AGREEMENT COMPUTATION IN SENTENCE PRODUCTION: CONCEPTUAL AND TEMPORAL FACTORS

A dissertation presented
by
Maureen Gillespie

to
The Department of Psychology

In partial fulfillment of the requirements for the degree of
Doctor of Philosophy

in the field of
Psychology

Northeastern University
Boston, Massachusetts
July, 2011
AGREEMENT COMPUTATION IN SENTENCE PRODUCTION:
CONCEPTUAL AND TEMPORAL FACTORS

by

Maureen Gillespie

ABSTRACT OF DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Psychology
in the Graduate School of Arts and Sciences of
Northeastern University, July, 2011
Abstract

The study of language production is concerned with how speakers translate non-verbal thoughts into meaningful, grammatical utterances. In all languages, information in one part of an utterance may depend on information that appears in another part of the utterance, and the distance between these parts of the utterance may span an arbitrary amount of intervening material. Subject-verb agreement is an example of a long-distance syntactic dependency in that the number information encoded in the subject and the verb of a sentence must agree. Agreement errors are likely to occur when a singular head noun (e.g., key) of a subject noun phrase (NP) is followed by a plural local noun (e.g., cabinets) in a modifier (e.g., the key to the cabinets *ARE), and this pattern is referred to as the mismatch effect (Bock & Miller, 1991). This dissertation follows the tradition of studies examining subject-verb agreement errors as a way to gain insight into the syntactic planning process.

Most theories of agreement production have suggested that syntactic structure is particularly important for computing agreement relations (e.g., Bock & Cutting, 1992; Eberhard, Cutting, & Bock, 2005; Franck, Vigliocco, & Nicol, 2002). The scope of planning account of agreement (Gillespie & Pearlmutter, 2011) does not rely on syntactic structure and instead suggests that it is the timing of planning of local nouns relative to their head that is important for correct agreement computation. This dissertation examines two major themes. Experiments 1-5 investigated whether syntactic structure constrains agreement computation independent of other lexical properties, and how one particular conceptual property, imageability, modulates the strength of the mismatch effect. Experiments 6 and 7 examined how another conceptual property, semantic integration (Solomon & Pearlmutter, 2004b), affects
the timing of planning of elements within complex NPs, and how differences in timing of planning affect agreement computation.

Earlier studies supporting the scope of planning account did not test the influence of clausal structure on agreement, and there is evidence that subject-verb agreement errors are more likely to follow a complex NP with a prepositional phrase (PP) modifier (e.g., *the editor of the history books*) than complex NPs with relative clause (RC) modifiers (e.g., *the editor who rejected the books*; Bock & Cutting, 1992; Solomon & Pearlmutter, 2004b). Experiment 1 examined whether hierarchical syntactic structure constrains subject-verb agreement computation, by contrasting subject NPs with PP and RC modifiers that were matched in many properties that varied in previous experiments (e.g., overall meaning). The mismatch effect was replicated; however, there was no interaction of local noun number and structure. These results suggest that structure is not involved in subject-verb agreement computation during production.

Experiments 2 and 3 ruled out two word-level properties (frequency and transitivity of the embedded verb in RCs) as sources of the difference between Experiment 1 and earlier studies (Bock & Cutting, 1992; Solomon & Pearlmutter, 2004b) that had found support for structural effects. Experiment 4 examined the effects of imageability and syntactic structure to determine whether imageability was a factor contributing to previously observed structural effects. Agreement errors were more common when stimuli were less imageable, but structure had no influence on agreement error rates, replicating Experiment 1. Experiment 5 was a regression meta-analysis of Experiments 1–4 and Solomon and Pearlmutter (2004b; Exp. 5) which showed that error rates increased with decreasing imageability.

Semantic integration, the degree of conceptual linkage between elements within an utterance, has been hypothesized to influence the timing of planning of elements within a phrase, such that highly semantically integrated elements are planned with more overlap than less integrated elements. But the evidence that integration affects timing during production is
indirect, as it largely comes from speech error data (Gillespie & Pearlmutter, 2011; Pearlmutter & Solomon, 2007; Solomon & Pearlmutter, 2004b). Experiment 6 examined the effect of semantic integration on temporal separation of words in recordings of stimuli from a subject-verb agreement elicitation study in Solomon and Pearlmutter (2004b). Linear mixed-effect regression modeling showed that temporal separation between the head and local noun decreased with increasing semantic integration, consistent with and providing the first direct evidence for Solomon and Pearlmutter’s claim that more integrated elements are planned with more overlap.

Experiment 7 examined how semantic integration affects utterance planning and subject-verb agreement error rates. Participants described picture displays with NP PP subject noun phrases and then completed the descriptions as full sentences. Semantic integration (Solomon & Pearlmutter, 2004b) was manipulated, and speech onset times and agreement error rates were recorded. Speakers were faster to initiate speech when the head and local noun were integrated than when they were unintegrated. Agreement errors were more likely when the local noun was plural than when it was singular. Supporting the scope of planning account of agreement (Gillespie & Pearlmutter, 2011), speakers who were slower to initiate speech produced more agreement errors, suggesting that when speakers do more advance planning they are more likely to experience interference during agreement computation.
Dedicated to Dr. Stella R. Arambel,

Gregory M. Kochman,

and my teachers
Acknowledgments

My brain hurt like a warehouse, it had no room to spare
I had to cram so many things to store everything in there
And all the fat-skinny people, and all the tall-short people
And all the nobody people, and all the somebody people
I never thought I’d need so many people...

– David Bowie, “Five Years”

First and foremost, I want to thank my advisor, Dr. Neal Pearlmutter. It has been such a privilege to work with you. Your attention to detail, sense of humor, and extensive knowledge of theoretical issues have taught me so much about what it means to be a successful researcher. I appreciate that every time I have come into your office with a “quick question” you have not only taken the time answer it, but you have also helped me come up with new research questions, study ideas, and ways to analyze my data. But above all else, I want to thank you for all the fun it has been to nerd out with you about language, statistics, and nearly everything else for the past five years. I am looking forward to a career’s worth of collaboration and conference dinners. I must also extend my thanks to Jodie Silverman and Ruby Edith Pearlmutter for all the nights Neal was meeting with me about this dissertation or some other project instead of heading home at a reasonable hour.

I have the utmost appreciation and respect for the members of my dissertation committee. I am constantly in awe of how Dr. Joanne Miller manages research and administrative duties, while still being intimately invested and involved in the development of our graduate students’ research programs and careers. Thank you for your dedication and the example you set. I would like to thank Dr. Kay Bock for all the fun that I have had discussing my research with her – especially when we have disagreed! I am excited to continue my work
on language production with you at Illinois next year. I thank Dr. Steven Harkins for his willingness to venture into the realm of psycholinguistic research. I know we psycholinguists are a strange (and neurotic) bunch, so I am grateful for the patience, insight, and calm you brought to every committee meeting.

I feel as though I have found a home in my field of research. This is largely a result of the friendships I have formed with the brilliant, inspiring, and fun-loving people I have met while attending conferences, workshops, and summer schools. I am extremely privileged to be a part of a field where I respect my colleagues as scientists and enjoy their company as friends. Thank you all for making research fun.

I have appreciated the supportive and intellectually challenging atmosphere of the Department of Psychology at Northeastern University, and I especially thank the members of the Language and Cognition group for five years of feedback, encouragement, and (of course) snacks. I cannot imagine being prepared to defend this dissertation without the support of my wonderful colleagues in L&C. In addition, the classes, lunches, drinks, laughs, frustrations, wedding celebrations, and defense preparation woes I have shared with Danielle Blanch-Hartigan and Maria Carrillo have added so much to my time at Northeastern. I have been so lucky to have such amazing, funny, and intelligent women as classmates and friends.

I would like to thank the past and present members of the Sentence Processing Laboratory for all the hard work, interesting conversations, and λαβ Halloween and Valentine’s Day Extravaganza sugar highs. Thank you for a great five years Amy DiBattista, Lauren Hovey, Devin St. John, Adrienne Davis Anderson, Athulya Arivand, Addya Bhomick, Abigail Cinamon, Danielle Cohen, Shreya Divata, Alyson Pospisil Doughrety, Jessica Eleman, Nathan Fowler, Ranya Gebara, Jade Goldsmith, Laura Goodman, Claire Gross, Jenesse Kaitz, Sandra Lechner, Keith Levin, Jacqueline Smith Palchik, Linda Raibert, Kathrin Ritter, Carolyn Schulz, Shaina Silverman, Courtney Tallarico, Daryl Velez, Emily Wallace, and Mariah Warren.
I have benefitted from the dedication and support of so many of my teachers throughout the years, and I want to thank those who truly made a difference by inspiring me to try harder, dig in deeper, and share my knowledge with others. Thank you for everything Sandy Sundquist, Anne Killeen, Elaine Merrifield, Joanne Crowder, Kathy Roberts, Anne Nichols, June Donegan, Susan Scuito, Linda Minickiello, Edward Flemming, Jed Butterfield, Dr. J. Robert Loftis, Professor Paul Graham, Dr. Naomi Nagy, Professor Andrew Merton, Dr. Rochelle Lieber, Dr. David Pillemer, Dr. Michael Ferber, and Dr. Rhea Eskew. I would especially like to thank Dr. John Limber and Dr. Stefanie Shattuck-Hufnagel for helping me to think about my research in new ways – every conversation has been an absolute pleasure. Your undying enthusiasm, curiosity, and passion is truly inspiring. I would also like to thank the students who I had the chance to teach; I know that I was the instructor, but I felt like I was the one who was learning so much.

The support of my family and friends has been constant. While I am thankful for many things that my parents Karen and Patrick Gillespie have done for me, I am most thankful that they allowed me the freedom to make my own decisions about my life and my plans. You instilled in me a sense of independence that has been a crucial aspect of any success I have ever had or that I am likely to have in the future. My sincerest gratitude is also extended to my aunt, Joanne Sears, for always supporting me, regardless of how far apart we have lived and how infrequently we have seen each other; your support has not gone unnoticed. The camaraderie I share with my brother, SSG Sean P. Gillespie, has kept me sane through some of the most crazy and stressful times in my life, and the dissertation has been no exception. Thank you for being there since the very beginning.

I have been fortunate enough to have life-long friends that have been as close as family, and I cannot imagine where I would be without them. I do not know where to begin to thank Michelle Zitta Manley, Molly Thieme Compton, and Joshua Zitta. You know where I came from, who I am now, and who I am striving to become; I cannot imagine better
people with whom to share a lifetime. I also express my gratitude to the Zitta-Manley-Brown, Thieme-Wyman-Compton-Kerylow, and Maguire-Considine clans for welcoming me into their families and for always supporting me, whether that support took the form of cheering at races and concerts, letting me stay the night, or greeting me as “Dr. Moe.”

My time in Boston has been defined by my friendships with Lyndsey Rawson, Lauren Hovey, Glen Coppersmith, Steven Piantadosi, Benjamin Ellis, and David Gray. Lyndsey agreed to move to Boston with me (basically on whim!) and we never looked back. It was so fun to explore, and get lost in, our new city together. During my first two years at Northeastern, I always looked forward to my daily chats, ABP runs, and Uno’s afternoons with Lauren and Glen. Northeastern was a different place after you moved on. Anytime I was working too hard or too long during the past few years, Steve, Ben, and David always found a way to make me take a break. Thank you for all the conversations that have started with statistics, science, philosophy, or logic, but have ended with horror movies, a barbecue, or recording ridiculous covers of even more ridiculous songs. It has been a blast living, working, and playing with you all. Thank you for making Boston feel like home from day one.

There are some friendships that allow us to appreciate all that life has to offer. My friendships with Alex Fine, Judith Degen, Florian Jaeger, Masha Fedzechkina, Peter Graff, and Gwendoline Fox have brought me places that I never dreamed possible. We have run a half marathon; hiked in the Swiss Alps; gone on death marches in the Adirondacks; sang karaoke until 3AM in a New York City bar; mustered the courage to fly in a six-seater plane; wandered the streets of Boston, New York, San Francisco, Montréal, Paris, Barcelona, Edinburgh, and York; and spent countless hours discussing research, statistics, and life. These experiences have made the past five years some of the most fun, interesting, and challenging times of my life. I am so fortunate to have you as my colleagues as well as my friends. Wegen eurer Freundschaft, bin ich mein besseres Ich.
Finally, I must acknowledge the two people to whom this dissertation is dedicated. I miss them both dearly, but their memory has inspired and motivated me through the most difficult aspects of this process.

Gregory Kochman once told me that we would learn a lot from each other. I will never know what he thought he learned from me, but I wish I could tell him how much I learned from him. Losing Greg taught me the ultimate lesson in the virtues of resilience and compassion. Without knowing him, I would have never had the heart nor the strength to become a researcher and teacher.

This dissertation simply would not exist if not for the guidance of the person who first introduced me to the field of psycholinguistics, Dr. Stella R. Arambel. Because of the short six months that I knew her, my life has taken the direction that it has. Dr. Arambel was a joy to know and continues to be such an inspiration in my life. She showed me that a female scientist can be intelligent, passionate, successful, and fiercely independent, while still being kind, funny, beautiful, and a wonderful mother. Every step of this journey has been taken with her in mind. I hope that I would have made her proud.
Contents

Abstract 3

Dedication 6

Acknowledgments 7

Part 1. Introduction 15

Chapter 1. Introduction 16
  1. Structural Accounts of Morphing 22
  2. Timing of Planning Accounts of Morphing 26

Part 2. Clause Boundedness in Agreement Production 35

Chapter 2. Evidence against structural constraints 36
  1. Experiment 1 36

Chapter 3. Exploring lexical alternatives to the clause-boundedness effect 51
  1. Experiment 2 55
  2. Experiment 3 66
  3. Experiment 2 and 3 Discussion 78

Chapter 4. The effect of imageability on agreement computation 81
  1. Experiment 4 84
  2. Experiment 5 95
  3. Experiment 4 and 5 Discussion 109

Part 3. Timing of Planning During Sentence Production 113
Chapter 5. Prosody reflects planning processes in production
   1. Experiment 6

Chapter 6. Effects of semantic integration and advance planning on grammatical
   encoding and agreement computation
   1. Experiment 7

Part 4. Discussion

Chapter 7. Discussion
   3. Revised Model of Agreement Computation
   4. Conclusions

Appendices

Appendix A. Experiment 1 Stimuli

Appendix B. Empirical Logit Weighted Linear Regression

Appendix C. Experiment 2a Stimuli

Appendix D. Semantic Integration Rating Instructions

Appendix E. Experiment 2b Stimuli

Appendix F. Experiment 3a Stimuli

Appendix G. Experiment 3b Stimuli

Appendix H. Full Subject Noun Phrase Imageability Rating Instructions

Appendix I. Content Word Imageability Rating Instructions

Appendix J. Experiment 4 Stimuli
Part 1

Introduction
CHAPTER 1

Introduction

The study of language production is concerned with how speakers translate non-verbal thoughts into meaningful, grammatical utterances. While this seems like a fairly effortless task that requires little conscious consideration on behalf of the speaker, the nature of the processes that underlie this task are not well understood. Models of language production generally separate the planning process into different levels. Bock and Levelt (1994) separate the planning process into three main levels: the message level, which represents the speaker’s intended meaning; the grammatical encoding level, which translates the meaning into a sequence of words; and the phonological encoding level, which translates the sequence of words into the sounds required to produce the utterance. The studies in this dissertation focus on the grammatical encoding process.

In all languages, information in one part of an utterance may depend on information that appears in another part of the utterance, and the distance between these parts of the utterance may span an arbitrary amount of intervening material. Because of this fact, languages are considered to have long-distance dependencies. Subject-verb agreement is an example of a long-distance syntactic dependency in that the number information encoded in the subject and the verb of a sentence must agree. This dissertation follows the tradition of studies examining subject-verb agreement errors as a way to gain insight into the syntactic planning process (e.g., Bock & Cutting, 1992; Bock & Miller, 1991; Franck, Lassi, Frauenfelder, & Rizzi, 2006; Franck, Vigliocco, & Nicol, 2002; Gillespie & Pearlmutter, 2011; Hartsuiker, Anton-Mendez, & Zee, 2001; Solomon & Pearlmutter, 2004b; Vigliocco & Nicol, 1994, 1998).

Bock and Miller (1991) conducted the first study that elicited subject-verb agreement errors in the laboratory. Bock and Miller used sentence preambles that were composed of a
head noun followed by a modifier containing a local noun. Subject-verb agreement errors are often produced in sentences containing subject noun phrases (NPs) of this structure because the verb erroneously agrees with the local noun. This effect is referred to as attraction or proximity concord. Experimental items in Bock and Miller manipulated the number marking of the head and local nouns to form four number conditions as seen in (1). (1b) and (1c) were considered mismatch conditions because the head noun (key(s)) and local noun (cabinet(s)) had different number markings, while (1a) and (1d) were considered match conditions because the head noun and the local noun had the same number marking. Preambles were presented auditorily, and participants were required to repeat them and then complete them as full sentences.

(1) a. The key to the cabinet\textsuperscript{a} (SS) 
b. The key to the cabinets (SP) 
c. The keys to the cabinet (PS) 
d. The keys to the cabinets (PS)

\textsuperscript{a}The codes following the preambles refer to the number marking on each of the nouns in the subject noun phrase. S = Singular; P = Plural.

The majority of agreement errors (> 90%) occurred in the mismatch conditions (1b and 1c). This error pattern is referred to as the mismatch effect and has been replicated in essentially all studies examining subject-verb agreement (e.g., Bock & Cutting, 1992; Bock & Eberhard, 1993; Bock, Eberhard, Cutting, Meyer, & Schriefers, 2001; Bock, Nicol, & Cutting, 1999; Eberhard, 1999; Franck et al., 2002, 2006; Gillespie & Pearlmutter, 2011; Hartsuiker et al., 2001; Vigliocco & Nicol, 1994). Within the mismatch conditions, agreement errors were almost exclusively produced when the head noun was singular and the local noun was plural (1b) rather than when the head noun was plural and the local noun was singular (1c). Nearly all studies examining subject-verb agreement errors have replicated this pattern of results (Bock & Eberhard, 1993; Bock et al., 1999, 2001; Eberhard, Cutting, & Bock, 2005; Franck et al., 2002; Vigliocco & Nicol, 1994, 1998). The interference of plural local nouns,
and relative lack of interference of singular local nouns, on subject-verb agreement provides support for the hypothesis that plural noun forms are marked with a plural feature, while singular nouns are unmarked (Berent, Pinker, Tzelgov, Bibi, & Goldfarb, 2005; Bock & Eberhard, 1993; Bock & Miller, 1991; Eberhard et al., 2005; Vigliocco & Nicol, 1994, 1998). Because the mismatch effect is only present in conditions with singular head nouns and plural local nouns, it is clear that attraction is not merely caused by interference from the noun that is closest to the verb. This has led researchers to examine message-level (conceptual) and grammatical encoding-level (syntactic) explanations for the mismatch effect.

In English, there are words that are grammatically singular but conceptually plural (e.g., *fleet, army, choir*), and grammatically plural but conceptually singular (e.g., *scissors, glasses, binoculars, pants*). These words provide an interesting disconnect between the message level and grammatical encoding level of language production. Researchers have used words with conflicting grammatical and conceptual number as local nouns within preambles to uncover the influence of message and grammatical properties on subject-verb agreement. In these studies, local nouns consistently elicited subject-verb agreement errors when they were grammatically plural and conceptually singular, rather than when they were conceptually plural and grammatically singular, suggesting that subject-verb agreement is largely determined at the grammatical encoding level (Bock et al., 1999, 2001; Bock & Eberhard, 1993; Bock, Eberhard, & Cutting, 2004; Eberhard et al., 2005).

An additional way researchers have examined the possibility that conceptual number may influence subject-verb agreement is by manipulating preambles in a way that allows for a distributive reading. A subject NP such as *the key to the cabinets* refers to a single thing (i.e., a single key that opens multiple cabinets) while a phrase such as *the picture on the postcards* can be interpreted as referring to multiple things (i.e., the same picture that appears on multiple postcards). Only the latter NP has a conceptually plural or distributive interpretation. When preambles were concrete, highly imageable NPs, subject-verb agreement errors were more common for distributive preambles than for non-distributive
preambles (Eberhard, 1999; Hartsuiker, Kolk, & Huinck, 1999; Vigliocco, Butterworth, & Garrett, 1996; Vigliocco, Hartsuiker, Jarema, & Kolk, 1996). These findings indicate that some information from the message level can influence subject-verb agreement in sentence production.

Eberhard et al.’s (2005) marking and morphing model (M&M) explains these seemingly conflicting results with an account of agreement production that combines the conceptual number of the full subject NP and the number of each individual lexical item. Generally, in this model, the marking process assigns the subject NP a number value reflecting the number from the conceptual representation. The process of morphing assigns and combines number values from the individual lexical items within the subject NP to represent the grammatical number information. The marking and morphing processes proceed separately, and finally the marking number value is combined with the morphing number value to yield the final number of the subject NP. This number can be assigned to the verb or a pronoun1 in the M&M model.

In Eberhard et al.’s model, the number values assigned by the marking and morphing processes are called SAP (Singular and Plural) values. SAP values range from −1 to +1. NPs that refer to specific individuals (e.g., proper names, NPs beginning with the individuating quantifier one) are assigned a SAP value of −1 and NPs that refer to clear pluralities (e.g., conjunctions) are assigned a SAP value of +1. Grammatically singular subject NPs without a distributive reading are assigned marking SAP values of 0. Grammatically singular subject NPs with a distributive or collective reading (e.g., the picture on the postcards, the gang of boys) are assigned marking SAP values of .48. SAP values for each of the lexical items in the subject NP to be combined during morphing vary depending on noun type. Nouns with plural inflections (e.g., dogs, boys) and invariant plurals (e.g., scissors, pants) are assigned the maximum SAP value of 1. Uninflected collective nouns (e.g., gang, group) are assigned

---

1The M&M model also predicts the production of pronouns. The pronoun model allows marking to influence the subject NP’s final SAP value more strongly than the verb model (see Bock et al., 1999, 2001, 2004, 2006; Eberhard et al., 2005).
a SAP value of .07. All uninflected, singular count and mass nouns (e.g., postcard, milk) are assigned a SAP value of 0.

In the model, the subject NP node receives its SAP value to pass to the verb from the combination of the SAP values from marking and morphing, as shown in Figure 1.1. Marking directly passes its SAP value to the subject NP node. The SAP values from each of the individual lexical items are passed to the subject NP node during morphing. In Eberhard et al. (2005), the SAP value from the head noun exerts more influence on the subject NP SAP value than the local noun’s SAP value because the head is the controller of agreement and it is hierarchically proximal to the subject NP node (Bock & Cutting, 1992; Franck et al., 2002; Hartsuiker et al., 2001; Vigliocco & Hartsuiker, 2002).

According to the M&M model, the distributivity effect is a result of the entire subject NP receiving a SAP value closer to 1 during marking (because it is conceptually plural), and thus the verb is passed a higher SAP value resulting in more plural verb forms being produced for distributive subject NPs. In addition, when the head noun of a subject NP is a collective (e.g., the gang near the motorcycle), plural verbs are more likely to be produced because the whole subject NP has a more plural interpretation and is assigned a marking SAP value close to 1 than subject NP with a singular head noun (e.g., the biker near the motorcycle). The lack of attraction errors for preambles containing conceptually plural local nouns in studies that used words with conflicting grammatical and conceptual number can be explained by the morphing process. Words that are conceptually plural and grammatically singular (e.g., army) have SAP values close to 0 (i.e., singular) in the morphing process because morphing is most concerned with grammatical number. If both nouns within a subject NP have SAP values near 0, as would be the case in preambles containing a grammatically singular head noun and a grammatically singular but conceptually plural local noun (e.g., the general of the army), then the morphing process should reconcile the number information contained

---

SAP values for collective nouns and distributive NPs were determined by model fitting procedures described in Eberhard et al. (2005).

SAP values for individual lexical items are weighted by the proportion of the time the noun appeared in the singular form relative to the proportion of the time it appeared in a plural form in a written corpus.
Figure 1.1. Diagram of the number reconciliation process outlined in Marking and Morphing model adapted from Eberhard et al. (2005). The subject NP node receives its SAP value from two sources: marking and morphing. The SAP value from marking is assigned based on conceptual number. During Morphing the SAP value from the head noun and local noun are passed to the subject NP node. The head noun SAP has more influence on the subject NP node’s SAP value, and is shown with a thick arrow. The local noun SAP has less influence on the subject NP node’s SAP value and is shown with a thin arrow. NP = Noun phrase; PP = Prepositional phrase; P = Preposition.

within the subject NP easily, and the SAP value from the morphing process (near 0) should combine with the SAP value from the marking process (also near 0, as the subject NP as a whole is conceptually singular) to transmit the subject NP node’s SAP value to the VP, resulting in a singular verb being produced.

There has been a considerable amount of work examining effects of conceptual number on subject-verb agreement (e.g., Bock & Eberhard, 1993; Bock & Miller, 1991; Eberhard, 1999; Hartsuiker et al., 1999). One intriguing line of research has suggested that the degree to which the conceptual number is accessible during agreement computation determines
how strong distributivity effects (described above) will be. Highly imageable subject NPs with distributive readings produce stronger mismatch effects than less imageable distributive subject NPs (Eberhard, 1999). Hartsuiker et al. (1999) found that agrammatic aphasics do not show stronger mismatch effects for distributive items and attributed this to their reduced processing capacities, which prevents them from accessing the conceptual number of the NP during agreement computation. Taken together, these studies have shown that when speakers have access to the conceptual number of a subject NP, they are more likely to produce verbs that agree with the conceptual number.

In addition, researchers have considered two ways of how number information from the lexical items combine during morphing, and how this process gives rise to attraction effects. Eberhard et al.’s (2005) model suggests that structural properties of the subject noun phrase determine how morphing proceeds; however, there is also evidence that the timing of planning of elements within the subject NP determines interference patterns during morphing. The evidence for these two accounts is discussed below.

1. Structural Accounts of Morphing

Bock and Cutting (1992) examined the role of structural relations in language production and hypothesized that the unit of syntactic planning is the clause. They formulated this hypothesis because Bock and Miller (1991) found preliminary evidence suggesting that attraction errors were less frequent when local nouns were within embedded clauses than when they were in phrasal modifiers. Bock and Cutting suggested that in clauses are autonomous units language production, and information in one clause is unlikely to interfere with information in another clause. Using Bock and Miller’s paradigm, Bock and Cutting conducted a study in which preambles containing a head noun and local noun varied in number as well as the type of construction in which the local noun appeared, as in (2). In (2a-d) the local
noun appeared in a prepositional phrase (PP), and in (2e-h) the local noun appeared in a relative clause (RC) containing the same number of syllables⁴.

(2) a. The editor of the history book (SS)
b. The editor of the history books (SP)
c. The editors of the history book (PS)
d. The editors of the history books (PS)
e. The editor who rejected the book (SS)
f. The editor who rejected the books (SP)
g. The editors who rejected the book (PS)
h. The editors who rejected the books (PS)

The majority of errors occurred in mismatch conditions where the head noun was singular (2b, f), replicating the mismatch and plural markedness effects observed in Bock and Miller (1991). The proportion of errors was greater in PPs than in RCs, supporting the hypothesis that attraction errors are more likely to occur when a local noun is contained within the same clause as the head noun than when they are in separate clauses, and that a syntactic planning unit in language production is the clause.

Other researchers have proposed that the reason clause-boundedness effects have been observed is because the syntactic distance between the head and local noun, as defined by nodes in a syntactic tree, modulates the size of the mismatch effect (Franck et al., 2002, 2006; Hartsuiker et al., 2001; Negro, Chanquoy, Fayol, & Louis-Sidney, 2005; Vigliocco & Nicol, 1994, 1998). According to the syntactic distance hypothesis, cases in which attraction errors occur are the result of the plural feature of the local noun percolating too far up the syntactic tree, overwriting the singular subject NP (Franck et al., 2002; Vigliocco & Nicol, 1994). The likelihood of a local noun’s plural feature percolating all the way up to the subject NP node is inversely related to the number of nodes the plural feature would have

⁴When pronounced in casual speech, the number of syllables may not have been exactly matched within items. For example, history is likely to be pronounced with two syllables in casual speech, rather than three as specified in a dictionary.
to pass through to reach the subject NP node in the syntactic tree. Vigliocco and colleagues (Franck et al., 2002; Vigliocco & Nicol, 1994) showed that the difference in error rates for clausal and phrasal modifiers can be explained in terms of syntactic distance rather than clause-boundedness as the plural feature of a local noun has a shorter distance to travel in PPs than in RCs; therefore, errors would be less likely in RCs than PPs.

Franck et al. (2002) sought to develop the syntactic distance hypothesis by examining the influence of multiple local nouns within PP modifiers on agreement errors. Preambles like (3) were composed of a head noun (helicopter, N1) and two PP modifiers, each containing a noun (flight(s), N2; canyon(s), N3). The first PP modified the head noun while the second PP modified the noun in the first PP (flight(s), N2). The four singular head noun conditions of each item were created by manipulating the number marking of N2 and N3.

\[(3) \quad \text{a. The helicopter for the flight over the canyon} \quad (SSS) \]
\[(3) \quad \text{b. The helicopter for the flights over the canyon} \quad (SPS) \]
\[(3) \quad \text{c. The helicopter for the flight over the canyons} \quad (SSP) \]
\[(3) \quad \text{d. The helicopter for the flights over the canyons} \quad (SPP) \]

Franck et al.’s (2002) experiments tested linear distance, clause-boundedness, and syntactic distance hypotheses. According to the linear distance hypothesis, a plural N3 would be more likely to create agreement errors because it is linearly closer to the verb than N2. According to the clause-boundedness hypothesis (Bock & Cutting, 1992), N2 and N3 should have an equal chance of creating agreement errors because they are both part of the matrix clause. The syntactic distance hypothesis suggests that N2 should have a greater influence on agreement errors because it is situated higher in the syntactic tree than N3. As shown in Figure 1.2, the plural feature on N3 has a greater distance to travel up the syntactic tree than the plural feature of N2. Because N3’s syntactic distance from the verb is greater than N2’s, fewer errors should be observed when N3 is plural than when N2 is plural.

\[5\text{Four plural head conditions were also used, but results were mixed and difficult to interpret.}\]
Figure 1.2. Syntactic path a plural feature must travel to interfere with agreement in Franck et al.'s (2002) stimuli. The route for a feature from N2 is shown with solid arrows; the route for a feature from N3 includes the route from N2 as well as the dashed arrows, so additional feature-passing errors would have to occur before N3’s plural feature could influence verb number, predicting fewer subject-verb agreement errors when N3 is plural compared to when N2 is.

Error rates were higher for the conditions in which N2 was plural (SPS, SPP) than for the conditions in which N2 was singular (SSS, SSP), yet the conditions in which N3 was plural did not differ significantly from their otherwise matched singular conditions (SSS = SSP; SPS = SPP). Franck et al. (2002) interpreted these results as support for the syntactic distance hypothesis because N2’s plurality had a greater influence on erroneous agreement than N3’s plurality. Franck et al. ruled out the linear distance and clause-boundedness hypotheses and proposed that the attraction phenomenon can be explained by erroneous hierarchical feature passing during the grammatical encoding phase of language production when words are hierarchically organized but the order in which they are to be produced is not yet determined (see also Vigliocco & Nicol, 1994, 1998).

Eberhard et al. (2005) incorporated these findings into the weights used to transmit the individual lexical items’ SAP values to the subject NP node during morphing. In the
model, the local noun’s SAP value has less of an influence on the subject NP node’s SAP value because the local noun’s SAP value has a larger hierarchical distance to travel to the subject NP node than the head noun’s SAP value does (cf. Franck et al., 2002; Vigliocco & Nicol, 1994). Eberhard et al. only model subject-verb agreement for NP PP subject NPs; however, if the M&M model were to be extended to model NP PP PP subject NPs (like those in Franck et al.), the SAP values of local nouns situated deeper in the syntactic tree (e.g., canyon(s) in (3)) would have less of an influence on the subject NP node’s SAP value than the SAP values of hierarchically proximal local nouns (e.g., flight(s) in (3)). Assigning weights based on syntactic distance also allows M&M to account for the reduced mismatch effect for subject NPs with RC modifiers (Bock & Cutting, 1992; Solomon & Pearlmutter, 2004b) and low attraction errors from plural sentential objects in languages that allow the object to appear before the verb (Franck et al., 2006; Franck, Soare, Frauenfelder, & Rizzi, 2010; Hartsuiker et al., 2001).

2. Timing of Planning Accounts of Morphing

While structural accounts of the mismatch effect can explain a range of studies in the literature, there is accumulating evidence that properties thought to influence the timing of planning of elements within an utterance also influence interference effects during agreement computation.

Solomon and Pearlmutter (2004b) investigated whether the degree to which phrases are linked at the conceptual level affects subject-verb agreement error rates by influencing timing of planning. Solomon and Pearlmutter termed the degree of conceptual linkage semantic integration and hypothesized that highly semantically integrated elements within an utterance are more likely to be planned simultaneously, thus allowing their features to interfere with each other. They also hypothesized that the effect of semantic integration is separate from the effect of hierarchical distance.
Solomon and Pearlmutter (2004b) manipulated the degree of semantic integration in preambles consisting of a head noun and a PP modifier. Four versions of each stimulus item were created by varying the degree of semantic integration between the head and local noun and the local noun’s number marking as shown in (4). In integrated conditions (4a-b), the head noun and local noun were linked with the preposition *for* and were in a functional relationship, and in the unintegrated conditions (4c-d) the head noun and local noun were linked with *with* and were in an accompaniment relationship. Critically, the hierarchical distance of the local noun was equal in both conditions. Solomon and Pearlmutter predicted that error rates should be higher in integrated conditions than in unintegrated conditions because of the increased degree of planning overlap. Solomon and Pearlmutter found that the mismatch effect was larger in the integrated conditions than in the unintegrated conditions. They suggested that the mechanism of the integration effect was the relative timing of planning of the nouns within the preamble, and suggested that the degree to which phrases are planned or activated together increases subject-verb agreement error rates. These findings suggest that conceptual-level factors that influence timing of planning of elements within phrases affect error rates, independent of structural properties.

(4) a. The chauffeur for the actor (SS)
b. The chauffeur for the actors (SP)
c. The chauffeur with the actor (SS)
d. The chauffeur with the actors (SP)

Solomon and Pearlmutter (2004a) rated Franck et al.’s (2002) NP PP PP preambles (see (3)) for semantic integration between the head noun and each of its local nouns. Semantic integration was found to be confounded with syntactic distance in these preambles. N1 and N2 were significantly more integrated than N1 and N3; thus, semantic integration provides an

---

6Semantic integration ratings of these items confirmed that the versions with a functional relationship between the head and local noun (e.g., *for* conditions) were more integrated than versions with an accompaniment relationship between the head and local noun (e.g., *with* conditions). For more details on the semantic integration rating procedure, see (Solomon & Pearlmutter, 2004b).
alternative explanation to the hierarchical feature passing hypothesis. Semantic integration would predict a larger N2 mismatch effect than N3 mismatch effect because N2 was more integrated with N1 than N3, and this was the pattern observed in Franck et al.. In addition, because the semantic integration account suggests that timing of planning of elements relative to the head noun of the subject NP is critical in explaining interference effects, it is also possible that the order in which local nouns are produced affects timing of planning. There is evidence that language production is highly incremental; the order in which items are to be produced has been hypothesized to be the order in which they are planned (Bock & Levelt, 1994; Brown-Schmidt & Konopka, 2008; Brown-Schmidt & Tanenhaus, 2006; Griffin, 2001; Wheeldon & Lahiri, 2002). If language production is incremental, and timing of planning relative to the head determines interference effects, it is possible that a local noun’s linear distance from the head noun may have an effect on agreement production. Thus, linear distance would also predict Franck et al.’s results, as N2 was closer to N1 than N3 was and should have been planned closer in time to N1 than N3 if elements are planned in the order of production.

Gillespie and Pearlmutter (2011) extended Solomon and Pearlmutter’s (2004b) hypothesis that the degree of overlap in planning determines the size of mismatch effects in subject-verb agreement error production. They hypothesized that a local noun’s planning distance from the head noun due to linear proximity and semantic integration would determine the size of interference effects, and that these effects should be independent of hierarchical structure.

To examine hierarchical distance directly and address the semantic integration confound in Franck et al.’s (2002) stimuli, Gillespie and Pearlmutter (2011) compared two sets of NP PP PP preambles, which controlled semantic integration while manipulating hierarchical distance and local noun number, as in (5) and (6). Preambles like (5) had a descending hierarchical structure, such that the PP containing N3 (on the leather strap(s)) modified N2 (buckle(s)), as in Figure 1.2 and in Franck et al. (2002). Preambles like (6) had a flat structure, such that both PPs (to the western suburb(s), with the steel guardrail(s)) modified

\footnote{Gillespie and Pearlmutter’s stimuli also had an adjective or noun modifier of N2 and N3.}
Figure 1.3. Syntactic structure for Gillespie and Pearlmutter’s (2011) flat stimuli, in which both PPs attach to the first NP.

N1 (highway), as illustrated in Figure 1.3. Mean semantic integration of the N1-N2 pair was matched across structures, as was mean semantic integration of the N1-N3 pair.

(5) a. The backpack with the plastic buckle on the leather strap (SSS)
   b. The backpack with the plastic buckles on the leather strap (SPS)
   c. The backpack with the plastic buckle on the leather straps (SSP)

(6) a. The highway to the western suburb with the steel guardrail (SSS)
   b. The highway to the western suburbs with the steel guardrail (SPS)
   c. The highway to the western suburb with the steel guardrails (SSP)

Gillespie and Pearlmutter (2011) suggested that if hierarchical distance has an independent effect on agreement computation, the difference between the N2 and N3 mismatch effects should be smaller for the flat preambles than the descending preambles, because the distance N2’s plural feature would have to travel to affect agreement computation is matched across structures, while the distance N3’s plural feature would have to travel is shorter in flat structures than in descending structures. On the other hand, if the effects Franck et al. (2002) attributed to hierarchical distance were instead due to linear distance back to the
head noun, then both flat and descending structures should show the pattern Franck et al. found for their descending structures — a higher mismatch error rate for N2 mismatches than for N3 mismatches — with no interaction between structure and mismatch position. Finally, if the effects Franck et al. found were due to their stimuli’s semantic integration confound alone, then no interaction should be observed, and the N2 and N3 mismatch effects should differ only to the extent that integration between each of those nouns and N1 varied.

Gillespie and Pearlmutter (2011) found that the N2 mismatch effect was larger than the N3 mismatch effect, and there was no interaction with structure. They interpreted their results as strong evidence in favor of a linear distance to the head account of agreement errors and against a hierarchical distance account: The linear distance account correctly predicted a larger mismatch effect for N2 than for N3 mismatches for both descending and flat structures, and it correctly predicted that the size of this difference would be equal for the two structures.

Gillespie and Pearlmutter (2011) conducted a second experiment that tested whether linear distance to the head noun and semantic integration combine to influence agreement error production, using flat stimuli like (6) that controlled hierarchical distance of N2 and N3. Under a combined linear distance and semantic integration account, the likelihood of interference would be a function of whether the interfering element was within the scope of planning of the head noun; only local nouns planned close enough in time to the head would create mismatch effects, with both decreased head-local linear distance and increased head-local semantic integration increasing the chance of overlap in planning.

Each preamble had a NP PP PP structure, and the number of N2 and N3 was varied, as in (7). One PP (e.g., *with the torn page(s)*) was strongly integrated with the head noun, while the other (e.g., *by the red pen(s)*) was weakly integrated. Linear distance was manipulated by having two local nouns in different positions (i.e., N2 and N3), and switching the order of

---

8In Gillespie and Pearlmutter’s (2011) stimuli, the N1-N3 pair had significantly higher integration ratings than the N1-N2 pair, so the direction of the semantic integration effect would be opposite that of the effect of linear distance.
the PPs allowed control of semantic integration at each linear position. The stimuli equated hierarchical distance between the head noun (and thus also the verb) and each of the two local nouns by ensuring that both PPs modified the head, resulting in flat structures as in Figure 1.3.

(7) a. The book with the torn page by the red pen (SSS)
   b. The book with the torn pages by the red pen (SPS)
   c. The book with the torn page by the red pens (SSP)
   d. The book with the torn pages by the red pens (SPP)
   e. The book by the red pen with the torn page (SSS)
   f. The book by the red pens with the torn page (SPS)
   g. The book by the red pen with the torn pages (SSP)
   h. The book by the red pens with the torn pages (SPP)

If only linear distance to the head noun is responsible for agreement error rates, the N2 mismatch effect should be larger than the N3 mismatch effect in both the early- and late-integrated versions (7a–d and 7e–h, respectively), because N2 would always be planned closer to N1 than N3 would be.

Alternatively, if agreement error effects in these items depended only on semantic integration, then, collapsing over integration version, the N2 and N3 mismatch effects should be equal, because the mismatch effect of a given noun should be the same regardless of where the noun appears linearly within the stimulus. However, the N2 and N3 mismatch effects within each integration condition should differ: The N2 mismatch effect should be larger than the N3 mismatch effect for early-integrated cases (7a–d), because the more integrated noun (N2) should be planned with N1, while N3 should be planned later. For late-integrated cases (7e–h), the pattern should reverse, because N3 is the noun more tightly integrated with N1.

The scope of planning alternative, a combined effect of linear distance to the head and semantic integration (i.e., the two factors thought to influence the relative timing of planning
of local nouns in this experiment), predicts that linear distance to the head noun would partially determine the timing of planning of nouns within the phrase, and semantic integration should shift the relative planning time of more and less integrated nouns. Figure 1.4A shows the timing of planning of nouns according to the order in which they are to be produced (this corresponds to the predictions of linear distance to the head alone), and Figure 1.4B includes the shifting of the timing of planning due to semantic integration. In early-integrated cases (7a–d), because N2 was more integrated with N1, N2 would be planned closer together in time to N1; and because N3 was less integrated with N1, N3 would be planned later, predicting a large N2 mismatch effect and a very small N3 mismatch effect. In late-integrated cases (7e–h), because N2 was not very integrated with N1, N2 would be planned later; and because N3 was more integrated with N1, it would be planned sooner. Thus N2 and N3 should be planned close together in time and relatively late after N1. This predicts that the N2 and N3 mismatch effects should be about equal and fairly small.
Error rates in Gillespie and Pearlmutter (2011) showed a pattern reflecting a combination of linear distance to the head and semantic integration, which is consistent with the scope of planning account. The N2 mismatch effect was larger than the N3 mismatch effect in the early-integrated conditions, while the N2 and N3 mismatch effects were equal in the late-integrated conditions and both smaller than the N2 mismatch effect for the early integrated condition. These results suggest that structural relations are not necessary to explain mismatch effects within complex subject NPs with phrasal modifiers. Instead, the results point toward an account of agreement production that relies on scope of planning such that plural local nouns that are planned close in time to the head (due to linear distance and semantic integration) are the most likely to cause interference during agreement computation.

While Gillespie and Pearlmutter’s (2011) results provide fairly strong evidence that hierarchical distance is not a factor affecting agreement computation, there is still evidence in the literature that structure does affect error rates when semantic integration and linear distance are controlled. As mentioned above, Bock and Cutting (1992) found that plural local nouns in PP modifiers produced larger mismatch effects than local nouns in RC modifiers. Solomon and Pearlmutter (2004b) also found this effect when they compared mismatch effects from local nouns in length- and integration-matched PP and RC modifiers. In addition, they confirmed that Bock and Cutting’s PP and RC versions were integration-matched. Thus, it is possible that clause-boundedness may still play a role in the morphing process as the scope of planning account would not predict a difference in error rates between length- and integration-matched PPs and RCs (cf. Nicol, 1995).

Part 2 investigates whether clausal structure constrains agreement computation by removing potential confounds in studies that observed structure effects, and Part 3 further develops timing accounts of morphing.
Part 2

Clause Boundedness in Agreement

Production
CHAPTER 2

Evidence against structural constraints

1. Experiment 1

Theories of agreement production concerned with the mechanisms underlying the implementation of agreement have suggested that syntactic structure is particularly important for computing agreement relations. The clause-boundedness hypothesis suggests that agreement computation is only sensitive to information within the current clause (Bock & Cutting, 1992), and the hierarchical feature-passing hypothesis posits that agreement features are passed along the syntactic tree to their targets (Franck et al., 2002). Alternatively, the scope of planning hypothesis (Gillespie & Pearlmutter, 2011) explains agreement computation through processing that encodes the features of the agreement source and then retrieves them during the planning of the agreement target; however, the studies supporting the scope of planning hypothesis did not test the influence of clausal structure on agreement, nor the effects of hierarchical feature-passing over a limited amount of planned structure. The current study simultaneously tests these possibilities.

The first finding suggesting hierarchical structure is a component of agreement production was the clause-boundedness effect. As described in Chapter 1, Bock and Cutting’s (1992) Experiment 1 compared prepositional phrase (PP) modifier preambles (8a) to corresponding length-matched relative clause (RC) modifier preambles (8b), using an agreement error elicitation task in which participants recited each preamble aloud and completed it as a sentence. Subject-verb agreement error rates were larger when the local noun was plural (e.g., books) than when the local noun was singular (e.g., book), the standard “mismatch effect” (Bock & Miller, 1991); but this difference was larger for PP than for RC cases (see also Solomon & Pearlmutter, 2004b, Exp. 5). Bock and Cutting suggested that this was
because clauses are planned independently, so elements within separate clauses are unlikely
to interfere with each other.

(8) a. The editor of the history book(s) (PP)
b. The editor who rejected the book(s) (RC)

An alternative explanation for the clause-boundedness effect comes from the hierarchical
feature-passing hypothesis, which provides a structure-based mechanism for implementing
agreement (Eberhard et al., 2005; Franck et al., 2002; Hartsuiker et al., 2001; Vigliocco &
Hartsuiker, 2002). On this view, agreement is computed using the syntactic tree structure of
a sentence, with number features being passed up through the subject NP and then to the
verb phrase. Mismatch effects occur when a plural feature is inadvertently passed too far up
the tree, overwriting the number from the head noun with the number from a local noun.
Franck et al. provided the most direct test of hierarchical feature-passing, using subject
NP preambles containing two PPs, as in (3). Franck et al. found that the N2 mismatch
effect was larger than the N3 mismatch effect in both English and French, and argued for
a hierarchical feature-passing account of subject-verb agreement over an account in which
interference increases with linear proximity to the verb. Hierarchical feature-passing explains
the clause-boundedness effect because local nouns in PPs are hierarchically closer to the verb
than local nouns in RCs (Franck et al.; see Solomon & Pearlmutter, 2004b, for discussion).

In addition to structural properties, semantic and temporal properties that influence
timing of planning also seem to affect agreement computation. Solomon and Pearlmutter
(2004b) hypothesized that semantic integration (i.e., the degree to which elements within
a phrase are linked at the message level) affects the timing of planning of elements within
a phrase, such that elements of more semantically integrated phrases are more likely to be
planned overlappingly. Solomon and Pearlmutter manipulated local noun number in NP PP
stimuli and compared integrated cases (e.g., The pizza with the yummy topping(s)) to corre-
responding unintegrated ones (e.g., The pizza with the tasty beverage(s)). Across a series of
experiments, they found larger mismatch effects for integrated than for unintegrated conditions. They interpreted these findings as support for the hypothesis that increased semantic integration increases the chance of overlap in planning, which leads to increased interference during agreement computation (for evidence from exchange errors see Pearlmutter & Solomon, 2007).

Gillespie and Pearlmutter (2011) noted that stimuli in Franck et al. (2002) had a semantic integration confound: The head noun (N1; *helicopter* in (3)) and N2 were more semantically integrated than N1 and N3, so semantic integration might explain Franck et al.’s results. In addition, Franck et al. did not discuss the possibility that a local noun’s linear proximity to N1 might increase error rates, which could also explain the results they attributed to hierarchical distance. Gillespie and Pearlmutter conducted two studies using NP PP PP preambles (see Chapter 1 for more detail), and found a combined effect of semantic integration and linear distance on error rates, with no evidence that structural distance was a factor affecting agreement computation. Gillespie and Pearlmutter proposed a scope of planning account of agreement production, predicting more agreement errors when a plural local noun is planned within the scope of (i.e., close in time to) a singular head noun, with semantic integration and linear order combining to influence planning time, independent of hierarchical distance.

This scope of planning account can explain Franck et al.’s (2002) results and many other effects reported in the agreement literature without a need for hierarchical feature-passing, which raises the question of whether agreement computations are constrained directly by structure at all. But while Gillespie and Pearlmutter (2011) argued against a hierarchical account of existing agreement data and suggested an alternative mechanism, they could not rule out two possibilities: that hierarchical feature-passing is the mechanism underlying all agreement computation, but that its effects are constrained by scope of planning (errant feature-passing cannot occur from within as-yet-unplanned constituents); or that feature-passing applies only to or around clause boundaries. Clause-boundedness itself, the other main proposed structural constraint (cf. Franck et al., 2010), also cannot be explained.
by scope of planning: Bock and Cutting’s (1992) PP and RC stimuli were matched for length in syllables (linear distance from head to local noun), and Solomon and Pearlmutter (2004b) showed they were also matched for semantic integration; Solomon and Pearlmutter also replicated the clause-boundedness effect with their own set of integration- and length-matched stimuli. Thus, either or both of hierarchical feature-passing and clause-boundedness might at least constrain agreement computation in addition to the scope of planning.

Experiment 1 investigated this question by re-examining the clause-boundedness effect. While PPs and RCs in previous studies were matched on length and semantic integration, they differed in at least three other potentially relevant ways: First, the RCs linked the head and local noun with a content word (a semantically-rich verb), whereas the PPs used a function word (the preposition). Second, the number of adjectives modifying the local noun differed, because adjectives were added to the PPs to equate overall length. Either of these differences might have influenced the timing of planning of the local noun relative to the head noun, which would be relevant under a scope of planning account (see also Lewis & Badecker, 2010). Third, the PPs and RCs differed in overall meaning; and various conceptual properties have been shown to influence agreement error rates, either directly (e.g., distributivity, noun conceptual number; Eberhard et al., 2005; Vigliocco, Butterworth, & Garrett, 1996; Vigliocco, Butterworth, & Semenza, 1995; Vigliocco, Hartsuiker, et al., 1996) or indirectly (e.g., concreteness, Eberhard, 1999).

The goal of this study was to test the predictions of clause-boundedness and hierarchical feature-passing while controlling for semantic integration, linear distance, and the three properties discussed above. The only difference between the PPs and RCs used in this experiment was that the PPs contained the preposition with, in its attribute/possessive sense, while the RCs contained the verb had, in its relatively semantically-light possessive sense. RCs also always contained the complementizer that, making RCs exactly one word longer than PPs. Thus, the extra word was always one syllable long, making RCs one syllable longer than PPs. Thus, the PPs and RCs were matched in number of adjectives, properties
of the linking word, and general meaning; however, they differed in clausal structure and the local noun’s hierarchical distance to the subject NP node (see Franck et al., 2002; Solomon & Pearlmutter, 2004b, Exp. 5). According to most syntactic theories, clauses introduce extra functional nodes (e.g., CPs, TPs) to the syntactic tree which increases the structural distance between the head and local noun (e.g., Carnie, 2005; Pollard & Sag, 1994; Radford, 2006). If the difference in error rates between PPs and RCs in previous studies was due to structure instead of any of the other factors that varied, the PP mismatch effect here should be greater than the RC mismatch effect.

A secondary goal was to compare the two commonly-used versions of the agreement elicitation task: (1) recall tasks, which require speakers to listen to or read preambles, hold them in memory, then repeat them to complete a full sentence; and (2) no-recall tasks, which require speakers to read preambles aloud and then complete them as full sentences (see Bock et al., 2006). Both tasks have shown structural effects (cf. Bock & Cutting, 1992; Solomon & Pearlmutter, 2004b), but timing of planning may be different across them due to differing memory demands or differences in the influence of comprehension processes during production. Task was manipulated between participants to examine these possibilities.

1.0.1. Method

Participants. Fifty-nine Northeastern University undergraduates participated in the no-recall task, but one participant was excluded for being unable to read the preambles before they disappeared. Sixty-four Northeastern University students participated in the recall task, but two participants were excluded because they began speaking before the signal tone on nearly every trial. All participants were native American English speakers and received course credit for their participation; no participant provided data for more than one part of the experiment.
Table 2.1. Stimuli and Semantic Integration Ratings by Condition

<table>
<thead>
<tr>
<th>Modifier</th>
<th>Noun Number</th>
<th>Example</th>
<th>Semantic Integration</th>
</tr>
</thead>
<tbody>
<tr>
<td>PP</td>
<td>SP</td>
<td>The pizza with the missing slices</td>
<td>5.56 (1.22)</td>
</tr>
<tr>
<td>SS</td>
<td></td>
<td>The pizza with the missing slice</td>
<td>5.65 (1.20)</td>
</tr>
<tr>
<td>RC</td>
<td>SP</td>
<td>The pizza that had the missing slices</td>
<td>5.58 (1.28)</td>
</tr>
<tr>
<td>SS</td>
<td></td>
<td>The pizza that had the missing slice</td>
<td>5.53 (1.35)</td>
</tr>
</tbody>
</table>

Note. The semantic integration rating scale was 1 (loosely linked) to 7 (tightly linked); standard deviations are in parentheses. PP = prepositional phrase; RC = relative clause; SP = singular head, plural local noun; SS = singular head, singular local noun.

Materials and design. Twenty-four stimulus sets like that shown in Table 2.1 were constructed. Each began with a head NP (e.g., The pizza) followed by a modifier containing a local noun (e.g., slice(s)). The head noun was always singular, and the four different versions of an item were created by varying modifier type and (local) noun number. The modifier was either a PP or an RC and described an attribute of the head noun. PP modifiers began with the preposition with and were followed by a local NP consisting of a determiner, adjective, and noun. RC modifiers began with the complementizer that and the verb had, followed by the same local NP. As a result, the RCs were always exactly one word (or syllable) longer than the corresponding PPs. The complete list of critical stimuli is shown in Appendix A.

In addition to the critical items, 88 fillers were included. Twenty-four of the fillers had structures like the critical items but had plural heads. The rest had a variety of structures varying in head noun number and were similar in length and complexity to the critical items. The critical items and fillers were combined in four counterbalanced lists, each containing all fillers and exactly one version of each of the critical items. Each list was seen by 14–15 participants in the no-recall task and by 15–16 participants in the recall task.
Stimulus norming. The 24 critical stimuli were normed for semantic integration by 51 participants (2 more were excluded for failing to follow instructions). The 4 different versions of each of the 24 items, along with 24 fillers intended to cover the full rating scale, were rated using a 1 (loosely linked) to 7 (tightly linked) scale. The instructions included example phrases the ketchup or the mustard and the bracelet made of silver and indicated that although ketchup and mustard are similar in meaning, they are not closely related in the particular example phrase, in contrast to bracelet and silver, which are closely related in the example phrase\(^1\). The 4 versions of each item were counterbalanced across 4 lists such that exactly one version of each stimulus item appeared in each list, and 12–14 integration ratings were obtained for each version of an item\(^2\).

Table 2.1 shows the mean integration ratings and standard deviations by condition for the critical stimuli. A linear mixed-effect regression (Baayen, Davidson, & Bates, 2008) on these data (random factors: participant, item; fixed effects: local noun number, modifier, and their interaction) revealed no main effects nor an interaction (all \(|t|s < 1.2, ps > .23\))\(^3\).

Apparatus and procedure. Each participant was run individually in the main experiment using either the no-recall or recall task. In the no-recall task, participants read each visually-presented preamble aloud as soon as it appeared and added an ending that formed a complete sentence. In the recall task, participants read each visually-presented preamble silently as soon as it appeared, and then, after a tone, repeated the preamble aloud and added an ending that formed a complete sentence. Participants were not instructed as to how they should formulate a completion, only that they should form a complete sentence.

\(^1\)See (Solomon & Pearlmutter, 2004b) for the full text of the instructions used for semantic integration ratings in Experiment 1.

\(^2\)Similar norming studies were completed for Experiments 2–4 in this dissertation and will be discussed in Chapter 3.

\(^3\)All regression analyses in this dissertation were performed in R (R Development Core Team., 2011), using the languageR package (Baayen, 2008). Models were fit using the lme4 package (v. 0.999375-37), and \(p\)-values were obtained using the MCMC sampling function in the coda package (v. 0.14-2).
In both tasks, on each trial, a fixation cross appeared at the left edge of the display for 1000 ms, followed by the preamble. Each preamble was presented for the longer of 1000 ms or 50 ms/character. After the preamble disappeared, the screen was blank for 2000 ms, followed by a prompt to begin the next trial. In the recall task, a tone was presented immediately after the preamble disappeared to indicate that participants could begin speaking. A PC running the MicroExperimental Laboratory software package (Schneider, 1988) controlled stimulus presentation, and participants’ responses were recorded to CD for analysis, using a Shure SM58 microphone connected to a Mackie 1202-VLZ Pro mixer/preamp and an Alesis Masterlink ML-9600 (OS v2.20) CD recorder. Five practice items preceded the 112 trials. If at any point the participant’s speech rate slowed, the experimenter encouraged the participant to speak more quickly.

**Scoring.** All responses were transcribed and assigned to one of four coding categories: (1) correct, if the participant repeated the preamble correctly exactly once, produced an inflected verb immediately after the preamble, and used a verb form that was correctly marked for number; (2) error, if all the criteria for correct responses were met, but the verb form failed to agree in number with the subject; (3) uninflected, if all the criteria for correct responses were met, but the verb was uninflected; and (4) miscellaneous, if the participant made an error repeating the preamble, if a verb did not immediately follow the preamble, if participants began speaking before the tone in the recall task, or if the response did not fall into any of the other categories. Trials in which a participant made no response were excluded from all analyses. If the participant produced a disfluency (e.g., pauses, coughs) during or immediately after producing the preamble and went on to produce a correct, error, or uninflected response, the scoring category and the disfluency were recorded. On miscellaneous trials, disfluencies were not scored.
Table 2.2. Experiment 1 Response Counts by Task and Condition

<table>
<thead>
<tr>
<th>Task</th>
<th>Modifier</th>
<th>Noun Number</th>
<th>Error</th>
<th>Correct</th>
<th>Uninflected</th>
<th>Misc</th>
<th>No Resp</th>
</tr>
</thead>
<tbody>
<tr>
<td>No-Recall</td>
<td>PP</td>
<td>SP</td>
<td>18 (4)</td>
<td>196 (42)</td>
<td>95 (29)</td>
<td>37</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SS</td>
<td>0 (0)</td>
<td>230 (50)</td>
<td>84 (20)</td>
<td>31</td>
<td>3</td>
</tr>
<tr>
<td>RC</td>
<td>SP</td>
<td>SP</td>
<td>21 (1)</td>
<td>210 (49)</td>
<td>66 (16)</td>
<td>50</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SS</td>
<td>1 (0)</td>
<td>215 (53)</td>
<td>88 (23)</td>
<td>43</td>
<td>1</td>
</tr>
<tr>
<td>Recall</td>
<td>PP</td>
<td>SP</td>
<td>18 (3)</td>
<td>195 (11)</td>
<td>91 (12)</td>
<td>67</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SS</td>
<td>2 (0)</td>
<td>222 (27)</td>
<td>98 (9)</td>
<td>46</td>
<td>4</td>
</tr>
<tr>
<td>RC</td>
<td>SP</td>
<td>SP</td>
<td>8 (3)</td>
<td>186 (21)</td>
<td>90 (15)</td>
<td>86</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SS</td>
<td>0 (0)</td>
<td>207 (24)</td>
<td>107 (14)</td>
<td>57</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>68 (11)</td>
<td>1661 (277)</td>
<td>719 (138)</td>
<td>417</td>
<td>15</td>
</tr>
</tbody>
</table>

Note. Disfluency counts are in parentheses. PP = prepositional phrase; RC = relative clause; SP = singular head, plural local noun; SS = singular head, singular local; Misc = Miscellaneous; No Resp = No Response.

1.0.2. Results

Table 2.2 shows the counts of each response type by task, modifier, and noun number, with the number of responses containing a disfluency in parentheses. Separate analyses were performed for error rates (the proportion of error responses out of error plus correct responses), uninflected rates (the proportion of uninflected responses out of total scorable responses), and miscellaneous rates (the proportion of miscellaneous responses out of total scorable responses). The reported error and uninflected analyses included disfluencies, and unless otherwise noted, the patterns were identical if disfluency cases were excluded.

Following Barr (2008), the data were analyzed using empirical logit weighted linear regression, aggregating separately over participants and items. For the details and a discussion of this method see Appendix B. By-participant and by-item weighted linear regressions on transformed error, miscellaneous, and uninflected rates were performed; with noun number, modifier, task, and all interactions as sum-coded fixed effects (t-tests of parameter estimates...
### Table 2.3. Experiment 1 Agreement Error Rate Results (Weighted Empirical Logit Linear Regression)

<table>
<thead>
<tr>
<th>Effect</th>
<th>By Participants</th>
<th>By Items</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β</td>
<td>SE</td>
</tr>
<tr>
<td>Noun number (SP)</td>
<td>0.53</td>
<td>.08</td>
</tr>
<tr>
<td>Modifier (RC)</td>
<td>−0.03</td>
<td>.08</td>
</tr>
<tr>
<td>Noun number × Modifier</td>
<td>−0.13</td>
<td>.15</td>
</tr>
<tr>
<td>Task (Recall)</td>
<td>0.01</td>
<td>.13</td>
</tr>
<tr>
<td>Task × Noun number</td>
<td>−0.14</td>
<td>.15</td>
</tr>
<tr>
<td>Task × Modifier</td>
<td>−0.18</td>
<td>.15</td>
</tr>
<tr>
<td>Task × Noun number × Modifier</td>
<td>−0.06</td>
<td>.30</td>
</tr>
</tbody>
</table>

*Note.* The level shown in parentheses for each variable was sum-coded +0.5 and the other level −0.5, so βs estimate the difference between the two levels of the variable in log-odds space. SP = singular head, plural local noun; RC = relative clause.

***p < .001. †p < .10.

are identified as $t_1$ for the by-participant analysis and as $t_2$ for the by-item analysis. We also computed corresponding ANOVAs on arcsine-transformed proportions (J. Cohen & Cohen, 1983), including all 58 participants from the no-recall task and 56 of the 62 participants from the recall task (6 were excluded because they were missing data in one or more cells). ANOVA result patterns were identical to those in the regression analyses unless otherwise noted.

**Agreement errors.** Figure 2.1 shows untransformed error rates by condition collapsed over task, and Table 2.3 shows the weighted linear regression effect estimates. Errors were more likely when the local noun was plural than when it was singular, but there was no main effect of modifier and, critically, no interaction of noun number and modifier.

The tasks did not differ in the main analyses, but when disfluencies were excluded, there were more errors in the no-recall task than in the recall task (significant by participants, marginal by items). The interaction of task and modifier was marginal by participants and
by items, such that in the no-recall task RCs yielded more errors than PPs, but in the recall task PPs yielded more errors than RCs; however, separate analyses on each task showed no main effects of modifier. The interaction of task and modifier reached significance in the ANOVA by participants when disfluencies were excluded. The interaction of task and noun number only reached significance in the ANOVA by items when disfluencies were excluded, indicating that the mismatch effect was larger in the no-recall task than in the recall task. There was no hint of a three-way interaction in any analysis.
Uninflected rates. The interaction of task and modifier was significant by items ($t_1 = 1.45, p = .21; t_2 = 2.14, p < .05$), with higher uninflected rates for RCs than PPs in the recall task, and higher rates for PPs than RCs in the no-recall task. The interaction of local noun number and modifier was marginal by items ($t_1 = -1.66, p = .14; t_2 = -3.18$), with RCs yielding higher uninflected rates for singular than for plural local noun cases, and PPs yielding nearly equal uninflected rates for the two; this effect was non-significant when disfluences were excluded. Also, when disfluences were excluded, uninflected responses were more likely for singular local nouns ($t_1 = -1.70, p = .09; t_2 = -3.67, p < .05$). No other main effects nor interactions approached significance (all $|t|s < 2.6, ps > .11$).

Miscellaneous rates. Miscellaneous responses were more likely for plural than singular local nouns ($t_1 = 3.74, t_2 = 4.06, ps < .01$), for RCs than for PPs ($t_1 = 2.94, t_2 = 3.54, ps < .05$), and in the recall task than in the no-recall task ($t_1 = 3.15, t_2 = 6.02, ps < .001$). There were no interactions (all $|t|s < 1.9, ps > .11$).

1.0.3. Discussion

The large noun number effect replicates essentially all studies examining mismatch effects: With singular heads, agreement error rates are larger when the local noun is plural than when it is singular (e.g., Bock & Miller, 1991; Eberhard, 1997). However, unlike similar previous studies (Bock & Cutting, 1992; Solomon & Pearlmutter, 2004b), modifier and noun number did not interact, indicating equal mismatch effects for PPs and RCs. Thus, the current study provided no evidence for structural effects on agreement when other differences between PPs and RCs were minimized. Because this experiment directly tested predictions of clause-boundedness and hierarchical feature-passing by manipulating clausal structure, thus varying the number of syntactic nodes a local noun’s plural feature would have to pass through in order to influence agreement computation, these findings argue for an account of agreement production that does not involve a hierarchical component.
This result goes beyond Gillespie and Pearlmutter’s (2011), which had left open the possibility that hierarchical feature-passing could be the mechanism underlying agreement production, as long as the set of feature sources for feature-passing was constrained by the scope of planning. This possibility would allow a hierarchical feature-passing theory a second explanation for clause-boundedness effects (in addition to hierarchical distance differences), if local nouns in RCs were less likely than those in PPs to be planned overlappingly with the head. But the presence of equal interference from local nouns in PPs and in RCs in the current study indicates that the relevant scope of planning did not vary across modifier type, ruling out a hierarchical explanation in these terms as well. Thus, this finding is incompatible with even a highly constrained use of feature-passing in agreement production.

A secondary goal of Experiment 1 was to determine if task affected error rates, but there was no three-way interaction of task, modifier, and noun number. Because the recall task requires speakers to hold the preamble in memory prior to repeating it and completing a sentence, interference could have arisen during retrieval of the preamble. This would have led to more agreement errors overall in the recall task if number information was susceptible to this interference. The recall task did increase miscellaneous errors; but agreement error rates tended to be higher in the no-recall than in the recall task, suggesting that recall was not responsible for agreement error production. One limitation to this study is that both tasks involve comprehension, which could be another source of interference (for discussion see Gillespie & Pearlmutter, 2011). While the comprehension component differed between tasks (concurrent with production in the no-recall task, prior to production in the recall task), mismatch effects were equal. Overall, the current findings indicate that the two tasks produce very similar error patterns, and number interference effects are likely to arise during the planning for overt production and not during retrieval or comprehension (also see Bock et al., 2006).

One general concern with this study is that the conclusions depend on the absence of an interaction; but this cannot be due to a lack of power: First, the noun number effect
was clear, replicating earlier results. Second, the clause-boundedness pattern was robust in previous studies, and with the current study’s greater number of participants per list (29+, compared to 10 for both Bock & Cutting, 1992, and Solomon & Pearlmutter, 2004b) and equal or greater number of items per condition (6 for the current study and for Solomon and Pearlmutter, 4 for Bock and Cutting), it should have been replicable.

Thus while structure-based theories cannot account for the current results, Gillespie and Pearlmutter’s (2011) scope of planning account does, and it may be able to explain previous clause-boundedness effects, as long as one or more of the factors controlled in this study but varying in the prior clause-boundedness experiments (meaning, properties of the linking word, number of adjectives) affects the relative timing of planning of the head and local noun. While this study alone cannot determine how controlling each of the additional factors may have contributed to eliminating the clause-boundedness effect, we explore some possible explanations below and these alternatives are discussed in more detail in Chapters 3–4.

First, the differences in meaning between the PPs and RCs in previous studies may have affected the processing required to compute the semantic relation between the head and local noun. Verbs require activation of their particular argument structure (e.g., Bock & Levelt, 1994), and the variety of verbs with complex argument structures used in previous studies may have resulted in less planning overlap of the head and local nouns in RCs than in PPs. In the current study, both the PPs and RCs always specified an attribute relation between the head and local noun, which may have led to equal planning overlap in the two conditions. As discussed in Chapter 1, conceptual number also affects agreement computation, so matching meaning may have eliminated relevant differences unrelated to timing of planning as well. This possibility is explored further in Chapter 4. Second, the verbs used in previous studies’ RCs were much lower in frequency than the prepositions used in their PPs. In general, low-frequency words are processed more slowly than high-frequency words (e.g., Oldfield & Wingfield, 1965), and this could have led to later planning of the local nouns in the RCs, placing them more often outside the scope of planning of the
head. In Experiment 1, this difference in frequency (for *had* vs. *with*) was relatively small (see Chapter 3). A final possible explanation for the clause-boundedness effect observed in Bock and Cutting (1992) and Solomon and Pearlmutter (2004b) comes from the fact that these studies always matched length between PP and RC conditions by adding an adjective modifying the local noun in the PP conditions. In the current study, the PP and RC versions of an item each had the same single adjective modifying the local noun. Bock and Cutting (Exp. 3) found that PPs lengthened by an extra adjective produced larger mismatch effects; therefore, the clause-boundedness effect observed in previous studies could have been due to the extra adjective, which would predict more errors for PPs relative to RCs.

In sum, Experiment 1 provides evidence that interference can arise from elements within separate clauses, and thus that the scope of planning in language production can extend beyond the current clause. The results also argue against hierarchical feature-passing as the mechanism of agreement computation in production, suggesting that structure does not independently constrain the computation. Gillespie and Pearlmutter’s (2011) scope of planning account does explain these results, suggesting that agreement interference depends on the timing of availability of interfering elements. Earlier clause-boundedness effects may have been confounded with differences in meaning and with other properties that may affect timing of planning during production. Chapter 3 explores how linking word frequency differences and differences in embedded verb structural biases may be responsible for the conflicting findings in Experiment 1 and other studies that reported clause boundedness effects (Bock & Cutting, 1992; Solomon & Pearlmutter, 2004b) as these properties have been shown to affect timing in other language production and comprehension phenomena.
CHAPTER 3

Exploring lexical alternatives to the clause-boundedness effect

The findings discussed in Chapter 2 suggest that clausal structure does not strongly constrain agreement computation, as local nouns in PPs and RCs showed equal mismatch effects when a variety of properties were matched between modifier conditions. However, it was unclear which properties that were matched in Experiment 1 but varied in previous studies (Bock & Cutting, 1992; Solomon & Pearlmutter, 2004b) were responsible for eliminating the clause-boundedness effect.

A great deal of research has examined the effects of structural properties on agreement computation (e.g., Bock & Cutting, 1992; Franck et al., 2002, 2006, 2010; Hartsuiker et al., 2001; Solomon & Pearlmutter, 2004b); however, studies that have found the clause-boundedness effect have varied lexical properties along with structure (Bock & Cutting, 1992; Solomon & Pearlmutter, 2004b). In those studies, phrasal conditions linked the head and local noun with a small number of prepositions, while in clausal conditions the head and local noun were linked with a wide variety of content verbs. Some studies have examined properties of individual lexical items in agreement computation, such as the relative frequency of plural and singular forms of local nouns (e.g., Barker & Nicol, 2000a; Barker, Nicol, & Garrett, 2000), the relative concreteness of the head and local noun (e.g., Bock & Miller, 1991), and the conceptual number of head and local nouns (e.g., Bock et al., 2001, 2004, 2006; Humphreys & Bock, 2005). However, no studies to date have examined properties of the content verbs used in subject NPs with relative clause modifiers, even though they have varied in many ways. Experiments 2 and 3 examined two properties of the content verbs that varied in previous studies that were held constant in Experiment 1: frequency and verb bias.
Many types of information are encoded in each lexical item and nearly all theories of lexical access claim that semantic and syntactic information is separately represented from the phonological information (e.g., Caramazza, 1997; though there is much disagreement about the nature of these representations). The representation that encodes the semantic and syntactic information is called the *lemma* and the representation that encodes the phonological information is called the *lexeme* (Bock & Levelt, 1994). Word frequency and syntactic bias (e.g., how likely a word is to appear in a particular syntactic structure) are both types of information thought to be encoded in representations of individual lexical items (e.g., Dell, 1990; Garnsey, Pearlmutter, Myers, & Lotocky, 1997; Jescheniak & Levelt, 1994). Most evidence suggests that frequency is encoded at the lexeme level (Jescheniak & Levelt, 1994; Jescheniak, Meyer, & Levelt, 2003, cf. Caramazza, Miozzo, & Bi, 2001; Caramazza, Bi, Costa, & Miozzo, 2004), while syntactic biases are encoded at the lemma level (Garnsey et al., 1997; Jaeger, 2010). Thus, Experiments 2 and 3 examined lexical influences on agreement computation that may arise from different stages of lexical access.

The verbs used in the RCs in the former studies were fairly low-frequency content words (when compared to prepositions in the phrasal modifier conditions), while in Experiment 1, the verb in the RCs (always possessive *had*) was relatively high frequency and a homophone of a high-frequency function word (auxilliary *had*). In general, high-frequency words are processed and named more quickly than low-frequency words (e.g., Oldfield & Wingfield, 1965). There is evidence that high-frequency function word and lower-frequency content word homophones (e.g., *be* and *bee*) are named equally quickly because they share a phonological representation (Jescheniak & Levelt, 1994; Jescheniak et al., 2003, cf. Caramazza et al., 2001, 2004), and that content words are susceptible to the same frequency effects as their function word homophones in phonological speech errors (Dell, 1990). Thus, in Experiment 1, *had* may have been processed like other high-frequency function words (e.g., prepositions).

Because speakers in agreement elicitation studies are required to produce (and thus, phonologically access) each word in the preamble, the large frequency difference between RC
verbs and prepositions in previously published studies and the comparatively small frequency
difference between RC verbs and prepositions in Experiment 1 may have been responsible
for the difference in error patterns between studies: When semantic integration and linear
distance are controlled, fast planning of high-frequency linking words (e.g., function words
and their homophones) may cause N2 to be planned closer in time to the head than when the
linking word is low-frequency (e.g., content verbs), thus causing a greater chance of interfer-
ence when linking words are high-frequency than when they are low-frequency. Experiment 2
tests the hypothesis that RC error rates in previous studies (Bock & Cutting, 1992; Solomon
& Pearlmutter, 2004b) were artificially reduced because the low-frequency content verb de-
layed the planning of the local noun relative to the PP condition. This frequency difference
between PPs and RCs was not present in Experiment 1, which may have resulted in the
equal error rates across structure conditions.

As mentioned above, word frequency is not the only lexical property that has been shown
to affect language processing in real time. Constraint-based theories of sentence comprehen-
sion suggest that as a word is recognized, several types of information become activated,
and these sources of information can constrain the interpretation of the upcoming material
within a sentence. One source of information shown to affect sentence comprehension is verb
bias, which is the relative frequency with which a verb is followed by a particular syntactic
structure (e.g., Garnsey et al., 1997). For example, the verb read can appear in a variety
of syntactic structures, but it is most often followed by a direct object (e.g., She read the
book. vs. She read that the book was written by a woman.); thus, the verb read’s bias assigns
higher probability to an upcoming noun being a direct object than other alternatives. In
contrast, the verb believe has a strong bias to be followed by an embedded clause (e.g., I
believed that the girl was nice. vs. I believed the girl.). Thus, the verb believe’s bias assigns
higher probability to the upcoming noun being a subject of a sentential complement than
other possible alternatives. In sentence comprehension, words that are predictable in a given
context are processed faster than less predictable words (Garnsey et al., 1997; Levy, 2008;
Trueswell, Tanenhaus, & Kello, 1993). There is some evidence speakers are sensitive to verb bias during production as verb bias has been shown to influence production choices (for evidence from optional that-mentioning, see e.g., V. S. Ferreira & Dell, 2000; Jaeger, 2010; for evidence from phrasal ordering preferences based on dative alternation biases, see Stallings, MacDonald, & O’Seaghdha, 1998).

Because verbs varied in more ways than prepositions in previous studies examining the clause-boundedness effect, it is likely that the prepositions and verbs used in these studies differed in the degree to which they biased speakers toward expecting an upcoming noun (i.e., the local noun). In English, prepositions are almost always followed by a noun phrase (Huddleston & Pullum, 2002), while verbs have a range of transitivity biases (Connine, Ferreira, Jones, Clifton, & Fraizer, 1984; V. S. Ferreira & Dell, 2000; Gahl & Garnsey, 2004; Gahl, Jurafsky, & Roland, 2004; Garnsey et al., 1997; Pearlmutter & MacDonald, 1995; Trueswell et al., 1993). If highly probable/expected words are processed faster than less probable/expected words, as has been shown in comprehension, then it is possible that a preceding structure that strongly predicts the presence of an upcoming local noun will cause the head and local noun to be processed closer in time to each other, thus increasing the chance for interference. Because had, the only verb used in RCs in Experiment 1, had a similar bias toward being followed by a noun phrase as prepositions, this may have eliminated the clause-boundedness observed in other studies that used a variety of verbs in the RC versions (Bock & Cutting, 1992; Solomon & Pearlmutter, 2004b).

As discussed above, language users are sensitive to and keep track of probabilistic information from lexical items as they comprehend and produce speech. If language users use this information to predict upcoming material within an unfolding sentence during comprehension and to structure their utterances during production, it is possible that these expectations could also influence the agreement computation process (Thornton & MacDonald, 2003). There is some evidence that probabilistic information is being tracked and used during agreement computation (e.g., Barker & Nicol, 2000a; Solomon & Pearlmutter, 2005;
Staub, 2009, 2010), but these findings have come from work on agreement comprehension, as similar work in production has been limited and produced mixed results (Barker et al., 2000).

The goal of the following studies was to provide a direct test of lexical alternatives to the clause-boundedness effect by matching linear distance and semantic integration, while varying only lexical properties. Experiment 2 examined the effect of the embedded verb’s frequency on agreement error rates, and Experiment 3 examined the effect of the embedded verb’s transitivity bias on agreement error rates. If either of these lexical properties affect agreement computation, they could explain the differing findings between Bock and Cutting (1992) and Solomon and Pearlmutter (2004b) on one hand and Experiment 1 on the other.

1. Experiment 2

Previous work examining the effect of lexical item frequency on agreement computation has focused on the frequency of the local nouns. Barker and Nicol (2000a) found that local nouns that appeared in their plural form more often were more likely to interfere with agreement computation during comprehension than local nouns that were more likely to appear in their singular form; however, these effects were not replicated in similar production studies (Barker & Nicol, 2000b; Barker et al., 2000).

One reason why agreement computation in production may not be affected by word frequency is because word frequency is thought to be a property that is encoded in the lexeme representation of the word (Jescheniak & Levelt, 1994; Jescheniak et al., 2003, cf. Caramazza et al., 2001, 2004), and the lexeme is accessed after grammatical encoding and during phonological encoding. If agreement computation is a process that is affected by the scope of planning during grammatical encoding (Gillespie & Pearlmutter, 2011), it is possible that frequency effects occur too late in the production process (i.e., post-grammatical encoding) to have an effect on agreement.
Because no studies to date have examined the frequency of the embedded verb in RC modifiers, Experiment 2 tested the frequency hypothesis using NP RC preambles and varying the frequency of the embedded verb in the RC, while holding semantic integration and linear distance constant. If the speed at which the verb in the relative clause is processed affects the timing of planning of the local noun, more errors should be produced when plural local nouns are within RCs containing high-frequency verbs than when plural local nouns are within RCs containing low-frequency verbs because of the increased chance of planning overlap of the head and local nouns when the processing of the verb is speeded. Experiment 2a used NP RC stimuli that referred to concrete subjects, and Experiment 2b used NP RC stimuli that referred to more abstract subjects.

1.1. Experiment 2a

1.1.1. Method

Participants. Eight-one Northeastern University undergraduates participated in the experiment for course credit. Two participants who were not native English speakers were excluded as well as one participant who reported a language disability. Data were lost from one participant due to a recording failure. All remaining participants (N = 77) were native American English speakers.

Materials and design. Thirty-six stimulus sets like that shown in Table 3.1 were constructed. Each began with a head NP (e.g., The farmer) followed by a modifier containing a local noun (e.g., goat(s)). The head noun was always singular, and the four different versions of an item were created by varying verb frequency and local noun number. The modifier was always an RC, and although the verbs differed, care was taken to match the overall meaning when possible. Modifiers began with the complementizer that and one verb (e.g., poked, pushed), followed by the same local NP. Verb lemma log frequencies were obtained.
Table 3.1. Experiment 2a Stimuli and Semantic Integration Ratings by Condition

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Number</th>
<th>Noun Example</th>
<th>Semantic Integration</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>SP</td>
<td>The farmer that pushed the stubborn goats</td>
<td>3.88 (1.88)</td>
</tr>
<tr>
<td></td>
<td>SS</td>
<td>The farmer that pushed the stubborn goat</td>
<td>3.94 (2.01)</td>
</tr>
<tr>
<td>Low</td>
<td>SP</td>
<td>The farmer that poked the stubborn goats</td>
<td>3.99 (2.01)</td>
</tr>
<tr>
<td></td>
<td>SS</td>
<td>The farmer that poked the stubborn goat</td>
<td>4.19 (1.99)</td>
</tr>
</tbody>
</table>

Note. The semantic integration rating scale was 1 (loosely linked) to 7 (tightly linked); standard deviations are in parentheses. SP = singular head, plural local noun; SS = singular head, singular local noun.

from a subset of the CELEX database (Baayen, Feldman, & Schreuder, 2006). To ensure that verb frequency was manipulated as desired, a linear mixed effect model predicting verb frequency with frequency condition as a fixed effect factor, and item intercept as a random factor, was conducted. Verbs in the high frequency condition had significantly higher log frequencies than verbs in the low frequency condition ($t = 5.39, p < .001$). All verbs were one or two syllables long, and mean length did not differ across frequency conditions ($t = -1.20, p = .23$). The complete list of critical stimuli is shown in Appendix C.

In addition to the critical items, 84 fillers were included. Thirty-six of the fillers had structures like the critical items but had plural heads. The rest had a variety of structures varying in head noun number and were similar in length and complexity to the critical items. The critical items and fillers were combined in four counterbalanced lists, each containing all fillers and exactly one version of each of the critical items. Each list was seen by 18–20 participants.

Semantic integration rating survey. A large semantic integration survey was conducted on Amazon Mechanical Turk, an online service that allows workers to complete surveys and tasks for monetary compensation. Seventy-eight participants provided data for
the survey. All items from Experiment 1, Experiment 2a, Experiment 2b, Experiment 3a, Experiment 3b, and Experiment 4, as well as the PP and RC versions of Solomon and Pearlmutter (2004b; Experiment 5) and Bock and Cutting (1992; Experiment 2) were rated using nearly identical instructions to those used in Solomon and Pearlmutter and Gillespie and Pearlmutter (2011). The rating instructions that were used are presented in Appendix D. A quarter of the stimulus items from each of the experiments listed above were included in four separate versions of the survey, for a total of 58 items per version (6–9 items per study were included in each version).

In each version, there were four lists (each with 3 randomized orders) and each list contained each item in exactly one of its four conditions. Thus, twelve ratings per condition of each item were obtained. The head noun and the local noun in each item appeared in capital letters (e.g., the FARMER that pushed the stubborn GOAT) and participants were instructed to rate how conceptually linked the capitalized items in the phrase were on a 1 (not linked) to 7 (tightly linked) scale by clicking a radio button on the screen indicating their choice.

Participants could participate in each of the four versions of the survey (as each item only appeared in one version), but were not allowed to submit data for more than one list per version. Each version of the survey took approximately six minutes to complete and participants were paid $0.34 for each one that they completed.

The 36 critical stimuli in Experiment 2a were normed for semantic integration by 78 participants using the procedure described above. Table 3.1 shows the mean integration ratings and standard deviations by condition for the critical stimuli. A linear mixed-effect regression analysis on these data with local noun number, frequency, and their interaction as fixed effect factors, and participant and item intercepts as random factors, revealed no

---

1There was an error in the instructions in which the word bracelet was substituted for the word ring when referring to an example text in all versions of the survey (see Appendix D). We assume that this error was unlikely to cause any substantial problems; but, it may create noise, resulting in somewhat less stable ratings.
significant results (all $|t|s < 1.33, ps > .18$), indicating that semantic integration was matched across conditions, as desired.

**Apparatus and procedure.** Each participant was run individually in the main experiment. Exactly as in the no-recall task described in Experiment 1, participants read each visually-presented preamble aloud as soon as it appeared and added an ending that formed a complete sentence. Participants were not instructed as to how they should formulate a completion, only that they should form a complete sentence.

On each trial, a fixation cross appeared at the left edge of the display for 1000 ms, followed by the preamble. Each preamble was presented for the longer of 1000 ms or 50 ms/character. After the preamble disappeared, the screen was blank for 2000 ms, followed by a prompt to begin the next trial. The same apparatus was used as in Experiment 1. Five practice items preceded the 120 trials. If at any point the participant’s speech rate slowed, the experimenter encouraged the participant to speak more quickly.

**Scoring.** Scoring was identical to the scoring for the no-recall procedure in Experiment 1.

**1.1.2. Results and Discussion**

Table 3.2 shows the counts of each response type by frequency and noun number, with the number of responses containing a disfluency in parentheses. Subject-verb agreement errors were very rare (only 22 errors occurred in the entire experiment), but those errors tended to occur in SP conditions, as in previous studies (e.g., Bock & Cutting, 1992; Bock & Miller, 1991; Solomon & Pearlmutter, 2004b). Previous studies examining agreement error rates in RC preambles with content embedded verbs often produced mismatch effects between 5% and 10% (Bock & Cutting, 1992; Solomon & Pearlmutter, 2004b); however, subtracting the overall SS error rate from the SP error rate in Experiment 2a indicated that the mismatch effect was less than 2%. Thus, no inferential statistics were computed because the error
Table 3.2. Experiment 2a Response Counts by Condition

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Noun Number</th>
<th>Error</th>
<th>Correct</th>
<th>Uninflected</th>
<th>Misc</th>
<th>No Resp</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>SP</td>
<td>9 (2)</td>
<td>386 (68)</td>
<td>225 (39)</td>
<td>73</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>SS</td>
<td>1 (0)</td>
<td>342 (58)</td>
<td>269 (46)</td>
<td>78</td>
<td>3</td>
</tr>
<tr>
<td>Low</td>
<td>SP</td>
<td>9 (2)</td>
<td>343 (65)</td>
<td>256 (46)</td>
<td>84</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>SS</td>
<td>3 (1)</td>
<td>339 (62)</td>
<td>253 (44)</td>
<td>94</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>22 (5)</td>
<td>1380 (253)</td>
<td>1003 (175)</td>
<td>329</td>
<td>8</td>
</tr>
</tbody>
</table>

*Note.* Disfluency counts are in parentheses. SS = singular head, singular local; SP = singular head, plural local; Misc = Miscellaneous; No Resp = No Response.

rates were too low to examine the frequency by local noun number interaction, which was the critical test of the hypothesis.

1.2. Experiment 2b

One difference between the stimuli in Experiment 2a and previous studies that have observed the clause-boundedness effect was that the RCs in previous stimuli tended to refer to fairly abstract subject NPs (e.g., *the proof that changed the original ruling(s)*; Solomon & Pearlmutter, 2004b; cf. Experiment 1) while the stimuli in Experiment 2a were very concrete. The stimuli in Experiment 2b were designed to be more abstract in order to better match stimuli to previous studies and increase error rates.

1.2.1. Method

**Participants.** One hundred eighteen Northeastern University undergraduates participated in the experiment for course credit. Four participants who were not native English speakers were excluded, as well as two participants who reported a language disability. Two
participants were excluded for not completing the majority of sentences. Data were lost from one participant due to a recording failure. All remaining participants (N = 109) were native American English speakers.

**Materials and design.** Twenty-four stimulus sets like that shown in Table 3.3 were constructed. Each began with a head NP (e.g., *The evidence*) followed by a modifier containing a local noun (e.g., *witness(es)*). Plurality was manipulated the same way as in Experiment 2a. The modifier was always an RC and although the verbs differed, care was taken to match the overall meaning when possible. Modifiers began with the complementizer *that* and one verb (e.g., *confused, puzzled*), followed by the same local NP. To ensure that verb frequency was manipulated as desired, a linear mixed effect model predicting verb frequency with frequency condition as a fixed effect factor, and item intercept as a random factor, was conducted. Verbs in the high frequency condition had significantly higher log frequencies than verbs in the low frequency condition (t = 9.31, p < .001). Across languages higher frequency words are generally shorter than lower frequency words (e.g., Zipf, 1949). When choosing stimuli, the length difference across frequency conditions was minimized as much as possible given other constraints; however, in the end, verbs in the high frequency condition were shorter than the verbs in the low frequency condition (2.6 syllables vs. 2.3 syllables; t = -2.65, p < .05). This length difference will be discussed in relation to the results. The head nouns and verbs used in Experiment 2b were more abstract than those in Experiment 2a. Half of the verbs were psychological state verbs (B. Levin, 1993) and the other half were verbs of description (e.g., *described, reported*). The complete list of critical stimuli is shown in Appendix E.

In addition to the critical items, 80 fillers were included. Twenty-four of the fillers had structures like the critical items but had plural heads. The rest had a variety of structures varying in head noun number and were similar in length and complexity to the critical items. The critical items and fillers were combined in four counterbalanced lists, each containing
Table 3.3. Experiment 2b Stimuli and Semantic Integration Ratings by Condition

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Noun Number</th>
<th>Example</th>
<th>Semantic Integration</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>SP</td>
<td>The evidence that confused the witnesses</td>
<td>3.37 (1.90)</td>
</tr>
<tr>
<td></td>
<td>SS</td>
<td>The evidence that confused the witness</td>
<td>3.46 (1.89)</td>
</tr>
<tr>
<td>Low</td>
<td>SP</td>
<td>The evidence that puzzled the witnesses</td>
<td>3.65 (1.92)</td>
</tr>
<tr>
<td></td>
<td>SS</td>
<td>The evidence that puzzled the witness</td>
<td>3.48 (1.74)</td>
</tr>
</tbody>
</table>

Note. The semantic integration rating scale was 1 (loosely linked) to 7 (tightly linked); standard deviations are in parentheses. SP = singular head, plural local noun; SS = singular head, singular local noun.

all fillers and exactly one version of each of the critical items. Each list was seen by 27–28 participants.

**Stimulus norming.** The 24 critical stimuli were normed for semantic integration by 78 participants in the survey described in Experiment 2a. Table 3.3 shows the mean integration ratings and standard deviations by condition for the critical stimuli. A linear mixed-effect regression on these data with local noun number, frequency, and their interaction as fixed effect factors, and participant and item intercepts as random factors, revealed no significant results (all $|t|s < 1.33$, $ps > .19$), indicating that semantic integration was matched across conditions, as desired.

**Apparatus and procedure.** The apparatus and procedure were identical to Experiment 2a, except there were 104 trials.

**Scoring.** Scoring was identical to Experiment 2a.
Table 3.4. Experiment 2b Response Counts by Condition

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Noun Number</th>
<th>Error</th>
<th>Correct</th>
<th>Uninflected</th>
<th>Misc</th>
<th>No Resp</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>SP</td>
<td>27 (10)</td>
<td>450 (64)</td>
<td>80 (17)</td>
<td>96</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>SS</td>
<td>3 (2)</td>
<td>487 (67)</td>
<td>101 (17)</td>
<td>62</td>
<td>1</td>
</tr>
<tr>
<td>Low</td>
<td>SP</td>
<td>33 (5)</td>
<td>447 (70)</td>
<td>79 (19)</td>
<td>92</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>SS</td>
<td>1 (0)</td>
<td>497 (90)</td>
<td>90 (21)</td>
<td>63</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>64 (17)</td>
<td>1881 (291)</td>
<td>350 (74)</td>
<td>313</td>
<td>8</td>
</tr>
</tbody>
</table>

*Note. Disfluency counts are in parentheses. SS = singular head, singular local; SP = singular head, plural local; Misc = Miscellaneous; No Resp = No Response.*

1.2.2. Results

Table 3.4 shows the counts of each response type by frequency and noun number, with the number of responses containing a disfluency in parentheses. Identical to Experiment 1, separate analyses were performed for error rates (the proportion of error responses out of error plus correct responses), uninflected rates (the proportion of uninflected responses out of total scorable responses), and miscellaneous rates (the proportion of miscellaneous responses out of total scorable responses). The reported error and uninflected analyses included disfluencies, and unless otherwise noted, the patterns were identical if disfluency cases were excluded. The data were analyzed using empirical logit weighted linear regression, aggregating separately over participants and items. By-participant and by-item weighted linear regressions on transformed error, miscellaneous, and uninflected rates were performed; with noun number, frequency, and their interaction as sum-coded fixed effect factors (*t*-tests of parameter estimates are identified as *t*₁ for the by-participant analysis and as *t*₂ for the by-item analysis).

We also computed corresponding ANOVAs on arcsine-transformed proportions (J. Cohen & Cohen, 1983), including 104 participants (5 were excluded because they were missing data
in one or more cells). ANOVA result patterns were identical to those in the regression analyses unless otherwise noted.

**Agreement errors.** Figure 3.1 shows untransformed error rates by condition, and Table 3.5 shows the weighted linear regression effect estimates. Errors were more likely when the local noun was plural than when it was singular, but there was no main effect of frequency and, critically, no interaction of noun number and frequency.

**Uninflected rates.** There were no main effects and no interaction (all $|t|s < 1.75, ps > .17$). In the ANOVA analysis, if disfluencies were excluded, uninflected responses were more
Table 3.5. Agreement Error Rate Results for Experiment 2b
(Weighted Empirical Logit Linear Regression)

<table>
<thead>
<tr>
<th>Effect</th>
<th>By Participants</th>
<th>By Items</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β</td>
<td>SE</td>
</tr>
<tr>
<td>Noun number (SP)</td>
<td>0.52</td>
<td>.08</td>
</tr>
<tr>
<td>Frequency (High)</td>
<td>−0.02</td>
<td>.08</td>
</tr>
<tr>
<td>Noun number × Frequency</td>
<td>−0.17</td>
<td>.16</td>
</tr>
</tbody>
</table>

*Note.* The level shown in parentheses for each variable was sum-coded +0.5 and the other level −0.5, so βs estimate the difference between the two levels of the variable in log-odds space. SP = singular head, plural local noun.

**Miscellaneous rates.** Miscellaneous responses were more likely when the local noun was plural than when it was singular (t_1 = 4.07, p < .001; t_2 = 5.96, p < .01), but there was no main effect of frequency and no interaction of noun number and frequency. In the ANOVA analysis by participants, the effect of local noun number was significant (F_1 = 12.21, p < .001; F_2 = 1.81, p = .19).

1.2.3. Discussion

In Experiment 2b, there was a reliable mismatch effect, suggesting that reducing the imageability of the head nouns and verbs used in the preambles may have led to an increase in error rates (compare Experiment 2b’s error rates to those in Experiment 2a). However, there was no effect of verb frequency, nor an interaction of noun number and frequency. These results indicate that verb frequency does not affect agreement computation for subject NPs with RC modifiers. Although the low-frequency verbs in Experiment 2b were slightly longer
(in syllables) than the high-frequency verbs, this difference in length did not affect agreement error rates. Thus, the verb frequency difference between Experiment 1 and previous studies (Bock & Cutting, 1992; Solomon & Pearlmutter, 2004b) was unlikely to have been responsible for the different error patterns observed.

The only other previous work examining the effect of word frequency on agreement computation showed fairly mixed results (Barker & Nicol, 2000a). While the word frequency effect examined in Experiment 2 was different from those previously studied, the results are consistent with those findings and suggest that word frequency does not affect agreement computation in production. As mentioned above, one reason why this may be the case is because frequency is assumed to be a property of the lexeme - the phonological form of a word (Jescheniak & Levelt, 1994; Jescheniak et al., 2003, cf. Caramazza et al., 2001, 2004) and it may be access too late in the production process to affect grammatical encoding.

While verb frequency did not have an effect on agreement computation, other properties of content verbs may influence agreement error rates. Experiment 3 tested whether agreement computation is affected by the embedded verb’s transitivity bias.

2. Experiment 3

Previous studies of the clause-boundedness effect have not controlled for possible differences in verb bias nor differences between verb and preposition bias, though a variety of verbs were used in the RC modifiers (Bock & Cutting, 1992; Solomon & Pearlmutter, 2004b). In Experiment 1, only the verb possessive *had* was used in the RC modifiers, and *had* has a strong bias toward being followed by a noun phrase, much like prepositions in the PP modifiers.

Experiment 3 used NP RC preambles and varied the degree to which the verb within the RC modifier was biased toward being followed by a direct object. To manipulate this property, the verb’s transitivity biases was used (i.e., how often a verb occurs with a direct object). Verbs that have a strong transitivity bias should make local nouns more predictable
than verbs with a weak transitivity bias. According to an account of agreement production that suggests that the overlap in the timing of planning of the head noun and a local noun determines the degree to which a plural local noun will interfere with agreement computation, in conjunction with findings suggesting that the language production and comprehension systems are sensitive to probabilistic properties of upcoming material, the stronger the transitivity bias, the more likely errors should be. High transitivity verbs should speed the processing of the highly predictable local noun, increasing the chance of planning overlap and interference when the local noun is plural.

In previous studies that observed the clause boundedness effect (Bock & Cutting, 1992; Solomon & Pearlmutter, 2004b), the verbs in the RC conditions varied in the extent they predicted that a noun phrase would follow, while the prepositions in the PP conditions nearly always required a following noun phrase. In Experiment 1 there was no difference in the predictability of the upcoming local noun in the PP and RC conditions because the preposition *with* and the verb *had* are equally likely to be followed by a noun phrase. Thus, the difference between Experiment 1 and studies that observed a clause boundedness effect could have been due to differences in predictability of the local noun across PP and RC conditions.

Experiment 3 examined whether the transitivity bias of the embedded verb in RC modifiers has an effect on agreement computation. The stimuli used in Experiment 3a referred to concrete subjects, and the stimuli used in Experiment 3b referred to abstract subjects.

### 2.1. Experiment 3a

Experiment 3 examined whether verb bias affected agreement computation using NP RC preambles and varying the transitivity of the verb in the RC while holding linear distance constant.
2.1.1. Method

Participants. Ninety-two Northeastern University undergraduates participated in the experiment for course credit. Two participants who were not native English speakers were excluded as well as one participant who reported a language disability. Error rates in the study were extremely low, thus responses from 31 of the 92 participants were not transcribed or coded. In total, 58 participants’ data were included in the following analyses.

Materials and design. Twenty-four stimulus sets like that shown in Table 3.6 were constructed. Each began with an animate head NP (e.g., The actor) followed by an RC modifier containing a local noun (e.g., line(s)). The head noun was always singular, and the four different versions of an item were created by varying verb transitivity and local noun number. Modifiers began with the complementizer who and one verb (e.g., quoted, yelled), followed by the same local NP. Although the verbs differed across conditions, care was taken to match the overall meaning of the preambles whenever possible. Verbs were selected from the norming lists provided in Gahl et al. (2004). Only verbs that had high or low transitivity biases across a range of corpora were used (see Gahl et al., Table 4). The complete list of critical stimuli is shown in Appendix F.

In addition to the critical items, 88 fillers were included. Thirty-six of the fillers had structures like the critical items but had plural heads. The rest had a variety of structures varying in head noun number and were similar in length and complexity to the critical items. The critical items and fillers were combined in four counterbalanced lists, each containing all fillers and exactly one version of each of the critical items. Each list was seen by 22–23 participants.

Stimulus norming. The 24 critical stimuli were normed for semantic integration by 78 participants in the survey described in Experiment 2a. Table 3.6 shows the mean integration ratings and standard deviations by condition for the critical stimuli. A linear mixed-effect
Table 3.6. Experiment 3a Stimuli and Semantic Integration
Ratings by Condition

<table>
<thead>
<tr>
<th>Transitivity</th>
<th>Noun</th>
<th>Example</th>
<th>Semantic Integration</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>SP</td>
<td>The actor who quoted the lines</td>
<td>4.18 (1.83)</td>
</tr>
<tr>
<td></td>
<td>SS</td>
<td>The actor who quoted the line</td>
<td>4.41 (2.05)</td>
</tr>
<tr>
<td>Low</td>
<td>SP</td>
<td>The actor who yelled the lines</td>
<td>3.87 (1.95)</td>
</tr>
<tr>
<td></td>
<td>SS</td>
<td>The actor who yelled the line</td>
<td>4.02 (2.06)</td>
</tr>
</tbody>
</table>

*Note.* The semantic integration rating scale was 1 (loosely linked) to 7 (tightly linked); standard deviations are in parentheses. SP = singular head, plural local noun; SS = singular head, singular local noun.

regression (Baayen et al., 2008) on these data with local noun number, transitivity, and their interaction as fixed effect factors, and participant and item intercepts as random factors, revealed that high transitivity items were more integrated than low transitivity items ($t = -4.20, p < .001$). No other results were significant (all $|t|s < 1.29, ps > .19$). Though the integration rating difference between transitivity conditions was small (High = 4.30, Low = 3.95) compared to other studies that have manipulated integration (e.g., integration rating differences between integrated and unintegrated versions of stimuli in Solomon and Pearlmutter (2004b) were over 1 point on the same scale), this unexpected difference will be discussed in relation to the results of the main experiment.

**Apparatus and procedure.** The apparatus and procedure were identical to Experiment 2a, except there were 112 trials.

**Scoring.** Scoring was identical to Experiment 2a.
Table 3.7. Experiment 3a Response Counts by Condition

<table>
<thead>
<tr>
<th>Transitivity Number</th>
<th>Noun</th>
<th>Error</th>
<th>Correct</th>
<th>Uninflected</th>
<th>Misc</th>
<th>No Resp</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>SP</td>
<td>5 (0)</td>
<td>185 (14)</td>
<td>111 (14)</td>
<td>46</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>SS</td>
<td>2 (0)</td>
<td>174 (18)</td>
<td>130 (18)</td>
<td>41</td>
<td>1</td>
</tr>
<tr>
<td>Low</td>
<td>SP</td>
<td>6 (2)</td>
<td>162 (19)</td>
<td>117 (14)</td>
<td>62</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>SS</td>
<td>1 (0)</td>
<td>149 (13)</td>
<td>147 (9)</td>
<td>50</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>14 (2)</td>
<td>670 (64)</td>
<td>505 (55)</td>
<td>199</td>
<td>4</td>
</tr>
</tbody>
</table>

*Note.* Disfluency counts are in parentheses. SS = singular head, singular local; SP = singular head, plural local; Misc = Miscellaneous; No Resp = No Response.

2.1.2. Results and Discussion

Table 3.7 shows the counts of each response type by frequency and noun number, with the number of responses containing a disfluency in parentheses. Similar to Experiment 2a, subject-verb agreement errors were very rare (only 14 errors occurred in the entire experiment), but those errors tended to occur in SP conditions, as in previous studies (e.g., Bock & Miller, 1991; Bock & Cutting, 1992; Solomon & Pearlmutter, 2004b). Even the small difference in integration between high and low transitivity conditions did not serve to increase error rates. Because of the low error rates, no inferential statistics were computed.

2.2. Experiment 3b

As in Experiment 2, the more imageable preambles in Experiment 3a did not produce reliable error rates. Thus, the preambles in Experiment 3b were designed to be more abstract than those in Experiment 3a to increase the chance of errors occurring. Semantic integration ratings of the high and low transitivity conditions were matched in Experiment 3b, removing the potential confound that was present in Experiment 3a.
2.2.1. Method

Participants. Sixty-two Northeastern University undergraduates participated in the experiment for course credit. One participant was excluded for being a native speaker of British English and one participant was excluded for failing to complete the task. All remaining participants (N = 60) were native American English speakers.

Materials and design. Twenty-four stimulus sets like that shown in Table 3.8 were constructed. Each began with a head NP (e.g., *The memoir*) followed by an RC modifier containing a local noun (e.g., *transgression(s)*). Plurality was manipulated the same way as in Experiment 3a. The RC modifiers began with the complementizer *that* and one verb (e.g., *denied, confessed*), followed by the same local NP. Verbs were selected from the norming lists provided in Gahl et al. (2004) as in Experiment 3a. The nouns and verbs used were less imageable than in Experiment 3a, based on intuitions of the author². The complete list of critical stimuli is shown in Appendix G.

In addition to the critical items, the same 88 fillers from Experiment 3a were included. The critical items and fillers were combined in four counterbalanced lists, each containing all fillers and exactly one version of each of the critical items. Each list was seen by 14–16 participants.

Stimulus norming. The 24 critical stimuli were normed for semantic integration by 78 participants in the survey described in Experiment 2a. Table 3.8 shows the mean integration ratings and standard deviations by condition for the critical stimuli. A linear mixed-effect regression (Baayen et al., 2008) on these data with local noun number, transitivity, and their interaction as fixed effect factors, and participant and item intercepts as random factors, revealed no significant results (all |t|s < 1, ps > .34).

²Imageability ratings of all stimuli will be discussed in more detail in Chapter 4.
Table 3.8. Experiment 3b Stimuli and Semantic Integration Ratings by Condition

<table>
<thead>
<tr>
<th>Transitivity</th>
<th>Noun Number</th>
<th>Example</th>
<th>Semantic Integration</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>SP</td>
<td>The memoir that denied the transgressions</td>
<td>3.80 (1.88)</td>
</tr>
<tr>
<td></td>
<td>SS</td>
<td>The memoir that denied the transgression</td>
<td>3.93 (1.84)</td>
</tr>
<tr>
<td>Low</td>
<td>SP</td>
<td>The memoir that confessed the transgressions</td>
<td>3.85 (1.68)</td>
</tr>
<tr>
<td></td>
<td>SS</td>
<td>The memoir that confessed the transgression</td>
<td>3.71 (1.87)</td>
</tr>
</tbody>
</table>

*Note.* The semantic integration rating scale was 1 (loosely linked) to 7 (tightly linked); standard deviations are in parentheses. SP = singular head, plural local noun; SS = singular head, singular local noun.

**Apparatus and procedure.** The apparatus and procedure were identical to Experiment 2a, except there were 112 trials.

**Scoring.** Scoring was identical to Experiment 2a.

**2.2.2. Results**

Table 3.9 shows the counts of each response type by frequency and noun number, with the number of responses containing a disfluency in parentheses. Analyses were identical to those performed in Experiment 2b, except transitivity (instead of frequency) and its interaction with local noun number were included as fixed effect factors. Corresponding ANOVAs on arcsine-transformed proportions (J. Cohen & Cohen, 1983), were computed including 59 participants (1 participant was excluded because of missing data in two cells). ANOVA result patterns did not differ from those in the regression analyses.

**Agreement errors.** Figure 3.2 shows untransformed error rates by condition, and Table 3.10 shows the weighted linear regression effect estimates. Errors were more likely when
Table 3.9. Experiment 3b Response Counts by Condition

<table>
<thead>
<tr>
<th>Transitivity</th>
<th>Noun Number</th>
<th>Error</th>
<th>Correct</th>
<th>Uninflected</th>
<th>Misc</th>
<th>No Resp</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>SP</td>
<td>15 (4)</td>
<td>222 (27)</td>
<td>47 (11)</td>
<td>74</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>SS</td>
<td>4 (0)</td>
<td>235 (28)</td>
<td>53 (6)</td>
<td>66</td>
<td>2</td>
</tr>
<tr>
<td>Low</td>
<td>SP</td>
<td>21 (2)</td>
<td>209 (25)</td>
<td>54 (7)</td>
<td>74</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>SS</td>
<td>1 (0)</td>
<td>228 (20)</td>
<td>48 (8)</td>
<td>80</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>41 (6)</td>
<td>894 (100)</td>
<td>202 (32)</td>
<td>294</td>
<td>9</td>
</tr>
</tbody>
</table>

Note. Disfluency counts are in parentheses. SS = singular head, singular local; SP = singular head, plural local; Misc = Miscellaneous; No Resp = No Response.

Table 3.10. Agreement Error Rate Results for Experiment 3b
(Weighted Empirical Logit Linear Regression)

<table>
<thead>
<tr>
<th>Effect</th>
<th>By Participants</th>
<th>By Items</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\beta$</td>
<td>SE</td>
</tr>
<tr>
<td>Noun number (SP)</td>
<td>0.47</td>
<td>.12</td>
</tr>
<tr>
<td>Transitivity (High)</td>
<td>−0.06</td>
<td>.12</td>
</tr>
<tr>
<td>Noun number $\times$ Transitivity</td>
<td>−0.26</td>
<td>.23</td>
</tr>
</tbody>
</table>

Note. The level shown in parentheses for each variable was sum-coded +0.5 and the other level −0.5, so $\beta$s estimate the difference between the two levels of the variable in log-odds space. SP = singular head, plural local noun.

***$p<.001$, †$p = .10$.

the local noun was plural than when it was singular, but there was no main effect of transitivity. The interaction of noun number and transitivity was marginal by items, indicating a stronger mismatch effect in the low transitivity conditions. This interaction reached significance by items when disfluencies were excluded ($t_1 = −1.26, p = .12; t_2 = −2.02, p < .05$).
Uninflected and miscellaneous rates. There were no significant main effects nor interactions in the uninflected and miscellaneous analyses (all $|t|s < 1.88$, $ps > .30$).

2.2.3. Discussion

The mismatch effect was only observed when preambles were abstract subjects, just as in Experiment 2. The scope of planning hypothesis (Gillespie & Pearlmutter, 2011), coupled with a hypothesis that transitivity biases should affect the relative timing of planning of local nouns, predicted that the mismatch effect would be larger for preambles containing high
transitivity verbs than low transitivity verbs because local nouns would be more predictable and planned overlappingly with the head noun in the high transitivity conditions. The results of Experiment 3b show that the mismatch effect was numerically larger for low transitivity conditions than high transitivity conditions, opposite the pattern predicted by a hypothesis that verb transitivity biases affect the relative timing of planning of the head and local nouns. Thus differences in verb transitivity bias were not likely to have been responsible for the conflicting results from Experiment 1 and previous studies that observed a clause-boundedness effect (Bock & Cutting, 1992; Solomon & Pearlmutter, 2004b).

These results do not necessarily invalidate the scope of planning hypothesis (Gillespie & Pearlmutter, 2011), but they do suggest that one or more assumptions about how a verb’s transitivity bias may affect timing of planning may be incorrect. For example, while a high transitivity bias may make the upcoming noun phrase easier to process, it may not necessarily increase the chance of planning the head and local noun closer together in time. Instead, the effect may be quite local, and only affect the local noun’s planning time relative to the verb.

Interestingly, the transitivity bias of the embedded verb did show a trend toward affecting agreement error production for subject NPs with RC modifiers, which had never been investigated before. One possible explanation for this pattern is related to how transitivity biases are usually defined and computed. A verb’s transitivity bias refers to how often it is followed by a noun phrase serving as a direct object of the verb; however, many different types of structures can follow a given verb, some of which can begin with a noun that is not the direct object (e.g., Gahl et al., 2004; Garnsey et al., 1997). For example, the verb *decided* has a low transitivity bias because it is most often followed by a sentential complement (e.g., *John decided (that) books were a good choice*). If the complementizer (*that*) is omitted, a noun directly follows a low-transitivity verb, but it is not the direct object of the verb. Thus, a noun following a verb with a high transitivity bias is generally the direct object of the sentence, while a noun following a verb with a high sentential complement (SC) bias is
generally the subject of an embedded sentential complement. The direct object/sentential complement ambiguity has been used in numerous comprehension studies examining the effect of verb bias on the processing of upcoming material, and comprehenders were more likely to interpret a NP following a verb as a direct object when it followed a verb with a high transitivity bias, than when it followed a verb with a high SC bias (e.g., Fine, Qian, Jaeger, & Jacobs, 2010; Garnsey et al., 1997). Thus, comprehenders make predictions about the role an element within a sentence will play based on biases introduced by individual lexical items.

According to sentence completion norms for a number of English verbs (Argaman & Pearlmutter, 2002), seven of the low-transitivity verbs used in Experiment 3b had strong SC biases. Most speakers completed the critical items using the local noun as the direct object of an RC\(^3\); however, that does not rule out the possibility that speakers at least temporarily interpreted the local noun as the subject of a sentential complement in the strongly SC-biased items. If this were the case, speakers would have had two nouns simultaneously active during grammatical encoding with both of them potentially assigned the role of subject within a clause, which may have increased verb agreement with the local noun due to its similarity to the head noun of the subject NP (for discussion of the effect of multiple potential subjects on agreement error production, see Badecker & Lewis, 2007; Bock & Miller, 1991, Experiment 3; Haskell & MacDonald, 2005; Lewis & Badecker, 2010). In fact, of the 21 mismatch errors observed in the low-transitivity condition, 13 occurred in the seven items with high SC-biased verbs. This distribution of errors is significantly different than what would be expected if errors were distributed evenly among the items (\(\chi^2 = 10.89, p < .001\)).

In addition, there is evidence that agreement errors are more likely when a local noun is not explicitly marked as not being the subject of a sentence. Hartsuiker et al. (2001) examined the effect of object attraction in Dutch subject-verb agreement computation. Because

\(^3\)Any cases in which speakers completed the sentence with an SC (e.g., *The council that doubted the ideas were useful met yesterday*) were included in the miscellaneous scoring category because the verb following the preamble was not the verb of the matrix clause.
the word order in Dutch subordinate clauses is subject-object-verb (SOV), objects appear before the verb and can interfere with agreement computation in a similar way that local nouns do within complex subject NPs. In Dutch, full NPs do not have overt case marking, so they have the same form when they appear as subjects or objects of sentences; however, the form of some pronouns varies depending on whether they appear as subjects or objects of a sentence. Hartsuiker et al. found that plural object full NPs with no overt case marking and plural pronominal objects that have the same form regardless of sentence role interfered with agreement computation more than plural pronominal objects that had an unambiguous form indicating that they were not the subject of the sentence (for more evidence that ambiguous case marking affects agreement see Hartsuiker, Schriefers, Bock, & Kikstra, 2003).

Because English does not have overt case marking, there are situations in which the role an NP will play within the sentence is ambiguous (e.g., the local nouns in the preambles in Experiment 3b); however, a verb’s bias can make a particular interpretation of an upcoming ambiguous noun more or less likely. Given the preponderance of errors in the items with strongly SC-biased verbs in Experiment 3b, it is possible that the ambiguous role of the local noun was a factor that influenced agreement computation. Thus, an interesting avenue for future research would be to directly test the possibility that the predictability of the role of upcoming nouns can affect agreement computation. Specifically, plural local nouns that are likely to be interpreted as subjects of SCs should be more likely to interfere with agreement than plural local nouns that are likely to be interpreted as direct objects due to similarity-based interference. This would suggest that lexical biases can affect function assignment during language production (though possibly only by way of comprehension), and that nouns assigned similar grammatical functions as the head noun of the subject NP are more likely to interfere with agreement computation (cf. Badecker & Lewis, 2007; Lewis

---

4This is similar to English where full NPs (the cat) have identical forms whether they appear as sentential subjects (e.g., The cat chased the mouse) or as sentential objects (The dog chased the cat), but pronouns differ depending on what role they play in the sentence (e.g., compare He/She chased the mouse and The dog chased him/her).
Similarity-based interference is, to a large degree, a separate source of interference that arises through retrieval, unlike the interference effects predicted by a scope of planning account; however, for two similar elements to interfere with each other, both would need to be simultaneously active (see Badecker & Kuminia, 2007).

When the low-transitivity items with SC-biased verbs were excluded from analyses, errors were equal for SP versions (High Transitivity: 6.6%, Low Transitivity: 6.6%). Thus, when the potential confounding items were removed, there was no suggestion that transitivity bias affected error rates, nor was responsible for conflicting findings concerning the clause-boundedness effect.

3. Experiment 2 and 3 Discussion

Experiments 2 and 3 examined possible lexical alternatives to the clause-boundedness effect. Experiment 2 varied the frequency of the embedded verb of an RC modifier and Experiment 3 varied the transitivity bias of the embedded verb to determine whether either of these properties affected agreement error rates. Experiment 2 showed no effect of verb frequency on agreement computation, as high frequency and low frequency verbs produced similar error patterns. Experiment 3 showed a trend toward an effect of transitivity bias on agreement computation, with low transitivity conditions producing a slightly larger mismatch effect than high transitivity conditions. This pattern was opposite to that predicted by the scope of planning account combined with inferences about how verb transitivity biases may affect timing of planning. Instead, these findings could be attributed to interference due to similarity-based interference from ambiguous role assignment (Badecker & Lewis, 2007; Lewis & Badecker, 2010). In sum, neither of these experiments provided evidence that properties of the verbs tested here that differed between Experiment 1 and Solomon and Pearlmutter

\[ Badecker, 2010 \]. Simon et al

5 Barker, Nicol, and Garrett (2001) found evidence that the mismatch effect was larger when the head and local noun shared many semantic features (e.g., *The canoe by sailboat(s)*) than when they shared fewer semantic features (e.g., *The canoe by the cabin(s)*), which also suggests that similarity-based interference can influence agreement computation.
(2004b; Experiment 5) and Bock and Cutting (1992) can explain the different results observed between Experiment 1 and other studies that reported a clause-boundedness effect (Bock & Cutting, 1992; Solomon & Pearlmutter, 2004b).

Even though verb frequency and verb transitivity bias could not explain the conflicting results, this does not preclude the possibility that other lexical factors may be responsible for Experiment 1’s findings. For example, the mere presence of a semantically-rich content verb may be enough to reduce agreement error rates enough in RC conditions to create a clause-boundedness effect. According to slot-and-frame models of language production, sentence frames containing function words are planned and content words are retrieved from the mental lexicon and inserted into slots within these frames (e.g., Dell, 1986; Dell, Chang, & Griffin, 1999; Dell, Juliana, & Govindjee, 1993; Dell & O’Seaghdha, 1992). Because Experiment 1 used only the semantically light verb had, it may have been planned as part of a frame, while in previous studies the content verbs may have been retrieved from the lexicon separately. This difference may have either led to differences in planning time of the head and local noun or the complexity of the planning process in the relevant RC conditions, and could explain the differing results in Experiment 1 and previous studies (Bock & Cutting, 1992; Solomon & Pearlmutter, 2004b).

Experiments 2 and 3 suggest that imageability of the subject noun phrase is an important factor affecting agreement computation. Errors were only reliably elicited in Experiments 2 and 3 when the preambles were less imageable and referred to fairly abstract entities. Increased imageability has been shown to increase the size of the distributivity effect in NP PP preambles (Eberhard, 1999); however, no study to date has examined the effect of imageability on non-distributive complex subject NPs with RC modifiers. Given the result patterns from Experiments 2 and 3, it will be important to examine the effect of imageability in other constructions (see Chapter 4).

6Vigliocco and Hartsuiker (2002; fn. 11) suggest that agreement errors are more common when preambles are less imageable; but, it is unclear what evidence they have in mind to support this claim.
Considering the findings of Experiment 1, it is clear that structure does not entirely constrain agreement computation as agreement errors were equally likely when local nouns appeared in RC modifiers as when they appeared in PP modifiers. Experiments 2 and 3 provided a first attempt at determining which lexical properties of embedded verbs in RC modifiers may affect agreement computation, and how these lexical properties maybe be responsible for low error rates in studies that examined preambles with syntactically complex structures (e.g., Bock & Cutting, 1992; Hartsuiker et al., 2001; Solomon & Pearlmutter, 2004b). Variation in verb frequency and transitivity bias cannot explain the clause-boundedness effect observed in previous studies, but it seems clear that properties that are thought to affect processing occurring concurrent or prior to grammatical encoding (e.g., imageability, subject selection) have the best chance of affecting agreement computation. Thus, one possibility that deserves further research is that properties of lemmas may be more likely to influence agreement computation because lemma access is a process that occurs as part of grammatical encoding\(^7\), while properties of lexemes will not readily affect agreement computation because lexeme selection occurs as part of phonological encoding (Bock & Levelt, 1994).

Experiments 2 and 3 showed an effect of imageability on mismatch effects for non-distributive NP RC preambles containing content verbs. There has been no systematic examination of the effects of structure and imageability during agreement computation, thus it is important to understand how imageability differences may have contributed to apparent structural effects previously reported (Bock & Cutting, 1992; Solomon & Pearlmutter, 2004b), as imageability was not experimentally controlled in these studies. Experiments 4 and 5 examine how imageability affects agreement computation across a wide range of stimuli with differing structural and lexical properties.

\(^7\)Experiment 3 showed that verb bias, a property of lemmas, did not reliably affect agreement computation, though there was a trend toward RCs containing verbs with high SC-biases being more likely to elicit errors.
The effect of imageability on agreement computation

There are accounts of grammatical agreement from the linguistic and psycholinguistic literature that suggest that grammatical features such as person, number, and gender are copied or passed from the source to the target of agreement (e.g., Chomsky, 1965; Franck et al., 2006, 2010, 2002; Kempen & Hoenkamp, 1987). These accounts assume that agreement is largely a grammatical process that is constrained and controlled by aspects of the syntactic structure of the utterance. However, there are linguistic accounts of agreement that suggest that speakers often use conceptual information during the agreement process, even when the conceptual information is in conflict with grammatical information (Pollard & Sag, 1988; Quirk, Greenbaum, Leech, & Svartvik, 1972). Additionally, there is accumulating evidence that speakers often use conceptual information when they compute agreement during real-time language production (for evidence from gender agreement see Vigliocco & Franck, 1999, 2001; for evidence from number agreement see Hartsuiker & Barkhuysen, 2006; Hartsuiker et al., 1999; Haskell & MacDonald, 2003; Humphreys & Bock, 2005; Vigliocco, Butterworth, & Garrett, 1996; Vigliocco et al., 1995; Vigliocco & Hartsuiker, 2002).

Many languages have a grammatical rule specifying that the number-marking of the head of the subject NP and the verb must agree (Bock et al., 2006; Hartsuiker & Barkhuysen, 2006; Vigliocco, Hartsuiker et al., 1996); however, there has been conflicting evidence as to whether conceptual information regularly plays a role in subject-verb number agreement computation during language production. Speakers of Dutch, French, Italian, and Spanish showed a larger mismatch effect for subject NPs that had conflicting grammatical and conceptual number because of a distributive reading (e.g., the label on the bottles) than for subject NPs with singular grammatical and conceptual number (e.g., the bridge to the
islands; Vigliocco, Butterworth, & Garrett, 1996; Vigliocco, Hartsuiker, et al., 1996). In early studies examining the effect of distributivity on English subject-verb number agreement production, comparable effects were not observed, as English speakers did not show reliable distributivity effects (Bock & Miller, 1991; Vigliocco, Butterworth, & Garrett, 1996; Vigliocco, Hartsuiker, et al., 1996). Vigliocco, Butterworth, and Garrett (1996) suggested that this cross-linguistic variability in the influence of conceptual number on subject-verb agreement could be explained by how often a language explicitly specifies agreement features of the subject on its verbs. Romance languages (e.g., Spanish, French, Italian) almost always unambiguously mark agreement features of subjects on their verbs, while this is much less common in English. They suggested that because the majority of verb forms in English do not overtly specify number (exceptions being to be, and third-person present tense forms of regular verbs), English relies solely on grammatical number to number mark verbs, while in languages with richer verbal morphology, conceptual number is also considered during the number marking of the verb (cf. Berg, 1998; for additional discussion, see Eberhard, 1999).

To provide a unified account of the influence of conceptual number on subject-verb agreement, Eberhard (1999) examined whether the degree of imageability of a subject NP (i.e., concreteness) determined how available the conceptual number of the subject NP was during agreement computation. Using the test case of distributivity in English number agreement production, Eberhard found that highly imageable distributive referents, like (9a), showed larger mismatch effects than less imageable distributive referents, like (9b). In addition, imageability rating studies suggested that cross-linguistic differences in conceptual number agreement previously observed (e.g., Bock & Miller, 1991; Vigliocco, Butterworth, & Garrett, 1996; Vigliocco, Hartsuiker, et al., 1996) could be explained by differences in imageability of the distributive preambles across studies (cf. Humphreys & Bock, 2005; Vigliocco & Hartsuiker, 2002). Thus, Eberhard’s account suggests that when the conceptual number is readily available during agreement computation, it can exert an influence on agreement production (cf. Hartsuiker et al., 1999).
Currently the Marking and Morphing (M&M) model of agreement computation does not have a parameter for imageability, so it cannot directly model Eberhard’s results (Eberhard et al., 2005). There are at least two ways to account for Eberhard’s imageability effect in the M&M framework. Eberhard et al. suggest that “less imageable referents have less discernible internal boundaries, [and] they may tend to be construed as units even when they are potentially ambiguous in notional number” (p. 537). Thus, less imageable preambles would be associated with lower error rates. An alternative possibility is that the degree of imageability determines how much of an influence the marking process has on the subject NP’s singular and plural (SAP) value. If increased imageability makes the conceptual number of the subject NP more available during agreement computation, verbs should be more likely to agree with the conceptual number of highly imageable subject NPs than less imageable subject NPs (see Eberhard, 1999). Thus, for subject NPs that are grammatically and conceptually singular, increased imageability should reduce the overall number of agreement errors produced.

Imageability has been shown to affect language processing and memory processes in a variety of ways. Imageable sentences and words are recalled more accurately than less imageable ones (e.g., Begg, 1972; Begg & Pavio, 1969; Groot & Keijzer, 2000; Nelson & Schreiber, 1992; Pavio, 1965; Sadoski, Goetz, & Fritz, 1993) and concrete nouns are more likely to appear as sentence subjects (Bock & Warren, 1985; Christianson & Ferreira, 2005; McDonald, Bock, & Kelly, 1993). Effects of imageability on production are thought to arise due to better access to the conceptual representation of concrete referents during message-level planning and lemma access (Bock & Levelt, 1994; Bock & Warren, 1985; Levelt, 1989). Recent accounts of agreement computation have suggested that agreement production is a process that is inherently tied to memory encoding and retrieval processes (e.g., Badecker

---

1The high imageability and low imageability preambles were given the same SAP values for marking in the implemented model.
& Kuminiak, 2007; Badecker & Lewis, 2007; Gillespie & Pearlmutter, 2011; Hartsuiker & Barkhuysen, 2006; Lewis & Badecker, 2010; Solomon & Pearlmutter, 2004b); thus, imageability may affect number agreement computation even in cases in which grammatical and conceptual number do not conflict. For example, in Experiments 2 and 3 (Chapter 3), significant mismatch effects were only produced when non-distributive NP RC preambles were low imageability. These findings were unexpected, but can possibly be explained by Eberhard’s (1999) hypothesis that verbs following highly imageable subject NPs are more likely to receive a number marking consistent with the conceptual number of the NP than verbs following less imageable subject NPs, due to increased accessibility to the conceptual number for highly imageable subject NPs.

Experiments 4 and 5 were designed to directly test the effect of imageability on the production of subject-verb agreement errors. Experiment 4 tested whether imageability and structure interact during agreement computation by examining the clause-boundedness effect in highly imageable and less imageable subject NPs with prepositional phrase (PP) and relative clause (RC) modifiers. The meta-analyses in Experiment 5 examined the effects of measures of imageability on agreement error production in Experiments 1–4 and the PP and RC versions of Solomon and Pearlmutter’s (2004b) Experiment 5, while controlling for other factors known or hypothesized to affect error rates (e.g., semantic integration and clausal structure, Bock & Cutting, 1992; Eberhard et al., 2005; Gillespie & Pearlmutter, 2011; Solomon & Pearlmutter, 2004b).

1. Experiment 4

In subject-verb agreement error elicitation studies using preambles with PP modifiers, the mismatch effect is usually significant, even when the preambles generally refer to imageable, concrete subjects (e.g., Solomon & Pearlmutter, 2004b). However, in Experiments 2 and 3 (which only included preambles with RC modifiers), the mismatch effect was only significant when the preambles referred to low imageability, abstract subjects. Thus, it is possible
that imageability plays a different role in agreement computation when subject NPs have phrasal modifiers than when they have clausal modifiers. This difference could have led to the conflicting results obtained in Experiment 1 versus the studies that observed the clause-boundedness effect, as the preambles in Solomon and Pearlmutter (2004b) and Bock and Cutting (1992) were less imageable than the preambles in Experiment 1.

In Experiment 1, all of the preambles were highly imageable, and no clause-boundedness effect was observed. In contrast, the preambles in Solomon and Pearlmutter (2004b) and Bock and Cutting (1992) were fairly unimageable and produced a clause-boundedness effect. M&M stipulates that the morphing process relies on the structural relation between the head and local noun of a subject NP, with plural local nouns situated deeper in the syntactic tree being less likely to influence the subject NP node’s SAP value (Eberhard et al.). Gillespie and Pearlmutter (2011) and Experiment 1 provided evidence against strong structural constraints on agreement computation; however, the results of these experiments do not rule out the possibility that morphing has a structural component that only factors into agreement computation when the influence of the marking SAP value is weak due to decreased conceptual number accessibility (i.e., low imageability).

To test the influence of imageability and structure on agreement error production, Experiment 4 used non-distributive (i.e., grammatically singular and conceptually singular) subject NP preambles with PP and RC modifiers matched in meaning (see Experiment 1). Half of the preambles were highly imageable and the other half were less imageable. No study to date has examined the combined effects of imageability and structure on agreement computation. If the only effect increased imageability has on agreement computation is decreasing the overall error rate due to increased access to the conceptual number, then a main effect of imageability should be observed (see Vigliocco & Hartsuiker, 2002). If structural effects on agreement computation are only seen when access to the conceptual number of the subject

---

2 An imageability rating study conducted for Experiment 5 (described below), confirmed that there was a significant difference in overall imageability between the stimuli in these studies. The stimuli in Bock and Cutting (1992) were less imageable than the stimuli in Experiment 1 ($t = -4.20, p < .001$), as were the stimuli in Solomon and Pearlmutter (2004b) ($t = -12.05, p < .001$).
NP is weak, there should be an interaction of imageability, local noun number, and structure: The mismatch effect for highly imageable items should be equal across structure, as access to the conceptual number should be strong (replicating the findings of Experiment 1), while low imageability items should show different mismatch effect sizes across structures, with PPs producing a larger mismatch effect than RCs (replicating the clause-boundedness findings observed in Bock & Cutting, 1992; Solomon & Pearlmutter, 2004b).

1.0.4. Method

Participants. Ninety-five Northeastern University undergraduates participated in the experiment for course credit. Two participants were excluded because they were non-native speakers and two participants were excluded for reporting a language disability. All remaining participants (N = 91) were native speakers of American English.

Materials and design. Thirty-two stimulus sets like that shown in Table 4.1 were constructed. Each began with a head NP (e.g., The sweater, The report) followed by a modifier containing a local noun (e.g., pocket(s), mistake(s)). The head noun was always singular, and the four different versions of an item were created by varying modifier type and local noun number. The modifier was either a PP or an RC and described an attribute of the head noun. PP modifiers began with the preposition with and were followed by a local NP consisting of a determiner, adjective, and noun. RC modifiers began with the complementizer that and the verb had, followed by the same local NP. As a result, the RCs were always exactly one word (or syllable) longer than the corresponding PPs. Half of the items were highly imageable and half of the items were less imageable. None of the preambles had a distributive reading. The complete list of critical stimuli is shown in Appendix K.

In addition to the critical items, 88 fillers were included. Thirty-six of the fillers had structures like the critical items but had plural heads. The rest had a variety of structures varying in head noun number and were similar in length and complexity to the critical items.
Table 4.1. Experiment 4 Stimuli and Semantic Integration Ratings by Condition

<table>
<thead>
<tr>
<th>Imageability</th>
<th>Modifier</th>
<th>Noun Number</th>
<th>Example</th>
<th>Semantic Integration</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>PP</td>
<td>SP</td>
<td>The sweater with the yellow pockets</td>
<td>5.21 (1.83)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SS</td>
<td>The sweater with the yellow pocket</td>
<td>5.11 (1.88)</td>
</tr>
<tr>
<td></td>
<td>RC</td>
<td>SP</td>
<td>The sweater that had the yellow pockets</td>
<td>5.33 (1.71)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SS</td>
<td>The sweater that had the yellow pocket</td>
<td>4.93 (1.92)</td>
</tr>
<tr>
<td>Low</td>
<td>PP</td>
<td>SP</td>
<td>The report with the silly mistakes</td>
<td>4.13 (1.85)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SS</td>
<td>The report with the silly mistake</td>
<td>3.98 (1.98)</td>
</tr>
<tr>
<td></td>
<td>RC</td>
<td>SP</td>
<td>The report that had the silly mistakes</td>
<td>4.43 (1.84)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SS</td>
<td>The report that had the silly mistake</td>
<td>4.25 (1.86)</td>
</tr>
</tbody>
</table>

Note. The semantic integration rating scale was 1 (loosely linked) to 7 (tightly linked); standard deviations are in parentheses. PP = prepositional phrase; RC = relative clause; SP = singular head, plural local noun; SS = singular head, singular local noun.

The critical items and fillers were combined in four counterbalanced lists, each containing all fillers and exactly one version of each of the critical items. Each list was seen by 22–24 participants.

Stimulus norming. The 32 critical stimuli were normed for semantic integration by 78 participants in the survey described in Experiment 2a. Table 4.1 shows the mean integration ratings and standard deviations by condition for the critical stimuli. A linear mixed-effect regression (Baayen et al., 2008) on these data with imageability, modifier, local noun number, and their interactions as fixed effect factors, and participant and item intercepts as random factors, revealed that high imageability items were rated as more semantically integrated than low imageability items ($t = 4.64, p < .001$)\(^3\). Thus, the semantic integration rating pattern predicts that agreement errors would be more likely in the high imageability conditions than the low imageability conditions, opposite the predictions derived from the apparent effect of imageability on error rates observed in Experiments 2 and 3. We will return to these

\(^3\)See Part 4 for discussion of how imageability and semantic integration are related.
unexpected results below, with the discussion of the results of the production experiment. There were no other significant main effects nor interactions (all $|t| < 1.52$, $p > .11$).

**Apparatus, scoring, and procedure.** The apparatus, scoring, and procedure were identical to Experiment 2a.

### 1.0.5. Results

Table 4.2 shows the counts of each response type by imageability, structure, and noun number, with the number of responses containing a disfluency in parentheses. Separate analyses were performed for error rates (the proportion of error responses out of error plus correct responses), uninflected rates (the proportion of uninflected responses out of total scorable responses), and miscellaneous rates (the proportion of miscellaneous responses out of total scorable responses). The reported error and uninflected analyses included disfluencies, and unless otherwise noted, the patterns were identical if disfluency cases were excluded.

The data were analyzed using empirical logit weighted linear regression, aggregating separately over participants and items. By-participant and by-item weighted linear regressions on transformed error, miscellaneous, and uninflected rates were performed; with noun number, imageability, modifier, and their interactions as sum-coded fixed effect factors ($t$-tests of parameter estimates are identified as $t_1$ for the by-participant analysis and as $t_2$ for the by-item analysis).

We also computed corresponding ANOVAs on arcsine-transformed proportions (J. Cohen & Cohen, 1983), including 73 participants (18 were excluded because they were missing data in one or more cells). ANOVA result patterns were identical to those in the regression analyses unless otherwise noted; however, because approximately 20% of participants had to be excluded to compute these ANOVAs, it is unclear to what extent the comparisons of the statistical patterns from the ANOVAs and regression analyses will be meaningful.
Table 4.2. Experiment 4 Response Counts by Task and Condition

<table>
<thead>
<tr>
<th>Imageability</th>
<th>Modifier</th>
<th>Noun Number</th>
<th>Error</th>
<th>Correct</th>
<th>Uninflected</th>
<th>Misc</th>
<th>No Resp</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>PP</td>
<td>SP</td>
<td>22 (3)</td>
<td>253 (17)</td>
<td>60 (8)</td>
<td>29</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SS</td>
<td>0 (0 )</td>
<td>274 (20)</td>
<td>57 (5)</td>
<td>30</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>RC</td>
<td>SP</td>
<td>16 (3)</td>
<td>255 (15)</td>
<td>53 (8)</td>
<td>37</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SS</td>
<td>0 (0 )</td>
<td>273 (24)</td>
<td>52 (4)</td>
<td>37</td>
<td>2</td>
</tr>
<tr>
<td>Low</td>
<td>PP</td>
<td>SP</td>
<td>14 (3)</td>
<td>213 (26)</td>
<td>88 (16)</td>
<td>47</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SS</td>
<td>0 (0 )</td>
<td>248 (32)</td>
<td>78 (20)</td>
<td>37</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>RC</td>
<td>SP</td>
<td>27 (8)</td>
<td>217 (30)</td>
<td>55 (9)</td>
<td>63</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SS</td>
<td>4 (0 )</td>
<td>237 (32)</td>
<td>68 (17)</td>
<td>53</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>83 (17)</td>
<td>1970 (196)</td>
<td>511 (87)</td>
<td>333</td>
<td>15</td>
</tr>
</tbody>
</table>

Note. Disfluency counts are in parentheses. PP = prepositional phrase; RC = relative clause; SP = singular head, plural local noun; SS = singular head, singular local; Misc = Miscellaneous; No Resp = No Response.

Agreement errors. Figure 4.1 shows untransformed mean error rates by imageability and structure condition. Only versions with singular heads and plural local nouns are shown as the error rate for singular head and singular local noun versions was 0% in three of the four conditions (see Table 4.2). Table 4.3 shows the weighted linear regression effect estimates. Errors were more common when the local noun was plural than when it was singular. Errors were less likely when the preamble was high imageability than when it was low imageability. This effect was only marginal in the analysis by items when disfluencies were included, but when disfluencies were excluded it reached significance ($t_1 = 2.31, p < .001; t_2 = 1.36, p < .05$). There was a significant interaction of imageability and modifier with a larger difference in errors between RCs and PPs for low imageability preambles than high imageability ones. Paired tests showed that imageable PPs and RCs produced equivalent error rates ($t_1 = -0.64, p = .41; t_2 = -0.78, p = .40$), while errors were more likely for low imageability RCs than low imageability PPs ($t_1 = 1.35, p = .06; t_2 = 2.33, p < .05$). In addition, errors were less likely when RCs were high imageability than when they were low
Figure 4.1. Untransformed agreement error rates for plural local noun conditions as a function of imageability and modifier (PP = Prepositional phrase, RC = Relative clause) for Experiment 4. Error bars indicate ±1 SEM, computed by items.

imageability ($t_1 = -2.19, p < .01; t_2 = -1.92, p < .05$), but errors were equally likely when PPs were high imageability and low imageability ($t_1 = -0.01, p = .90; t_2 = -0.03, p = .51$). When disfluencies were excluded, the interaction of modifier and imageability did not reach significance ($t_1 = -1.12, p = .11; t_2 = -1.88, p = .06$). Critically, there was no interaction of imageability, modifier, and noun number. No other main effects nor interactions were significant (all $|t|s < 1.17, ps > .24$).
Table 4.3. Experiment 4 Agreement Error Rate Results (Weighted Empirical Logit Linear Regression)

<table>
<thead>
<tr>
<th>Effect</th>
<th>By Participants</th>
<th>By Items</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\beta$</td>
<td>$SE$</td>
</tr>
<tr>
<td>Imageability (imageable)</td>
<td>-0.11</td>
<td>.07</td>
</tr>
<tr>
<td>Modifier (RC)</td>
<td>0.04</td>
<td>.07</td>
</tr>
<tr>
<td>Noun number (SP)</td>
<td>0.42</td>
<td>.07</td>
</tr>
<tr>
<td>Imageability $\times$ Modifier</td>
<td>-0.20</td>
<td>.13</td>
</tr>
<tr>
<td>Imageability $\times$ Noun number</td>
<td>0.04</td>
<td>.13</td>
</tr>
<tr>
<td>Modifier $\times$ Noun number</td>
<td>-0.04</td>
<td>.13</td>
</tr>
<tr>
<td>Imageability $\times$ Modifier $\times$ Noun number</td>
<td>-0.15</td>
<td>.27</td>
</tr>
</tbody>
</table>

Note. The level shown in parentheses for each variable was sum-coded +0.5 and the other level −0.5, so $\beta$s estimate the difference between the two levels of the variable in log-odds space. RC = relative clause; SP = singular head, plural local noun.

***$p < .001$. *$p < .05$. †$p < .10$.

Results were similar for the ANOVA analyses; however, the imageability main effect did not reach significance ($F_1(1,72) = 1.16, p = .29; F_2(1,30) = 1.01, p = .30$). The main effect of modifier was marginal, with more errors produced in the RC conditions ($F_1(1,72) = 3.58, p = .06; F_2(1,30) = 3.20, p = .08$). There was a significant three-way interaction of imageability, modifier, and local noun number that was only significant by subjects ($F_1(1,72) = 4.77, p < .05; F_2(1,30) = 1.95, p = .17$). The mismatch effect was equal for PPs and RCs for the imageable preambles, while the RC mismatch effect was larger than the PP mismatch effect for the low imageability preambles. No other patterns differed from the regression analyses.

Uninflected rates. Uninflected responses were more likely for low imageability than high imageability preambles ($t_1 = -3.32, p < .001; t_2 = -2.41, p = .14$), and for PPs than RCs ($t_1 = -1.89, p < .05; t_2 = -3.47, p = .07$), though these effects only reached significance in the analysis by subjects. When disfluencies were excluded, the effect of modifier was marginal by subjects ($t_1 = -1.45, p = .08; t_2 = -2.40, p = .17$).
The ANOVA analyses were identical, except the effect of modifier reached significance by items (including disfluencies: $F_1(1, 72) = 8.23, p < .01; F_2(1, 30) = 7.13, p < .01$; excluding disfluencies: $F_1(1, 72) = 3.19, p = .08; F_2(1, 30) = 4.80, p < .05$).

**Miscellaneous rates.** Miscellaneous responses were more likely for low imageability than high imageability preambles ($t_1 = -3.77, p < .001; t_2 = 4.23, p < .01$), and for RCs than PPs ($t_1 = 2.34, p < .01; t_2 = 3.46, p = .05$). There were no other main effects nor interactions (all $|t|$s $< 1.73, ps > .14$).

In the ANOVA analyses, the effect of imageability was only significant by subjects ($F_1(1, 72) = 10.16, p < .001; F_2(1, 30) = 1.97, p = .17$), and the modifier effect was marginal by subjects and items ($F_1(1, 72) = 2.76, p = .10; F_2(1, 30) = 2.95, p = .10$). No other patterns differed from the regression analyses.

**1.0.6. Discussion**

There was a large mismatch effect (e.g., Bock & Miller, 1991), indicating that speakers reliably produced agreement errors. Errors were more likely for low imageability than high imageability preambles, supporting the hypothesis that increased imageability increases access to the conceptual number of the subject NP during agreement computation. This pattern is opposite that predicted by the semantic integration differences observed, suggesting that, if anything, the effect of imageability on agreement error rates could actually be larger than the effect observed in Experiment 4. Importantly, Experiment 4 replicated the findings of Experiment 1, as there was no effect of structure on the size of mismatch effects. To the extent that any effect of structure was observed, it was that low imageability RCs produced more errors than low imageability PPs; this effect is in the opposite direction of that predicted by any structural account. These results support the conclusions in Experiment 1, and suggest that agreement computation is not strongly constrained by syntactic
structure (cf. Bock & Cutting, 1992; Eberhard et al., 2005; Franck et al., 2002; Solomon & Pearlmutter, 2004b; Vigliocco & Hartsuiker, 2002).

There was an interaction of imageability and modifier in the error analyses, but the pattern was in an unexpected direction. The results for the imageable preambles in Experiment 4 were nearly identical to those in Experiment 1, suggesting that there is no effect of structure on agreement computation. However, the low imageability RCs produced more errors than the low imageability PPs, which is opposite the predictions that any structural account of agreement computation would make (Bock & Cutting, 1992; Eberhard et al., 2005; Franck et al., 2002; Vigliocco & Hartsuiker, 2002). One reason this effect may have emerged is because the low imageability PP condition elicited the most uninflected responses of any condition in the study (see Table 4.2), which may have made the error rate estimate less reliable due to fewer inflected responses (i.e., corrects, errors) in this condition. It would be interesting to use these preambles in an experiment using a paradigm that encourages speakers to use inflected verbs (e.g., a completion task requiring speakers to use a target adjective in their completions; Barker et al., 2001; Haskell & MacDonald, 2003; Meyer & Bock, 1999; Vigliocco & Nicol, 1998) to determine if the imageability by modifier interaction is replicated.

Miscellaneous errors were more likely when preambles were low imageability than when they were high imageability, as were agreement errors. These patterns can be taken to reflect the difficulty of creating sentences with abstract subject NPs. This difficulty may arise due to the relative inaccessibility of the conceptual number (or other conceptual properties) of the subject NP when subject NPs are less imageable. Miscellaneous errors were more common for RCs than PPs; however, this main effect of modifier was not observed in the agreement error results. In addition, miscellaneous errors were equally likely for preambles with singular or plural local nouns. These results suggest that the agreement error patterns observed were not solely due to difficulty in processing the preambles.

One aspect of the error results that complicates the emerging evidence that imageability affects agreement computation is that there was a significant mismatch effect observed
for imageable RCs in Experiment 4, which was not the case for Experiments 2a and 3a in Chapter 3. This difference cannot be due to differences in overall imageability across experiments, as the imageable preambles in Experiment 4 were significantly more imageable than the preambles in Experiments 2a and 3a. Thus, regardless of imageability condition, a mismatch effect was observed for both structure conditions in Experiment 4, raising the possibility that agreement errors are particularly likely when (1) the local noun and head noun are linked with RCs containing semantically-light verbs (e.g., *had*), or (2) when the head and local noun are in an attribute or possession relationship. The PP versions of stimuli in previous studies that observed the clause-boundedness effect (Bock & Cutting, 1992; Solomon & Pearlmutter, 2004b) often created an attribute or possession relationship between the head and local noun (e.g., The laboratory with the analog computer(s); The speculation of the secret and elite committee(s)), while very few of the corresponding RC versions did so (e.g., The laboratory that invented the computer(s); The speculation that disgusted the elite committee). In contrast, the PP and RC versions of the stimuli in Experiment 4 and Experiment 1 always created an attribute relationship between the local and head noun, and these studies did not show a reliable clause-boundedness effect.

The possibility that agreement errors are more likely to occur when the head and local noun are in an attribute-like relationship could explain many of the error patterns observed in Gillespie and Pearlmutter (2011) and Solomon and Pearlmutter (2004b). In those studies local nouns that were attributes of the head noun (e.g., The sweater with the tiny hole(s)) produced more interference than local nouns that were in other types of relationships with the head noun (e.g., The sweater with the clean skirt(s)). Future work will be necessary to determine whether attribute and possession relations between the head and local noun in

---

4 Using overall subject NP imageability ratings obtained in a survey described in Experiment 5 (below), a linear mixed effect model predicting imageability from experiment (a three-level treatment-coded categorical predictor) confirmed that overall imageability ratings were higher for the imageable items in Experiment 4 than the preambles in Experiment 2a (*t* = −2.77, *p* < .01) and the preambles in Experiment 3a (*t* = −3.08, *p* < .001).

5 Note that in the example *The speculation of the committee*, it is the local noun (*committee*) that possesses the head noun (*speculation*). The majority of the head and local nouns in the PP versions in Solomon and Pearlmutter’s (2004b) Experiment 5 were in this type of relationship with each other.
an RC modifier always increases error rates and eliminates the clause-boundedness effect\(^6\). For example, it will be important to determine if RCs containing lexical verbs that create an attribute relationship (e.g., *the book that contained the torn pages*) produce similar error rates to PPs reflecting the same attribute relation (e.g., *the book with the torn pages*), as it could simply be the presence of a lexical verb (i.e., a verb with more semantic content) that reduces agreement error rates (for more discussion of this possibility, see Chapter 3 and Chapter 7).

2. Experiment 5

The results of Experiments 2–4 showed that imageability appears to have an effect on agreement computation, with agreement errors being more likely when preambles were less imageable. The effect of imageability has been examined in a number of agreement elicitation studies (e.g., Bock & Eberhard, 1993; Eberhard, 1999; Hartsuiker et al., 1999; Humphreys & Bock, 2005). In Experiments 2–4, only overall imageability was manipulated, but the effect of imageability of nouns within complex subject NPs has also been examined (e.g., Bock & Miller, 1991). Experiment 5 is a meta-analysis that examined the effect of different measures of imageability on agreement computation. Table 4.4 shows imageability ratings for mismatch versions (i.e., singular head, plural local noun) of preambles by condition and experiment for stimuli included in these analyses.

Eberhard (1999) found that the mismatch effect was larger for more imageable distributive preambles (see 9), and suggested that this effect arose due to eased accessibility of the subject NP’s conceptual number when the referent was more imageable. According to Eberhard’s theory, imageability should provide a “protective” effect against agreement error production in the studies included in these meta-analyses because all of the preambles were grammatically and conceptually singular: Increased imageability should ease access to

---

\(^6\)The PP condition is generally assumed to be a baseline to which RC error rates are compared, because significant mismatch effects are almost always observed for PP conditions.
Table 4.4. Imageability Ratings by Condition for Stimuli Included in Experiment 5

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Link Type</th>
<th>Verb Type</th>
<th>Overall</th>
<th>Head</th>
<th>Verb</th>
<th>Local</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Struc</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experiment 1</td>
<td>PP A</td>
<td>A</td>
<td>5.17 (0.85)</td>
<td>5.76 (0.96)</td>
<td>4.78 (0.96)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RC A F</td>
<td></td>
<td>5.38 (0.80)</td>
<td>5.76 (0.96)</td>
<td>2.70 (0.00)</td>
<td>4.78 (0.96)</td>
</tr>
<tr>
<td>Experiment 2a</td>
<td>RC O C</td>
<td></td>
<td>4.82 (1.23)</td>
<td>4.52 (0.82)</td>
<td>4.98 (1.32)</td>
<td>4.53 (0.74)</td>
</tr>
<tr>
<td>Experiment 2b</td>
<td>RC O C</td>
<td></td>
<td>2.97 (0.97)</td>
<td>4.63 (0.69)</td>
<td>3.38 (1.02)</td>
<td>4.62 (0.80)</td>
</tr>
<tr>
<td>Experiment 3a</td>
<td>RC O C</td>
<td></td>
<td>4.78 (0.85)</td>
<td>4.62 (0.91)</td>
<td>4.42 (1.51)</td>
<td>4.38 (0.75)</td>
</tr>
<tr>
<td>Experiment 3b</td>
<td>RC O C</td>
<td></td>
<td>2.57 (0.74)</td>
<td>3.31 (1.37)</td>
<td>3.22 (1.32)</td>
<td>3.36 (1.32)</td>
</tr>
<tr>
<td>Experiment 4</td>
<td>PP A</td>
<td>A</td>
<td>5.83 (0.74)</td>
<td>5.99 (0.69)</td>
<td>5.55 (0.88)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RC A F</td>
<td></td>
<td>5.64 (0.92)</td>
<td>5.99 (0.69)</td>
<td>2.70 (0.00)</td>
<td>5.55 (0.88)</td>
</tr>
<tr>
<td></td>
<td>PP A</td>
<td></td>
<td>2.95 (0.93)</td>
<td>3.20 (1.48)</td>
<td>2.68 (1.10)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RC A F</td>
<td></td>
<td>2.84 (0.56)</td>
<td>3.20 (1.48)</td>
<td>2.70 (0.00)</td>
<td>2.68 (1.10)</td>
</tr>
<tr>
<td>S&amp;P (2004)</td>
<td>PP O</td>
<td></td>
<td>2.86 (0.80)</td>
<td>4.59 (0.92)</td>
<td>4.55 (0.80)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RC O C</td>
<td></td>
<td>2.85 (0.68)</td>
<td>4.59 (0.92)</td>
<td>3.48 (1.17)</td>
<td>4.55 (0.80)</td>
</tr>
</tbody>
</table>

Note. Means presented are for singular head, plural local noun versions. Standard deviations are in parentheses. S&P = Solomon & Pearlmutter (2004b; Exp. 5); Struc = Structure; PP = prepositional phrase; RC = relative clause; A = attribute relation; O = other; F = function (i.e., had); C = content.

the conceptual number, which should result in correctly inflected verbs (i.e., singular). We, therefore, examined the effect of overall subject NP imageability on agreement error rates.

Bock and Miller (1991) examined whether the relative concreteness of the head and local noun could affect agreement computation. Specifically, they tested whether errors would be more likely when the local noun was more concrete than the head noun, because a relatively concrete local noun may have more influence in assigning the subject NP’s number marking than a relatively abstract head noun. Bock and Miller found a marginal effect of local-head noun relative concreteness, with preambles with concrete local nouns and abstract head nouns (e.g., The nomad of the mountain(s)) producing larger mismatch effects than preambles with abstract local nouns and concrete head nouns (The mountain of the
nomad(s)); however, this effect was not observed when the study was replicated to correct a counterbalancing error. To reexamine the effect that the nouns’ relative concreteness may have on agreement computation, a measure of local-head noun relative imageability was computed for the stimuli in the experiments listed above. If Bock and Miller’s hypothesis is correct, preambles with relatively more imageable local nouns than head nouns should elicit more agreement errors.

Preambles in the experiments included in these analyses not only varied in imageability, but in other properties as well. We included control predictors in all models for other factors that may affect agreement computation.

Nearly all theories of agreement computation contain a structural component (Eberhard et al., 2005; Vigliocco & Hartsuiker, 2002), and suggest local nouns that are structurally proximal to the head noun are more likely to interfere with agreement computation. Thus we included a predictor to control for the effect syntactic structure may have on agreement error rates. Experiment 1, Experiment 4, and Solomon and Pearlmutter (2004b; Experiment 5) used NP PP and NP RC preambles, while the rest of the experiments included in these analyses only used NP RC preambles.

Some of the studies included in these analyses tightly controlled some conceptual and lexical factors while others did not. Experiments 1 and 4 intentionally matched PP and RC versions on a variety of properties: In both studies the local noun was an attribute of the head noun in all versions, and both studies used NP RC preambles that linked the head and local noun with the semantically light verb *had*. In the remainder of the experiments, RCs contained lexical verbs and the local nouns were not generally attributes of their heads. Solomon and Pearlmutter’s (2004b) Experiment 5 had a structure manipulation and used RCs containing lexical verbs but no attempt was made to match meaning across RC and PP versions of stimulus items. As mentioned in the Discussion of Experiment 4, it is possible that agreement errors are particularly likely when a local noun is (1) an attribute of its head noun, or (2) is linked to the head noun with a semantically light, function-like word (e.g.,

97
had, with), because no structural effects nor a reliable imageability effect were observed in Experiments 1 and 4 where PPs and RCs were matched on these properties. To control for effects of the type of link created between the head and local noun, a link type factor was included in the model. Stimuli from Experiments 1 and 4 were coded as having attribute links (attribute) between the head and local noun, while all other studies were coded as having “other” links between the head and local noun. All RCs that were coded as having attribute links contained the function-like verb had, while all RCs that were coded as having “other” links contained content verbs.

Because previous studies have found that the degree of semantic integration of the head and local noun affects agreement error rates (Gillespie & Pearlmutter, 2011; Solomon & Pearlmutter, 2004b), semantic integration was included as a control predictor. In all studies included in the meta-analysis there was no direct manipulation of semantic integration, and in the majority of these studies semantic integration was matched across conditions, so we do not expect a large effect of semantic integration. However, semantic integration ratings did vary item-by-item, so it is possible that the more integrated items elicited more errors than the less integrated items.

Because of the exploratory nature of Experiments 2–4, imageability was manipulated based on intuitions. Thus the degree of imageability of the preambles likely varied across items and experiments. To determine how different measures of imageability predict agreement error patterns across a range of stimuli, we conducted surveys to obtain overall imageability ratings and individual content word imageability ratings for the stimuli used in Experiment 1, Experiment 2a, Experiment 2b, Experiment 3a, Experiment 3b, Experiment 4, and the PP and RC versions from Solomon and Pearlmutter (2004b; Experiment 5). The imageability rating surveys will be discussed in more detail below. We then conducted two sets of regression analyses predicting error rates from these measures of imageability while controlling for other factors known or hypothesized to affect agreement computation. The first set of analyses examined overall and local-head noun relative imageability across the
full dataset, and the second set of analyses examined overall and individual content word imageability of only the RC conditions.

2.0.7. Method

**Error rate data.** By-item error rates (including disfluencies) for plural local noun versions of each preamble from Experiment 1, Experiment 2a, Experiment 2b, Experiment 3a, Experiment 3b, Experiment 4, and the PP and RC versions from Solomon and Pearlmutter’s (2004b) Experiment 5 were obtained. The recall and no-recall versions of Experiment 1 were treated as separate experiments in all of the following analyses. Singular local noun versions uniformly produced extremely low error rates and thus were not included in the analyses.

**Full subject noun phrase survey.** Participants were instructed to rate how easy it was to create a mental image of the phrase using a 1 (difficult to image) to 7 (easy to image) scale by clicking a radio button on the screen indicating their choice. The instructions were based on the concreteness rating instructions in Pavio, Yuille, and Stephen (1968), and are presented in Appendix I. A quarter of the stimulus items from each experiment listed above were included in four separate versions of the survey (i.e., no item appeared in more than one survey), for a total of 58 items per version. In each version, there were four lists (each with three randomized presentation orders) and each list contained each item in exactly one of its four conditions. Four participants rated each randomized list, thus up to 16 ratings per condition of each item were obtained. Participants could participate in each of the four versions of the survey because each item only appeared in a single version of the survey, but were not allowed to submit data for more than one list per version. Eighty-two participants completed at least one version of the survey. After excluding data where participants did not provide a rating, each condition of each item received 12–14 ratings. Each version of the survey took approximately 6 minutes to complete and participants were paid $0.34 for each version they completed.
Content word surveys. All of the content words (nouns and verbs) from the studies listed above were rated in separate surveys with instructions and examples modified from Pavio et al. (1968). These instructions are presented in Appendix J. Participants were instructed to rate how easy it was to create a mental image of the phrase (i.e., the noun preceded by *the* or the verb preceded by *to*) using a 1 (difficult to image) to 7 (easy to image) scale by clicking a radio button on the screen indicating their choice.

All nouns rated were preceded by the determiner *the* and only the singular version of each noun was presented. The nouns from the studies listed above were rated in five separate versions of the noun survey with 92-93 items. Verbs were rated in the infinitival form (e.g., *to push*). The verbs from the studies listed above were rated in four separate versions of the verb survey with 89–90 items. In each version, there was a single list with five randomized presentation orders. Four participants rated each randomized list, thus up to 20 ratings per item could be obtained. In one version of the noun survey, the final two items of each list did not display properly, so only 16 ratings per word were obtained for those items. All words received between 16 and 20 imageability ratings.

Participants could participate in each of the seven versions of the surveys because each word only appeared in a single version of the survey, but were not allowed to submit data for more than one list per version. Each experiment took approximately 8 minutes to complete and participants were paid $0.75 for each version they completed.

2.0.8. Results and Discussion

Full dataset analysis. To examine the effect of imageability on agreement error rates two measures of imageability were included in a regression model. The full subject NP imageability ratings were included to determine if overall imageability affected agreement computation. Using the ratings from the noun surveys described above, head noun imageability was subtracted from local noun imageability for each preamble to obtain a measure of the local-head noun relative imageability (i.e., positive values indicated that the local noun
was more imageable than the head noun). Semantic integration ratings for the preambles from the survey described in Experiment 2a (Chapter 3) were used to control for any effects of semantic integration. Structure and link type were entered into the model as sum-coded binary categorical predictors.

Following Barr (2008), the error data were analyzed using empirical logit weighted linear regression, aggregating over items, resulting in 436 data points to be modeled. Overall imageability, relative noun imageability, structure (PP vs. RC), link type (attribute vs. other), semantic integration, and the two-way interactions of each imageability measure and structure and link type were included as fixed effect factors, and item and experiment intercepts were included as random factors. All continuous predictors were centered to reduce collinearity. After centering, integration and overall imageability were still moderately positively correlated \( (r = .4) \), thus the residuals of a model predicting overall imageability from semantic integration were included in the model as the overall imageability predictor. This residualization choice was conservative as the predictor of interest in this model was overall imageability. Thus, the estimate of the overall imageability predictor is a reflection of how much variance in error rates overall imageability predicts beyond what can be explained by semantic integration.

Table 4.5 shows the weighted linear regression effect estimates. Errors were more likely when preambles were less imageable and when preambles contained prepositional phrase (PP) modifiers. There was an interaction of overall imageability and link type. When the head and local noun were linked with function words creating an attribute relation (Experiments 1 and 4), the effect of overall imageability on agreement error production was smaller than the effect of overall imageability when the head and local noun were linked in other ways. No other main effects nor interactions were significant.

\[7\] Model comparison confirmed that a more complex random effect structure was not warranted, as no more complex model significantly improved model fit.

\[8\] This correlation is not entirely unexpected as the semantic integration rating instructions suggest forming a mental image of the phrase to determine how closely linked the parts of the utterance are. Thus, more imageable preambles are likely to elicit higher integration ratings as this strategy for rating integration would be more effective (for more details on this confound, see Part 4).
Table 4.5. Results of Weighted Empirical Logit Linear Regression Model Predicting Error Rates from Overall and Local-Head Noun Relative Imageability

<table>
<thead>
<tr>
<th>Effect</th>
<th>$\beta$</th>
<th>$SE$</th>
<th>$t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall imageability</td>
<td>−0.18</td>
<td>.05</td>
<td>−3.87***</td>
</tr>
<tr>
<td>Local-head imageability</td>
<td>−0.01</td>
<td>.04</td>
<td>−0.08</td>
</tr>
<tr>
<td>Structure (PP)</td>
<td>0.20</td>
<td>.09</td>
<td>2.15*</td>
</tr>
<tr>
<td>Link type (attribute)</td>
<td>0.10</td>
<td>.29</td>
<td>0.35</td>
</tr>
<tr>
<td>Semantic integration</td>
<td>−0.01</td>
<td>.08</td>
<td>−0.19</td>
</tr>
<tr>
<td>Overall imageability $\times$ Structure</td>
<td>−0.01</td>
<td>.07</td>
<td>−0.15</td>
</tr>
<tr>
<td>Overall imageability $\times$ Link type</td>
<td>−0.22</td>
<td>.09</td>
<td>−2.29*</td>
</tr>
<tr>
<td>Local-head imageability $\times$ Structure</td>
<td>−0.03</td>
<td>.07</td>
<td>−0.39</td>
</tr>
<tr>
<td>Local-head imageability $\times$ Link type</td>
<td>0.01</td>
<td>.09</td>
<td>0.22</td>
</tr>
</tbody>
</table>

Note. For categorical predictors, the level shown in parentheses was sum-coded +0.5 and the other level −0.5, so $\beta$s estimate the difference between the two levels of the variable in log-odds space. PP = Prepositional phrase.

***$p < .001$. *$p < .05$. †$p = .10$.

This analysis found a significant effect of overall imageability on error rates, with more imageable preambles eliciting fewer agreement errors than less imageable preambles. All the preambles included in this analysis were grammatically and conceptually singular, which indicates that speakers were more likely to produce agreement consistent with the subject NP’s conceptual number when it was more imageable. This finding is similar to that of Eberhard (1999), where highly imageable distributive subject NPs showed stronger distributivity effects than less imageable distributive subject NPs. Bock and Miller (1991) hypothesized that local nouns that are relatively more concrete than the head noun should have more influence on the number marking of the subject NP, and should be more likely to interfere with agreement computation. The relative imageability of the head and local noun did not prove to be a significant predictor of error rates in these studies, replicating the findings of Bock and Miller. Thus, while overall imageability affects agreement computation, the relative imageability of individual lexical items does not seem to play an independent role.
There was an interaction of imageability and link type, with imageability playing less of a role in agreement computation when the head and local noun were in an attribute relationship that was created with a semantically light, function-like linking word (Experiments 1 and 4). This pattern of results raises the possibility that imageability does not strictly constrain agreement computation when the head and local noun are linked with a function word (e.g., prepositions, the light verb *had*). As has been shown in numerous agreement error elicitation tasks, significant mismatch effects were observed regardless of preamble imageability when the head noun and local noun were linked with a preposition (e.g., Bock & Cutting, 1992; Bock & Miller, 1991; Eberhard, 1997, 1999; Solomon & Pearlmutter, 2004b).

Contrary to the findings from Experiments 1 and 4 and consistent with previous studies reporting a clause-boundedness effect (Bock & Cutting, 1992; Solomon & Pearlmutter, 2004b), there was a significant effect of structure, with PPs eliciting more errors than RCs. While this effect seems to support the role of structural constraints in agreement computation, it is likely that it only emerged because of the nature of the experiments and stimuli included in this analysis. The clause-boundedness effect was only observed in Solomon and Pearlmutter’s (Experiment 5), which was also the study with the highest overall error rates. Running an identical model to the one reported above but excluding data from Solomon and Pearlmutter’s (Experiment 5) yielded a significant effect of overall imageability (*t* = −3.21, *p* < .001), but the effect of structure was no longer significant (*t* = −0.73, *p* = .69), nor was the imageability by link type interaction (*t* = 2.53, *p* = .26). In fact, the structure effect was in the opposite numeric direction when Solomon and Pearlmutter’s (Experiment 5) stimuli were not included. Thus, it appears that the structure effect was driven by a single study that did not match PPs and RCs on a number of conceptual properties (e.g., link type, meaning).

Previous examinations of the semantic integration effect have shown no reliable correlation between an individual item’s integration rating and its error rate (for further discussion
see the meta-analysis in Solomon & Pearlmutter, 2004b). There were no direct manipulations of semantic integration in the studies included in this analysis, so it is not surprising that there was not a significant effect of semantic integration.

Taken together, the results of this analysis indicate that increased overall imageability of the subject NP affects agreement computation, with more imageable subject NPs producing fewer errors than less imageable subject NPs across a wide range of grammatically and conceptually singular subject NPs. This pattern indicates that speakers are more likely to produce verbs that agree with the conceptual number of the subject when the subject NP is more imageable. This conclusion is incompatible with Eberhard et al.’s (2005) hypothesis that less imageable subjects are likely to be conceived of as units and produce singular verbs. The relative imageability of the local and head noun did not affect agreement error rates, suggesting that relatively imageable local nouns do not influence the number marking of the subject NP more than relatively low imageability head nouns (cf. Bock & Miller, 1991). These findings suggest that imageability affects the marking process by modulating access to the conceptual number of the subject NP as a whole, rather than affecting the morphing process by requiring a measure of the imageability of individual lexical items (Eberhard et al.).

**RC dataset analyses.** In the experiments included in these analyses, preambles with RC modifiers varied in more ways than the preambles with PP modifiers. The RC conditions had an extra content word (i.e., the verb), and the head and local nouns were linked in a variety of ways because the embedded verb often introduced more semantic content to the relation than the prepositions did. In addition, Experiments 1–4 suggest that agreement error rates for subject NPs with RC modifiers may vary more than error rates for subject NPs with PP modifiers. For example, in Experiments 2a and 3a, agreement errors rarely occurred; however, in Experiments 2b and 3b, NP RC preambles that were similar in complexity to their high-imageability counterparts produced significant mismatch effects. Because of
the complexity and variety of RC modifiers in these experiments, additional analyses were conducted examining the effect of a number of measures of imageability on the RC conditions alone. The goals of the following two analyses were to examine how overall and individual word imageability affect agreement error rates in subject NPs with RC modifiers.

To confirm that the effect of overall imageability holds for RC conditions, the error data for NP RC preambles were analyzed using empirical logit weighted linear regression, aggregating over items. The model included a measure of overall imageability (described above), semantic integration (described in Chapter 3), link type (attribute vs. other), and the interaction of overall imageability and link type. The continuous predictors were centered to reduce collinearity. Link type was entered into the model as a sum-coded binary predictor. After centering, integration and overall imageability were still moderately positively correlated, thus the residuals of a model predicting overall imageability from semantic integration were included in the model as the overall imageability predictor (i.e., the same residualization technique as described above).

The results of the overall imageability model are shown in Table 4.6. Higher overall imageability was associated with significantly lower agreement error rates, indicating that overall imageability was a significant predictor of error rates for preambles with RC modifiers. No other main effects nor interactions reached significance.

While the nouns’ relative imageability did not have an effect on agreement error rates in the full dataset, it is still possible that the imageability of individual lexical items influences agreement computation, as local-head noun relative imageability and individual content word imageability are independent. Given the set of experiments in these analyses, it is likely that there was more variability in the individual content word imageability than there was in local-head noun relative imageability because the stimuli were written to be overall imageable or not, so head and local noun imageability co-varied to some extent ($r = .40$). Thus, an additional analysis was conducted examining whether the imageability of each individual content word in NP RC preambles influenced agreement error rates.
Table 4.6. Results of Weighted Empirical Logit Linear Regression Model Predicting Error Rates from Overall Imageability for Preambles with RC Modifiers

<table>
<thead>
<tr>
<th>Effect</th>
<th>$\beta$</th>
<th>SE</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall imageability</td>
<td>−0.15</td>
<td>.05</td>
<td>−2.98***</td>
</tr>
<tr>
<td>Semantic integration</td>
<td>−0.05</td>
<td>.06</td>
<td>−0.71</td>
</tr>
<tr>
<td>Link type (attribute)</td>
<td>0.29</td>
<td>.24</td>
<td>1.19</td>
</tr>
<tr>
<td>Overall imageability $\times$ Link type</td>
<td>−0.15</td>
<td>.11</td>
<td>−1.42</td>
</tr>
</tbody>
</table>

*Note.* The level shown in parentheses for link type was sum-coded +0.5 and the other level $−0.5$, so the link type $\beta$ estimates the difference between the two levels of the variable in log-odds space.

***$p < .001$.***

The model included a measure of each individual content word’s imageability (head noun, verb, local noun), semantic integration of each preamble, and link type of the preamble. The results of the individual word imageability model are shown in Table 4.7. Only increased verb imageability significantly reduced agreement error rates, although the same numerical pattern was observed for head noun imageability and local noun imageability. No other effects were significant.

The overall imageability of a subject NP is likely to be at least partially determined by the individual lexical items that comprise it; however, it is possible that overall imageability and lexical imageability play independent roles during agreement computation. Using model comparison techniques, the relative contributions of overall imageability and lexical imageability were examined.

To determine whether the full effect of imageability can be captured by measures of the imageability of individual content words within the preambles, a superset model containing all the predictors from the overall and individual word models was run and compared to the

---

9The interaction of link type and verb imageability could not be included in the model because there was no variability in the verb imageability ratings in the attribute conditions as the verb was always *had.*
Table 4.7. Results of Weighted Empirical Logit Linear Regression Model Predicting Error Rates from Content Word Imageability for Preambles with RC Modifiers

<table>
<thead>
<tr>
<th>Effect</th>
<th>$\beta$</th>
<th>$SE$</th>
<th>$t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head noun imageability</td>
<td>−0.08</td>
<td>0.05</td>
<td>−1.50</td>
</tr>
<tr>
<td>Verb imageability</td>
<td>−0.10</td>
<td>0.04</td>
<td>−2.40*</td>
</tr>
<tr>
<td>Local noun imageability</td>
<td>−0.05</td>
<td>0.05</td>
<td>−0.91</td>
</tr>
<tr>
<td>Semantic integration</td>
<td>−0.02</td>
<td>0.06</td>
<td>−0.26</td>
</tr>
<tr>
<td>Link type (attribute)</td>
<td>0.14</td>
<td>0.23</td>
<td>0.60</td>
</tr>
</tbody>
</table>

Note. The level shown in parentheses for link type was sum-coded +0.5 and the other level −0.5, so the link type $\beta$ estimates the difference between the two levels of the variable in log-odds space.

*p < .05.

individual word model\textsuperscript{10}. This model comparison revealed that the more complex superset model containing the overall imageability predictor (main effect of overall imageability and its interaction with link type) did not significantly increase model fit ($\chi^2(2) = 3.39, p = .18$). To determine if overall imageability captured all the variance that the imageability of the individual lexical items did, the superset model was compared to the overall imageability model. This model comparison also revealed that the more complex superset model did not significantly improve model fit ($\chi^2(3) = 5.07, p = .17$).

The fact that the imageability of all the individual lexical items had a similar effect on error rates and the overall imageability model and the individual content word imageability model were statistically indistinguishable suggests that overall imageability was the main factor that influenced agreement error rates for subject NPs with RC modifiers in these experiments. These results also support the general claim of M&M that only grammatical properties, and not conceptual properties, of the local nouns are likely to affect agreement computation (Eberhard et al., 2005; also see Bock et al., 2001, 2004).

\textsuperscript{10}Model comparison was conducted using the \textit{anova} function in the \textit{stats} package (v. 2.12.1) in R (Chambers & Hastie, 1992).
Because only the imageability of the verb was a significant predictor of agreement error rates in the individual word model, it is important to determine if verb imageability is a significant predictor of agreement error rates after accounting for overall imageability. Including the verb imageability predictor in the overall imageability model significantly improved model fit ($\chi^2(1) = 4.32, p < .05$). A model adding the head noun imageability predictor to the overall imageability model did not produce a significant increase in model fit, nor did a model adding the local noun imageability predictor ($\chi^2 s < 1$).

One reason that verb imageability may have affected agreement error rates beyond what could be explained by overall imageability is that the verb with the lowest imageability in the studies included in the meta-analyses was had, which was the verb linking all head and local nouns in the RC versions used in Experiments 1 and 4. Verb imageability was still a significant predictor of agreement error rates ($t = -2.37, p < .05$) in a model that only included data from RCs that did not contain had (i.e., those coded as having the “other” link type). Given that the verb imageability effect was not driven by the verb had, one possible reason for the pattern is that verb imageability reflects the amount of semantic content encoded in the embedded verb of the RC, and increased semantic content is associated with lower error rates. This is consistent with the pattern of results from Experiments 1 and 4, where errors were produced in RC conditions and the light-verb had added very little semantic content to the preamble. It is unclear, however, to what extent verb imageability reflects the semantic content of a larger range of content verbs and how one could quantify semantic content (cf. Gordon & Dell, 2003; Jackendoff, 1990).

Because the size of the effect of verb imageability was not much different from the size of the verb imageability effect (see Table 4.7) and the imageability rating instructions were different for nouns and verbs (see Appendix J), the verb imageability effect should be interpreted with caution. Even if the verb imageability effect is real, positing a special role for verb imageability in agreement computation is problematic. Any model of agreement computation that predicts production behavior for a wide range of subject NPs would need...
to account for subject NPs that have PP modifiers and do not include a verb. But, a measure of additional semantic content introduced by a linking word (be it a verb or a preposition) would be applicable to all complex subject NPs and could explain why subject NPs with RC modifiers containing the light-verb *had* produce similar error patterns to subject NPs with PP modifiers.

In general, these patterns are consistent with the results from the model containing NP PP data because they show that overall imageability was an important predictor of agreement error rates. It is unclear to what extent the imageability of each individual content word contributes to the agreement computation process. The fact that including verb imageability significantly improved a model containing only overall imageability presents an intriguing possibility that linking word imageability may affect agreement computation for complex subject NPs. However, it is possible that this effect was observed because verb imageability is highly correlated with other factors that may be important during agreement computation (e.g., semantic content/weight/complexity of linking words).

3. Experiment 4 and 5 Discussion

In Experiments 4 and 5, overall imageability affected error rates, with higher imageability associated with fewer agreement errors across a range of studies with non-distributive subject NPs with different structural, semantic, and lexical properties. In contrast, Eberhard (1999) reported larger mismatch effects for highly imageable distributive subject NPs compared less imageable distributive subject NPs. Thus, in Experiments 4 and 5, increased imageability decreased error rates, while in Eberhard (1999) increased imageability increased error rates. Instead of assigning the marking SAP value of a subject NP based on imageability (see Eberhard et al., 2005), a weighting parameter could be introduced to the marking process, as it appears as though the degree to which the subject NP’s conceptual number is accessible is a major factor affecting agreement computation.
In M&M, the marking SAP value is copied to the subject NP node. We propose that the effect of marking’s SAP value on the subject NP node’s SAP value could be weighted based on overall imageability, with higher weights assigned to more imageable subject NPs. Introducing this weighting parameter to the M&M model, would account for the higher error rates observed for highly imageable distributive NPs in Eberhard (1999) and the lower error rates observed for imageable preambles in Experiments 2a, 3a, and 4.

The introduction of this weighting parameter would allow marking to exert a stronger influence on agreement production when the conceptual number of the subject NP is highly accessible due to increased imageability than when it is less imageable. Given that the marking SAP values and morphing SAP values combine to assign the subject NP its number, less imageable subject NPs would be more susceptible to interference arising during the morphing process. Further research will be necessary to determine the relative contributions of marking and morphing processes during agreement computation.

In no analysis did measures of the imageability of the head and local nouns significantly impact agreement computation. These findings suggest that individual noun imageability does not need to be included in the morphing process (Eberhard et al., 2005). These findings are consistent with the claim that morphing is not sensitive to conceptual properties (Bock et al., 2001; Eberhard et al., 2005). Only the imageability of the embedded verb in RC modifiers had a significant effect on agreement error rates. While this result has a variety of potential causes (see above), it may hint at the possibility that a measure of linking word imageability or semantic weight may need to be incorporated into the weighting parameter \( w \) in morphing. Gillespie and Pearlmutter (2011) proposed replacing \( w \), a weighting parameter based on syntactic distance in Eberhard et al., with the combined effects of semantic integration and linear distance. If linking word imageability (or semantic content) affects agreement error rates, it would be sensible to include it as an additional factor contributing to \( w \) in the morphing process. Future work will be necessary to determine how conceptual properties of the linking word affect agreement computation.
Most importantly, the results of these experiments are consistent with a theory of agree-
ment computation that suggests speakers regularly use conceptual information to generate
the number marking of the verb during sentence production (Eberhard, 1999; Eberhard et
al., 2005; Humphreys & Bock, 2005; Pollard & Sag, 1988; Quirk et al., 1972). When the sub-
ject NP’s conceptual number is readily available to the speaker due to increased imageability,
speakers are more likely to agree with the conceptual number (Eberhard). These results also
show that when grammatical and conceptual number do not conflict, grammatical attraction
effects are more likely to arise when the influence of the conceptual number is not strong.

The influence of conceptual number on agreement computation has also been shown to
be affected by brain damage. Hartsuiker et al. (1999) found evidence that Broca’s apha-
sics were sensitive to grammatical interference, and produced number attraction effects that
were similar to non-disordered speakers. However, Broca’s aphasics showed smaller distribu-
tivity effects than non-disordered speakers, even when subject NPs were highly imageable.
Hartsuiker et al. suggested that Broca’s aphasics are less likely to use conceptual informa-
tion during agreement computation because of limitations in computational resources that
characterize Broca’s aphasia (e.g., Caplan, Baker, & Dehaut, 1985; Caplan & Waters, 1995,
research would be to examine whether individuals vary in the extent to which they use con-
ceptual information during agreement computation. It is possible that speakers with lower
working memory capacities or under memory load may show smaller imageability effects
than speakers with larger working memory capacities or under no memory load.

Taken together, the results of Experiments 4 and 5 suggest that agreement computation
is a complex process involving input from the conceptual representation of the subject NP as
well as grammatical information from individual lexical items. These results are consistent
with a version of the M&M model of agreement computation, but suggest that imageability
needs to be incorporated into the marking process to explain the range of agreement error
patterns observed in this dissertation.
Unfortunately, the experiments in Part 2 do not provide an explanation for why Experiment 1 and Experiment 4 showed no effect of structure while Bock and Cutting (1992) and Solomon and Pearlmutter (2004b) did, though Experiments 2 and 3 did rule out some confounds that could have been responsible for the conflicting results. Future work will be necessary to determine how other properties of the link between the head and local noun (e.g., imageability, function/content) affect agreement computation and may be responsible for structural effects that have been observed.
Part 3

Timing of Planning During Sentence Production
CHAPTER 5

Prosody reflects planning processes in production

1. Experiment 6

The study of language production is concerned with how speakers translate their thoughts into the sounds required to produce an utterance. According to Bock and Levelt’s (1994) model of language production, this process has three main stages — message planning, grammatical encoding, and phonological encoding. No theory of language production assumes that speakers plan entire utterances prior to speech onset. Instead, planning in language production has been hypothesized to be cascading, so as elements are being planned at one level, planning at other levels can proceed simultaneously. Experiment 6 examines two broad issues. First, how and where do message-level properties affect language production, and second, how ongoing planning during language production can be examined.

In this experiment, the prosodic phrasing of utterances was examined. Prosody is the “linguistic representation at which the acoustic-phonetic properties of an utterance vary independently of its lexical items” (M. Wagner & Watson, 2010, p. 905), and generally refers to the rhythm, pitch, and timing of speech. One particularly interesting aspect of prosody is that it is affected by information at all levels of the production process. The intended meaning of a string of words can be conveyed through its intonational phrasing, suggesting that message level properties are reflected in prosodic structure. For example, if John went to the store is produced with a drop in pitch on the last few words, this sentence will be interpreted as a statement; however, if the pitch rises at the end of the sentence, it is likely to be interpreted as question (Selkirk, 1984). Intonational breaks and lengthening are likely to appear at major syntactic boundaries (Cooper & Paccia-Cooper, 1980), suggesting prosody is sensitive to aspects of grammatical encoding, though syntactic units are often confounded...
with message units (F. Ferreira, 1988). Additionally, prosody is affected by phonological encoding, as the intrinsic length of phonemes within a word affects the relationship between the duration of that word and its surrounding pauses (F. Ferreira, 1993). In addition, lexical stress may shift depending upon the stress patterns of the words surrounding it to avoid lapses in stress or adjacent stressed elements (Selkirk, 1984). There is increasing evidence that prosody reflects real-time linguistic planning (for a review see M. Wagner & Watson, 2010). Linguistic elements that are predictable within their context are shorter than less predictable elements (e.g., Aylett & Turk, 2004; Bell, Brenier, Gregory, Girand, & Jurafsky, 2009; Gahl & Garnsey, 2004; Jaeger & Post, 2010; Piantadosi, Tily, & Gibson, 2011; Tily et al., 2009). There is also evidence that speakers lengthen words to buy themselves time to plan upcoming material that is difficult to access (Clark & Wasow, 1998; Fox Tree & Clark, 1997; Jaeger & Kidd, 2008). In addition, intonational phrases have been hypothesized to reflect planning units in production. Intonational phrases are units of prosodic grouping that are perceptually salient and the breaks between intonational phrases co-occur with duration changes, intensity changes, and pauses (Lehiste, 1973; Selkirk, 1984; M. Wagner & Watson, 2010). Previous work has suggested that speakers are unlikely to separate constituents that are related in meaning (F. Ferreira, 1993; Selkirk, 1984).

Following up on previous work, Watson, Breen, and Gibson (2006) developed a model of intonational break placement that assumed the purpose of intonational boundaries is to provide speakers with time to recover from expending resources after planning and producing long constituents, and time to plan properties of upcoming material (see also Breen, Gibson, & Watson, in press; Watson & Gibson, 2004). Watson et al. cite evidence that planning units in production may be based on semantic relationships, as elements that are semantically related often appear in the same intonational phrase (e.g., Selkirk, 1984; Watson & Gibson, 2004). Watson et al. found that speakers were less likely to place intonational breaks between the head of a verb phrase and its syntactically obligatory arguments (e.g., investigated the crash) than its nonobligatory adjuncts (e.g., arrived after the crash). Watson et al.
(2006) suggested that heads and their following constituents are more likely to be planned simultaneously when the following constituent is obligatory because heads and their obligatory arguments share a tight semantic link (see also Breen et al.). Watson and colleagues’ models of intonational break placement, which suggest an incremental planning system with planning units partially specified by the semantic content of the utterance, predict intonational break placement as well as, and in some cases better than, older models that relied on syntactic structure (Cooper & Paccia-Cooper, 1980; Gee & Grosjean, 1983).

Watson et al. (2006) suggested that the semantic integration effects observed in Solomon and Pearlmutter (2004b) are further evidence that semantic properties affect planning in production. Solomon and Pearlmutter claimed that elements that are more conceptually linked are planned with more temporal overlap during grammatical encoding, increasing their chance of interfering with each other resulting in subject-verb agreement errors (for similar evidence from exchange errors, see Pearlmutter & Solomon, 2007).

A critical aspect of language production that must be explained is how far in advance speakers plan parts of their utterance, and how this planning unfolds as an utterance is produced. Unfortunately, existing methods for examining planning in language production have some limitations. Traditionally, the way to study planning in production has been to examine distributions of errors elicited experimentally or through an observed corpus (e.g., Bock & Miller, 1991; V. S. Ferreira & Humphreys, 2001; Fromkin, 1973; Garrett, 1975, 1980). But it is well known that there may be biases in observation when collecting a corpus (Ferber, 1991; MacKay, 1980), and eliciting errors in the lab can be difficult (see Chapter 2). Even when errors are obtained in a laboratory setting they may be so rare that the sparse distributions can be troublesome for many methods of data analysis (for discussion see Chapter 2; Appendix B). Another fairly standard way of examining how far in advance speakers plan parts of their utterances is using speech onset times (e.g., Smith & Wheeldon, 1999; Wheeldon & Lahiri, 2002); however, speech onset times only reflect the planning that happens prior to initiating speech. Given that planning in language production
is thought to be at least somewhat incremental, these speech onset time measures do not take into consideration planning that happens as an utterance is unfolding. Finally, eyetracking has recently been used to study ongoing planning during production, but nearly all studies to date have shown evidence of radical incrementality (i.e., words are planned one-by-one, with little or no simultaneous planning) in the production system (e.g., Brown-Schmidt & Konopka, 2008; Brown-Schmidt & Tanenhaus, 2006; Griffin, 2001; Griffin & Spieler, 2006), which seems unlikely given that speakers often produce speech errors where they exchange words and sounds across fairly large distances during spontaneous production (Garrett, 1975, 1980).

Experiment 6 used prosodic phrasing, specifically word durations, to examine ongoing planning during language production. Solomon and Pearlmutter (2004b) hypothesized that semantic integration affected the relative timing of planning of elements within phrases, with more conceptually linked elements being planned closer together in time than less conceptually linked elements. Currently, the only evidence for this hypothesis comes from speech error data (Gillespie & Pearlmutter, 2011; Pearlmutter & Solomon, 2007), which do not directly reflect timing. Thus, the goal of this study is to determine if there is direct evidence that semantic integration affects timing during production.

If semantic integration affects timing of planning in language production, there may be evidence of more temporal separation between less integrated elements than between tightly integrated elements. When speakers have difficulty formulating a linguistic message or are about to produce unpredictable material, they often lengthen the words preceding the difficult or unpredictable elements (e.g., Fox Tree & Clark, 1997; Jaeger & Kidd, 2008), presumably to buy themselves time to plan this difficult material. Thus, one way prosodic separation due to differences in semantic integration may arise is through a process similar to strategic lengthening, which would predict that word durations should be affected. Another possible way prosodic separation of unintegrated elements could arise is through the
placement of intonational breaks, with breaks being less likely to be placed between tightly semantically linked elements (see Watson et al., 2006).

Experiment 6 examined if the duration of words linking unintegrated elements is longer than the duration of words linking integrated elements using recordings of responses from Solomon and Pearlmutter’s Experiment 4. Three types of control predictors were included in the models: (1) speech rate, (2) phonological context, and (3) predictability/availability. The exact controls included in each model are shown in Table 5.2 and described in detail below. The placement of intonational breaks was examined for a subset of tokens using the ToBI prosodic labeling system (Silverman et al., 1992), but breaks were extremely rare, and these results will not be discussed further.

1.0.9. Method

Materials. Twenty-four stimulus sets like that shown in Table 5.1 were created for an agreement error elicitation task reported in Solomon and Pearlmutter (2004b; Experiment 4). Each began with a head NP (e.g., The sweater) followed by a modifier containing a local noun (e.g., hole(s); skirt(s)). The modifier was always a PP beginning with the preposition with. The integrated versions indicated an attribute relationship between the head and local noun, and the unintegrated versions indicated an accompaniment relationship between the head and local noun. The head noun was always singular, and the four different versions of an item were created by varying semantic integration and local noun number. Each word position within the preamble was labeled (see Table 5.1).

Participants in Solomon and Pearlmutter’s (2004b) Experiment 4 read preambles aloud off a computer screen and then went on to complete the preambles as full sentences in any way that they chose (the experiments in Chapter 3 used an identical procedure). Recordings of all 24 critical items from 16 speakers (4 from each presentation list, 4 tokens per version of each item) were obtained, resulting in 384 total tokens. Twenty-nine tokens were excluded because of preamble repetition errors (19) and disfluencies (10), leaving 355 analyzed tokens.
Table 5.1. Solomon and Pearlmutter’s (2004b; Experiment 4) Stimuli and Semantic Integration Ratings by Condition

<table>
<thead>
<tr>
<th>Integration</th>
<th>Noun Number</th>
<th>Example</th>
<th>Semantic Integration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integrated</td>
<td>SP</td>
<td>The sweater with the tiny holes</td>
<td>5.53 (.63)</td>
</tr>
<tr>
<td></td>
<td>SS</td>
<td>The sweater with the tiny hole</td>
<td>5.44 (.80)</td>
</tr>
<tr>
<td>Unintegrated</td>
<td>SP</td>
<td>The sweater with the clean skirts</td>
<td>3.28 (.79)</td>
</tr>
<tr>
<td></td>
<td>SS</td>
<td>The sweater with the clean skirt</td>
<td>3.25 (.89)</td>
</tr>
</tbody>
</table>

Note. The semantic integration rating scale was 1 (loosely linked) to 7 (tightly linked); standard deviations are in parentheses. SP = singular head, plural local noun; SS = singular head, singular local noun; Det1 = Determiner1; Prep = Preposition; Det2 = Determiner2; Adj = Adjective.

Word duration measure. Word boundaries for each word position within the preambles were hand-marked in Praat (Boersma & Weenik, 2011). Using the guidelines established in Turk, Nakai, and Sugahara (2006), reliable acoustic landmarks of phonetic segments beginning and ending words were located. The time between the beginning and ending markers was used as the word duration measure. Durations were labeled by two coders (the author and a research assistant blind to the hypothesis). Correlating the word duration measures of both labelers revealed high agreement ($r = .97$), thus only the data from the naive coder was used.

Rate control predictor. Faster speech rate is characterized by shorter word durations and shorter pauses (e.g., Aylett & Turk, 2004; Bell et al., 2009; Jaeger & Kidd, 2008; Jaeger & Post, 2010) and is associated with the production of more reduced word forms (e.g., Frank & Jaeger, 2008; Jurafsky, Bell, Gregory, & Raymond, 2001). Thus it is important to control for a speaker’s average speech rate as it is likely to affect their word durations.
### Table 5.2. Control predictors for each individual word model in Experiment 6

<table>
<thead>
<tr>
<th>Control Type</th>
<th>Predictor</th>
<th>Det1</th>
<th>Noun1</th>
<th>Prep</th>
<th>Det2</th>
<th>Adj</th>
<th>Noun2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate</td>
<td>Speech rate</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Phonological Context</td>
<td>Length($w$)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Length($w - 1$)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>$w + 1$ Vowel Initial</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>$w - 1$ Vowel Final</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Pause following $w$</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Availability/Predictability</td>
<td>Length($w + 1$)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>F(Noun1)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>F(Adj)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>F(Noun2)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>P($w + 1$</td>
<td>$w - 1$, $w$)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>P($w + 1$</td>
<td>$w$)</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>P($w$</td>
<td>$w - 1$, $w + 1$)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>P($w$</td>
<td>$w - 1$)</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>P($w$</td>
<td>$w + 1$)</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

*Note.* Det1 = Determiner1; Prep = Preposition; Det2 = Determiner2; Adj = Adjective; $w$ = current word; $w - 1$ = previous word; $w + 1$ = next word; $F$ = log frequency; $P$ = log probability; Length = length in phonemes; Vowel Initial = begins with a vowel; Vowel Final = ends with a vowel.

**Speech rate.** Recordings of six filler trials from Solomon and Pearlmutter (2004b, Experiment 4) were obtained for each speaker. These fillers had a variety of structures and these structures were not identical to the critical items. Word durations for each word within the filler preambles were obtained using the same technique as was used for critical trials. The total duration of the preamble in seconds was computed and was divided by the number of syllables in the preamble. Each speaker’s log speech rate in syllables per second was included as a control predictor (for other duration models including a speech rate control predictor, see Bell et al., 2009; Jaeger & Kidd, 2008; Jaeger & Post, 2010).
Phonological context control predictors. The phonological context in which a word appears can affect its duration. For example, determiners appearing before adjectives and nouns that begin with a vowel are often produced in their full form (e.g., the rhymes with pea), while determiners preceding words that begin with a consonant are usually produced in their reduced form (e.g., the rhymes with sofa; Bell et al., 2009; Fox Tree & Clark, 1997; Rhodes, 1996). Word durations can also be affected by where in a prosodic phrase a word appears. Intonational breaks often appear at the edge of major syntactic constituents (Watson et al., 2006; Watson & Gibson, 2004) and pauses often follow intonational breaks (Breen et al., in press; Turk & Shattuck-Hufnagel, 2007). Durations are likely to be longer when a word appears at the edge of an intonational boundary as words preceding intonational breaks often undergo phrase-final lengthening (Turk & Shattuck-Hufnagel, 2007; Silverman et al., 1992).

While the preambles in Experiment 6 were fairly well-matched in phonological properties, content words did vary across items and versions. In addition, speakers often placed pauses after the preamble. Thus, it is important to control for differences in duration that could be caused by phonological properties of the utterances.

Length of current and previous content word. The content words in the preambles differed among items and versions. Thus, the models predicting the duration of the content words had a control for the length of the word in phonemes, as words with more phonemes are likely to have longer durations. This predictor is referred to as $\text{Length}(w)$ in Table 5.2. The length of the immediately previous word in phonemes was included in models where the previous word was a content word, and this predictor is referred to as $\text{Length}(w - 1)$ in Table 5.2.

Following word begins with vowel. As noted above, determiners are more likely to be produced in their full form prior to a word beginning with a vowel. In addition, acoustic landmarks of vowels can be fairly difficult to locate (Turk et al., 2006). Thus, this two-level
categorical predictor was included to control for the possible differences in pronunciation when a word precedes a vowel, and is referred to as $w + 1$ Vowel Initial in Table 5.2.

*Previous word ends with vowel.* As mentioned above, acoustic landmarks indicating vowel offsets can be difficult to locate. Thus, this predictor was included in the models to control for possible variation of the acoustic landmarks indicating the offset of the vowel. This two-level categorical predictor is referred to as $w - 1$ Vowel Final in Table 5.2.

*Pause following preamble.* While only items with no overt pauses during the preambles (i.e., Det1 – Noun2) were used in the analyses, speakers often paused after reading the preamble and prior to producing the sentence completion. In the Noun2 model, the presence of a pause following Noun2 was included as a two-level categorical predictor to control for lengthening effects that may be present due to intonational break placement. This predictor is referred to as Pause following $w$ in Table 5.2.

**Availability and predictability control predictors.** The principle of immediate mention (V. S. Ferreira & Dell, 2000) states that speakers are likely to produce available (i.e., active) words as early in an utterance as possible (for evidence from optional that-mentioning see V. S. Ferreira and Dell; Jaeger, 2010; for evidence from structural choices see Bock, 1986; Bock, Loebell, & Morey, 1992; Bock & Warren, 1985; Christianson & Ferreira, 2005; V. S. Ferreira, 1996; V. S. Ferreira & Yoshita, 2003), while less available elements are produced as late as possible in the utterance. Speakers often lengthen words or produce disfluencies (e.g., *um, uh*, restarts) if an upcoming word is not available to buy themselves time to plan the upcoming elements (Clark & Wasow, 1998; Fox Tree & Clark, 1997). Taken together, these findings suggest that speakers are sensitive to the availability of upcoming material and produce less available elements as late as possible. The predictability of a word given its surrounding context has also been shown to affect structural choices (e.g., Frank &
Jaeger, 2008; Gómez Gallo & Jaeger, 2009; Jaeger) and word duration (e.g., Aylett & Turk, 2004; Bell et al., 2009; Jaeger & Kidd, 2008; Jaeger & Post, 2010; Jurafsky et al., 2001; Tily et al., 2009) in natural speech, with more predictable elements showing reduced syntactic forms and shorter durations than less predictable words.

While availability and predictability have been shown to have independent effects on production (e.g., Jaeger, 2010; Jaeger & Kidd, 2008; Jaeger & Post, 2010), it is not the goal of Experiment 6 to determine how availability and predictability affect word duration. Instead, the following controls were included in the models to ensure that differences in word duration predicted by the degree of semantic integration were not due to overall differences in lexical availability and predictability in Solomon and Pearlmutter’s (2004b, Experiment 4) items.

Length of following word. The length of upcoming material has been shown to affect word duration. Speakers lengthen words before longer, more complex, and more difficult to access material (e.g., Fox Tree & Clark, 1997; Jaeger & Kidd, 2008). This predictor was included in the models to control for differences in accessibility due to word length, and is referred to as $\text{Length}(w + 1)$ in Table 5.2.

Frequency of content words. Words that are more frequent in the language are named more quickly than lower-frequency words (e.g., Oldfield & Wingfield, 1965), suggesting that frequency affects lexical accessibility (e.g., Grainger, 1990; Jescheniak & Levelt, 1994). In addition, lower frequency words tend to be longer than higher frequency words (Zipf, 1949, cf. Piantadosi et al., 2011). Thus, the log frequency of each content word obtained from the SUBTLEXus database (Brysbaert & New, 2009) was included to control for duration differences associated with lexical availability and length. These predictors are referred to as $F(\text{Noun1})$, $F(\text{Adj})$, and $F(\text{Noun2})$ in Table 5.2.
Availability of following word. Words are often lengthened if they precede unavailable material (Fox Tree & Clark, 1997). To control for availability of the word following the modeled word given its previous context, the log probability of the following word \((w + 1)\) given the two preceding words \((w - 1, w)\) was calculated from the Google n-gram corpus (Brants & Franz, 2006). This predictor is referred to as \(P(w + 1 \mid w - 1, w)\) in Table 5.2. For the Determiner1 model, there was no previous context, thus availability of the following word was estimated using only Determiner1 as the previous context. This predictor is referred to as \(P(w + 1 \mid w)\) in Table 5.2.

Predictability of word in its context. A measure of predictability included in the models was the probability of a word occurring given the words it appeared between. The log probability of a word given the word that preceded and followed it was calculated from the Google n-gram corpus (Brants & Franz, 2006). This predictor is referred to as \(P(w \mid w - 1, w + 1)\) in Table 5.2. Because Noun2 was not followed by another word in the preamble, predictability was estimated using only the previous word as context. This predictor is referred to as \(P(w \mid w - 1)\) in Table 5.2. Because Determiner1 was not preceded by any previous context, the predictability measure for the Determiner1 model was estimated using the following word as its immediate context. This predictor is referred to as \(P(w \mid w + 1)\) in Table 5.2.

Semantic integration predictor. The main predictor of interest was semantic integration. Semantic integration ratings for each version of each item obtained from Solomon and Pearlmutter (2004b) were included in the model. Semantic integration was rated on a 1 to 7 scale with ratings closer to 7 indicating a tight conceptual link between the head and local noun and ratings closer to 1 indicating a weak conceptual link between the head and local noun. The mean semantic integration ratings for each version of the stimulus items are shown in Table 5.1. In accordance with Solomon and Pearlmutter’s timing of planning
hypothesis, higher integration ratings were predicted to be associated with shorter word
durations.

1.0.10. Results

To determine if semantic integration affected word durations differently at each word posi-
tion, two linear mixed-effect models predicting log word durations were run and subjected to
model comparison. A model that included integration, word position, and their interaction
as fixed effect predictors, and participant and item intercepts as random factors, was com-
pared to an identical model with only main effects of integration and word position. The
model with the interaction term provided a significantly better fit ($\chi^2 = 19.94, p < .01$),
indicating that the size of the integration effect differed across word position.

To determine the unique effect of semantic integration on each individual word position,
separate mixed-effect models were conducted for each word, controlling for a number of
factors, and including random participant and item intercepts. In all models, predictors
were standardized (Gelman & Hill, 2007), and fixed-effect correlations were minimized by
centering and residualizing. If the semantic integration predictor was collinear with any
predictors, it was residualized against the other predictors and the semantic integration
residuals were entered into the model. After residualization, the fixed-effect correlations
with the semantic integration predictor were low ($r < .27$). The specific control predictors
included for each of the individual models are shown in Table 5.2. The results of all word
models, including control predictors, are shown in Appendix L.

The rule of thumb in regression modeling is to have at least 15 data points per predictor
to avoid overfitting (Harrell, 2001). None of the following models violate this rule; however,
any number of predictors can be added to a model and improve model fit and reduce the
generalizability of the model (Baayen, 2008). To address the possibility of overfitting, Baayen

---

1The word duration models had 355 data points, which allows up to 23 predictors per model. When including
fixed effect factors, random effect factors, and the intercept, no full duration model had any more than 13
predictors.
suggests running a model with all desired control predictors but without the predictor(s) of interest. Only control predictors whose coefficients are larger than their standard error are then included in a final model. When this procedure was followed for all the individual duration models in Experiment 6, result patterns were statistically identical to the full models reported below.

**Determiner1 model.** After controlling for other factors, semantic integration did not have a significant effect on Determiner1 duration ($t = 0.43, p = .66$).

**Noun1 model.** After controlling for other factors, semantic integration did not have a significant effect on Noun1 duration ($t = -0.57, p = .61$).

**Preposition model.** After controlling for other factors, higher semantic integration was associated with shorter Preposition durations ($t = -2.83, p < .01$). The average Preposition duration was 134ms, and with all else being equal, at the maximum semantic integration rating the model predicted a 127ms preposition while at the minimum rating the model predicted a 142ms preposition. Thus the range in semantic integration in these preambles predicted a 15ms change in Preposition duration (11% of its average duration).

**Determiner2 model.** After controlling for other factors, higher semantic integration was associated with shorter Determiner2 durations ($t = -2.88, p < .01$). The average Determiner2 duration was 99ms, and with all else being equal, at the maximum semantic integration rating the model predicted a 89ms Determiner2 while at the minimum rating the model predicted a 111ms Determiner2. Thus the range in semantic integration observed in these preambles predicted a 22ms change in Determiner2 duration (22% of its average duration).
Adjective model. After controlling for other factors, higher semantic integration was associated with shorter Adjective durations \((t = -4.33, p < .001)\). The average Adjective duration was 311ms, and with all else being equal, at the maximum semantic integration rating the model predicted a 286ms Adjective while at the minimum rating the model predicted a 340ms Adjective. Thus the range in semantic integration observed in these preambles predicted a 54ms change in Adjective duration (17% of its average duration).

Noun2 model. The results of the Noun2 model are shown in Table K.6. After controlling for other factors, semantic integration did not have a significant effect on Noun2 duration \((t = -0.58, p = .65)\).

Summary. Figure 5.1 shows the standardized semantic integration coefficient in each of the individual models. Determiner2 showed the largest proportional change in duration as a function of integration, followed by the next largest on the Adjective, and the smallest significant effect on the Preposition. Thus the entire preamble duration was not affected by semantic integration; instead, only the region between Noun1 and Noun2 (the head noun and local noun) was affected, with decreased temporal separation between the head and local noun when the head and local noun were more integrated.

1.0.11. Discussion

These findings show that semantic integration, a property that has been hypothesized to affect timing of planning, is reflected in the prosodic phrasing of an utterance. The temporal separation of less integrated elements was larger than the temporal separation of more integrated elements, and this effect was most pronounced on Determiner2, but present for the entire region between the head and local nouns (Preposition, Determiner2, Adjective).

It is well documented that upcoming material that is less available or less predictable due to frequency effects during lemma/lexeme access or grammatical encoding properties such
as subcategorization preferences is associated with lengthening. This lengthening is realized in word durations or by the inclusion of optional words (e.g., V. S. Ferreira & Dell, 2000; Fox Tree & Clark, 1997; Gahl & Garnsey, 2004; Jaeger, 2010; Tily et al., 2009). Importantly, effects of lexical and grammatical availability and predictability cannot explain the pattern observed in Experiment 6, as these properties were either controlled statistically or did not differ in the preambles. Thus, the results of Experiment 6 are consistent with the hypothesis that lengthening may also be observed when conceptual-level factors modulate accessibility of upcoming material.

There are a few possible reasons why the largest proportional change in duration due to semantic integration effect was observed on Determiner2. First, because participants in
Solomon and Pearlmutter (2004b) were reading preambles aloud, it is possible that the eye-voice span during reading aloud may have contributed to the effect. When reading aloud, fixations are approximately two words ahead of the word being articulated (H. Levin & Buckler-Addis, 1979). This would mean that as speakers in Solomon and Pearlmutter’s study were reading the word or words that determined the semantic integration relationship (Adjective and Noun2), they would have most likely been articulating Determiner2; thus, the most lengthening may be observed in this position. A second possible reason is that determiners are good candidates for lengthening. This has been suggested because determiners are high frequency, accessible, function words (Bell et al., 2009; Fox Tree & Clark, 1997; Jaeger & Kidd, 2008). Because determiners are easily accessible, appear before content words which vary in many ways and may require additional processing resources, and determiners tend to have reduced and unreduced forms, it is likely that determiners may adjust duration easily in response to processing needs.

These findings provide more evidence that prosodic phrasing reflects planning processes in sentence production. Watson et al. (2006) showed that syntactically and semantically related elements are more likely to appear in the same intonational phrase and hypothesized that this was because speakers are unlikely to place boundaries between elements that are more likely to be planned simultaneously. Experiment 6 did not show evidence of intonational break placement being modulated by semantic integration (see Introduction); however, it does provide evidence of increased temporal separation between less semantically linked elements, which suggests duration measures may reflect planning time and may capture more subtle properties that intonational breaks may not.

What these results have shown is that semantic integration, a conceptual-level factor, has an influence at multiple levels of production. Solomon and Pearlmutter (2004b) and Gillespie and Pearlmutter (2011) demonstrated that high integration was associated with increased interference during grammatical encoding which results in speech errors (also see Pearlmutter & Solomon, 2007). Experiment 6 provides evidence that semantic integration affects the
temporal separation between words at the phonological encoding level. Whether this is a
direct influence of conceptual factors on a separate prosodic representation or phonological
encoding, or a result of top-down cascading processing, is still an open question. Even though
the source of these effects cannot be determined from this experiment, these results provide
stronger support for the hypothesis that semantic factors influence timing of planning in
utterance formulation. Additionally, this work provides a first step as to how researchers can
combine prosodic analysis with other methods to address the issues of planning in production.

One concern with all the experiments in this dissertation is that of the tasks required a
comprehension component; speakers read the preambles prior to completing them as full sen-
tences. Thus, any of the effects observed may have been due to aspects of the comprehension
process (for discussion of this concern see Gillespie & Pearlmutter, 2011; Haskell & Mac-
Donald, 2003, 2005). Experiment 7 was designed to test the effect of semantic integration
on agreement computation while removing the influence of comprehension and obtaining a
measure of timing of planning in addition to agreement error rates.
CHAPTER 6

Effects of semantic integration and advance planning on grammatical encoding and agreement computation

1. Experiment 7

Planning in language production is thought to proceed incrementally; the system takes advantage of the sequential nature of production, allowing speech to be initiated before an entire utterance is prepared for articulation. Planning in language production proceeds quickly and accurately the majority of the time; however, speech errors do occasionally occur. If the scope of advance planning at a given level is large, multiple items are likely to be simultaneously available, which increases the chance of interference and speech errors occurring (Garrett, 1975, 1980).

Gillespie and Pearlmutter (2011) claimed that the mismatch effect in subject-verb agreement production is likely to arise from the same source as other speech errors. Specifically, Gillespie and Pearlmutter proposed that interference from local nouns is determined by the scope of advance planning in language production: Local nouns planned overlappingly with the head noun are more likely to interfere with agreement computation.

One way the head and local noun may be planned with more overlap is if they occur in a close semantic relationship (e.g., Breen et al., in press; Oppermann, Jescheniak, & Schriefers, 2008; Pearlmutter & Solomon, 2007; Solomon & Pearlmutter, 2004b; Watson et al., 2006). Solomon and Pearlmutter (2004b) found that agreement errors were more likely to occur in highly integrated cases (e.g., the chauffeur for the actor(s)) than less integrated cases (e.g., the chauffeur near the actor(s)). They interpreted these findings as evidence that semantic
integration can affect the timing of planning of elements within a phrase, independent of syntactic structure (also see Pearlmutter & Solomon, 2007).

Another way the head and local noun may be planned with more overlap is if they appear close together within the utterance. Gillespie and Pearlmutter (2011) showed that plural local nouns that are linearly closer to the head noun are more likely to create interference during agreement computation than local nouns appearing more distally, when controlling structural distance. These results suggest that planning proceeds in the order in which items are to be produced, with semantic integration shifting the relative timing of planning. Thus, given Gillespie and Pearlmutter’s findings, the degree to which speakers plan local nouns in advance should affect agreement error rates.

There is experimental evidence that the production system does at least some amount of advance planning (e.g., Allum & Wheeldon, 2007; Oppermann, Jescheniak, & Schriefers, 2010; Smith & Wheeldon, 1999; V. Wagner, Jescheniak, & Schriefers, 2010; Wheeldon & Lahiri, 2002). Smith and Wheeldon examined the scope of conceptual and grammatical encoding during language production using a picture description task. They designed visual arrays of three pictures in which one, two, or all three pictures moved, and participants had to describe the movement. The movements were designed to elicit sentences where the subject NP was either complex (e.g., the dog and the foot) or simple (e.g., the dog). They found that speech onset times for single-clause sentences with complex subject NPs were longer than for single-clause sentences with simple subject NPs. They argued from this that the entirety of the subject NP must be grammatically encoded prior to speech onset, and that the scope of grammatical encoding is the first full phrase of an utterance.

V. Wagner et al. (2010) found evidence that the scope of advance planning of lexical items is flexible rather than determined by grammatical structure. In a series of studies, participants described picture displays using simple sentences (e.g., the frog is next to the mug) and complex sentences (e.g., the red frog is next to the blue mug). An auditory distractor word that was semantically related to either the first noun or second noun was played
at varying stimulus onset asynchronies. In general, when a semantically related distractor was co-present with a picture to be named, speech onset times were longer than when the distractor was unrelated to the picture, due to both lexical items being active and in competition for production. This effect is known as semantic interference. V. Wagner et al. found a reliable semantic interference effect for the second noun in the simple sentences and a weaker semantic interference effect for complex sentences, suggesting that speakers plan the second noun in their utterance farther in advance when cognitive demands are low. In addition, they found that speakers who initiated speech faster (those with shorter average speech onset times) showed reduced semantic interference effects relative to speakers who were slower to initiate speech, indicating the possibility of individual differences in planning scope. These findings suggest that individuals’ planning scopes may influence how susceptible they are to agreement errors.

Experiment 7 examined the effect of advance planning on agreement error rates, by combining methods from the scope of planning literature and subject-verb agreement error elicitation studies. All previous studies of agreement production have had a comprehension component: Participants have either read or listened to preambles prior to using them to complete a full sentence. As mismatch effects are also observed in comprehension studies (e.g., Pearlmutter, Garnsey, & Bock, 1999), it is possible that interference from the comprehension system could be causing some effects in production due to the two systems sharing at least some representations (e.g., Garrod & Pickering, 2004; Pickering & Garrod, 2004, 2007). Using a presentation method that requires speakers to formulate their utterance from the message level should better approximate the natural production process and avoid potential confounds from comprehension.

In Experiment 7 participants described picture displays using complex subject NPs (subject head nouns followed by PP modifiers) and then went on to complete these descriptions (preambles) as full sentences. Semantic integration was manipulated by varying the preposition used to link the head and local noun in the display. Subject-verb agreement errors were
recorded, and speech onset times, which can reflect advance planning in language production (e.g., Smith & Wheeldon, 1999), were measured. This study aimed to replicate the mismatch and semantic integration effects on error rates from earlier studies, but using pictorial stimuli. Furthermore, because Solomon and Pearlmutter (2004b) suggested that semantically integrated nouns are planned with more overlap than unintegrated pairs of nouns, planning of the subject NP may be completed more quickly when the head noun and local noun are semantically integrated than when they are unintegrated, predicting shorter speech onset times for integrated conditions than for unintegrated conditions.

1.0.12. Method

Participants. One hundred fifty-three Northeastern University undergraduates participated in the experiment for course credit. Due to a counterbalancing error, extra participants were run in one list, so 10 randomly selected participants’ data in that list were excluded. Four participants were excluded because they were non-native speakers of English and four participants were excluded because they reported having language/cognitive disorders. Eight participants were unable to follow the instructions for a majority of the trials and were excluded. Three participants’ data were excluded due to a recording failure. After exclusions, 127 participants’ data were included in the error analyses. Speech onset time (SOT) data were available for 118 of these participants.

Materials and design. Twenty-four picture pairs like that shown in Figure 6.1 were created. The computer display screen was split in half vertically with one or two pictures arranged vertically on each side. A picture of the head noun, enclosed by a colored outline, was presented on one side of the screen, and a picture of the local noun was presented on

---

1These participants found this task very difficult and did not reliably produce NP PP descriptions of the picture displays (e.g., the apple near the pie...). Instead these participants often separated the head noun and local noun with an inflected verb (e.g., the apple is near the pie). Thus, these responses were unscorable because the local noun did not appear before the verb of the sentence.
the other side of the screen. In half the trials the head noun appeared on the left side of the screen (as in Figure 6.1), and in the other half the head noun appeared on the right. Local noun plurality was manipulated by presenting either a single picture (as in Figure 6.1a, 6.1c) or a pair of identical pictures on the same side of the screen (as in Figure 6.1b, 6.1d). Participants were required to construct NP PP preambles from the pictures presented. Participants linked the head noun and local noun with one of two prepositions that manipulated semantic integration; the color of the outline around the head noun determined which preposition was to be used. A blue outline indicated for, the preposition that created a more integrated relationship between the head and local noun; and a green outline indicated near, the preposition that created a less integrated relationship between the nouns (see Solomon & Pearlmutter, 2004b). Examples of the preambles to be constructed for the displays in Figure 6.1 are provided in the second column. Local noun number, linking word, and head noun position were varied within picture pair, creating eight conditions. Each of eight counterbalanced lists included 24 fillers and exactly one version of each of the 24 experimental items. All fillers had the same basic structure as the critical items, but with plural head nouns.

**Picture norming.** One hundred twenty-two black-and-white line drawings of inanimate objects were initially created. Seventy-six participants not involved in any other part of the experiment were presented with each picture for 2000ms and were required to provide a one-word label for the picture as quickly as possible. Data from two participants were excluded because they were not native speakers of American English. Names and naming times were recorded for each picture. Due to a recording failure, naming times were lost for two participants.

The 48 pictures used in the critical items were highly codable (i.e., the picture only elicited one or two nearly synonymous labels). Critical items were given the intended label (or a synonym) 96% ($SD = 4\%$) of the time. These labels were singular count nouns that
had regular plural forms. The average naming latency for each picture was calculated for each critical item’s dominant response. The average naming time for the critical items was 914ms ($SD = 94$ms), and all naming times were less than 1200ms. Pictures for filler and practice trials were also taken from this set, but no measures were taken to ensure that these pictures were highly codable (i.e., had a dominant label).

**Semantic integration norming.** A separate set of 123 participants produced the preamble aloud and rated the preamble for semantic integration on a 7-point scale with higher
numbers indicating a tighter conceptual link. Data from seven participants were excluded due to participants being non-native speakers of American English or recording failure. As in Solomon and Pearlmutter (2004b), the for versions were referred to as the integrated condition and the near versions were referred to as the unintegrated condition. Only trials for which participants correctly constructed the preamble with the expected noun labels (according to picture norming) were included, and these remaining data were analyzed in a linear mixed-effect model with head position, integration, local noun number, and their interactions as fixed effects; and with participant and item intercepts as random effects. Items linked with for were rated as more integrated than items linked with near ($t = 13.12, p < .001$). Items with plural local nouns were rated as more integrated than items with singular local nouns ($t = 2.13, p < .05$). An interaction of linking word and local noun number was marginal ($t = 1.94, p = .06$), with the effect of plurality stronger for for items than near items. The semantic integration ratings for the plural local noun conditions are the most critical for the integration manipulation as only plural local noun conditions should show significant error effects. Thus, integration in critical cases was manipulated as desired. No other main effects nor interactions were significant.

**Procedure.** Each participant was run individually in the main experiment. On each trial, a fixation cross was presented in the center of the screen for 1000ms to focus attention. The picture display immediately followed. 500ms later the colored outline appeared, indicating which picture was the head noun. Participants were required to name the head noun picture (including the determiner the), produce the linking word indicated by the color of the outline, name the local noun picture (including the determiner the), and then produce an ending that formed a complete sentence. The pictures disappeared after 3000ms, and a prompt to continue to the next trial was presented. Participants were not instructed as to how they should formulate a completion, only that they should form a complete sentence. Speech onset times (SOTs) were measured from the onset of the picture display until the
speaker triggered the voice key. Nine practice items preceded the 48 trials in the main experiment.

**Apparatus.** A Macintosh computer running the PsyScope (v. 1.2.5) software package (J. D. Cohen, MacWhinney, Flatt, & Provost, 1993) controlled stimulus presentation, with a button box to record SOTs. Responses were recorded to CD for analysis, using a Shure SM58 microphone connected to a Mackie 1202-VLZ Pro mixer/preamp and an Alesis Masterlink ML-9600 (OS v2.20) CD recorder.

**Response scoring.** All responses were transcribed and assigned to one of four coding categories: (1) correct, if the participant uttered the preamble correctly, said it exactly once, produced an inflected verb immediately after the preamble, and used a verb form that was correctly marked for number; (2) error, if all the criteria for correct responses were met, but the verb form failed to agree in number with the subject; (3) uninflected, if all the criteria for correct responses were met, but the verb was uninflected; and (4) miscellaneous, if the participant made an error in the preamble, if a verb did not immediately follow the preamble, or if the response did not fall into any of the other categories. Trials in which a participant made no response were excluded from all analyses. If the participant produced a disfluency (e.g., pauses, coughs) before, during, or immediately after producing the preamble and went on to produce a correct, error, or uninflected response, the scoring category and the disfluency were recorded.

### 1.0.13. Results

**Speech onset time analyses.** A linear mixed-effect model was used to analyze untransformed SOT data from 118 participants (Baayen et al., 2008). Analyses were also conducted for log-transformed SOTs; statistical patterns were identical to the untransformed data unless otherwise noted. Fixed effect predictors were head position, integration, noun number,
and their interactions; random effects were participant and item intercepts. Only trials for which participants correctly constructed the preamble with the expected noun labels (according to norming) were included. Trials where the SOT was more than ±2.5 SDs from the participant’s mean SOT were not included, nor were trials where the voice key was triggered before the appearance of the outline.

Mean SOTs for each condition are shown in Figure 6.2. Participants were faster to initiate speech when the head noun appeared on the left side of the screen than when it appeared on the right \((t = -4.87, p < .001)\). Speech was initiated faster for integrated than unintegrated versions \((t = -2.20, p < .05)\)^2. No other main effects nor interactions were significant \(|t|s < 1.63, ps > .13\). A second linear mixed effect model was run with a by-subject random integration slope, to determine if speakers were differentially sensitive to the integration manipulation. According to model comparison, there was a significant improvement to model fit \((\chi^2 = 7.97, p < .05)\), indicating there was significant variation in speakers’ sensitivity to the integration manipulation; however, the fixed-effect result patterns were identical to the model without the random slope.

If SOTs reflect advance planning of upcoming material, the speakers who were slower to initiate speech should show stronger integration effects than speakers who were faster to initiate (see V. Wagner et al., 2010). Given that there was significant variation in speakers’ sensitivity to the integration manipulation, two linear mixed-effect models (with only random subject and item intercepts) were run. One model was run on the data from the fastest half of participants \((N = 59)\) and the other was run on the data from the slowest half of participants \((N = 59)\)^3. The fastest half of participants were faster to respond when the head noun appeared on the left than when it appeared on the right \((t = -4.68, p < .001)\), but no other main effects nor interactions were significant. The slowest half of participants also showed the head position effect \((t = -2.94, p < .01)\). In addition, they were faster

---

^2This effect did not reach significance for log-transformed SOTs, \(t = -1.60, p = .11\).

^3Identical analyses with random by-subject integration slopes did not significantly improve the fit of either model \((\chi^2s < 1.5, ps > .45)\).
to respond to integrated than unintegrated versions ($t = -2.77$, $p < .01$). The interaction of integration and local noun number was marginal ($t = 1.89$, $p = .08$). The integration by local noun number interaction for the slower initiators is displayed in Figure 6.3. In the integrated condition, slower participants were slower to respond to the plural local noun condition than the singular local noun condition; this difference was not observed for the unintegrated condition. A linear mixed-effect model equivalent to a $t$-test comparing the local noun number conditions in the integrated version showed that plurals were marginally slower than singualrs ($t = 2.31$, $p = .052$).

**Verb response analyses.** Analyses were performed for error rates (the proportion of error responses out of error plus correct responses), which included disfluencies (the patterns were identical if disfluency cases were excluded). Following Barr (2008), the data were
analyzed using empirical logit weighted linear regression, aggregating separately over participants and items (see Chapter 2; Appendix B). By-participant and by-item weighted linear regressions on transformed error rates were performed with head position, linking word, noun number, and all interactions as sum-coded fixed effects (t-tests of parameter estimates are identified as $t_1$ for by-participant analyses and as $t_2$ for by-item analyses).

Response counts are shown in Table 6.1, and untransformed mean error rates are shown in Figure 6.4. Agreement errors were more likely when the local noun was plural than when it was singular ($t_1 = 6.81, p < .01; t_2 = 10.2, p < .01$), but there were no other main effects nor interactions ($|t|s < 1$).
Table 6.1. Experiment 7 Response Counts by Condition

<table>
<thead>
<tr>
<th>Integration</th>
<th>Head Position</th>
<th>Noun Number</th>
<th>Error</th>
<th>Correct</th>
<th>Uninflected</th>
<th>Misc</th>
<th>No Resp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integrated</td>
<td>Left</td>
<td>SP</td>
<td>25 (7)</td>
<td>254 (56)</td>
<td>28 (7)</td>
<td>73</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SS</td>
<td>1 (0)</td>
<td>255 (60)</td>
<td>35 (3)</td>
<td>90</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Right</td>
<td>SP</td>
<td>21 (7)</td>
<td>247 (48)</td>
<td>30 (6)</td>
<td>81</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SS</td>
<td>2 (1)</td>
<td>267 (65)</td>
<td>31 (10)</td>
<td>76</td>
<td>11</td>
</tr>
<tr>
<td>Unintegrated</td>
<td>Left</td>
<td>SP</td>
<td>28 (11)</td>
<td>240 (45)</td>
<td>28 (10)</td>
<td>82</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SS</td>
<td>1 (1)</td>
<td>273 (53)</td>
<td>32 (9)</td>
<td>74</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Right</td>
<td>SP</td>
<td>21 (10)</td>
<td>216 (50)</td>
<td>38 (9)</td>
<td>101</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SS</td>
<td>2 (1)</td>
<td>266 (58)</td>
<td>32 (6)</td>
<td>80</td>
<td>7</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>101 (38)</td>
<td>2018 (435)</td>
<td>254 (60)</td>
<td>657</td>
<td>66</td>
</tr>
</tbody>
</table>

Note. Disfluency counts are in parentheses. SP = singular head, plural local noun; SS = singular head, singular local; Misc = Miscellaneous; No Resp = No Response.

To determine if speakers’ average SOTs predicted their error rates, a second by-participant error rate model was run with data from participants whose SOTs were available (N =118). Fixed effect predictors were local noun number, participant mean speech onset latency (obtained from the SOT analyses), and their interaction; the only random effect was the participant intercept. Agreement errors were more likely when the local noun was plural than when it was singular (t = 6.68, p < .001). The likelihood of producing agreement errors increased with speech onset latency (t = 1.11, p < .05). There was a marginal interaction of noun number and speech onset latency, such that the mismatch effect was larger for speakers with longer average SOTs (t = 1.40, p = .06).

For uninflected and miscellaneous responses, the only significant result was that there were more miscellaneous errors when the head appeared on the right than on the left in the by-subject analysis (t₁ = -2.01, p < .05; t₂ = -2.50, p = .19).
Figure 6.4. Mean error rates as a function of integration and local noun number for Experiment 7. Error bars represent ±1 SEM calculated from the by-item analysis.

1.0.14. Discussion

This agreement error elicitation paradigm allowed speech onset time (SOT) to be measured in addition to error rates, while approximating the natural production process. The SOT results reflect the nature of the grammatical encoding process in this task. Speakers were faster to initiate speech when the head noun appeared on the left. This result is possibly due to the speakers in this task being biased to attend to the left side of the screen because English is written left to right, and speakers prefer to begin their utterances with the first item they look at (Griffin & Bock, 2000). Speakers were faster to initiate their responses when the head and local noun were integrated (linked with for) than when they were unintegrated (linked with near). This result suggests that speakers were sensitive to the semantic integration manipulation and completed planning of the subject NP more quickly when the preposition
created a tighter semantic link between the head and local noun (cf. Gillespie, Pearlmutter, & Shattuck-Hufnagel, 2010; Oppermann et al., 2008).

In Experiment 7, the fastest half of initiators had a mean SOT of 1045ms. These SOTs were not much slower than the average naming latency for the pictures used as head nouns in the picture norming study (946ms), although the visual scene was far more complex. The slower initiators had a mean SOT of 1485ms, which was over half a second slower than the average head noun naming time from the picture norming study. These patterns suggest that as speech initiation latency decreased (i.e., shorter SOTs) speakers may have been more incremental in planning and producing the preamble by essentially planning each word individually. As initiation times increased (i.e., longer SOTs) speakers may have been more likely to plan more of their utterance before initiating speech. One avenue for further research would be to experimentally impose a planning strategy on participants by introducing time pressure or explicit instruction and examine whether speakers who were instructed to plan more of their utterance in advance were more likely to show sensitivity to material later in the utterance than those who were instructed to plan more incrementally (e.g., Griffin & Spieler, 2000).

Interestingly, the integration effect was only present for the half of the data that included the slowest initiators, while the half of the data that included the faster initiators showed no trend toward an integration effect, which is consistent with the hypothesis that slower initiators did more advance planning. The slower initiators also showed a trend toward a local noun number effect in the integrated condition, with no such effect in the unintegrated condition. If slower initiation times reflect planning of the full subject NP prior to speech onset (see Smith & Wheeldon, 1999), the head and local noun should be planned with some degree of temporal overlap, though this may be modulated by semantic integration. When the head and local nouns mismatched in number (plural local noun conditions), the number mismatch could introduce competition or interference, increasing SOTs. Thus, the speeded planning of the integrated condition may result in the head and local noun being planned with
some overlap which would introduce competition in the number mismatch cases, predicting a local noun number effect. In the slower-to-plan unintegrated condition, the head and local noun may be planned with relatively little overlap, resulting in little or no interference in the number mismatch cases, resulting in no local noun number effect.

The scope of planning account of agreement computation (Gillespie & Pearlmutter, 2011) predicts that differences in timing of planning due to semantic integration (as observed in SOTs) should lead to a difference in error rates, but the integration effect (Solomon & Pearlmutter, 2004b) was not replicated in the error data. It is possible that the overall timing difference between integrated and unintegrated cases was not large enough to have a reliable effect on error production. Or, speakers may have been less sensitive to the semantic integration manipulation than they were in previous studies. In previous studies, there were fillers that contained a variety of structures which could have forced participants to pay attention to the message-level properties expressed in the preambles; in the current task, the preambles had identical structures that only varied in the preposition, possibly allowing the speakers to do less conceptual processing. Notably, speakers in this study often produced sentences that reflected a tight semantic link between the head and local noun even in unintegrated conditions (e.g., The apple near the pies should be used in them), which also could have decreased sensitivity to the semantic integration manipulation.

Even though the semantic integration effect was not replicated in the error analyses, these results suggest that this new paradigm can reliably reproduce the mismatch effect observed in many other agreement error elicitation studies relying on text or auditory stimuli (e.g., Bock & Miller, 1991). Interestingly, as mean initiation time increased, the mismatch effect increased. These results are consistent with the scope of planning account of agreement production (Gillespie & Pearlmutter, 2011): Speakers who plan more of their utterance in advance are more likely to experience interference from plural local nouns during agreement computation.
Of course it is possible that the speakers who were slower to initiate speech were just not very good at the task, which could lead to agreement errors as well as other types of speech errors, such as disfluencies and preamble errors. To determine if this was the case, two linear regression analyses were performed predicting disfluency rates and preamble error rates (both empirical-logit transformed) from a participant’s mean SOT. There was no reliable relationship between mean SOT and disfluency rate \((t(112) = -0.22, p = .82)\); however, the numerical pattern showed that speakers who were faster to initiate speech were more likely to be disfluent. V. Wagner et al. (2010) found a similar numerical pattern, with speakers who did less advance planning producing more disfluencies. In general, speakers tend to produce disfluencies before elements that are not yet available for production (Clark & Fox Tree, 2002; Fox Tree & Clark, 1997; Clark & Wasow, 1998), suggesting that disfluencies should be more common when speakers employ a more incremental planning strategy. Thus, the pattern of disfluencies observed in this study is consistent with the SOT and error data, suggesting that faster initiators did less advance planning that slower initiators.

There was a nonsignificant positive relationship between SOT and preamble error rates \((t(112) = 1.10, p = .28)\), with slower initiators more likely to produce errors when creating the preambles, suggesting that the slower initiators may have had more difficulty than the faster initiators. These preamble errors were of many types, which also included exchange errors (e.g., the pies for apple... I mean the apple for the pies) and anticipations (e.g., the apples for the pies, no... the apple for the pies). These types of errors have been hypothesized to arise due to multiple items being simultaneously planned and prepared for production (Bock & Levelt, 1994; Dell, 1986; Garrett, 1975, 1980; Pearlmutter & Solomon, 2007). Unfortunately, unambiguous exchange and anticipation errors were very rare in this experiment (10 out of 592 preamble errors), thus there was not enough data to determine if slower initiators were also more susceptible to these types of speech errors, as would be predicted by Gillespie and Pearlmutter’s (2011) scope of planning account.
Finally, an interesting future direction for this work would be to examine the prosodic phrasing of the sentences produced in this experiment. Experiment 7 only examined SOTs; however, Experiment 6 showed that effects of semantic integration can be seen on the duration of words linking the head and local noun (see Chapter 5; Gillespie et al., 2010). It is possible that the word durations of slower and faster initiators may be differentially affected by the semantic integration manipulation. While the faster initiators did not show any effect of integration at speech onset, there may be an effect of integration on word durations, as observed in Experiment 6, because the faster initiators delay planning the preposition until after speech onset (i.e., concurrent with the production of early elements of the preamble). In contrast, it is possible that because slower initiators showed an effect of semantic integration at speech onset, there would be no effect of integration on word durations because the processing required to create the integration link was completed prior to speech onset. As a follow up to the current analyses, a subset of participants’ data will be analyzed using similar techniques to those described in Experiment 6 to determine how the degree of advance planning affects subsequent word durations.

In sum, this study has established a new paradigm to study grammatical encoding and agreement computation. These results also support previous research indicating that conceptual properties affect the planning time of elements within an utterance (Chapter 5; Gillespie et al., 2010; Oppermann et al., 2008; Solomon & Pearlmutter, 2004b) and that the grammatical scope of planning influences agreement error rates (Gillespie & Pearlmutter, 2011).
Part 4

Discussion
CHAPTER 7

Discussion

Nearly every language has some form of agreement, whether it be number, gender, case, or some combination. In English number agreement is computed approximately every six seconds in running speech (Bock et al., 2006). Thus, an account of number agreement computation is an important aspect of any theory of language production. The current work adds to a body of literature that examines agreement computation in a variety of languages (Bock & Miller, 1991; Bock et al., 2004; Eberhard, 1999; Eberhard et al., 2005; Franck et al., 2002; Gillespie & Pearlmutter, 2011; Vigliocco & Hartsuiker, 2002). Specifically, this dissertation examined two main questions. Part 2 examined whether syntactic structure constrains agreement computation at all and examined if lexical and conceptual properties could be responsible for previously observed structural effects. Part 3 examined how the relative timing of planning of elements within an utterance is affected by a conceptual property (semantic integration) and whether these alterations in the timing of planning of nouns within a complex subject NP affect subject-verb agreement computation.


The results of Experiment 1 and Experiment 4 suggest that agreement computation is not strongly constrained by syntactic structure. Subject NPs with PP and RC modifiers that were matched in meaning, length, and semantic integration produced equivalent mismatch effects. These patterns contrast with those observed by Solomon and Pearlmutter (2004b) and Bock and Cutting (1992), even though their preambles were also matched in length and semantic integration.
Experiments 2 and 3 examined how properties of the lexical verbs used in previous studies may have contributed to the clause-boundedness effect observed in Bock and Cutting (1992) and Solomon and Pearlmutter (2004b). Experiment 2 showed that the embedded verb’s frequency did not affect agreement error rates. Experiment 3 showed that a verb’s transitivity bias did affect error rates, with preambles containing low-transitivity embedded verbs producing more errors than preambles containing high-transitivity embedded verbs. This pattern was tentatively attributed to the strong preference for many of the low-transitivity verbs to take sentential complements as arguments. The local nouns in the low-transitivity conditions could have been interpreted as the subjects of sentential complements, disrupting agreement computation because two subject-like nouns would be simultaneously active (see Badecker & Lewis, 2007; Lewis & Badecker, 2010). Given these results, the frequency and transitivity bias of the embedded verb in the RCs were not likely to have been responsible for the conflicting results between Experiments 1 and 4 on the one hand and studies that observed a clause-boundedness effect on the other (Bock & Cutting, 1992; Solomon & Pearlmutter, 2004b).

An interesting outcome of Experiments 2 and 3 was that agreement errors were only reliably produced when the preambles were abstract and low imageability (Experiments 2b and 3b). There have been a few studies that address the effect of imageability on agreement computation, but these studies also examined distributivity effects and none experimentally manipulated imageability within an experiment using only text stimuli (Eberhard, 1999; Humphreys & Bock, 2005). Experiment 5 examined the effect of imageability (and other factors) on agreement computation and showed that less imageable subject NP preambles produced stronger interference effects from plural local nouns than more imageable subject NPs, resulting in higher agreement error rates. These results suggest that imageability affects how accessible the conceptual number of a subject NP is during agreement computation and influences how influential the SAP value computed during marking is during agreement computation (cf. Eberhard, 1999; Eberhard et al., 2005).
This dissertation provides considerable evidence that agreement computation is not largely constrained by syntactic structure as has been assumed in most of the literature (e.g. Bock & Cutting, 1992; Eberhard et al., 2005; Franck et al., 2002; Vigliocco & Hartsuiker, 2002). In fact, structure may not be a factor that influences agreement computation at all. Experiments 2, 3, and 5 provide evidence that imageability has a strong influence on whether agreement errors are likely to occur at all when local nouns are in RCs, and Experiments 1 and 4 suggest that when the meaning relationship between the head and local noun is attribute-like, errors are likely to occur regardless of structure or imageability manipulations. Thus a complex picture is emerging where structure and imageability seem to play different roles under different conditions. An attempt to reconcile these seemingly conflicting patterns is discussed below.

As mentioned above, imageability seems to be a strong factor determining interference effects during agreement computation. The findings from this dissertation and Eberhard (1999) suggest that imageability affects how accessible the conceptual number of the subject NP is during agreement computation. Increased imageability being is associated with verb agreement that reflects the conceptual number of the subject NP. Thus, future models of agreement computation will need to take imageability into account. One possible way to alter the Marking and Morphing (M&M) model (Eberhard et al., 2005) to account for imageability effects is to introduce a weighting parameter to the marking component, with high imageability increasing the influence of marking on the SAP value of the subject NP node.

One result that is difficult to explain even with the proposed modifications to the M&M model is that agreement errors were reliably produced when subject NPs with RCs were imageable in Experiment 1 and Experiment 4; however, similarly highly imageable NP RC preambles failed to produce reliable mismatch effects in Experiment 2a and Experiment 3a. One difference between Experiments 1 and 4 and Experiments 2 and 3 is that the former studies only included preambles where the local noun was an attribute of the head noun,
while in the latter studies the head and local noun were in a variety of different types of relationships. This raises the question of whether there is a greater likelihood of interference during agreement computation when an attribute relationship exists between the head and local noun.

The fairly abstract concept of semantic integration is likely tapping into the degree to which one of the nouns is likely to be interpreted as an attribute, possessor, or component of the other noun\(^1\). In richly-inflected languages with case-marking systems, nouns in these types of relationships are often marked with the genitive case (Blake, 2001), suggesting that these relationships between nouns may be treated differently than other modified noun phrases. It will be interesting to determine if a measure of conceptual linkage which focuses on attribute, component, and possession relationships (which are likely to be related to the current operational definition of semantic integration) would be a better predictor of error rates across a wide range of stimuli.

An additional reason to re-examine the underlying relationships that semantic integration ratings attempt to capture is that semantic integration ratings tended to be confounded with imageability ratings in the studies in this dissertation as more imageable preambles were generally rated as more integrated than less imageable preambles\(^2\). However, as studies in this dissertation have shown, agreement errors tend to be more likely when preambles are low imageability. This pattern may be an artifact of the instructions for semantic integration rating, as the rating instructions suggested that participants create a mental image of the items within the preamble to determine how linked they were (see Appendix D; Chapter 3). Thus, a measure of how conceptually inseparable the head and local noun are could be less

\(^1\)Preambles like *The translator for the ambassador* produce mismatch effects similar to preambles like those in Experiments 1 and 4 where the local noun is an attribute of the head (see also Solomon & Pearlmutter, 2004b); however, the relationship between the translator and the ambassador may be semantically similar to an attribute relationship (e.g., *the ambassador who had/used the translator*).

\(^2\)High imageability preambles were rated as significantly more integrated than low imageability preambles in Experiment 2 (\(t = 2.45, p < .01\)) and Experiment 4 (\(t = 4.64, p < .001\)), and there was a numerical trend in this direction for Experiment 3 (\(t = 1.34, p = .18\)).
confounded with imageability, as it may be possible to rate this type of conceptual linkage without needing to create a mental image of the relation.

Another potential way to explain why highly imageable NP RC preambles in Experiments 1 and 4 produced significant mismatch effects while equally imageable NP RC preambles in Experiments 2a and 3a did not, is related to the effect of verb imageability on agreement error rates. Experiment 5 showed that increased verb imageability was associated with lower error rates for NP RC preambles. As mentioned in Chapter 4, it is possible that verb imageability was correlated with the amount of semantic content encoded in the verb. Quantifying the amount of semantic content is likely to be a difficult task, but there are theories of lexical semantics that attempt to describe the conceptual structure of a particular lexical item (e.g., B. Levin, 1993; B. Levin & Pinker, 1991; Jackendoff, 1990; Parsons, 1990). Much of the work in this area has focused on the lexical semantics of verbs, but there is some work on prepositions (e.g., Vandeloise, 1994). Jackendoff’s theory of lexical semantics suggests that a verb’s meaning can be decomposed into separate predicates and these predicates can be quantified. An interesting avenue for further research would be to determine if the number of predicates needed to encode a verb’s meaning predicts agreement error rates for NP RC preambles, and how this could be extended to examine NP PP preambles.

In sum, the results discussed in Part 2 rule out some of the potential confounds between studies observing a clause-boundedness effect (Bock & Cutting, 1992; Solomon & Pearlmuter, 2004b) and those that did not (Experiment 1, Experiment 4). However, it is still an open question as to why NP RC preambles that include the embedded verb had produce similar agreement error patterns to those with meaning-matched prepositions. These studies have suggested at least two possible reasons for this results. First, the presence of a lexical verb within an RC modifier may be enough to reduce error rates for preambles with RC modifiers because the lexical verb introduces additional semantic content to the subject NP. Second, errors may be more likely to occur when the head noun and local noun are linked in an attribute-like relationship, and lexical embedded verbs are less likely to create this sort
of relation between the head and local noun than prepositions. Both of these explanations suggest that syntactic structure is not a major factor constraining agreement computation, and instead rely on conceptual and lexical factors to explain error patterns. To reconcile these findings with Gillespie and Pearlmutter’s (2011) scope of planning account, further work will be necessary to determine the degree to which these conceptual and lexical factors may influence the timing of planning of elements within the subject NP.


Conceptual factors, such as imageability, have been shown to influence word order and structural choices in language production (e.g. Bock et al., 1992; Bock & Warren, 1985; Christianson & Ferreira, 2005), but these results provide fairly indirect measures of timing of planning during production. Similarly, semantic integration, another conceptual factor, has been hypothesized to affect timing of planning, but evidence for this hypothesis has largely been obtained from speech errors (Gillespie & Pearlmutter, 2011; Solomon & Pearlmutter, 2004b; cf. Pearlmutter & Solomon, 2007). The findings of Experiment 6 and Experiment 7 indicate that semantic integration directly affects timing during language production, with evidence from prosodic phrasing (Experiment 6) and speech onset times (Experiment 7).

Experiment 6 examined the prosodic phrasing of utterances and showed that elements of the utterance that were more semantically integrated were less prosodically separated than less integrated elements (i.e., durations of words separating the highly integrated elements were shorter). These findings are consistent with the predictions set forth by Solomon and Pearlmutter (2004b), and provide the first direct evidence that semantic integration modulates timing during language production. Methodologically, Experiment 6 shows that one can combine prosodic analyses with other production methods (e.g., error elicitation) to examine the relative timing of planning of elements during language production. The findings of this study suggest that prosodic phrasing and word durations are a rich source
of information that can nicely complement response data often used in production research (for more discussion see Gillespie et al., 2010; M. Wagner & Watson, 2010).

Experiment 7 developed a new paradigm to examine agreement computation in production that does not rely on the comprehension system. Speakers created sentence preambles from picture displays and completed them as full sentences. Speech onset times and agreement errors were recorded. This study suggested that speakers can vary in the degree to which they plan parts of their utterance in advance, prior to articulation. The half of the speakers with the slowest initiation times were sensitive to properties of words that appeared later in the utterance (e.g., semantic relations, plurality), and these speakers also experienced more interference from plural local nouns during agreement computation when compared to the half of the speakers with the fastest initiation times. The speakers with the fastest initiation times showed no sensitivity to properties of words downstream, and possibly were only planning the head noun phrase prior to initiating speech. This evidence suggests that the slower initiators may have planned more of their utterance in advance, while the faster initiators may have adopted a more incremental planning strategy.

Thus far, there is very little evidence that planning scope in language production is variable (cf. Oppermann et al., 2010), yet Experiment 7 provided evidence that speakers may employ different planning strategies when producing sentences with complex subject noun phrases (cf. Smith & Wheeldon, 1999). Moreover, there is fairly limited research on individual differences in language production (cf. Belke, 2008; Slevc, 2007) when compared to the amount of individual difference research in language comprehension (see Caplan & Waters, 1999). Thus, an interesting avenue for further research will be to determine whether individual differences in general cognitive capacities (e.g., working memory, attentional resources) are associated with different planning strategies during language production and how this affects speech error production and fluency in general.

In sum, the studies in Part 3 provided some of the first direct evidence that semantic integration affects the timing of production and planning of elements in sentence production
(cf. Gillespie et al., 2010; Pearlmutter & Solomon, 2007). In addition, these studies suggest that agreement computation is sensitive to the timing of planning of elements within a phrase and support the scope of planning account proposed by Gillespie and Pearlmutter (2011).

3. Revised Model of Agreement Computation

The results of the studies in this dissertation are largely consistent with the M&M model of agreement computation (Eberhard et al., 2005). We found evidence that conceptual information regularly influences agreement computation and that attraction errors are likely to occur when plural local nouns are combined with the number information of the head noun. However, to account for the range of effects observed in this dissertation, we propose two revisions to the M&M model. First, we suggest the influence of the marking SAP value should be modulated by the imageability of the subject NP, and second, the weighting parameter in morphing based on syntactic structural distance should be replaced with a weighting based on temporal, conceptual, and lexical properties of the link between the head and local noun.

Given the findings in Part 2, it is clear that imageability needs to be included in the M&M model (Eberhard et al., 2005). Because Experiment 5 indicated that only subject NP overall imageability was a significant factor predicting agreement error production, we propose that the influence of marking should be weighted based on the imageability of the subject NP. This weighting would account for the increased effect of distributivity (i.e., agreement with the conceptual number) when multi-token NPs are highly imageable (Eberhard, 1999) as well as the extremely low error rates in Experiments 2a and 3a when single-token subject NPs were highly imageable.

Consistent with the findings of Gillespie and Pearlmutter (2011), Experiments 1 and 4 provided no evidence for strong structural constraints on agreement computation, as subject NPs with PP and RC modifiers produced equal mismatch effects. Thus, the weighting parameter $w$ in the current M&M model does not seem to be appropriate (Eberhard et al.,
2005). As suggested by Gillespie and Pearlmutter (2011), $w$ should represent the relative timing of planning of the head and local noun during the encoding of the subject NP. They suggested that $w$ should be based on a combination of semantic integration (Solomon & Pearlmutter, 2004b) and linear distance. Experiments 2 and 3 ruled out an influence of the frequency and transitivity bias of the linking word as ways to influence relative timing of planning of the head noun and a local noun in an RC modifier. However, because the imageability of the verb may play a role in agreement computation (see Experiment 5b) it is possible that the amount of semantic content encoded in the linking word could affect the relative timing of planning of the head noun and local noun it links. This factor may prove to be a necessary component of the weighting parameter in morphing. Finally, given the results of Experiment 7 it seems plausible that a measure of an individual speaker’s grammatical scope of planning, or whatever cognitive capacities constrain it, may be required to fully account for agreement error patterns.

Currently the M&M model includes a measure of each noun’s relative frequency of appearing as a plural or as a singular as a factor contributing to the morphing process (Eberhard et al., 2005; see also Barker & Nicol, 2000a). It is possible that other lexical factors may also need to be included; however individual noun imageability does not seem to affect agreement computation (see Experiment 5b). In addition, as suggested by Bock and Miller (1991), the relative imageability of the head and local noun does not seem to be a factor affecting agreement computation. Thus the imageability of the head and local noun do not need to be included in a revised M&M model.

Figure 7.1 shows the proposed revisions to the M&M model of agreement computation (Eberhard et al., 2005). This model includes a weighting parameter for marking ($w_{\text{Marking}}$) based on the overall imageability of the subject NP. It also includes a revised weighting parameter for morphing ($w_{\text{Morphing}}$) based on a combination of lexical, temporal, and conceptual factors shown to influence agreement computation. These minor revisions to the current model can account for a wide range of data, while possibly removing the need of a
structural component. For example, if the semantic weight of the linking word is a factor that influences agreement error rates and the relative timing of planning of the head and local noun, it could explain why the clause-boundedness effect was observed in Bock and Cutting (1992) and Solomon and Pearlmutter (2004b) even in the absence of semantic integration or length differences in the PP and RC versions of the preambles: The verbs used in the RC modifiers in those studies were lexical and introduced more semantic content to the subject NP than the prepositions in the PP modifiers.

4. Conclusions

Clearly subject-verb agreement, an ostensibly grammatical phenomenon, is not solely informed by the grammatical number of the subject NP. Thus, the results discussed in this dissertation and others found in the literature are difficult to explain adopting the claims of syntax-only theories of agreement computation (e.g., Franck et al., 2002, 2006, 2010). Instead, the studies in this dissertation support a constraint-based theory of agreement computation, as multiple sources of information are used and combined to determine the final number of the verb (e.g., Eberhard et al., 2005; Thornton & MacDonald, 2003). The current results indicate that subject-verb number agreement errors are particularly likely to occur when the conceptual number of the subject NP is fairly inaccessible and a plural interfering noun is planned overlappingly with the head noun, due to high semantic integration, a semantically-light linking word, linear proximity, or a wide grammatical scope of planning (also see Gillespie & Pearlmutter, 2011; Solomon & Pearlmutter, 2004b). Errors may also arise when the local noun is likely to play a similar syntactic role as the head noun (see Chapter 3; Bock & Miller, 1991, Experiment 3, cf. Barker et al., 2001).

Now that researchers have identified a number of factors that affect agreement computation, it will be important to determine how these sources of information combine to determine the number marking that is eventually produced on the verb. The M&M model does to some extent show how multiple sources of information combine during agreement
Figure 7.1. A revised Marking and Morphing model adapted from Eberhard et al. (2005). The subject NP node receives its singular and plural (SAP) value from two sources: marking and morphing. The SAP value from marking is assigned based on conceptual number and overall imageability determines the $w_{Marking}$ weight, which is multiplied by the marking SAP value to yield the overall SAP value provided by marking. When $w_{Marking}$ is large, marking exerts more influence on the SAP value of the subject NP node. The SAP values from the head noun and local noun are passed to the subject NP node during morphing. The head noun SAP has more influence on the subject NP node’s SAP value, and is shown with a thick arrow. The local noun SAP has less influence on the subject NP node’s SAP value and is shown with a thin arrow. The degree to which the local noun SAP value influences the subject NP’s SAP value is determined by weighting ($w_{Morphing}$) based on a combination of the degree of semantic integration between the head and local noun, the linear distance between the head and local noun, the amount of semantic content the word that links the head and local nouns contains, and a measure of the speaker’s planning scope. NP = Noun phrase.

computation; however, it is now clear that the model must be revised to incorporate the influence of imageability (Experiments 2–5; Eberhard, 1999) and the relative timing of planning of the head and local noun (Experiments 6–7; Gillespie & Pearlmutter, 2011; Solomon
& Pearlmutter, 2004b). It is also unclear what role, if any, syntactic structure will play in a comprehensive model of agreement processing (Experiment 1; Experiment 4; Gillespie & Pearlmutter, 2011).

Eberhard et al. (2005) are clear that their model is incomplete as it “takes no account of time, incrementality, or variations in syntactic complexity...” (p. 551).
Appendix A

Experiment 1 Stimuli

The singular PP versions of the stimuli from Experiment 1 are shown below. The RC versions were created by replacing with with that had. The plural local noun versions were created by making the second noun plural.

1. The pizza with the missing slice
2. The phone with the new keypad
3. The truck with the special bumper
4. The ship with the spacious deck
5. The desk with the sliding drawer
6. The shark with the strong fin
7. The shirt with the expensive fabric
8. The plant with the delicious root
9. The stereo with the tiny switch
10. The loaf with the exotic grain
11. The telescope with the polished lens
12. The television with the sharp image
13. The fan with the wide blade
14. The box with the dented corner
15. The statue with the imported stone
16. The beach with the sloping dune
17. The hotel with the luxury suite
18. The rollerblade with the metal axle
19. The concert with the rock band
20. The zoo with the controlled habitat
21. The movie with the famous scene
22. The episode with the surprise ending
23. The satellite with the integrated computer
24. The newsletter with the insightful article
APPENDIX B

Empirical Logit Weighted Linear Regression

Performing ANOVAs on categorical data is problematic and may produce spurious results. Therefore, it has been suggested that categorical data be analyzed using logit mixed-effect models or using regression techniques with appropriate transformations and weighting of cases (see Jaeger, 2008). The error proportions produced in subject-verb agreement studies are often extremely low, and in some conditions, errors may not occur at all (e.g., SS); thus, there are problems applying the logit link function because when probabilities are close to 0 or 1, the resulting log odds approach negative and positive infinity, respectively (Agresti, 2002; Barr, 2008). Because of the nature of the distribution of errors in our data, we will apply the empirical logit transform as shown in equation (1), where $y$ is the number of target responses (e.g., errors in the error analyses, uninflected in uninflected analyses), and $n$ is the total number of responses (e.g., inflected verbs in the error analyses, all scorable responses in other analyses), aggregating over participants and items. The empirical logit transforms the outcome variable into log odds space, and is a good approximation of the logit link function, but can handle probabilities near 0 and 1. After applying the transform, the data will be analyzed using weighted linear regression, with weights of $\frac{1}{v}$, where $v$ is calculated by equation (2). When $n$ is large, the value of $\frac{1}{v}$ increases, thus weighting cases with more observations more heavily in the regression model. Weighted empirical logit linear regression, though not as powerful as a logit mixed-effect model, avoids the problems associated with performing ANOVAs over proportions by transforming the outcome into the correct probability space and weighting cases based on the number of observations (for detailed discussion of the technique see Barr, 2008).
(1) \[ \ln \left( \frac{y + 0.5}{n - y + 0.5} \right) \]

(2) \[ v = \frac{1}{y + 0.5} + \frac{1}{n - y + 0.5} \]
APPENDIX C

Experiment 2a Stimuli

The singular versions of the stimuli from Experiment 2a are shown below. The verb in the high frequency versions is shown first and the verb in the low frequency versions is shown second. The plural local noun versions were created by making the final noun plural.

(1) The dog that bit/fetched the large bone
(2) The farmer that pushed/poked the stubborn goat
(3) The football that broke/smashed the car window
(4) The report that bored/tired the busy manager
(5) The waiter that served/dropped the disgusting meal
(6) The lawyer that tricked/bribed the incompetent judge
(7) The weed that killed/pinched the purple flower
(8) The speech that moved/shook the serious protestor
(9) The chef that cooked/fried the delicious pepper
(10) The priest that liked/praised the humble nun
(11) The pulley that raised/hauled the full bucket
(12) The accident that took/claimed the innocent life
(13) The tailor that fixed/mended the ripped suit
(14) The king that thanked/sought the brave knight
(15) The hammer that split/cracked the rotting board
(16) The lesson that pleased/shaped the young teacher
(17) The maid that washed/swept the dirty floor
(18) The bouncer that struck/punched the rowdy musician
(19) The shovel that touched/scratched the trash bag
(20) The story that helped/aided the lost soul
(21) The hobo that roamed/prowled the abandoned building
(22) The nurse that cured/healed the sick patient
(23) The gear that turned/twisted the water wheel
(24) The comment that hurt/mocked the childish secretary
(25) The girl that screamed/wailed the nasty insult
(26) The banker that doubted/hated the old customer
(27) The boat that bumped/thumped the wooden dock
(28) The warrant that named/charged the escaped prisoner
(29) The mother that cut/chopped the crisp carrot
(30) The cop that met/joined the smart investigator
(31) The heater that boiled/melted the frozen dinner
(32) The challenge that dared/tempted the bungee jumper
(33) The father that held/lifted the heavy box
(34) The student that feared/blamed the teaching assistant
(35) The knife that stabbed/nicked the ripe apple
(36) The program that strained/warned the discouraged worker
In this survey, we’d like you to read some phrases and decide how closely linked two capitalized words are within each phrase. Sometimes the capitalized words will be highly linked, whereas in other cases they will be more independent from each other. Your task is to choose the number that you think corresponds to how closely linked the capitalized words are within each particular phrase. One strategy that you can use when rating these items is to try to form a mental picture of each phrase.

The purpose of this survey is to rate how closely linked the capitalized words in a list of 58 phrases are within the phrase that they appear. To show you what we mean by ”closely linked,” here are two example phrases:

(1) the KETCHUP or the MUSTARD
(2) the RING made of SILVER

It is important that you pay close attention to the relationship between the capitalized words in each individual phrase because, although words like ketchup and mustard may be related in general meaning (i.e., they are both condiments), in the first example these words are not closely connected. In (1), the only information that you have is that there are two things – ketchup and mustard – but you do not know anything about how these objects are related to each other. So, in the first example you would most likely choose a relatively low number like 1 or 2. In the second example, unlike the first, the capitalized words are closely
linked because the object *bracelet* is actually made from *silver*. So, for (2), you would probably choose a rather high number like 6 or 7.

Remember that we’re just interested in your opinions here. Please, do not worry about right or wrong answers. Just make sure to take the time to read each phrase carefully before choosing a response.

---

1This text suggests that (2) contains the word *bracelet*, when it actually contains the word *ring*. This was unintended and is the mistake referred to in Chapter 3.
APPENDIX E

Experiment 2b Stimuli

The singular versions of the stimuli from Experiment 2b are shown below. The verb in the high frequency versions is shown first and the verb in the low frequency versions is shown second. The plural local noun versions were created by making the final noun plural.

(1) The fear that depressed/disheartened the student
(2) The evidence that confused/puzzled the witness
(3) The suggestion that interested/captivated the banker
(4) The point that worried/irked the teacher
(5) The threat that scared/angered the secretary
(6) The possibility that troubled/infuriated the father
(7) The situation that bored/alienated the senator
(8) The song that delighted/calmed the singer
(9) The remark that embarrassed/baffled the nurse
(10) The request that surprised/startled the lawyer
(11) The rumor that stunned/bewildered the priest
(12) The magazine that disgusted/revolted the mother
(13) The report that described/conveyed the problem
(14) The warrant that identified/disclosed the suspect
(15) The discovery that affected/defined the debate
(16) The letter that discussed/documentated the promotion
(17) The advice that revealed/outlined the risk
(18) The proposal that mentioned/prompted the change
(19) The warning that showed/signaled the obstacle
(20) The lesson that ensured/impacted the decision
(21) The guarantee that required/demanded the signature
(22) The article that recognized/publicized the product
(23) The memo that reported/sketched the issue
(24) The fax that detailed/itemized the project
APPENDIX F

Experiment 3a Stimuli

The singular versions of the stimuli from Experiment 3a are shown below. The verb in the high transitivity versions is shown first and the verb in the low transitivity versions is shown second. The plural local noun versions were created by making the final noun plural.

(1) The lawyer who discussed/advanced the bill
(2) The warden who guarded/hung the prisoner
(3) The chef who baked/started the meal
(4) The artist who painted/tempted the girl
(5) The soldier who attacked/escaped the village
(6) The child who provoked/doubted the boy
(7) The passenger who printed/ripped the ticket
(8) The priest who praised/helped the nun
(9) The defendant who denied/protested the claim
(10) The comedian who offended/worried the teacher
(11) The actor who quoted/yelled the line
(12) The therapist who comforted/relaxed the widow
(13) The president who established/permitteed the law
(14) The driver who killed/hurried the pedestrian
(15) The scientist who described/proved the theorem
(16) The cop who arrested/fought the robber
(17) The banker who accepted/refused the deposit
(18) The librarian who visited/cheered the grandmother
(19) The detective who reviewed/stopped the crime
(20) The singer who entertained/suggested the dancer
(21) The tailor who mended/moved the suit
(22) The official who appointed/knew the governor
(23) The maid who vacuumed/rolled the rug
(24) The biker who chased/raced the jogger
APPENDIX G

Experiment 3b Stimuli

The singular versions of the stimuli from Experiment 3b are shown below. The verb in the high transitivity versions is shown first and the verb in the low transitivity versions is shown second. The plural local noun versions were created by making the final noun plural.

(1) The scam that elected/hardened the official
(2) The comment that advocated/started the event
(3) The statement that comforted/excited the community
(4) The contest that included/frightened the spectator
(5) The job that killed/tired the insomniac
(6) The judgment that quoted/maddened the activist
(7) The probe that investigated/escaped the issue
(8) The council that praised/doubted the idea
(9) The rule that described/permitted the exception
(10) The memoir that denied/confessed the transgression
(11) The meeting that emphasized/continued the discussion
(12) The investigation that studied/expected the change
(13) The warning that offended/puzzled the organizer
(14) The doctrine that established/allowed the law
(15) The belief that gladdened/cheered the mourner
(16) The opinion that criticized/decided the debate
(17) The scandal that reviewed/started the inquiry
(18) The story that entertained/enthused the audience
(19) The analysis that copied/estimated the result
(20) The argument that attacked/persuaded the skeptic
(21) The discovery that confirmed/proved the theory
(22) The organization that understood/advanced the cause
(23) The donation that saved/helped the campaign
(24) The claim that provoked/obsessed the researcher
APPENDIX H

Full Subject Noun Phrase Imageability Rating Instructions

Phrases differ in their capacity to arouse mental images of things or events. Some phrases arouse a sensory experience, such as a mental picture or sound, very quickly and easily, whereas other phrases may do so only with difficulty (i.e., after a long delay) or not at all.

The purpose of this survey is to rate a list of 58 phrases as to the ease or difficulty with which they arouse mental images. Any phrase, that in your estimation arouses a mental image (i.e., a mental picture or sound, or other sensory experience) very quickly and easily, should be given a high imagery rating (at the upper end of the numerical scale). Any phrase that arouses a mental image with difficulty or not at all should be given a low imagery rating (at the lower end of the numerical scale). For example, think of the phrase “the portrait of the woman.” This phrase would probably arouse an image (e.g., a painting of a woman) relatively easily and would be rated as high. The phrase “the description of the issues” would probably do so with difficulty, or not at all, and would be rated as low imagery.

Because words in phrases tend to make you think of other words as associates, it is important that your ratings not be based on this, and that you judge only the ease with which you get a mental image of an object or event in response to each phrase.

Your imagery ratings will be made on a 1 to 7 scale. A value of 1 will indicate a low imagery rating, and a value of 7 will indicate a high imagery rating. Values of 2 to 6 will indicate intermediate ratings. When making your ratings, try to be as accurate as possible, but do not spend too much time on any one phrase.
APPENDIX I

Content Word Imageability Rating Instructions

(The examples given before each slash were used for the noun phrase surveys, while the examples given after the slash were used in the verb phrase surveys.)

Phrases differ in their capacity to arouse mental images of things or events. Some phrases arouse a sensory experience, such as a mental picture or sound, very quickly and easily, whereas other phrases may do so only with difficulty (i.e., after a long delay) or not at all.

The purpose of this survey is to rate a list of 92 phrases as to the ease or difficulty with which they arouse mental images. Any phrase, that in your estimation arouses a mental image (i.e., a mental picture or sound, or other sensory experience) very quickly and easily, should be given a high imagery rating (at the upper end of the numerical scale). Any phrase that arouses a mental image with difficulty or not at all should be given a low imagery rating (at the lower end of the numerical scale). For example, think of the phrase “the plant”/ “to jump.” This phrase would probably arouse an image (e.g., a house plant/ an animal or person jumping) relatively easily and would be rated as high. The phrase “the description”/ “to believe” would probably do so with difficulty, or not at all, and would be rated as low imagery.

Because words in phrases tend to make you think of other words as associates, it is important that your ratings not be based on this, and that you judge only the ease with which you get a mental image of an object or event in response to each phrase.

Your imagery ratings will be made on a 1 to 7 scale. A value of 1 will indicate a low imagery rating, and a value of 7 will indicate a high imagery rating. Values of 2 to 6 will indicate intermediate ratings. When making your ratings, try to be as accurate as possible, but do not spend too much time on any one phrase.
APPENDIX J

Experiment 4 Stimuli

The singular versions of the stimuli from Experiment 4 are shown below. The high imageability items are shown in 1-16 and the low imageability items are shown in 17-32. The RC versions were created by replacing with with that had. The plural local noun versions were created by making the final noun plural.

(1) The trophy with the fancy design
(2) The jacket with the hideous stripe
(3) The machine with the obnoxious whistle
(4) The island with the concrete bridge
(5) The birdcage with the squeaky hinge
(6) The mountain with the rusty cabin
(7) The necklace with the glass bead
(8) The document with the sloppy signature
(9) The window with the heavy curtain
(10) The playground with the rusty gate
(11) The cathedral with the towering steeple
(12) The cupcake with the birthday candle
(13) The banana with the brown bruise
(14) The computer with the broken monitor
(15) The flower with the delicious root
(16) The sweater with the yellow pocket
(17) The proposal with the strong claim
(18) The committee with the unpopular opinion
(19) The council with the deciding vote
(20) The proclamation with the weighty conclusion
(21) The description with the significant flaw
(22) The organization with the unfounded complaint
(23) The program with the annoying glitch
(24) The announcement with the important warning
(25) The option with the cute accessory
(26) The report with the silly mistake
(27) The route with the tricky obstacle
(28) The experiment with the fatal error
(29) The hoax with the sneaky objective
(30) The scam with the evil goal
(31) The story with the blatant lie
(32) The department with the new policy
APPENDIX K

Results of Models in Experiment 6

The full results of the six individual word duration models in Experiment 6, including all control predictors. Positive $\beta$s indicate that increasing values of the predictor were associated with longer word durations, while negative $\beta$s indicate that increasing values of the predictor were associated with shorter word durations. For categorical predictors, the level shown after the colon was sum-coded +0.5 and the other level −0.5. For all tables, significance levels were indicated as follows: $***p < .001$. **$p < .01$. *$p < .05$. †$p = .10$.

Table K.1. Results of Determiner1 Linear Mixed-Effect Regression Model

<table>
<thead>
<tr>
<th>Effect</th>
<th>$\beta$</th>
<th>$SE$</th>
<th>$t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>4.35</td>
<td>.14</td>
<td>30.70***</td>
</tr>
<tr>
<td>Speech rate</td>
<td>−0.05</td>
<td>.04</td>
<td>−1.37</td>
</tr>
<tr>
<td>$w + 1$ Vowel Initial: No</td>
<td>0.21</td>
<td>.14</td>
<td>1.51*</td>
</tr>
<tr>
<td>Length($w + 1$)</td>
<td>−0.07</td>
<td>.04</td>
<td>−1.67**</td>
</tr>
<tr>
<td>F(Noun1)</td>
<td>0.07</td>
<td>.11</td>
<td>0.68</td>
</tr>
<tr>
<td>F(Adj)</td>
<td>−0.03</td>
<td>.03</td>
<td>−1.02</td>
</tr>
<tr>
<td>F(Noun2)</td>
<td>−0.04</td>
<td>.02</td>
<td>−1.48†</td>
</tr>
<tr>
<td>P($w + 1 \mid w - 1, w$)</td>
<td>−0.05</td>
<td>.07</td>
<td>−0.70</td>
</tr>
<tr>
<td>P($w + 1 \mid w$)</td>
<td>0.10</td>
<td>.11</td>
<td>0.91</td>
</tr>
<tr>
<td>Semantic integration</td>
<td>0.01</td>
<td>.02</td>
<td>0.43</td>
</tr>
</tbody>
</table>
Table K.2. Results of Noun1 Linear Mixed-Effect Regression Model

<table>
<thead>
<tr>
<th>Effect</th>
<th>β</th>
<th>SE</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>5.79</td>
<td>.04</td>
<td>144.81***</td>
</tr>
<tr>
<td>Speech rate</td>
<td>−0.13</td>
<td>.05</td>
<td>−2.47**</td>
</tr>
<tr>
<td>Length(w)</td>
<td>0.22</td>
<td>.06</td>
<td>3.44***</td>
</tr>
<tr>
<td>F(Noun1)</td>
<td>−0.18</td>
<td>.06</td>
<td>−2.77***</td>
</tr>
<tr>
<td>F(Adj)</td>
<td>−0.01</td>
<td>.03</td>
<td>−0.17</td>
</tr>
<tr>
<td>F(Noun2)</td>
<td>0.02</td>
<td>.02</td>
<td>0.88</td>
</tr>
<tr>
<td>P(w</td>
<td>w + 1)</td>
<td>−0.03</td>
<td>.11</td>
</tr>
<tr>
<td>P(w</td>
<td>w − 1, w + 1)</td>
<td>−0.09</td>
<td>.07</td>
</tr>
<tr>
<td>Semantic integration</td>
<td>−0.01</td>
<td>.02</td>
<td>−0.57</td>
</tr>
</tbody>
</table>

Table K.3. Results of Preposition Linear Mixed-Effect Regression Model

<table>
<thead>
<tr>
<th>Effect</th>
<th>β</th>
<th>SE</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>4.90</td>
<td>.05</td>
<td>102.56***</td>
</tr>
<tr>
<td>Speech rate</td>
<td>−0.16</td>
<td>.08</td>
<td>−2.08**</td>
</tr>
<tr>
<td>w − 1 Vowel Final: No</td>
<td>−0.01</td>
<td>.03</td>
<td>−0.34</td>
</tr>
<tr>
<td>Length(w − 1)</td>
<td>−0.03</td>
<td>.05</td>
<td>−0.73</td>
</tr>
<tr>
<td>F(Noun1)</td>
<td>−0.04</td>
<td>.05</td>
<td>−0.95</td>
</tr>
<tr>
<td>F(Adj)</td>
<td>0.02</td>
<td>.02</td>
<td>0.79</td>
</tr>
<tr>
<td>F(Noun2)</td>
<td>0.03</td>
<td>.02</td>
<td>1.17</td>
</tr>
<tr>
<td>P(w + 1</td>
<td>w − 1, w)</td>
<td>0.05</td>
<td>.05</td>
</tr>
<tr>
<td>P(w</td>
<td>w − 1, w + 1)</td>
<td>0.04</td>
<td>.05</td>
</tr>
<tr>
<td>Semantic integration</td>
<td>−0.05</td>
<td>.02</td>
<td>−2.83**</td>
</tr>
</tbody>
</table>
Table K.4. Results of Determiner2 Linear Mixed-Effect Regression Model

<table>
<thead>
<tr>
<th>Effect</th>
<th>$\beta$</th>
<th>SE</th>
<th>$t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>4.59</td>
<td>.59</td>
<td>87.85 ***</td>
</tr>
<tr>
<td>Speech rate</td>
<td>-0.07</td>
<td>.07</td>
<td>-0.94</td>
</tr>
<tr>
<td>$w + 1$ Vowel Start: No</td>
<td>-0.05</td>
<td>.04</td>
<td>-1.40</td>
</tr>
<tr>
<td>Length($w + 1$)</td>
<td>-0.01</td>
<td>.05</td>
<td>-0.29</td>
</tr>
<tr>
<td>F(Noun1)</td>
<td>0.14</td>
<td>.06</td>
<td>2.11 *</td>
</tr>
<tr>
<td>F(Adj)</td>
<td>0.01</td>
<td>.05</td>
<td>0.24</td>
</tr>
<tr>
<td>F(Noun2)</td>
<td>-0.08</td>
<td>.04</td>
<td>-2.00 †</td>
</tr>
<tr>
<td>$P(w + 1 \mid w - 1, w)$</td>
<td>-0.10</td>
<td>.05</td>
<td>-2.13 †</td>
</tr>
<tr>
<td>$P(w \mid w - 1, w + 1)$</td>
<td>-0.08</td>
<td>.04</td>
<td>-1.85</td>
</tr>
<tr>
<td>Semantic integration</td>
<td>-0.11</td>
<td>.04</td>
<td>-2.88 *</td>
</tr>
</tbody>
</table>

Table K.5. Results of Adjective Linear Mixed-Effect Regression Model

<table>
<thead>
<tr>
<th>Effect</th>
<th>$\beta$</th>
<th>SE</th>
<th>$t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>5.74</td>
<td>.03</td>
<td>172.15 ***</td>
</tr>
<tr>
<td>Speech rate</td>
<td>-0.11</td>
<td>.04</td>
<td>-3.11 *</td>
</tr>
<tr>
<td>Length($w$)</td>
<td>0.19</td>
<td>.03</td>
<td>6.11 ***</td>
</tr>
<tr>
<td>Length($w + 1$)</td>
<td>-0.08</td>
<td>.02</td>
<td>-3.10 **</td>
</tr>
<tr>
<td>$w + 1$ Vowel Start: No</td>
<td>0.14</td>
<td>.02</td>
<td>6.25 ***</td>
</tr>
<tr>
<td>F(Noun1)</td>
<td>0.00</td>
<td>.04</td>
<td>-0.01</td>
</tr>
<tr>
<td>F(Adj)</td>
<td>-0.22</td>
<td>.02</td>
<td>-8.86 ***</td>
</tr>
<tr>
<td>F(Noun2)</td>
<td>-0.01</td>
<td>.02</td>
<td>-0.34</td>
</tr>
<tr>
<td>$P(w + 1 \mid w - 1, w)$</td>
<td>-0.06</td>
<td>.02</td>
<td>-3.02 **</td>
</tr>
<tr>
<td>$P(w \mid w - 1, w + 1)$</td>
<td>-0.05</td>
<td>.02</td>
<td>-2.22 *</td>
</tr>
<tr>
<td>Semantic integration</td>
<td>-0.08</td>
<td>.02</td>
<td>-4.33 ***</td>
</tr>
<tr>
<td>Effect</td>
<td>$\beta$</td>
<td>$SE$</td>
<td>$t$</td>
</tr>
<tr>
<td>-------------------------------------</td>
<td>---------</td>
<td>------</td>
<td>-------</td>
</tr>
<tr>
<td>Intercept</td>
<td>6.28</td>
<td>.04</td>
<td>177.43***</td>
</tr>
<tr>
<td>Speech rate</td>
<td>−0.12</td>
<td>.04</td>
<td>−2.66*</td>
</tr>
<tr>
<td>Length($w$)</td>
<td>0.21</td>
<td>.03</td>
<td>6.14***</td>
</tr>
<tr>
<td>Length($w − 1$)</td>
<td>−0.05</td>
<td>.04</td>
<td>−1.31</td>
</tr>
<tr>
<td>$w − 1$ Vowel Final: No</td>
<td>−0.02</td>
<td>.02</td>
<td>−1.16</td>
</tr>
<tr>
<td>F(Noun1)</td>
<td>0.03</td>
<td>.06</td>
<td>0.57</td>
</tr>
<tr>
<td>F(Adj)</td>
<td>0.04</td>
<td>.03</td>
<td>1.07</td>
</tr>
<tr>
<td>F(Noun2)</td>
<td>−0.26</td>
<td>.03</td>
<td>−8.61***</td>
</tr>
<tr>
<td>Pause following $w$</td>
<td>0.05</td>
<td>.01</td>
<td>4.35***</td>
</tr>
<tr>
<td>P($w \mid w − 1$)</td>
<td>−0.15</td>
<td>.03</td>
<td>−4.46***</td>
</tr>
<tr>
<td>Semantic integration</td>
<td>−0.01</td>
<td>.03</td>
<td>−0.58</td>
</tr>
</tbody>
</table>
References


of Memory and Language, 39, 392-417.


