenvalues: -2.7e-05 -6.5e-05 -5.9e-04

Random-Effects Principal Components Analysis.

Subject: first 6 components capture 100% of the random variance.

Item: first 9 components capture 100% of the random variance.

Dropping Variance Components.

z.VocabSize was removed from the both subject and item random effects.

```
model_L2_3_nocor <- list()
model_L2_3_nocor[[1]] <- model_L2
model_L2_3_nocor[[2]] <- update(model_L2_3_nocor[[1]], .~.-(1|SubjectID)-
(1|ItemID)+
Sentence.Typicality*Word.Typicality*z.VocabSize + Position +
(1+Sentence.Typicality+Word.Typicality + Position + z.RT.Sentence|SubjectID) +
(1+Sentence.Typicality+Word.Typicality + Position + z.RT.Sentence|ItemID))
```

Output.

Warning message:

Model failed to converge with 3 negative eigenvalues: -1.2e-06 -1.4e-04 -6.8e-04

Random-Effects Principal Components Analysis.

Subject: first 6 components capture 100% of the random variance.

Item: first 9 components capture 100% of the random variance.

Dropping Variance Components.

Position was removed from the subject random effect.

```
model_L2_4_nocor <- list()
model_L2_4_nocor[[1]] <- model_L2
model_L2_4_nocor[[2]] <- update(model_L2_4_nocor[[1]], .~.-(1|SubjectID)-
(1|ItemID)+
Sentence.Typicality*Word.Typicality*z.Voca
bSize + Position +
(1+Sentence.Typicality+Word.Typicality +
z.RT.Sentence||SubjectID) +
(1+Sentence.Typicality+Word.Typicality + P
osition + z.RT.Sentence||ItemID))
```

Output.

Warning message:

Model failed to converge with 4 negative eigenvalues: -8.0e-05 -1.6e-04 -1.8e-04 -1.9e-03

Random-Effects Principal Components Analysis.

Subject: first 5 components capture 100% of the random variance.

Item: first 9 components capture 100% of the random variance.

Dropping Variance Components.

Sentence.Typicality was removed from the subject random effect.

```
model_L2_5_nocor <- list()
model_L2_5_nocor[[1]] <- model_L2
model_L2_5_nocor[[2]] <- stats::update(model_L2_5_nocor[[1]], .~.-(1|Subje
ctID)-(1|ItemID) +
Sentence.Typicality*Word.Typicality*
z.VocabSize + Position +
(1+Word.Typicality + z.RT.Sentence||
SubjectID) +
```

```
(1+Sentence.Typicality+Word.Typicality + Position + z.RT.Sentence||ItemID))
```

Output.

Warning messages:

- 1: In checkConv(attr(opt, "derivs"), optpar, ctrl = controlcheckConv, : unable to evaluate scaled gradient
- 2: In checkConv(attr(opt, "derivs"), optpar, ctrl = controlcheckConv, : Model failed to converge: degenerate Hessian with 2 negative eigenvalues
- 3: Model failed to converge with 4 negative eigenvalues: -6.4e-05 -1.0e-04 -2.3e-04 -2.5e-04

Random-Effects Principal Components Analysis.

Subject: first 5 components capture 100% of the random variance.

Item: first 9 components capture 100% of the random variance.

Dropping Variance Components.

Word.Typicality was removed from the subject random effect.

```
model_L2_6_nocor <- list()
model_L2_6_nocor[[1]] <- model_L2
model_L2_6_nocor[[2]] <- stats::update(model_L2_6_nocor[[1]], .~.-(1|SubjectID)-(1|ItemID) +
                                     Sentence.Typicality*Word.Typicality*
z.VocabSize + Position +
                                     (1 + z.RT.Sentence||SubjectID) +
                                     (1+Sentence.Typicality+Word.Typicality + Position + z.RT.Sentence||ItemID))
```

Output.

Warning message: Model failed to converge with 5 negative eigenvalues: -2.0e-04 -2.5e-04 -3.8e-04 -5.5e-04 -9.3e-04

Random-Effects Principal Components Analysis.

Subject: first 2 components capture 100% of the random variance.

Item: first 8 components capture 100% of the random variance.

Dropping Variance Components.

z.RT.Sentence was removed from the item random effect.

```

model_L2_7_nocor <- list()
model_L2_7_nocor[[1]] <- model_L2
model_L2_7_nocor[[2]] <- stats::update(model_L2_7_nocor[[1]], .~. - (1|SubjectID) - (1|ItemID) +
                                     Sentence.Typicality*Word.Typicality*
z.VocabSize + Position +
                                     (1 + z.RT.Sentence||SubjectID) +
                                     (1+Sentence.Typicality+Word.Typicality||ItemID))

```

Output.

Warning messages:

- 1: In checkConv(attr(opt, "derivs"), optpar, ctrl = controlcheckConv, : unable to evaluate scaled gradient
- 2: In checkConv(attr(opt, "derivs"), optpar, ctrl = controlcheckConv, : Model failed to converge: degenerate Hessian with 1 negative eigenvalues
- 3: Model failed to converge with 4 negative eigenvalues: -2.3e-06 -3.5e-05 -4.5e-05 -3.2e-04

Random-Effects Principal Components Analysis.

Subject: first 2 components capture 100% of the random variance.

Item: first 8 components capture 100% of the random variance.

Dropping Variance Components.

Position was removed from the item random effect.

```

model_L2_8_nocor <- list()
model_L2_8_nocor[[1]] <- model_L2
model_L2_8_nocor[[2]] <- stats::update(model_L2_8_nocor[[1]], .~. - (1|SubjectID) - (1|ItemID) +
                                     Sentence.Typicality*Word.Typicality*
z.VocabSize + Position +
                                     (1 + z.RT.Sentence||SubjectID) +
                                     (1+Sentence.Typicality+Word.Typicality||ItemID))

```

Output.

Warning messages:

- 1: In checkConv(attr(opt, "derivs"), optpar, ctrl = controlcheckConv, : unable to evaluate scaled gradient
- 2: In checkConv(attr(opt, "derivs"), optpar, ctrl = controlcheckConv, : Model failed to converge: degenerate Hessian with 1 negative eigenvalues

Random-Effects Principal Components Analysis.

Subject: first 2 components capture 100% of the random variance.

Item: first 6 components capture 100% of the random variance.

Dropping Variance Components.

Word.Typicality was removed from the item random effect.

```
model_L2_9_nocor <- list()
model_L2_9_nocor[[1]] <- model_L2
model_L2_9_nocor[[2]] <- stats::update(model_L2_9_nocor[[1]], .~.-(1|SubjectID)-(1|ItemID)+
                                     Sentence.Typicality*Word.Typicality*
z.VocabSize + Position +
                                     (1 + z.RT.Sentence||SubjectID) +
                                     (1+Sentence.Typicality||ItemID))
```

Output.

Warning messages:

- 1: In checkConv(attr(opt, "derivs"), optpar, ctrl = controlcheckConv, : Model is nearly unidentifiable: large eigenvalue ratio - Rescale variables?
- 2: Model failed to converge with 2 negative eigenvalues: -3.8e-04 -4.3e-04

Random-Effects Principal Components Analysis.

Subject: first 2 components capture 100% of the random variance.

Item: first 3 components capture 100% of the random variance.

Dropping Variance Components.

z.RT.Sentence was removed from the subject random effect.

```
model_L2_10_nocor <- list()
model_L2_10_nocor[[1]] <- model_L2
model_L2_10_nocor[[2]] <- update(model_L2_10_nocor[[1]], .~.-(1|SubjectID)
-(1|ItemID)+ Sentence.Typicality*Word.Typicality*z.VocabSize + Position +
(1|SubjectID) +
(1+Sentence.Typicality||ItemID))
```

Output.

```
summary(model_L2_10_nocor[[2]])
## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
## method [lmerModLmerTest]
## Formula: log(RT.Stroop) ~ Pres.Order + z.RT.Sentence + Sentence.Typicality +
```

```

## Word.Typicality + z.VocabSize + Position + (1 | SubjectID) +
## (1 + Sentence.Typicality || ItemID) + Sentence.Typicality:Word.Typ
icality +
## Sentence.Typicality:z.VocabSize + Word.Typicality:z.VocabSize +
## Sentence.Typicality:Word.Typicality:z.VocabSize
## Data: L2.model
## Control: lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+
05),
## check.conv.singular = .makeCC(action = "ignore", tol = 1e-04))
##
## AIC BIC logLik deviance df.resid
## -187.7 -47.4 114.8 -229.7 5874
##
## Scaled residuals:
## Min 1Q Median 3Q Max
## -7.2062 -0.6517 -0.0817 0.5940 3.9325
##
## Random effects:
## Groups Name Variance Std.Dev. Corr
## ItemID Sentence.Typicalitytypical 0.0109301 0.10455
## Sentence.Typicalityatypical 0.0105359 0.10264 0.09
## ItemID.1 (Intercept) 0.0004731 0.02175
## SubjectID (Intercept) 0.0351157 0.18739
## Residual 0.0513386 0.22658
## Number of obs: 5895, groups: ItemID, 180; SubjectID, 36
##
## Fixed effects:
## Estimate Std. Error
## (Intercept) 6.683e+00 3.284e-
02
## Pres.Order -7.035e-04 6.017e-
05
## z.RT.Sentence 3.906e-02 3.512e-
03
## Sentence.Typicality2-1 1.548e-03 1.685e-
02
## Word.Typicality2-1 -2.567e-02 2.064e-
02
## Word.Typicality3-2 5.307e-02 2.065e-
02
## z.VocabSize 2.043e-03 3.146e-
02
## Position2-1 1.352e-02 1.686e-
02
## Sentence.Typicality2-1:Word.Typicality2-1 -1.056e-02 4.127
e-02
## Sentence.Typicality2-1:Word.Typicality3-2 -1.216e-02 4.130
e-02
## Sentence.Typicality2-1:z.VocabSize 5.063e-03 5.910
e-03
## Word.Typicality2-1:z.VocabSize -1.466e-02 7.225e-
03

```

```

## Word.Typicality3-2:z.VocabSize 1.285e-02 7.260e
-03
## Sentence.Typicality2-1:Word.Typicality2-1:z.VocabSize 1.456e-02 1.44
5e-02
## Sentence.Typicality2-1:Word.Typicality3-2:z.VocabSize -2.312e-03 1.45
2e-02
##
## df t value
## (Intercept) 4.304e+01 203.512
## Pres.Order 5.727e+03 -11.692
## z.RT.Sentence 5.810e+03 11.121
## Sentence.Typicality2-1 1.802e+02 0.092
## Word.Typicality2-1 1.801e+02 -1.244
## Word.Typicality3-2 1.807e+02 2.570
## z.VocabSize 3.592e+01 0.065
## Position2-1 1.805e+02 0.802
## Sentence.Typicality2-1:Word.Typicality2-1
6 1.801e+02 -0.25
## Sentence.Typicality2-1:Word.Typicality3-2
4 1.806e+02 -0.29
## Sentence.Typicality2-1:z.VocabSize
7 5.684e+03 0.85
## Word.Typicality2-1:z.VocabSize 5.684e+03 -2.030
## Word.Typicality3-2:z.VocabSize 5.684e+03 1.770
## Sentence.Typicality2-1:Word.Typicality2-1:z.VocabSize 5.684e+03 1.0
07
## Sentence.Typicality2-1:Word.Typicality3-2:z.VocabSize 5.684e+03 -0.1
59
##
## Pr(>|t|)
## (Intercept) <2e-16 ***
## Pres.Order <2e-16 ***
## z.RT.Sentence <2e-16 ***
## Sentence.Typicality2-1 0.9269
## Word.Typicality2-1 0.2152
## Word.Typicality3-2 0.0110 *
## z.VocabSize 0.9486
## Position2-1 0.4236
## Sentence.Typicality2-1:Word.Typicality2-1 0.7983
## Sentence.Typicality2-1:Word.Typicality3-2 0.7688
## Sentence.Typicality2-1:z.VocabSize 0.3917
## Word.Typicality2-1:z.VocabSize 0.0424 *
## Word.Typicality3-2:z.VocabSize 0.0768 .
## Sentence.Typicality2-1:Word.Typicality2-1:z.VocabSize 0.3137
## Sentence.Typicality2-1:Word.Typicality3-2:z.VocabSize 0.8735
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Correlation matrix not shown by default, as p = 15 > 12.
## Use print(x, correlation=TRUE) or
## vcov(x) if you need it

```

Random-Effects Principal Components Analysis.

Subject: first component capture 100% of the random variance.

Item: first 3 components capture 100% of the random variance.

Dropping Variance Components.

Sentence.Typicality was removed from the item random effect.

```
model_L2_11_nocor <- list()
model_L2_11_nocor[[1]] <- model_L2
model_L2_11_nocor[[2]] <- update(model_L2_11_nocor[[1]], .~.-(1|SubjectID)
-(1|ItemID)+ Sentence.Typicality*Word.Typicality*z.VocabSize + Position +
(1 + z.RT.Sentence||SubjectID) +
(1|ItemID))
```

Output.

```
summary(model_L2_11_nocor[[2]])
## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
## method [lmerModLmerTest]
## Formula: log(RT.Stroop) ~ Pres.Order + z.RT.Sentence + Sentence.Typicality +
## Word.Typicality + z.VocabSize + Position + (1 + z.RT.Sentence ||
## SubjectID) + (1 | ItemID) + Sentence.Typicality:Word.Typicality +
## Sentence.Typicality:z.VocabSize + Word.Typicality:z.VocabSize +
## Sentence.Typicality:Word.Typicality:z.VocabSize
## Data: L2.model
## Control: lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+
05),
## check.conv.singular = .makeCC(action = "ignore", tol = 1e-04))
##
##      AIC      BIC   logLik deviance df.resid
## -280.3  -153.3   159.2  -318.3    5876
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -6.6875 -0.6484 -0.0794  0.5837  4.0207
##
## Random effects:
## Groups      Name                Variance Std.Dev.
## ItemID      (Intercept)          0.011219 0.10592
## SubjectID   z.RT.Sentence        0.002081 0.04562
## SubjectID.1 (Intercept)         0.033565 0.18321
## Residual                                0.050034 0.22368
## Number of obs: 5895, groups: ItemID, 180; SubjectID, 36
##
```

```

## Fixed effects:
##
## (Intercept)
02 Estimate Std. Error
6.685e+00 3.218e-
## Pres.Order
05 -6.370e-04 6.024e-
## z.RT.Sentence
03 6.409e-02 8.801e-
## Sentence.Typicality2-1
-02 2.098e-03 1.684e
## Word.Typicality2-1
02 -2.622e-02 2.062e-
## Word.Typicality3-2
02 5.310e-02 2.064e-
## z.VocabSize
02 8.119e-03 3.080e-
## Position2-1
02 1.995e-02 1.693e-
## Sentence.Typicality2-1:Word.Typicality2-1
e-02 -1.176e-02 4.124
## Sentence.Typicality2-1:Word.Typicality3-2
e-02 -1.324e-02 4.127
## Sentence.Typicality2-1:z.VocabSize
e-03 5.448e-03 5.843
## Word.Typicality2-1:z.VocabSize
-03 -1.559e-02 7.143e
## Word.Typicality3-2:z.VocabSize
-03 1.306e-02 7.180e
## Sentence.Typicality2-1:Word.Typicality2-1:z.VocabSize 1.565e-02 1.43
0e-02
## Sentence.Typicality2-1:Word.Typicality3-2:z.VocabSize -2.831e-03 1.43
5e-02
##
## df t value
## (Intercept) 4.331e+01 207.744
## Pres.Order 5.732e+03 -10.575
## z.RT.Sentence 3.541e+01 7.283
## Sentence.Typicality2-1 1.802e+02 0.125
## Word.Typicality2-1 1.802e+02 -1.271
## Word.Typicality3-2 1.808e+02 2.573
## z.VocabSize 3.599e+01 0.264
## Position2-1 1.842e+02 1.178
## Sentence.Typicality2-1:Word.Typicality2-1 1.801e+02 -0.28
5
## Sentence.Typicality2-1:Word.Typicality3-2 1.806e+02 -0.32
1
## Sentence.Typicality2-1:z.VocabSize 5.660e+03 0.93
2
## Word.Typicality2-1:z.VocabSize 5.658e+03 -2.183
## Word.Typicality3-2:z.VocabSize 5.659e+03 1.819
## Sentence.Typicality2-1:Word.Typicality2-1:z.VocabSize 5.660e+03 1.0
95
## Sentence.Typicality2-1:Word.Typicality3-2:z.VocabSize 5.655e+03 -0.1

```

```

97
##
## (Intercept) Pr(>|t|)
## Pres.Order < 2e-16 ***
## z.RT.Sentence 1.55e-08 ***
## Sentence.Typicality2-1 0.9010
## Word.Typicality2-1 0.2052
## Word.Typicality3-2 0.0109 *
## z.VocabSize 0.7936
## Position2-1 0.2402
## Sentence.Typicality2-1:Word.Typicality2-1 0.7759
## Sentence.Typicality2-1:Word.Typicality3-2 0.7488
## Sentence.Typicality2-1:z.VocabSize 0.3512
## Word.Typicality2-1:z.VocabSize 0.0291 *
## Word.Typicality3-2:z.VocabSize 0.0689 .
## Sentence.Typicality2-1:Word.Typicality2-1:z.VocabSize 0.2737
## Sentence.Typicality2-1:Word.Typicality3-2:z.VocabSize 0.8436
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Correlation matrix not shown by default, as p = 15 > 12.
## Use print(x, correlation=TRUE) or
## vcov(x) if you need it

```

Model Comparisons.

A log likelihood ratio test showed that there was no difference in the AIC between the models. However, the model with z.RT.Sentence for subject random slope showed lower AIC than the model with Sentence.Typicality for item random slope.

```

anova(model_L2_10_nocor[[2]],model_L2_11_nocor[[2]])

## Data: L2.model
## Models:
## model_L2_11_nocor[[2]]: log(RT.Stroop) ~ Pres.Order + z.RT.Sentence +
Sentence.Typicality + Word.Typicality + z.VocabSize + Position + (1 + z.R
T.Sentence || SubjectID) + (1 | ItemID) + Sentence.Typicality:Word.Typica
lity + Sentence.Typicality:z.VocabSize + Word.Typicality:z.VocabSize + Se
ntence.Typicality:Word.Typicality:z.VocabSize
## model_L2_10_nocor[[2]]: log(RT.Stroop) ~ Pres.Order + z.RT.Sentence +
Sentence.Typicality + Word.Typicality + z.VocabSize + Position + (1 | Sub
jectID) + (1 + Sentence.Typicality || ItemID) + Sentence.Typicality:Word.
Typicality + Sentence.Typicality:z.VocabSize + Word.Typicality:z.VocabSiz
e + Sentence.Typicality:Word.Typicality:z.VocabSize
##          npar    AIC      BIC logLik deviance Chisq Df
## model_L2_11_nocor[[2]]  19 -280.30 -153.348 159.15  -318.30
## model_L2_10_nocor[[2]]  21 -187.68  -47.358 114.84  -229.68    0  2
##          Pr(>Chisq)

```

```
## model_L2_11_nocor[[2]]
## model_L2_10_nocor[[2]]      1
```

Dropping Variance Components.

z.RT.Sentence was removed from the subject random effect.

```
model_L2_12_nocor <- list()
model_L2_12_nocor[[1]] <- model_L2
model_L2_12_nocor[[2]] <- update(model_L2_12_nocor[[1]], .~. -(1|SubjectID)
-(1|ItemID)+
                                Sentence.Typicality*Word.Typicality*z.Voc
abSize + Position +
                                (1|SubjectID) +
                                (1|ItemID))
```

Output.

```
summary(model_L2_12_nocor[[2]])
## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
## method [lmerModLmerTest]
## Formula: log(RT.Stroop) ~ Pres.Order + z.RT.Sentence + Sentence.Typicality +
## Word.Typicality + z.VocabSize + Position + (1 | SubjectID) +
## (1 | ItemID) + Sentence.Typicality:Word.Typicality + Sentence.Typicality:z.VocabSize +
## Word.Typicality:z.VocabSize + Sentence.Typicality:Word.Typicality:z.VocabSize
## Data: L2.model
## Control: lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+05),
## check.conv.singular = .makeCC(action = "ignore", tol = 1e-04))
##
##      AIC      BIC   logLik deviance df.resid
## -193.7   -73.4   114.8   -229.7    5877
##
## Scaled residuals:
##      Min      1Q   Median      3Q      Max
## -7.2050 -0.6516 -0.0818  0.5951  3.9328
##
## Random effects:
## Groups   Name              Variance Std.Dev.
## ItemID   (Intercept) 0.01121  0.1059
## SubjectID (Intercept) 0.03512  0.1874
## Residual                0.05134  0.2266
## Number of obs: 5895, groups: ItemID, 180; SubjectID, 36
##
## Fixed effects:
##
##                                     Estimate Std. Error
## (Intercept)                        6.683e+00  3.284e-
```

```

02
## Pres.Order -7.035e-04 6.017e-
05
## z.RT.Sentence 3.906e-02 3.512e-
03
## Sentence.Typicality2-1 1.562e-03 1.685e
-02
## Word.Typicality2-1 -2.567e-02 2.064e-
02
## Word.Typicality3-2 5.307e-02 2.065e-
02
## z.VocabSize 2.043e-03 3.146e-
02
## Position2-1 1.349e-02 1.686e-
02
## Sentence.Typicality2-1:Word.Typicality2-1 -1.055e-02 4.127
e-02
## Sentence.Typicality2-1:Word.Typicality3-2 -1.216e-02 4.130
e-02
## Sentence.Typicality2-1:z.VocabSize 5.063e-03 5.910
e-03
## Word.Typicality2-1:z.VocabSize -1.466e-02 7.225e
-03
## Word.Typicality3-2:z.VocabSize 1.285e-02 7.260e
-03
## Sentence.Typicality2-1:Word.Typicality2-1:z.VocabSize 1.456e-02 1.44
5e-02
## Sentence.Typicality2-1:Word.Typicality3-2:z.VocabSize -2.316e-03 1.45
2e-02
##
## df t value
## (Intercept) 4.304e+01 203.512
## Pres.Order 5.727e+03 -11.691
## z.RT.Sentence 5.810e+03 11.121
## Sentence.Typicality2-1 1.803e+02 0.093
## Word.Typicality2-1 1.801e+02 -1.244
## Word.Typicality3-2 1.807e+02 2.570
## z.VocabSize 3.592e+01 0.065
## Position2-1 1.805e+02 0.800
## Sentence.Typicality2-1:Word.Typicality2-1 1.802e+02 -0.25
6
## Sentence.Typicality2-1:Word.Typicality3-2 1.807e+02 -0.29
4
## Sentence.Typicality2-1:z.VocabSize 5.684e+03 0.85
7
## Word.Typicality2-1:z.VocabSize 5.684e+03 -2.030
## Word.Typicality3-2:z.VocabSize 5.684e+03 1.770
## Sentence.Typicality2-1:Word.Typicality2-1:z.VocabSize 5.684e+03 1.0
08
## Sentence.Typicality2-1:Word.Typicality3-2:z.VocabSize 5.684e+03 -0.1
59
##
## Pr(>|t|)
## (Intercept) <2e-16 ***

```



```

## Pres.Order <2e-16 ***
## z.RT.Sentence <2e-16 ***
## Sentence.Typicality2-1 0.9263
## Word.Typicality2-1 0.2152
## Word.Typicality3-2 0.0110 *
## z.VocabSize 0.9486
## Position2-1 0.4245
## Sentence.Typicality2-1:Word.Typicality2-1 0.7985
## Sentence.Typicality2-1:Word.Typicality3-2 0.7687
## Sentence.Typicality2-1:z.VocabSize 0.3917
## Word.Typicality2-1:z.VocabSize 0.0424 *
## Word.Typicality3-2:z.VocabSize 0.0768 .
## Sentence.Typicality2-1:Word.Typicality2-1:z.VocabSize 0.3136
## Sentence.Typicality2-1:Word.Typicality3-2:z.VocabSize 0.8733
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Correlation matrix not shown by default, as p = 15 > 12.
## Use print(x, correlation=TRUE) or
## vcov(x) if you need it

```

Model Comparisons.

Although there was no statistical difference in the AIC, the model without the random slope for the item random effect showed lower AIC than the models including Sentence.Typicality for the item random slope. Additionally a log likelihood ratio test showed that including z.RT.Sentence for the subject random slope significantly reduced the AIC score. Therefore, the model with z.RT.Sentence for the subject random slope (model_L2_11_nocor) was chosen as the final model.

```

anova(model_L2_10_nocor[[2]],model_L2_12_nocor[[2]])

## Data: L2.model
## Models:
## model_L2_12_nocor[[2]]: log(RT.Stroop) ~ Pres.Order + z.RT.Sentence +
Sentence.Typicality + Word.Typicality + z.VocabSize + Position + (1 | Sub
jectID) + (1 | ItemID) + Sentence.Typicality:Word.Typicality + Sentence.T
ypicality:z.VocabSize + Word.Typicality:z.VocabSize + Sentence.Typicalit
y:Word.Typicality:z.VocabSize
## model_L2_10_nocor[[2]]: log(RT.Stroop) ~ Pres.Order + z.RT.Sentence +
Sentence.Typicality + Word.Typicality + z.VocabSize + Position + (1 | Sub
jectID) + (1 + Sentence.Typicality | ItemID) + Sentence.Typicality:Word.
Typicality + Sentence.Typicality:z.VocabSize + Word.Typicality:z.VocabSiz
e + Sentence.Typicality:Word.Typicality:z.VocabSize
##
##          npar      AIC      BIC logLik deviance Chisq Df

```

```

## model_L2_12_nocor[[2]] 18 -193.66 -73.382 114.83 -229.66
## model_L2_10_nocor[[2]] 21 -187.68 -47.358 114.84 -229.68 0.0216 3
##
## Pr(>Chisq)
## model_L2_12_nocor[[2]]
## model_L2_10_nocor[[2]] 0.9992

anova(model_L2_11_nocor[[2]],model_L2_12_nocor[[2]])

## Data: L2.model
## Models:
## model_L2_12_nocor[[2]]: log(RT.Stroop) ~ Pres.Order + z.RT.Sentence +
Sentence.Typicality + Word.Typicality + z.VocabSize + Position + (1 | Sub
jectID) + (1 | ItemID) + Sentence.Typicality:Word.Typicality + Sentence.T
ypicality:z.VocabSize + Word.Typicality:z.VocabSize + Sentence.Typicalit
y:Word.Typicality:z.VocabSize
## model_L2_11_nocor[[2]]: log(RT.Stroop) ~ Pres.Order + z.RT.Sentence +
Sentence.Typicality + Word.Typicality + z.VocabSize + Position + (1 + z.R
T.Sentence | SubjectID) + (1 | ItemID) + Sentence.Typicality:Word.Typica
lity + Sentence.Typicality:z.VocabSize + Word.Typicality:z.VocabSize + Se
ntence.Typicality:Word.Typicality:z.VocabSize
##
## npar AIC BIC logLik deviance Chisq Df
## model_L2_12_nocor[[2]] 18 -193.66 -73.382 114.83 -229.66
## model_L2_11_nocor[[2]] 19 -280.30 -153.348 159.15 -318.30 88.648 1
##
## Pr(>Chisq)
## model_L2_12_nocor[[2]]
## model_L2_11_nocor[[2]] < 2.2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Checking If Including Correlation Parameter Increases the Goodness-of-Fit

The correlation parameter was added to the model_L2_11_nocor and compare the AIC with the zero-correlation-parameter model.

```

model_L2_11_withcor <- list()
model_L2_11_withcor[[1]] <- model_L2
model_L2_11_withcor[[2]] <- update(model_L2_11_withcor[[1]], ~.(1|Subjec
tID)-(1|ItemID)+
                                Sentence.Typicality*Word.Typicality*z.V
ocabSize + Position +
                                (1 + z.RT.Sentence|SubjectID) +
                                (1|ItemID))

```

Model Comparisons.

A log likelihood ratio test showed that including the correlation parameter did not significantly reduce the AIC score. Therefore, the model without the correlation parameter was chosen as the final model.

```

anova(model_L2_11_nocor[[2]],model_L2_11_withcor[[2]])

## Data: L2.model
## Models:
## model_L2_11_nocor[[2]]: log(RT.Stroop) ~ Pres.Order + z.RT.Sentence +
Sentence.Typicality + Word.Typicality + z.VocabSize + Position + (1 + z.R
T.Sentence | SubjectID) + (1 | ItemID) + Sentence.Typicality:Word.Typica
lity + Sentence.Typicality:z.VocabSize + Word.Typicality:z.VocabSize + Se
ntence.Typicality:Word.Typicality:z.VocabSize
## model_L2_11_withcor[[2]]: log(RT.Stroop) ~ Pres.Order + z.RT.Sentence
+ Sentence.Typicality + Word.Typicality + z.VocabSize + Position + (1 +
z.RT.Sentence | SubjectID) + (1 | ItemID) + Sentence.Typicality:Word.Typi
cality + Sentence.Typicality:z.VocabSize + Word.Typicality:z.VocabSize +
Sentence.Typicality:Word.Typicality:z.VocabSize
##
##          npar      AIC      BIC logLik deviance  Chisq Df
## model_L2_11_nocor[[2]]      19 -280.30 -153.35 159.15  -318.30
## model_L2_11_withcor[[2]]     20 -278.51 -144.88 159.26  -318.51 0.2084
1
##
##          Pr(>Chisq)
## model_L2_11_nocor[[2]]
## model_L2_11_withcor[[2]]      0.648

finalmodel <- model_L2_11_nocor[[2]]

```

Results of the Final Model

Summary of the Final Model.

```

## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
## method [lmerModLmerTest]
## Formula: log(RT.Stroop) ~ Pres.Order + z.RT.Sentence + Sentence.Typicality +
Sentence.Typicality:z.VocabSize + Word.Typicality:z.VocabSize +
Sentence.Typicality:Word.Typicality:z.VocabSize
## Data: L2.model
## Control: lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+
05),
## check.conv.singular = .makeCC(action = "ignore", tol = 1e-04))
##
##      AIC      BIC  logLik deviance df.resid
## -280.3  -153.3   159.2  -318.3    5876
##
## Scaled residuals:
##      Min      1Q  Median      3Q      Max
## -6.6875 -0.6484 -0.0794  0.5837  4.0207
##
## Random effects:
## Groups      Name              Variance Std.Dev.

```

```

## ItemID      (Intercept)    0.011219 0.10592
## SubjectID   z.RT.Sentence 0.002081 0.04562
## SubjectID.1 (Intercept)   0.033565 0.18321
## Residual                                0.050034 0.22368
## Number of obs: 5895, groups: ItemID, 180; SubjectID, 36
##
## Fixed effects:
##
##                                     Estimate Std. Error
## (Intercept)                        6.685e+00  3.218e-
02
## Pres.Order                          -6.370e-04  6.024e-
05
## z.RT.Sentence                       6.409e-02  8.801e-
03
## Sentence.Typicality2-1              2.098e-03  1.684e
-02
## Word.Typicality2-1                 -2.622e-02  2.062e-
02
## Word.Typicality3-2                  5.310e-02  2.064e-
02
## z.VocabSize                         8.119e-03  3.080e-
02
## Position2-1                        1.995e-02  1.693e-
02
## Sentence.Typicality2-1:Word.Typicality2-1 -1.176e-02  4.124
e-02
## Sentence.Typicality2-1:Word.Typicality3-2 -1.324e-02  4.127
e-02
## Sentence.Typicality2-1:z.VocabSize     5.448e-03  5.843
e-03
## Word.Typicality2-1:z.VocabSize       -1.559e-02  7.143e
-03
## Word.Typicality3-2:z.VocabSize        1.306e-02  7.180e
-03
## Sentence.Typicality2-1:Word.Typicality2-1:z.VocabSize 1.565e-02  1.43
0e-02
## Sentence.Typicality2-1:Word.Typicality3-2:z.VocabSize -2.831e-03  1.43
5e-02
##
##                                     df t value
## (Intercept)                        4.331e+01 207.744
## Pres.Order                          5.732e+03 -10.575
## z.RT.Sentence                       3.541e+01  7.283
## Sentence.Typicality2-1              1.802e+02  0.125
## Word.Typicality2-1                  1.802e+02 -1.271
## Word.Typicality3-2                  1.808e+02  2.573
## z.VocabSize                         3.599e+01  0.264
## Position2-1                         1.842e+02  1.178
## Sentence.Typicality2-1:Word.Typicality2-1 1.801e+02 -0.28
5
## Sentence.Typicality2-1:Word.Typicality3-2 1.806e+02 -0.32
1
## Sentence.Typicality2-1:z.VocabSize     5.660e+03  0.93

```

```

2
## Word.Typicality2-1:z.VocabSize          5.658e+03  -2.183
## Word.Typicality3-2:z.VocabSize          5.659e+03   1.819
## Sentence.Typicality2-1:Word.Typicality2-1:z.VocabSize  5.660e+03   1.0
95
## Sentence.Typicality2-1:Word.Typicality3-2:z.VocabSize  5.655e+03  -0.1
97
##
##                                     Pr(>|t|)
## (Intercept)                            < 2e-16 ***
## Pres.Order                             < 2e-16 ***
## z.RT.Sentence                          1.55e-08 ***
## Sentence.Typicality2-1                  0.9010
## Word.Typicality2-1                      0.2052
## Word.Typicality3-2                      0.0109 *
## z.VocabSize                             0.7936
## Position2-1                             0.2402
## Sentence.Typicality2-1:Word.Typicality2-1  0.7759
## Sentence.Typicality2-1:Word.Typicality3-2  0.7488
## Sentence.Typicality2-1:z.VocabSize       0.3512
## Word.Typicality2-1:z.VocabSize          0.0291 *
## Word.Typicality3-2:z.VocabSize          0.0689 .
## Sentence.Typicality2-1:Word.Typicality2-1:z.VocabSize  0.2737
## Sentence.Typicality2-1:Word.Typicality3-2:z.VocabSize  0.8436
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Correlation matrix not shown by default, as p = 15 > 12.
## Use print(x, correlation=TRUE) or
##   vcov(x)           if you need it

```

Variance Inflation Factors (VIF).

```

## # Check for Multicollinearity
##
## Low Correlation
##
##          Term  VIF Increased SE Toleran
ce
##          Pres.Order  1.02          1.01    0.
98
##          z.RT.Sentence  1.02          1.01    0.
98
##          Sentence.Typicality  1.00          1.00
1.00
##          Word.Typicality  1.00          1.00    1.
00
##          z.VocabSize  1.00          1.00    1.
00
##          Position  1.00          1.00    1.
00
##          Sentence.Typicality:Word.Typicality  1.00          1.00

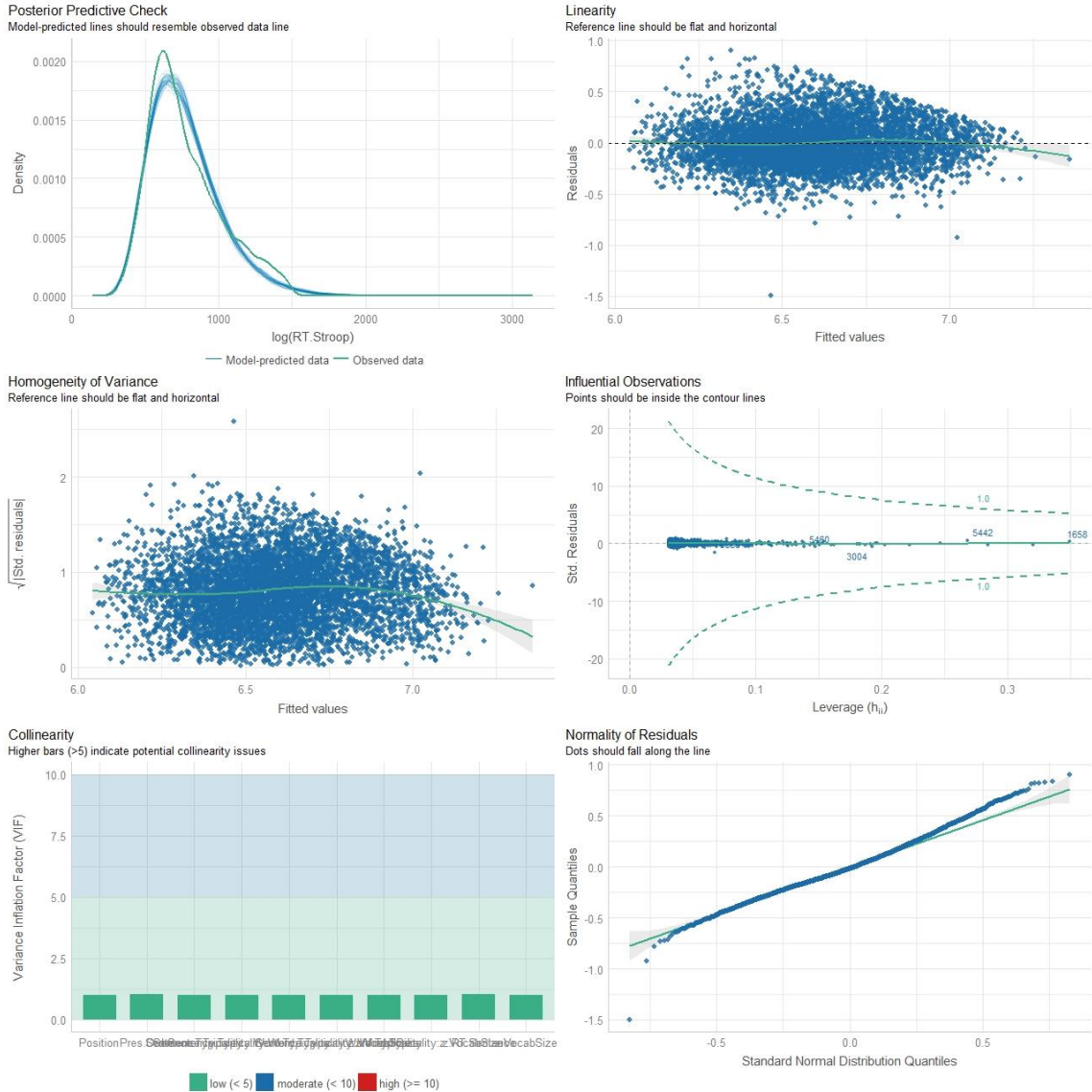
```

```

1.00
##          Sentence.Typicality:z.VocabSize 1.00      1.00
1.00
##          Word.Typicality:z.VocabSize 1.00        1.00
1.00
## Sentence.Typicality:Word.Typicality:z.VocabSize 1.00      1.00
1.00

```

Model Diagnosis.



The Model That Only Includes Significant Predictors

```

final_model_L2_3_nocor <- list()
final_model_L2_3_nocor[[1]] <- model_L2
final_model_L2_3_nocor[[2]] <- update(final_model_L2_3_nocor[[1]], .~.-(1|

```

```

SubjectID)-(1|ItemID)+ Word.Typicality*z.VocabSize +(1 + z.RT.Sentence||
SubjectID) +
(1|ItemID))

## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
## method [lmerModLmerTest]
## Formula: log(RT.Stroop) ~ Pres.Order + z.RT.Sentence + Word.Typicality
+
## z.VocabSize + (1 + z.RT.Sentence || SubjectID) + (1 | ItemID) +
## Word.Typicality:z.VocabSize
## Data: L2.model
## Control: lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+
05),
## check.conv.singular = .makeCC(action = "ignore", tol = 1e-04))
##
## AIC BIC logLik deviance df.resid
## -290.3 -210.1 157.2 -314.3 5883
##
## Scaled residuals:
## Min 1Q Median 3Q Max
## -6.6906 -0.6482 -0.0824 0.5894 4.0335
##
## Random effects:
## Groups Name Variance Std.Dev.
## ItemID (Intercept) 0.011341 0.10649
## SubjectID z.RT.Sentence 0.002059 0.04538
## SubjectID.1 (Intercept) 0.033591 0.18328
## Residual 0.050057 0.22373
## Number of obs: 5895, groups: ItemID, 180; SubjectID, 36
##
## Fixed effects:
## Estimate Std. Error df t value
## (Intercept) 6.685e+00 3.220e-02 4.336e+01 207.604
## Pres.Order -6.375e-04 6.024e-05 5.732e+03 -10.583
## z.RT.Sentence 6.396e-02 8.765e-03 3.545e+01 7.297
## Word.Typicality2-1 -2.630e-02 2.072e-02 1.801e+02 -1.269
## Word.Typicality3-2 5.319e-02 2.074e-02 1.807e+02 2.565
## z.VocabSize 8.023e-03 3.081e-02 3.598e+01 0.260
## Word.Typicality2-1:z.VocabSize -1.560e-02 7.144e-03 5.658e+03 -2.18
3
## Word.Typicality3-2:z.VocabSize 1.309e-02 7.182e-03 5.659e+03 1.82
3
## Pr(>|t|)
## (Intercept) < 2e-16 ***
## Pres.Order < 2e-16 ***
## z.RT.Sentence 1.47e-08 ***
## Word.Typicality2-1 0.2060
## Word.Typicality3-2 0.0111 *
## z.VocabSize 0.7960
## Word.Typicality2-1:z.VocabSize 0.0291 *
## Word.Typicality3-2:z.VocabSize 0.0683 .

```

```

## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##          (Intr) Prs.Or z.RT.S Wr.T2-1 Wr.T3-2 z.VcbS W.T2-1:
## Pres.Order  -0.171
## z.RT.Sentnc -0.015  0.147
## Wrd.Typc2-1  0.000  0.000 -0.001
## Wrd.Typc3-2  0.002 -0.009 -0.001 -0.501
## z.VocabSize -0.001  0.007  0.013  0.000  0.000
## Wr.T2-1:.VS  0.003 -0.009 -0.005  0.001 -0.002  0.000
## Wr.T3-2:.VS -0.002  0.006  0.006 -0.002  0.001  0.000 -0.502

```