NORTHWESTERN UNIVERSITY

# Verb Metaphors as Analogies: Patterns of Meaning Change and the Role of Structure-Mapping in Comprehension

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Cognitive Psychology

By

Daniel C. King

EVANSTON, ILLINOIS

June 2023

## Abstract

Metaphor is an important and pervasive phenomenon in language and cognition. The vast majority of psycholinguistic research on metaphor has focused on noun metaphors (e.g., *That surgeon is a butcher; That lawyer is a shark*), while relatively little has investigated the processing of verb metaphors (e.g., *The car limped down the road, The lizard worshipped*). The dearth of work in this area is unfortunate, as there is evidence that verbs are used metaphorically more frequently than nouns. The goal of this dissertation is to bridge that gap by producing a more thorough characterization of the nature of verb metaphor than is currently present in the literature. This research comprises two main lines of investigation: (1) describing the phenomena of verb metaphor (i.e., what patterns of meaning change occur when people interpret verb metaphors, as evidenced by their paraphrases), and (2) investigating the cognitive processes underlying comprehension that drives this behavior. In Chapter 1, we demonstrate *the verb mutability effect*, showing that people strongly prefer to interpret semantically-strained intransitive sentences (sentences where the verb is paired with an unexpected noun type) like *The motor complained* by changing the meaning of the verb while preserving the meaning of the noun (e.g., *The engine made a strange sound*). In Chapter 2, we delineate this pattern more specifically by showing that verbs follow a *minimal subtraction* pattern of meaning change, wherein the verb changes meaning only as far as necessary to accommodate the paired noun, preferentially altering domain-specific meaning components before more domain-general, abstract ones. In Chapter 3, we propose and test a novel process account of verb metaphor comprehension: that they are understood as analogical comparisons between the event denoted by the verb and an event activated by the noun, processed via structure-mapping. We aim to

show that viewing verb metaphor as a species of analogy serves to account for both the phenomena of verb metaphor described in Chapters 1 and 2, as well as connecting more broadly to findings from the analogical reasoning literature and language evolution.

## Acknowledgements

First and foremost, I thank my wife Erinn, without whom this dissertation almost certainly would not have happened. Thank you for your sacrifice, unwavering support and belief in me, and for showing me every day what relentless excellence looks like. You set the bar to which I try to rise. How lucky I am to be married to my greatest inspiration.

To my advisor, Dedre Gentner: what an incredibly special experience working with you over the past six years has been. Thank you for all your mentorship, and for your ceaseless dedication to making me into a cognitive scientist. What a privilege it's been to sit across the table from a giant in the field and discuss how the mind works, week after week! These are memories I will always treasure. Thank you for always taking my ideas seriously and the countless hours you've spent helping grow them into this thesis.

Thank you also to my other two committee members, Sid Horton and David Rapp, for your thoughtful feedback and also your kind encouragement. This thesis (and much prior work) has benefitted greatly from your time and expertise. I also owe a great deal of gratitude to my lab mates past and present, who have patiently listened to me present this work an ungodly number of times, without ever once complaining. Thank you Nina Simms, Sean Zheng, Apoorva Shivaram, and Fanyi Mo for many years of thoughtful feedback (and for making the lab a fun place to work). I hope you all will always remember that verbs change meaning more than nouns. And to the many hardworking undergraduate RAs who coded the data in this thesis: thank you for all the hours you spent making obscure judgments about bizarre sentences—the findings in this dissertation rest on your shoulders!

My time at Northwestern would not have been such an amazing experience without my fellow grad students. I am grateful to have gone through this experience alongside so many amazing and talented peers and friends whom I admire greatly. Thank you especially to my fellow cognitive peeps and friends Naomi Polinsky and Nikita Salovich, for all your support and advice, and to my mentor and friend, Natalie Gallagher, for your guidance and for always making me feel included.

Mom, Dad, Katie—as if a few sentences could ever capture my love and gratitude. Not everyone is so lucky as to have a family that believes in them, fights for them, and who always reminds them they can accomplish great things. Thank you for your unconditional love, support, and guidance. None of this would've happened without you.

# Dedication

To Erinn, my lighthouse in the storm

# Contents

Abstract ........ 2
Acknowledgements ........ 4
List of Figures ........ 13
List of Tables ........ 14
Overview of this dissertation ........ 15
Chapter 1 ........ 18
1 Introduction ........ 18
1.1 The verb mutability effect ........ 19
1.2 Processes underlying verb mutability ........ 22
1.3 Behavioral approaches to assessing semantic adjustment ........ 25
1.4 Using vector space word embedding models to assess semantic adjustment ........ 26
1.5 Sense selection vs online adjustment ........ 28
2 Experiment 1 ........ 29
2.1 Method ........ 29
2.1.1 Participants ........ 29
2.1.2 Materials & Design ........ 29
2.1.3 Design ........ 31
2.1.4 Procedure ........ 32
2.1.5 Coding ........ 32
2.1.6 Assessing Semantic Adjustment ........ 33
2.2 Results ........ 34
2.2.1 Obtaining direct ratings of semantic strain ........ 36
2.3 Discussion ........ 38
3 Experiment 2 ........ 41
3.1 Method ........ 42
3.1.1 Participants ........ 42
3.1.2 Materials ........ 42
3.1.3 Experimental Design ........ 43
3.1.4 Procedure ........ 44
3.1.5 Coding ........ 44

3.1.6 Assessing Semantic Adjustment ............ 44
3.2 Results ............ 44
3.3 Discussion ............ 47
3.3.1 Comparing the word2vec results with human judgments ............ 50
4 Experiment 3 ............ 51
4.1 Method ............ 52
4.1.1 Participants ............ 52
4.1.2 Materials & Design ............ 52
4.1.3 Procedure ............ 53
4.2 Results ............ 54
4.2.1 Scoring ............ 54
4.2.2 Analysis ............ 54
4.3 Discussion ............ 56
4.3.1 Qualitative differences in noun and verb changes ............ 57
5 General Discussion ............ 63
5.1 Verbs change more than nouns ............ 63
5.2 Online adjustment drives verb mutability ............ 64
5.3 Qualitative differences in noun and verb change ............ 65
5.4 Characterizing verb meaning change ............ 65
5.5 Mutability and meaning change over time ............ 67
5.6 Why do verbs change more than nouns? ............ 69
5.6.1 Syntactic influences: Word order ............ 69
5.6.2 Pragmatic influences: Predicate role ............ 69
5.6.3 Relationality and the predicate role combine to drive verb mutability and online adjustment ............ 72
5.7 Implications and Future Work ............ 73
6 Conclusion ............ 74
Chapter 2 ............ 75
1 Introduction ............ 75
1.1 Minimal Subtraction ............ 75
2 Experiment 1A ............ 79

2.1 Method ..... 79
2.1.1 Participants ..... 79
2.1.2 Materials ..... 80
2.1.3 Design ..... 81
2.1.4 Procedure ..... 81
2.1.5 Coding ..... 82
2.1.6 Assessing semantic adjustment ..... 83
2.2 Results ..... 84
2.3 Discussion ..... 87
2.3.1 Paraphrase exclusions ..... 89
2.3.2 Differing patterns of change by verb class ..... 90
3 Experiment 1B ..... 91
3.1 Method ..... 91
3.1.1 Participants ..... 91
3.1.2 Materials ..... 91
3.1.3 Design ..... 91
3.1.4 Procedure ..... 92
3.2 Results ..... 92
3.2.1 Strain ratings ..... 93
3.2.2 Materials ..... 95
3.2.3 Modeling word2vec score as a function of strain ..... 95
3.3 Discussion ..... 97
4 Experiment 2 ..... 97
4.1 Method ..... 98
4.1.1 Participants ..... 98
4.1.2 Materials ..... 98
4.1.3 Design ..... 100
4.1.4 Procedure ..... 101
4.2 Results ..... 105
4.3 Discussion ..... 109
5 Experiment 3A ..... 112

5.1 Method ... 113
5.1.1 Participants ... 113
5.1.2 Design ... 113
5.2 Results ... 115
5.3 Discussion ... 118
6 Experiment 3B ... 120
6.1 Method ... 120
6.1.1 Participants ... 120
6.1.2 Materials ... 120
6.1.3 Design ... 120
6.1.4 Procedure ... 121
6.2 Results ... 121
6.3 Discussion ... 123
7 General Discussion ... 124
7.1 Implications for verb mutability ... 126
8 Conclusion ... 127
Chapter 3 ... 129
1 Introduction ... 129
1.1 Verb metaphor processing ... 129
1.1.1 Structure-mapping in analogy ... 130
1.1.2 Structure-mapping in verb metaphor ... 131
1.2 Noun event schemas in literal sentence processing ... 134
1.3 Predictions of the structure-mapping model of verb metaphor ... 137
1.3.1 Prediction 1: The verb mutability effect ... 137
1.3.2 Prediction 2: Minimal subtraction ... 139
1.3.3 Prediction 3: The verb event generates inferences about the nature of the noun event 143
1.3.4 Prediction 4: Event similarity predicts processing time ... 144
1.3.5 Prediction 5: Prior activation of relevant event structure facilitates comprehension 145
1.4 Summary ... 146

2 Experiment 1 .......... 147
2.1 Method .......... 152
2.1.1 Participants .......... 152
2.1.2 Materials .......... 152
2.1.3 Design .......... 157
2.1.4 Procedure .......... 157
2.2 Results .......... 159
2.3 Discussion .......... 162
3 Experiment 2 .......... 168
3.1 Method .......... 171
3.1.1 Participants .......... 171
3.1.2 Materials .......... 171
3.1.3 Design .......... 178
3.1.4 Procedure .......... 179
3.2 Results .......... 180
3.2.1 Prime and paraphrase event coding .......... 183
3.3 Discussion .......... 189
3.3.1 Ambiguity of the primes .......... 190
3.3.2 Noun event schema selection .......... 191
3.3.3 Summary .......... 192
4 General Discussion .......... 193
4.1 Evidence for the structure-mapping model of verb metaphor .......... 194
4.1.1 Prediction 1: The verb mutability effect .......... 194
4.1.2 Prediction 2: Minimal subtraction .......... 195
4.1.3 Prediction 3: The verb event generates inferences about the nature of the noun event 197
4.1.4 Prediction 4: Event similarity predicts processing time .......... 198
4.1.5 Prediction 5: Prior activation of relevant event structure facilitates comprehension 199
4.1.6 Summary of predictions .......... 202
4.2 Theoretical implications .......... 203

4.2.1 A unified framework for different types of metaphors ........ 203
4.2.2 Language evolution and the Career of Metaphor ........ 205
5 Conclusion ........ 207
Summary of Chapters ........ 208
Chapter 1: The verb mutability effect ........ 208
Chapter 2: Minimal subtraction ........ 209
Chapter 3: Verb metaphors are analogies ........ 210
Parting words: As verb metaphor goes, so goes cognition? ........ 212
References ........ 214
Appendices ........ 230

# List of Figures

Figure 1. *Grid showing stimuli noun and verbs from Gentner & France (1988)* ........................ 31
Figure 2. *Noun and verb similarity scores from Experiment* ........................................................ 36
Figure 3. *Stimulus matrix for Chapter 1, Experiment 2* ................................................................ 43
Figure 4. *Fitted model plots showing the effect of strain and polysemy on word2vec scores for verbs and nouns in Chapter 1, Experiment 2.* ................................................................. 46
Figure 5. *Fitted models showing probability of resurfacing for verbs and nouns in Chapter 1, Experiment 3* ........................................................................................................................ 56
Figure 6. *Tallies for the metaphoric (analogous), associative (metonymic), and taxonomic categories for nouns and verbs from the Chapter 1 qualitative analysis* .................................... 62
Figure 7. *Fitted model plots from Chapter 2, Experiment 1A. Each panel represents the results for items from that verb type.* ..................................................................................................... 86
Figure 8. *Mean strain ratings by noun type and verb class from Chapter 2, Experiment 1B.* ..... 94
Figure 9. *Fitted models using the continuous strain measure.* ..................................................... 97
Figure 10. *Retrace probabilities by verb class for Chapter 2, Experiment 3A.* ......................... 117
Figure 11. *Retrace probabilities by verb class for Chapter 2, Experiment 3B.* ......................... 122
Figure 12. *Crossing noun event and verb event domains to create within- and cross-domain verb metaphors for Chapter 3, Experiment 1.* ................................................................................. 149
Figure 13. *Results from the pilot study..* ...................................................................................... 150
Figure 14. *Trial structure for Chapter 3, Experiment 1.* .............................................................. 159
Figure 15. *Reaction time results from Experiment 1 showing a significant effect of semantic strain on comprehension times.* .............................................................................................. 162
Figure 16. *Strain, RT, and word2vec results broken down by noun-verb event similarity.* ........ 167
Figure 17. *Trial structure for Chapter 3, Experiment 2.* .............................................................. 179
Figure 18. *Mean comprehension times from Chapter 3, Experiment 2 by prime type.* ............. 182
Figure 19. *Illustrating the event coding process from Chapter 3, Experiment 2* ....................... 185
Figure 20. *Results of event coding study from Chapter 3, Experiment 2* .................................. 187
Figure 21. *Experimental design for target items from the pilot and Chapter 3, Experiment 1 showing noun event domains crossed with verb event domains with example sentences for each condition in the cells.* .......................................................................................................... 244
Figure 22. *Pilot results.* ................................................................................................................ 245

# List of Tables

Table 1. *Example paraphrases from Chapter 1, Experiment 1* .................................................... 35
Table 2. *Example paraphrases from Chapter 1, Experiment 2* .................................................... 48
Table 3. *Number of resurfacings (hits) vs non-resurfacings (misses) for nouns and verbs from Experiment* ........................................................................................................................... 54
Table 4. *Codes used in the qualitative analysis.* ........................................................................ 60
Table 5. *Nouns and verbs used in Chapter 2, Experiment 1* ........................................................ 80
Table 6. *Results of the paraphrase coding task from Chapter 2, Experiment 1* ........................... 83
Table 7. *Results of Type III test of fixed effects for Chapter 2, Experiment 1A models*. .............. 86
Table 8. *Results of pairwise contrasts of word2vec scores for each noun type within each verb model* ............................................................................................................................... 87
Table 9. *Example paraphrases from Chapter 2, Experiment 1A* .................................................. 88
Table 10. *Pairwise contrasts of strain rating by noun type for each verb class* ........................... 94
Table 11. *Model summaries for each of the three verb classes*. .................................................. 96
Table 12. *Coding scheme used in Chapter 2, Experiment 2* ........................................................ 99
Table 13. *Example domain-general paraphrases from Chapter 2, Experiment 2* ...................... 104
Table 14. *Response tallies and percentages by verb class from Chapter 2, Experiment 2* ........ 105
Table 15. *Pairwise contrasts from Chapter 2, Experiment 2* ..................................................... 107
Table 16. *Overall response tallies and percentages from Chapter 2, Experiment 2* .................. 108
Table 17. *Retrace probabilities from Chapter 2, Experiment 3A* ............................................... 117
Table 18. *Contrasts by verb class from Chapter 2, Experiment 3A* ........................................... 118
Table 19. *Retrace Probabilities from Chapter 2, Experiment 3B* ............................................... 122
Table 20. *Model results from Chapter 2, Experiment 3B* .......................................................... 122
Table 21. *Example paraphrases from Chapter 2* ........................................................................ 132
Table 22. *Example paraphrases illustrating the minimal subtraction pattern* ........................... 141
Table 23. *Noun and verb stimuli for Chapter 3, Experiment 1* .................................................. 153
Table 24. *Sample of metaphoric items from Chapter 3, Experiment 1* ...................................... 154
Table 25. *Example paraphrases from Chapter 3, Experiment 2* ................................................ 162
Table 26. *Target items and primes from Chapter 3, Experiment 2* ............................................ 172
Table 27. *Top-ranked events from the event norming study* ...................................................... 177
Table 28. *Example paraphrases from Chapter 3, Experiment 2* ................................................ 180
Table 29. *Mean comprehension times by prime type for Experiment 2* ..................................... 182

## Overview of this dissertation

Metaphor is a pervasive phenomenon in language and cognition. Psycholinguistic work has made great advances in our understanding of metaphor (Blank, 1988; Bowdle & Gentner, 2005; Chiappe, Kennedy, & Smykowski, 2003; Gentner & Wolff, 1997, 2000; Gibbs, 1992; Giora, 1997; Glucksberg & Keysar, 1990; Glucksberg, McGlone, & Manfredi, 1997; Katz, 1989; Keysar et al., 2000; Ortony, 1979; Thibodeau & Durgin, 2011; Trick & Katz, 1986; Tourangeau & Rips, 1991; Wolff & Gentner, 2000, 2011), but this work has focused almost exclusively on noun metaphor—that is, on metaphors (and similes) of the form *X is (like) a Y*--e.g., *That surgeon is a butcher*. Relatively little work has examined how verb metaphors are processed (but see Cardillo et al., 2012; Chen et al., 2008; Desai et al., 2011; Ronderos et al., 2021; Stamenković et al., 2020). This is unfortunate, as there is evidence that verb metaphors are more common than noun metaphors (Jamrozik et al., 2013; Krennmayr, 2011). In addition, other important differences between nouns and verbs have been identified in the literature, such as that verbs more readily change meaning than nouns both in daily language (Gentner & France, 1988; Kersten & Earles, 2004), and also over historical time periods (Dubossarsky et al., 2016). Since metaphor is widely thought to be an important driver of language over time (e.g., Bowdle & Gentner, 2005; Hopper & Traugott, 2003; Xu et al., 2017), a better understanding verb metaphor processing has important implications for our understanding of both real-time sentence processing and language change over time.

The goal of this dissertation is to characterize the nature of verb metaphor. This consists of two main lines of investigation: that of outcome and that of underlying process. First, the phenomenology of verb metaphor is explored: when verbs extend their meanings metaphorically, what does that look like? Chapters 1 and 2 tackle this question by examining the patterns of meaning change that surface in people's interpretations of various kinds of intransitive verb metaphors of the form *The noun verb-ed* (e.g., *The motor complained, The wagon limped, The rumor trudged, The wisdom burped*). The results indicate that verbs follow a predictable pattern of meaning adaptation that depends intimately on the semantic relationship between the noun and the verb.

This delineation verb meaning change lays the foundation for the second line of investigation: understanding the cognitive processes underlying verb metaphor comprehension—i.e., that which ultimately produces this observed behavior. This is the goal of the third and final chapter of this dissertation, where a novel process model of verb metaphor comprehension is proposed and tested. The ultimate objective is to show that verb metaphors are a species of analogy and are processed as such, and that applying the analogical framework to verb metaphor comprehension (specifically, structure-mapping theory; Gentner, 1983) predicts the phenomena of verb metaphor described in Chapters 1 and 2 well.

In Chapter 1, we report three experiments that investigate the foundational phenomena of verb metaphor. First, we replicate and expand upon prior work that found evidence for a *verb mutability effect* (e.g., Gentner & France, 1988; Kersten & Earles, 2004), the phenomenon whereby people prefer to interpret sentences in which the verb receives an atypical argument type (as is the case for verb metaphors, e.g., *The motor complained*) by changing the meaning of the verb and preserving the meaning of the noun (e.g., *The engine made a strange sound*). Second, we find that this effect is driven largely by online adjustments to verb representations, rather than by sense selection from among preexisting senses in LTM. Third, we show that these verb meaning changes are predominantly metaphorical/analogical meaning extensions.

In Chapter 2, we report five experiments that explore more deeply how verbs change their meanings during metaphor comprehension. Specifically, we propose and find evidence for a set of principles that build upon Gentner and France's (1988) *minimal subtraction* hypothesis: that verbs adapt their meanings to their noun arguments in a fine-grained manner such that they are abstracted only to the extent necessary to accommodate the noun, and this abstraction follows a particular qualitative pattern that moves from the adjustment of domain-specific meaning components to more domain-general ones.

In Chapter 3, we synthesize our findings from Chapters 1 and 2 and propose a novel process model of verb metaphor comprehension: namely, that verbs are understood as analogical comparisons between the event denoted by the verb and an event that is activated by its noun argument, processed via structure-mapping (Gentner, 1983; Gentner & Markman, 1997). We discuss how this theory fits within the structure-mapping framework of analogy and predicts our findings from Chapters 1 and 2. We conclude with two experiments that further support this model by satisfying the online processing predictions of structure-mapping. Viewing verb metaphors as a type of analogy holds the promise of unifying the comprehension of several different types of metaphor (e.g., noun metaphors, extended metaphors) in a single framework, and has important implications for theories of language change over time.

# Chapter 1

## 1 Introduction

Metaphoric uses of verbs are frequent in everyday language. We use phrases like *surmounting a problem*, *eating our words*, or *stumbling on a solution* in ordinary conversation. Research in cognitive linguistics has also documented large systems of conventional metaphors that pervade language, and verb metaphors feature prominently among these (Clausner & Croft, 1997; Fauconnier & Turner, 1998; Gibbs, 2006; Lakoff & Johnson, 1980, 2008; see also Steen, 2007). For example, Lakoff and Johnson (2008) list many verb metaphors among the expressions that constitute the TIME IS MONEY metaphoric system:

*You're <u>wasting</u> my time.*

*This gadget will <u>save</u> you hours.*

*I don't <u>have</u> the time to <u>give</u> you.*

*How do you <u>spend</u> your time these days*?

*That flat tire <u>cost</u> me an hour.*

*I've <u>invested</u> a lot of time in her.*

Psychological research on metaphor processing has largely focused on noun-noun metaphors of the form *An X is a Y* (e.g., *my job is a jail, my lawyer is a shark*) (Blank, 1988; Bowdle & Gentner, 2005; Chiappe & Kennedy, 2003; Gentner & Wolff, 1997, 2000; Gibbs, 1992; Giora, 1997; Glucksberg & Keysar, 1990; Glucksberg, McGlone, & Manfredi, 1997; Jones & Estes, 2006; Katz, 1989; Keysar et al., 2000; Ortony, 1979; Shen, 1989; Thibodeau & Durgin, 2011; Trick & Katz, 1986; Tourangeau & Rips, 1991; Tourangeau & Sternberg, 1981, 1982; Wolff & Gentner, 2011). Psycholinguistic research on metaphoric uses of verbs is comparatively rare (but see Cardillo et al., 2010, 2012; Cardillo, Watson & Chatterjee, 2017; Gentner & France, 1988; Stamenković, Ichien, & Holyoak, 2019; Torreano et al., 2005).

The dearth of research on verb metaphor is unfortunate, as there is evidence that verb metaphors are more common than noun metaphors (Jamrozik et al., 2013; Krennmayr, 2011). Krennmayr (2011) conducted a corpus analysis over 186,688 words of text spanning multiple registers (news, academic, fictional, and conversational) and found that verb metaphors were more frequent than noun metaphors in all registers. Jamrozik et al. (2013) compared verbs and nouns in terms of what they called *metaphoric potential*—the likelihood that a word will be used metaphorically. For each word, the researchers randomly sampled 20 sentences from the Corpus of Contemporary American English (Davies, 2009) and asked judges to rate the metaphoricity of the selected word in the sentence. The results showed that, controlling for concreteness and imageability, verb uses were rated as significantly more metaphoric than noun uses.

## 1.1 The verb mutability effect

An early approach to studying verb metaphor in psychology was research on the *verb mutability effect* in sentence processing (Gentner, 1981; Gentner & France, 1988; Reyna, 1980). Verb mutability refers to the phenomenon whereby, under conditions of semantic strain, the verb is more likely to adapt its meaning to the noun than the reverse. Gentner and France (1988) investigated this effect by having participants paraphrase simple intransitive sentences that varied in semantic strain. They selected eight nouns and eight verbs and combined them factorially to generate 64 sentences (see Figure 1). The nouns and verbs were selected such that some combinations generated sentences in which the verb received its expected subject type, resulting in *semantically unstrained*, or literally interpretable, sentences (e.g., *The daughter agreed*), while other combinations generated sentences in which the noun violated the verb's expected subject type, resulting in *semantically strained* sentences that were not literally interpretable (e.g., *The car agreed*).

Gentner and France found that when paraphrasing, people altered the verb meanings more than the noun meanings overall, and that this effect increased with semantic strain. Thus, while participants generally preserved the standard meaning of both the noun and the verb when interpreting unstrained sentences (e.g., paraphrasing *The daughter agreed* as *The girl concurred*), there was a marked preference for changing the meaning of the verb, and not the noun, when interpreting strained sentences (e.g., paraphrasing *The car agreed* as *The automobile was easily controlled*). In other words, under conditions of semantic strain, people tended to interpret the verb metaphorically and the noun literally.

Further evidence for verb mutability in sentence comprehension comes from research on memory. Work going back decades has demonstrated that verbs are harder to remember than nouns, in both free recall and recognition tasks. (Clark, 1966; Earles et al., 1999; Earles & Kersten, 2000, 2017; Horowitz & Prytulak, 1969; Kersten & Earles, 2004). Earles et al., (1999) showed that in free recall tests of verb-noun pairs (e.g., *wave-hand*), participants were less able to recall the original verb than the original noun. Kersten and Earles (2004) tested memory for sentences and found the same pattern for recognition: verbs were recognized less well than were nouns overall. More specifically, they found that verbs were significantly less likely to be recognized when combined with a different noun at test than at encoding (e.g., when given *The quarter bounced* at encoding and *The ball bounced* at test). Nouns, however, were recognized equally well at test, regardless of whether the paired verb was the same or different as at encoding (e.g., *The quarter bounced* at encoding and *The quarter rolled* at test). Linking their results with Gentner's (1981) verb mutability hypothesis, Kersten and Earles interpreted their findings as evidence that verb encoding is more variable than noun encoding, with the noun providing a stable semantic context to which the verb's meaning is adapted.

Verb mutability has also been demonstrated in studies of meaning coercion imposed by syntactic constraints. For example, in *Art sneezed the foam off his beer*, the normally-intransitive verb *sneeze* acquires a transitive meaning by virtue of appearing in the transitive double-object construction (Goldberg, 1995). Kaschak and Glenberg (2000) showed that the interpretation of novel denominal verbs (nouns used in a novel way as verbs, see Clark & Clark, 1979) depends on the syntactic construction used. For example, when given the double-object construction *Lyn crutched Tom her apple to prove her point*, participants interpreted the verb to mean that Lyn conveyed her apple to Tom using a crutch. When given the transitive construction *Lyn crutched her apple to prove her point to Tom*, participants interpreted *crutched* as meaning simply that Lyn acted upon the apple in some way using the crutch. In either case, however, the verb's meaning is adjusted to the semantic context provided by construction and the surrounding nouns.

There is also indirect evidence for verb mutability from historical studies of language change over time (Dubossarsky, Weinshall, & Grossman, 2016; Sagi, 2019). Dubossarsky et al., (2016) compared rates of change for nouns, verbs, and adjectives from 1850 to 2000. They found that verbs changed meaning at a faster rate than both nouns and adjectives over the period of analysis. Dubossarsky et al. suggested that verbs' greater rate of change over time in language evolution might be driven by their greater mutability in processing, citing Gentner and France's (1988) findings.

## 1.2 Processes underlying verb mutability

Thus, there is evidence from studies of sentence processing, sentence memory, and diachronic meaning change that verbs have a greater propensity for semantic adjustment in context than do nouns. But how does this happen? In general, there are two prominent accounts of how meaning adjustments under semantic strain can take place: sense selection (often called *word sense disambiguation*) and online adjustment (also called *sense creation*), (e.g., Clark & Gerrig 1983; Frisson & Pickering, 2007; Gerrig, 1989; Gerrig & Bortfield, 1999; Lenat & Guha, 1989; Pritchard, 2019; Rapp & Gerrig, 1999; Vicente, 2018; Vicente & Falkum, 2017). There is little dispute that people often draw on existing word senses to resolve meaning when the typical literal interpretation of a word is contextually implausible. However, Gentner (1981; Gentner & France, 1988) interpreted their verb mutability findings as indicating that verbs are more likely to undergo *online adjustment* to their representations than are nouns. They noted that the online adjustment view provides a way to explain novel metaphoric extensions. For example, interpreting *The car agreed* as *The vehicle drove well* would seem to require online modification of the verb, as *agreed* lacks a conventional metaphoric sense that could be accessed from memory and applied to *car*.

The online adjustment view can also potentially explain the relationship between metaphor and language change. Metaphor is widely believed to be an important force in how words change meaning over time, including how words gain new senses (Bowdle & Gentner, 2005; Cardillo et al., 2012; Chatterjee, 2010; Dirven, 1985; Heine, 1997; Hopper & Traugott, 2003; Jamrozik et al., 2016; Joseph, Hock, & Joseph, 1996; Sweetser, 1990; Traugott, 1988; Wolff & Gentner, 2011; Xu, Malt, & Srinivasan, 2017). There is evidence suggesting that many conventional metaphoric senses originated as novel extensions of literal concepts. For example, *heart* referred literally to an organ before later gaining metaphoric senses such as *the center of things* (Dirven, 1985). Similarly, *bridge* originally referred only to a structure linking two physical locations, but now is frequently used metaphorically to mean anything that links two abstract situations (Zharikov & Gentner, 2002). Thus, online adjustment may be an important driving force behind polysemy.

However, before embracing the online adjustment account of verb mutability, we must first address an alternative explanation—namely, selection among existing word senses. There is evidence that, controlling for frequency, verbs are more polysemous than nouns (Gentner, 1981; Miller & Fellbaum, 1991). Thus, it could be that under semantic strain, it is easier on average to find an appropriate word sense for the verb than for the noun. On this account—the *sense selection* account of verb mutability—meaning adjustment occurs primarily by selecting among preexisting senses, rather than by deriving new meaning online. If sense-selection is the primary driver of verb mutability, it would suggest that the verb mutability effect in sentence processing is really a verb polysemy effect. Given this concern, we evaluated the polysemy of the stimuli used by Gentner and France (1988) and by Kersten and Earles (2004) by counting the number of senses listed for each word in WordNet (Miller, 1995). An independent-samples $t$-test showed that in both cases, the verbs used were significantly more polysemous than the nouns ($p$s < .05), leaving open the possibility that differences in polysemy could explain both studies' results.

Thus, sense selection may be the primary of verb mutability, instead of—or in addition to—online adjustment.

In this research, we investigate this question by systematically varying both noun and verb polysemy and semantic strain. As in Gentner and France's (1988) paradigm, participants paraphrased simple intransitive sentences that varied in semantic strain, which were then evaluated for the degree of noun and verb meaning change that occurred. Unlike Gentner and France, however, we selected nouns and verbs such that half were low-polysemy (1-2 senses) and half were high-polysemy (7-13 senses). Stimuli were generated by combining the nouns and verbs factorially so that across the full set of sentences, every possible combination of low- and high-polysemy nouns and verbs was realized. If sense selection drives verb mutability, we would expect (a) symmetrical patterns of change for nouns and verbs; and (b) a significant relationship between the polysemy of a word and the degree of change under strain. Alternatively, if online adjustment is the primary driver of verb mutability, we would expect (a) greater change in verbs than in nouns; (b) greater change in verb meaning as strain increases; and (c) relatively minor effects of polysemy on the degree of change.

Before describing the experiments, however, we must confront the issue of how to assess degree of semantic change. How does one objectively determine the relative degree of change in the noun vs. the verb when someone paraphrases, say, *The car agreed* as *The automobile was easily controlled*? Gentner and France (1988) approached this issue using three different behavioral measures. All three of them provided evidence for the verb mutability effect. However, each had significant drawbacks. We discuss these methods below.

## 1.3 Behavioral approaches to assessing semantic adjustment

In the *divide and rate* method, a group of raters was instructed to divide each paraphrase into the part that came from the noun and the part that came from the verb; they then rated the similarity of each part to the original noun and verb, respectively. The results indicated that the part that came from the verb tended to change more than the part that came from the noun. However, this method was time-consuming and labor-intensive. Worse, judges often could not agree on how to divide the sentences, resulting in a high amount of data loss. For example, in paraphrasing *The car limped* as *the badly-functioning vehicle struggled to drive*, the modifier *badly-functioning* and the verb phrase *struggled to drive* seem to owe their presence to *both* the original noun and verb, making it unclear how to divide them into noun- and verb-derived components.

Gentner and France devised two further methods that did not require dividing the paraphrases into parts: a *retrace* task and a *double paraphrase* task. In the retrace task, the paraphrases were given to a new group of judges, along with a list of either the initial stimulus nouns or the initial stimulus verbs. For each paraphrase, they indicated which noun or verb they thought had appeared in the original sentence. Participants were more accurate for nouns than for verbs, indicating that the initial noun meanings had changed less than the initial verb meanings. However, this method had the drawback that the lists of initial stimuli were not designed to test for degree of semantic change in either nouns or verbs.

In the double paraphrase task, the original paraphrases were given to a new set of participants to paraphrase. The rate at which the initial nouns and verbs resurfaced in the new paraphrases was taken to reflect the degree of meaning adjustment that had occurred: the greater the change in a word's meaning in the initial paraphrase, the less likely the word was to resurface again in the second paraphrase. Consistent with the verb mutability effect, nouns were more likely to resurface than verbs. The double paraphrase task had the advantage of being the most hands-off approach of the three; however, it too resulted in substantial data loss: only 19% of nouns and 4% of verbs resurfaced in this method.

In sum, all three of Gentner and France's methods indicated greater change of meaning in the verb than in the noun. However, none of them was ideal: the divide-and-rate and double-paraphrase techniques were liable to considerable data loss, and the retrace method was limited by the particular word sets chosen initially. Therefore, in the present work, we turn to new techniques for computing relatedness between texts that have since emerged out of work in computer science and computational linguistics: *vector space word embedding models.*

### 1.4 Using vector space word embedding models to assess semantic adjustment

In this research we use *word2vec* (Mikolov, Chen, et al. 2013; Mikolov, Sutskever, et al., 2013), a vector space word embedding model (WEM) to assess degree of semantic change between a stimulus word and its paraphrase. WEMs take as their foundation the notion that words are similar or related to the extent that they appear in similar contexts. WEMs are trained on a large corpus and derive a vector representation for each word (typically 100 to 300 dimensions) based on global distributions of co-occurrence patterns in the corpus. (An overview of the free parameter choices involved in training and using word2vec is included in the supplementary material] The similarity between any two word meanings is typically calculated by taking the cosine of the angle between their two associated vectors, resulting in a score between -1 and 1. Scores close to 1 are taken to indicate high levels of similarity, and scores close to 0 indicate low levels of similarity.

The logic of our approach is to use word2vec's cosine similarity scores to estimate the similarity between the paraphrase and the original verb (or noun), and therefore the degree of change under paraphrase. A high cosine score between a verb or noun and its paraphrase is taken to indicate that the meanings are highly similar, and therefore that the initial word's meaning was not much altered in that paraphrase. By the same logic, a low cosine similarity score is taken to indicate a high degree of semantic change (see details in Experiment 1).

We used pretrained word2vec vectors publicly available from Google, which were trained on a 100-billion word subset of the Google News corpus, resulting in a vocabulary of over 3 million words.[1] We chose word2vec over other WEMs based on a study by Pereira et al. (2016) which compared several prominent off-the-shelf WEMs, including word2vec, GloVe (Pennington, Socher, & Manning, 2014), and LSA (Landauer & Dumais, 1997). Word2vec and GloVe were the best performing of the test set, providing the highest correlations with human similarity judgments on almost all of the 17 datasets tested. We chose word2vec over GloVe because it is more widely used than any other WEM; its two foundational papers have been cited more than a combined 53,000 times. We used word2vec's set of pretrained vectors rather than training our own version because we wanted a general-purpose language corpus, not one aimed at verb (or noun) metaphor. Using pretrained vectors also minimizes the opportunity for inadvertently tailoring the space to fit the predicted results, and removes the need to make free-parameter choices. A further advantage is that the results can be more easily replicated and compared.

[1] Available at https://code.google.com/archive/p/word2vec/.

## 1.5 Sense selection vs online adjustment

Using word2vec, we investigated the question of sense selection vs online adjustment in verb mutability. If mutability is driven chiefly by sense selection, then polysemy should predict mutability: low-polysemy nouns and verbs should show little semantic change, while high-polysemy nouns and verbs will show substantial meaning change under semantic strain. We would further expect that when a high-polysemy noun is combined with a low-polysemy verb, the noun–and not the verb–should change meaning. Overall, when controlling for polysemy, there should be little or no difference in meaning change by syntactic class.

In contrast, the online adjustment view posits that verb mutability results primarily from online processes that alter the verb's typical representation to fit with the noun's meaning. In this case, we would expect by-class (rather than by-polysemy) differences. If online adjustment of verb meaning is the driver of verb mutability, verbs should change meaning more than nouns regardless of polysemy, and with noun meanings stable at both low- and high-polysemy. Of course, both processes may be involved, in which case we should find that high-polysemy words are more mutable than low-polysemy words, but that overall, verbs are more mutable than nouns.

The plan of this paper is as follows. In Experiment 1, we carried out a partial replication of Gentner and France's (1988) Experiments 1 and 2, but using word2vec to assess meaning change instead of human judges. The goal was to replicate the verb mutability effect and also to test the feasibility of using word2vec to assess semantic change in a sentence processing context. In Experiment 2, we compared the sense-selection and online adjustment accounts of mutability by testing polysemy as a predictor of meaning change, once again using word2vec to assess degree of change. In Experiment 3, we re-ran the paraphrases from Experiment 2, using human judges instead of word2vec to assess the degree of semantic change. The idea was to ascertain whether word2vec's results conform to human intuitions.

# 2 Experiment 1

In Experiment 1 we sought to replicate the findings of Gentner and France (1988) using word2vec to assess semantic adjustment. As in the original work, we asked participants to paraphrase sentences varying in semantic strain. We then used word2vec to assess the degree of change in the noun and in the verb, as described below.

## 2.1 Method

### *2.1.1 Participants*

121 university undergraduates completed the study in person in the laboratory. They received course credit in an introductory Psychology class for their participation. 5 were excluded for not being native English speakers, and 7 were excluded for failing the catch trial criteria, for a net of 109 participants.

### *2.1.2 Materials & Design*

Stimuli consisted of a subset of those used in Gentner and France's Experiments 1 and 2. Gentner and France generated stimulus sentences by combining 8 nouns with 8 intransitive verbs for a total of 64 different sentences, which can be visualized as forming a matrix (see Figure 1). The nouns consisted of two humans, two animals, two artifacts, and two abstract nouns. The verbs were matched to the nouns with respect to their preferred subject type. There were two verbs that prefer human subjects, two that prefer animals (or humans), two that prefer artifacts (or animals or humans), and two that prefer abstract nouns (or the other three categories). By arranging the nouns and verbs into a matrix, semantic strain can be varied systematically, as shown in Figure 1. When the noun meets the verb's selectional preference,[2] the result is a literal, unstrained sentence. But when the noun violates the verb's selectional preference, the result is a semantically strained (nonliteral) sentence. For example, *agree* prefers a human subject, so *The daughter agreed* is unstrained, but *The car agreed* is semantically strained.

We used Gentner and France's original stimuli with one modification: we excluded the abstract category, leaving 6 nouns and 6 verbs for a total of 36 of the original 64 sentences (see Figure 1). This was done for two reasons. First, it simplified and balanced the design such that each participant received an equal number of strained and unstrained sentences while seeing each stimulus noun and verb exactly once. Second, many of the original sentences involving abstract nouns seemed awkward (e.g., *The responsibility succeeded*). We were concerned that participants might not be able to provide meaningful interpretations of these sentences, which in turn might bias the results towards greater mutability (i.e., they might result in high numbers of meaningless but nevertheless semantically-distant adjustments). Removing the abstract category, therefore, provides a stricter test of the verb mutability effect.

| | | Human | | Animal | | Artifact | | Abstract | |
|---|---|---|---|---|---|---|---|---|---|
| | | agree | worship | shiver | limp | soften | cook | succeed | weaken |
| Human | daughter | | | | | | | | |

[2] By *selectional preference* we mean the semantic type(s) that conventionally act as the verb's subject—for example, some verbs typically require human subjects, while others allow a greater range of semantic types (e.g., Wilks, 1975). We use the term *selectional preference* rather than the term *selectional restriction* (e.g., Katz & Fodor, 1963), as it emphasizes the fact that verbs can accommodate arguments that don't match their preferred semantic types, as in the present research.

| | | | | | |
|---|---|---|---|---|---|
| | politician | *The daughter agreed* | *The daughter shivered* | *The daughter softened* | *The daughter succeeded* |
| Animal | mule<br>lizard | *The mule agreed* | *The mule shivered* | *The mule softened* | *The mule succeeded* |
| Artifact | car<br>lantern | *The car agreed* | *The car shivered* | *The car softened* | *The car succeeded* |
| Abstract | responsibility<br>courage | *The responsibility agreed* | *The responsibility shivered* | *The responsibility softened* | *The responsibility succeeded* |

Figure 1. Grid showing stimuli noun and verbs from Gentner & France (1988), with some examples of sentences generated from combining them. Shaded cells indicate semantically-strained combinations; unshaded cells indicate unstrained combinations. Noun-verb combinations used in Experiment 1 fall within the outlined box.

### *2.1.3 Design*

So that each participant saw each noun and verb exactly once, the 36 total stimulus sentences were divided into six different between-subject item groupings of six sentences each. Each grouping consisted of two strained and four unstrained sentences. Thus, the design was 6 (item grouping, between-subject) X 2 (item strain: strained vs. unstrained, within subject). Each participant saw each of the 6 nouns and 6 verbs exactly once. Two simple unstrained sentences were included as catch trials for checking attention and following directions; the criteria for excluding a subject was repeating a noun and/or verb in both of the catch trials or producing an obviously nonsensical answer in either. As each of the net 109 participants paraphrased 6 initial sentences, there were roughly 18 paraphrases per initial sentence.

### *2.1.4 Procedure*

Each participant was randomly assigned to one of the six item groupings. Participants completed the experiment individually, in person, on a computer. They first read instructions informing them that they would see a number of different sentences, and that they should provide a meaningful interpretation of each. They were explicitly instructed not to translate sentences mechanically (word-by-word), but rather to think of a plausible overall meaning for the sentence. To illustrate the difference between a mechanical and meaningful paraphrase, they were provided with an example of each. The full instructions can be found in Appendix A.

Sentences were presented one at a time in randomized order, and participants typed their responses. Once they had submitted a response for a sentence, they could not go back to previous responses.

### *2.1.5 Coding*

Two human coders, blind to the hypotheses, were used to exclude certain types of paraphrases from the analysis: blatantly noncompliant responses (e.g., paraphrasing *the daughter cooked* as *the child*), and responses that did not constitute a meaningful interpretation of the sentence. Two types of interpretations met this second criterion: (1) responses that described the context suggested by the initial sentence rather than actually interpreting it (e.g., paraphrasing *The mule shivered* as *It was a cold night*), and (2) mechanical, word-by-word paraphrases of strained sentences (e.g., paraphrasing *The lantern worshipped* as *The candle honored*). As noted above, participants had been explicitly instructed to try to interpret the intended meaning, not to deal with each word separately. Of course, for unstrained sentences (which are literally interpretable), a meaningful paraphrase is indistinguishable from a word-by-word paraphrase (e.g., paraphrasing *The daughter worshipped* as *The girl prayed*). Thus, coding for mechanical paraphrases was necessary only for the strained sentences; however, all paraphrases were coded for responses that described the situation and for noncompliant responses.

The two coders judged all 654 paraphrases. Each coder was presented with the original sentence and all corresponding paraphrases and indicated whether each paraphrase was meaningful, mechanical, describing the situation, or noncompliant. Coding was done in chunks wherein each judge coded a set of paraphrases independently, followed by a reconciliation session where the judges came to an agreement on any disparities. The judges were able to reach a final consensus on all items. Coding resulted in the exclusion of 128 paraphrases (91 mechanical, 24 describing the situation, and 13 noncompliant), leaving 526 of the original 654 paraphrases for the main analysis. Cohen's κ was run to determine interrater reliability. There was moderate initial agreement between the two judges, $\kappa = 0.66$, (95% CI, 0.57 to 0.75), $p < .001$. A summary of the results of the coding task is shown in Appendix B. After coding, an average of 14.61 paraphrases per item remained.

### *2.1.6 Assessing Semantic Adjustment*

For each paraphrase, word2vec was used to obtain two cosine similarity scores: a noun score and verb score, representing the amount of semantic adjustment the initial noun and verb underwent from original sentence to paraphrase, respectively. The scoring process was as follows. First, separate normalized vectors were obtained for each stimulus noun and verb. Next, a vector for each paraphrase was generated by averaging its normalized component word vectors.[3] The noun-change score was then computed by calculating the cosine similarity score between the original noun vector and the paraphrase vector; likewise, the verb-change score was calculated as the cosine similarity score between the original verb vector and the paraphrase vector.[4] Comparing the initial noun and verb to the entire paraphrase has the advantage of eliminating the need to divide paraphrases into components. For example, to assess the amount of meaning change that occurred for the noun and verb from the stimulus sentence *The lantern limped* to the paraphrase *The candle flickered*, the cosine of the angle between vector for *lantern* (the original noun) and the vector for *The candle flickered* (the participant paraphrase) was calculated, and likewise for *limped* (the original verb) and *The candle flickered*. The resulting noun and verb scores are .47 and .22, indicating that the verb's meaning changed more than the noun's in this paraphrase (recall that for WEMs, scores closer to 0 indicate a lower degree of similarity between items).

## 2.2 Results

To preview, the results bore out the two key findings necessary for a successful replication of the Gentner and France's (1988) findings: (1) overall, the change in meaning was greater for verbs than for nouns, and (2) this effect was greater for semantically strained sentences than for

[3] Approaches for representing componential meaning with word embeddings are currently limited (Finley, Farmer, & Pakhomov, 2017; Lenci, 2018) and embody a tension between ease of use and adequacy of meaning representation. The additive approach to compositional meaning we use here is the most widely-used (Blacoe & Lapata, 2012; Foltz et al., 1998; Lenci, 2018; Landauer & Dumais, 1997). It has the advantage of being simple and easy to implement; however, it ignores important complexities such as word order and other relational dependencies that affect word and phrase semantics. Despite these drawbacks, vector addition has been shown to perform as well as or better than more complex approaches (Blacoe & Lapata, 2012; Rimell et al., 2016; Lenci, 2018).

[4] Because the component vectors are normalized, the cosine of the angle between the noun vector and the stimulus sentence is equal to the cosine of the angle between the verb vector and the stimulus sentence. That is, *car* and *agreed* generate equal cosine similarity scores when compared to the vector for *The car agreed*. Thus, we interpret any difference between noun and verb score when compared to the *paraphrase* vector (e.g., *The candle flickered*) to represent a difference in degree of semantic change from stimulus to paraphrase.

unstrained sentences. As Figure 2a shows, verb meanings changed strongly in response to strain, while noun meanings remained stable. Table 1 shows example paraphrases for strained and unstrained sentences.

Table 1

*Example paraphrases from Experiment 1.*

| Condition | Stimulus Sentence | Paraphrase |
|---|---|---|
| Unstrained | *The daughter cooked* | *The girl made food* |
| | *The politician shivered* | *The statesman quivered* |
| | *The mule limped* | *The horse walked gingerly* |
| Strained | *The car agreed* | *The vehicle responded well to the driver* |
| | *The lantern limped* | *The candle flickered* |
| | *The lizard worshipped* | *The amphibian laid out in the sun* |

To test whether verb meanings changed more than noun meanings overall, a difference score for each paraphrase was calculated by subtracting the verb cosine score from the noun cosine score. Since lower word2vec scores indicate less relatedness between items, a positive difference score indicates greater verb change than noun change. Next, a linear mixed-effect model was fit, with difference score as the dependent measure, the intercept (representing the mean difference score) as the only fixed effect, and subjects and items as random effects.[5] The intercept was found to be significantly greater than 0, $B = .11$, SE = .02, $t = 5.35$, $p < .001$, indicating that, on average,

[5] In all mixed effect regressions described in this paper, the random effect structure was specified according to the procedure outlined in Bates et al. (2015). An initial model was fit with the maximal random effect structure, comprising random intercepts and slopes for subjects, and random intercepts for items (items were nested within condition). This structure was then simplified as far as possible via an iterative model comparison process. In most cases, an intercept-only model for subjects and items was sufficient. *P*-values for all models were obtained using Satterthwaite's approximation for degrees of freedom (see Luke, 2017).

verbs (M = 0.26, SD = 0.11) changed their meaning significantly more overall than nouns did (M = 0.38, SD = 0.15).

To test the effect of semantic strain on degree of meaning change, two additional models were fit: one for nouns and one for verbs. In both models, word2vec score was the dependent measure, strain (unstrained vs. strained) was the fixed effect, and subjects and items were included as random effects. For verbs, the effect of semantic strain was significant, $\beta = -.26$, SE = 0.08, $t = 3.09$, $p < .01$, indicating that verb meaning was adjusted to a greater extent in the strained condition (M = 0.21, SE = 0.02) than in the unstrained condition (M = 0.28, SE = 0.01). For nouns, there was no significant effect of semantic strain, $\beta = .07$, SE = 0.10, $t = 0.66$, $p = .51$. These results are shown in Figure 2a.

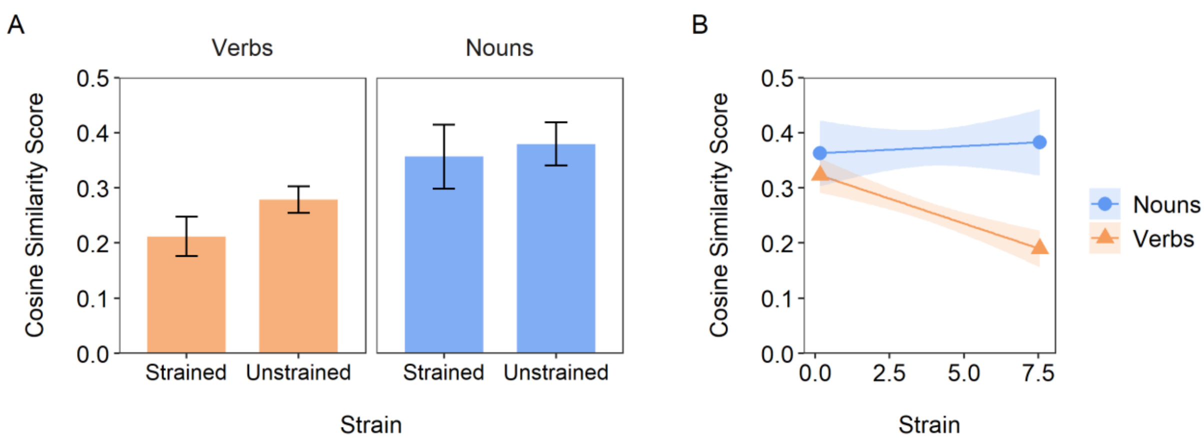



Figure 2. Noun and verb similarity scores from Experiment 1. Lower scores indicate greater semantic adjustment. Error bars/bands represent 95% confidence intervals. (A) Strain treated as a categorical predictor. (B) Strain as a continuous predictor, derived from the comprehensibility ratings described below.

### *2.2.1 Obtaining direct ratings of semantic strain*

In the analyses so far, we have followed Gentner and France's original procedure wherein strain was treated as a categorical predictor, with sentences categorized as either strained or unstrained based on whether the verb received its expected noun subject type (represented by the shaded squares in Figure 1). Although this provides a principled way to classify strained vs unstrained sentences, treating strain as a dichotomous predictor fails to capture the intuition that some sentences are more strained than others (e.g., consider *The mule agreed* vs *The lantern agreed*).

To provide a finer-grained continuous measure, we obtained direct ratings of sentence comprehensibility from a new group of 43 undergraduates. They were asked to rate, on a scale of 1 to 10, how easy or hard they thought it would be for a "typical person" to understand each of the stimulus sentences, with 1 meaning *very hard for most to understand* and 10 meaning *very easy for most to understand*. Each participant rated 12 of the 36 target items and 4 fillers, resulting in 11 ratings for each target item. On the assumption that high comprehensibility corresponds to low strain (and low comprehensibility to high strain), we inverted the scale so that a score of 0 corresponded to the least amount of strain possible, and a score of 9 corresponded to the maximum amount of strain possible. The mean ratings and standard errors for each item are provided in Appendix C.

Next, we reanalyzed the data from Experiment 1 using the new continuous measure of strain as the fixed effect. The results replicated the previous findings. There was a significant main effect of semantic strain for verbs, $\beta = -.38$, SE = 0.08, $t = 4.89$, $p < .001$, but not for nouns, $\beta = -.04$, SE = 0.11, $t = 0.39$, $p = .70$ (see Figure 2b). Notably, the value of standardized slope coefficient for verbs obtained using the continuous measure of strain (-.38) was larger than the parameter obtained in the categorical model (-.26), suggesting that the continuous measure of strain was indeed more sensitive than the categorical measure. Based on this finding, in the remaining experiments we followed the same procedure of obtaining direct strain ratings of the stimulus items and using the continuous predictor in the analyses.

## 2.3 Discussion

The results of Experiment 1 demonstrate a verb mutability effect, replicating Gentner and France's (1988) original findings. First, verb meanings were found to change significantly more than noun meanings overall. Second, semantic strain predicted verb change, but not noun change. In the categorical model, verbs in strained sentences changed more than verbs in unstrained sentences, while noun scores were nearly identical in the two conditions. In the continuous model, the degree of verb change increased linearly with the degree of semantic strain, while noun change remained flat. This shows that, as predicted, verbs changed their meaning more readily than nouns and were the locus of change in resolving semantically strained utterances. Table 1 shows example paraphrases of unstrained and strained sentences from Experiment 1.

In addition, the fact that the patterns of meaning change found using word2vec replicate Gentner and France's past results using human judges is encouraging evidence that word2vec is capable of capturing human intuitions regarding semantic adjustment in a sentence processing context. Of course, a more direct comparison between word2vec scores and human judgments is needed – we provide such a test in Experiment 3.

Nevertheless, two questions bear addressing before moving on. One concern is whether our results are confounded by a relationship between strain and paraphrase length. It may be that strained sentences require more words to interpret than unstrained sentences (e.g., compare *The* *daughter agreed* → *The girl concurred* vs. *The car agreed* → *The vehicle responded well to the driver*). This might artificially depress word2vec scores by making the noun or verb less similar to any single word in the paraphrase. A closer look at paraphrase lengths, however, alleviates this concern. The mean paraphrase length in Experiment 1 was fairly flat across strain; the average paraphrase length was 3.94 for the least-strained item and 4.25 for the most-strained item. A mixed effect linear regression confirmed no significant relationship between semantic strain and net paraphrase length (i.e., excluding stop words like *the* that were not included in the word2vec model), $\beta$ = -.01, SE = 0.05, $t$ = 0.24, $p$ = .82). In addition, for both nouns and verbs, there was no significant relationship between net paraphrase length and word2vec score. That is, the mean noun cosine similarity score of the longest paraphrases did not differ significantly from that of the shortest paraphrases ($\beta$ = -.06, SE = .04, $t$ = 1.47, $p$ = .14), and likewise for verbs ($\beta$ = -.004, SE = .04, $t$ = 0.10, $p$ = .93.). Thus, it does not appear that the observed effects of strain are attributable to paraphrase length.

A second concern is whether omitting mechanical paraphrases from the analysis could have distorted the findings. Some of the initial sentences (8 out of 36) had a high proportion of paraphrases that were coded as mechanical and were therefore not included in the analysis. The mean strain rating of these items was higher than the overall mean strain rating (5.74 vs. 3.81), meaning that there were many instances where participants did not produce meaningful interpretations of highly-strained items and instead provided a word-by-word transcription. This is not entirely surprising; we might expect strained sentences to be more difficult to interpret, and this may lead some participants to give up or to be unable to provide a meaningful paraphrase. However, the loss of data among the high-strain items is problematic.

To address this concern, we reran our analyses on the full dataset—that is, without excluding any mechanical or noncompliant paraphrases. The results were the same: there was a significant main effect of semantic strain for verbs but not for nouns.[6] Further, the word2vec scores for the eight items with high rates of paraphrase exclusion matched the overall pattern. The average cosine similarity score for these items was .20 for verbs and .32 for nouns, indicating that verbs changed more than nouns even among these items. Thus, the verb mutability effect appears to hold consistently across all items, including those with the highest rates of noncompliant paraphrases.

To summarize: in Experiment 1, we replicated Gentner and France's original finding of verb mutability, but using word2vec to assess change of meaning instead of human judges. The results bear out the key phenomena of the verb mutability hypothesis: (1) verbs changed more than nouns: and (b) this effect increased with semantic strain. Further, the fact that our results using word2vec parallel Gentner and France's original findings suggests that word2vec is a feasible method for assessing semantic adjustment under paraphrase.

We are now in a position to bear down on the key question: what are the processes underlying verb mutability? Since the verbs used in Experiment 1 (as in Gentner & France, 1988) were significantly more polysemous than the nouns, the results thus far cannot distinguish between sense selection and online adjustment as accounts of mutability. We next investigate whether the verb mutability will hold for sentences when polysemy is controlled, or whether the pattern of greater verb mutability disappears when verbs and nouns are matched for polysemy.

[6] Noun model: $\beta_{Strain} = .02$, SE = .09, $p = .87$; Verb model: $\beta_{Strain} = -.27$, SE = .07, $p < .001$

# 3 Experiment 2

To test whether verb meaning change is primarily driven by online adjustment or by sense selection, we followed the same procedure as in Experiment 1, but chose new nouns and verbs such that half were low polysemy (1-2 senses) and half were high-polysemy (7+ senses; see Figure 3 below). Polysemy was evaluated by counting the number of synsets for each word in WordNet (Miller, 1995), excluding any that referred to specific people or places (the WordNet entries for each word are included in the supplementary material). Nouns and verbs were combined factorially to form intransitive sentences that comprised every possible combination of low- and high- polysemy nouns and verbs.

The logic of Experiment 2 is as follows: if mutability is mainly driven by sense selection, then high-polysemy nouns and verbs will show a greater increase in meaning change than will low-polysemy nouns and verbs—resulting in a polysemy-by-strain interaction. Further, if sense selection is the sole driver of meaning change, then the pattern of meaning change should be similar for nouns and verbs. This pattern would be evidence that the verb mutability effect is driven primarily by differential polysemy. In contrast, if verb online adjustment is the main driver of meaning change, then we should find that degree of semantic strain predicts meaning change for *both* low- and high-polysemy verbs, but not for nouns. In this case, (a) there will be little if any effect of polysemy and (b) the pattern of meaning change will be different for verbs than for nouns.

## 3.1 Method

### *3.1.1 Participants*

262 university undergraduates completed the study in person in the laboratory, on a computer. They received course credit in an introductory psychology class for their participation. 1 participant was excluded for not being a native English speaker, and 11 were excluded for failing catch trial criteria, for a net of 250 participants.

### *3.1.2 Materials*

The 6 nouns and 6 verbs were combined to form 36 new intransitive sentences. Half the nouns and verbs were low-polysemy (N- and V-), and half were high-polysemy (N+ and V+; see Figure 3). Thus, across the 36 sentences, the four possible combinations of noun and verb polysemy occurred in equal numbers: 9 N+/V+ combinations, 9 N-/V- combinations, 9 N+/V- combinations, and 9 N-/V+ combinations. As in Experiment 1, participants saw each noun and verb exactly once, receiving 6 target sentences (2 strained, 4 unstrained) comprising an equal number of high- and low- polysemy nouns and verbs (3 N-, 3 V-, 3 N+, and 3 V+). The noun and verb categories were modified slightly from the previous experiment:  two nouns were human, two were dynamic artifacts (i.e., artifacts that are capable of performing an action) and two were static inanimate (inert) objects. The verb categories varied correspondingly, comprising two verbs that prefer human subjects, two that prefer dynamic artifacts (or humans) and two that accept all three noun categories as subjects (see Figure 4).

Following the same procedure described in Experiment 1, the 36 sentences were given to a separate group of 35 undergraduate raters who rated them for comprehensibility; the scale was then inverted to represent semantic strain (see Appendix C).

| | | | Human | | Dynamic Artifact | | Static Inanimate | |
|---|---|---|---|---|---|---|---|---|
| | | | complain | suffer | pause | fail | dry | burn |
| | | *# senses* | 2 | 11 | 2 | 13 | 2 | 15 |
| Human | professor | 1 | - / - | - / + | - / - | - / + | - / - | - / + |
| | queen | 10 | + / - | + / + | + / - | + / + | + / - | + / + |
| Dynamic Artifact | motor | 2 | - / - | - / + | - / - | - / + | - / - | - / + |
| | bell | 7 | + / - | + / + | + / - | + / + | + / - | + / + |
| Static Inanimate | tree | 2 | - / - | - / + | - / - | - / + | - / - | - / + |
| | box | 10 | + / - | + / + | + / - | + / + | + / - | + / + |

Figure 3. Stimulus matrix for Experiment 2. Shaded cells indicate combinations that result in strained sentences, following Gentner and France's (1988) approach. Pluses and minuseds indicate high or low polysemy, respectively. For example, - / + indicates a low-polysemy noun and high-polysemy verb combination (e.g., the motor suffered), while + / - indicates a high-polysemy noun and low-polysemy verb combination (e.g., The box complained).

### *3.1.3 Experimental Design*

The design was 6 (item grouping, between-subject) X 2 (item strain: strained vs. unstrained, within-subject) X 2 (polysemy: high vs. low, within-subject). Two simple unstrained sentences were included as catch trials for checking attention and following directions; the criteria for excluding a subject was repeating a noun and/or verb in both of the catch trials or producing an obviously nonsensical answer in either. As each of the net 250 participants paraphrased 6 initial sentences, there were roughly 41 paraphrases per initial sentence.

### *3.1.4 Procedure*

The procedure was identical to that of Experiment 1. The instructions to participants were the same, with the exception of a minor adjustment to the example provided to participants (see Appendix A).

### *3.1.5 Coding*

Using the same coding procedure as in Experiment 1, two coders that were blind to the hypotheses were used to remove mechanical paraphrases, paraphrases describing the situation, and noncompliant paraphrases. Of the 1493 total paraphrases obtained, 276 paraphrases were excluded based on these criteria (107 mechanical, 144 describing the situation, and 25 noncompliant), as well as 1 additional paraphrase that generated a null vector (containing no words recognized by word2vec), resulting in a net of 1216 paraphrases included in the analysis. Cohen's κ was run to determine interrater reliability. There was moderate agreement between the two judges, $\kappa = 0.63$, (95% CI, 0.58 to 0.69), $p < .001$. A summary of the results of the coding task is shown in Appendix B. After coding, an average of 33.77 paraphrases per item remained.

### *3.1.6 Assessing Semantic Adjustment*

Noun and verb cosine similarity scores were obtained for each paraphrase using the same procedure as in Experiment 1.

## 3.2 Results

To test whether verbs changed more than nouns overall, we followed the same procedure as in Experiment 1: for each paraphrase, a difference score was calculated by subtracting the verb cosine score from the noun cosine score and fit to an intercept-only linear mixed model, with subjects and items included as random effects. Once again, the intercept was significantly greater than 0, $B = .04$, SE = 0.02, $t = 2.62$, $p = .01$, indicating that verbs (M = .24, SD = .12) changed significantly more overall than nouns (M = .28, SD = .13).

Next, to test the extent to which polysemy and strain predicted semantic adjustment, two additional models were fit: one for nouns and one for verbs. In both models, word2vec score was the dependent measure, polysemy (high vs. low), semantic strain, and the interaction term were included as fixed effects, and subjects and items were included as random effects. The results are plotted in Figure 4.

For verbs, there was a significant main effect of semantic strain such that the degree of verb meaning change increased as strain increased, $\beta = -0.29$, SE = .08, $t = 3.51$, $p = .001$. There was also a significant main effect of polysemy, $\beta = -0.22$, SE = .08, $t = 2.70$, $p = .01$, with high-polysemy verbs (M = 0.21, SE = 0.01) changing meaning to a greater extent than low-polysemy verbs (M = 0.26, SE = 0.01). The interaction was not significant, $\beta = -.02$, SE = .08, $t = .28$, $p = .78$.

For nouns, a significant main effect of polysemy was found, $\beta = -0.16$, SE = .06, $t = 2.70$, $p = .01$, with high-polysemy nouns (M = 0.25, SE = 0.01) changing meaning to a greater extent than low-polysemy nouns (M = 0.30, SE = 0.01). There was no significant effect of semantic strain, $\beta = -0.03$, SE = .06, $t = 0.45$, $p = .65$, and the interaction was not significant, $\beta = 0.05$, SE = .06, $t = 0.85$, $p = .40$.

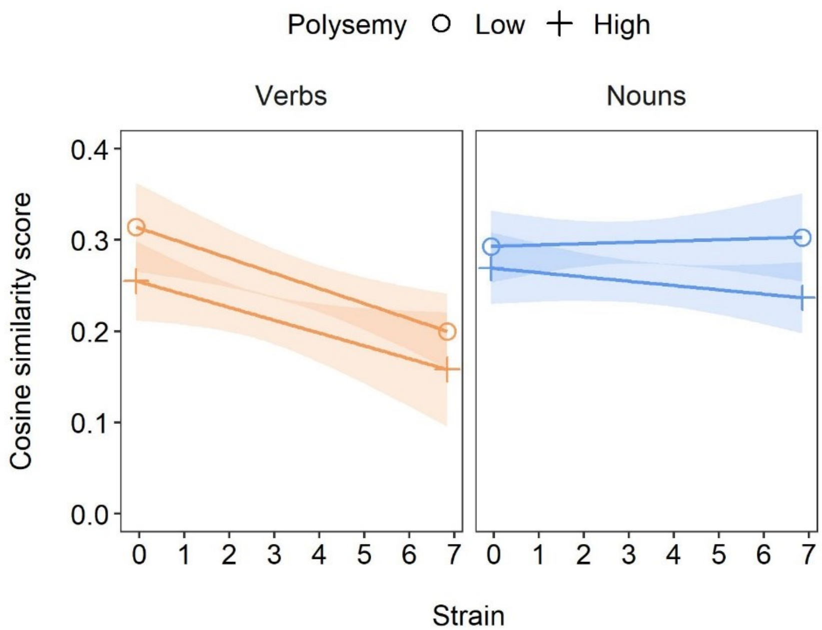



Figure 4. Fitted model plots showing the effect of strain and polysemy on word2vec scores for verbs and nouns in Experiment 2. Strain increases from left to right. Lower word2vec scores indicate greater meaning change. Shaded ribbons indicate 95% confidence bands.

As in Experiment 1, we tested for possible confounds between strain, paraphrase length, and word2vec scores. Once again, there was no significant relationship between semantic strain and net paraphrase length, $\beta$ = -.03, SE = 0.04, $t$ = 0.7, $p$ = .49, with the paraphrases of the least-strained item of roughly equal length (M = 4.30) to those of the highest-strain item (M = 4.25). As in Experiment 1, there was no significant relationship between net paraphrase length and noun word2vec scores ($\beta$ = -.02, SE = 0.03, $t$ = 0.58, $p$ = .56). For verbs, a small but significant relationship was found ($\beta$ = .11, SE = 0.03, $t$ = 4.19, $p$ < .001), such that verb similarity scores increased as paraphrase length increased. Note that this is in the opposite direction from that predicted by the concern discussed earlier (that verb similarity scores would be artificially depressed in longer paraphrases). Augmenting our original models with paraphrase length as a covariate resulted in nearly identical parameter estimates as in the original models.[7]

### 3.3 Discussion

The results of Experiment 2 point towards online adjustment as being the primary driver of verb mutability. Verbs changed more than nouns overall, and the degree of meaning change increased as a function of strain for *both* low- and high-polysemy verbs. In contrast, nouns showed no effect of strain: noun meaning change was flat from low- to high-strain contexts across both levels of polysemy. Thus, despite being matched for polysemy, nouns and verbs showed distinct patterns of semantic adjustment, with verbs being the locus of change in resolving semantic strain. This result replicates Experiment 1 and supports the verb mutability effect.

We also obtained a main effect of polysemy for both nouns and verbs, indicating that some sense selection was also occurring (though the effect in both cases appears smaller than the effect of strain on verb change). Importantly, however, this effect was orthogonal to both strain and word class: neither nouns nor verbs showed the interaction between polysemy and strain that is predicted by the sense selection view. Low-polysemy verbs changed at an equal rate as high-polysemy verbs, and low- and high-polysemy nouns were equally stable in meaning. Thus, sense selection fails to explain the asymmetry in patterns of meaning change observed between nouns and verbs, and cannot fully account for the verb mutability effect.

---

[7] Noun model: $\beta_{Strain} = -.03$, $\beta_{Polysemy} = -.16$, $\beta_{Str*Poly} = -.05$, $\beta_{Paraphrase\ Length} = -.02$; Verb model: $\beta_{Strain} = -.28$, $\beta_{Polysemy} = -.21$, $\beta_{Str*Poly} = -.03$, $\beta_{Paraphrase\ Length} = .11$

Examining the paraphrases revealed three patterns that underscore the importance of online adjustment in driving verb mutability (see Table 2 for examples). First, we found that low-polysemy verbs changed meaning even in sentences that comprised a high-polysemy noun paired with a low-polysemy verb, (e.g., *The bell complained* → *The alarm rang annoyingly;* 7 noun senses, 2 verb senses). If sense selection were the primary driver of mutability, we would expect unbalanced sentences like these to be most favorable towards noun adjustment and verb meaning preservation.

Second, many verb meaning adjustments resulted in novel metaphoric extensions, regardless of the verb's (or noun's) polysemy (e.g., *The box dried* → *All of the contents were eaten*; 10 noun senses, 2 verb senses). The third—and perhaps most striking—pattern was that these novel metaphoric extensions sometimes occurred even when a literal interpretation was available (i.e., when the sentence was unstrained) and even when the verb was low polysemy (and the noun was high polysemy). For example, some paraphrases of *The queen dried* (10 noun senses, 2 verb senses) included *The monarch aged, The monarch died, The monarch lost power,* and *The monarch grew cold and passionless*. Thus, even when conditions were most favorable towards noun change (e.g., low-polysemy verbs paired with high-polysemy nouns) or little change at all (unstrained sentences), verbs displayed a remarkable propensity for online adjustments to their meaning.

Table 2

*Example paraphrases from Experiment 2.*

| Polysemy |
| --- |

| | N | V | Stimulus | Paraphrase |
|---|---|---|---|---|
| | 7 | 2 | The bell complained | The alarm rang annoyingly |
| N+V- | 10 | 2 | The queen dried | The monarch aged |
| | 10 | 2 | The box dried | All of the contents were eaten |
| | 2 | 11 | The motor suffered | The engine sputtered |
| N-V+ | 2 | 13 | The tree failed | Someone who is usually reliable did not do their job |
| | 1 | 13 | The professor failed | The lecturer didn't get his message across |
| | 2 | 2 | The tree complained | The trunk creaked |
| N-V- | 2 | 2 | The motor paused | The car stalled |
| | 1 | 2 | The professor dried | The lecture became boring |
| | 10 | 15 | The queen burned | The ruler was enraged |
| N+V+ | 7 | 13 | The bell failed | The alarm stopped |
| | 10 | 11 | The box suffered | The container was crushed |

As in Experiment 1, there were some items with high rates of noncompliant paraphrases, although fewer than previously (5 out of 36 items had greater than one third of the paraphrases discarded, compared to 8/36 in Experiment 1). To test whether this influenced the results, we reran the analyses on the full dataset, including all noncompliant paraphrases (1491 paraphrases (i.e., 1493, less two paraphrases that generated null vectors). The results were the same: we found a significant main effect of semantic strain for verbs but not for nouns and a significant main effect of polysemy for both nouns and verbs (and no interaction).[8] Second, we confirmed that the word2vec scores for the five items with high rates of paraphrase exclusion matched the overall pattern. The mean cosine similarity scores were .19 and .17 for low- and high-polysemy

[8] Noun model: $\beta_{Strain}$ = -.04, SE = .06, $p$ = .49; $\beta_{Polysemy}$ = -.16, SE = .05, $p$ < .01; $\beta_{Str*Poly}$ = -.05, SE = .07, $p$ = .21; Verb model: $\beta_{Strain}$ = -.23, SE = .08, $p$ < .01; $\beta_{Polysemy}$ = -.22, SE = .08, $p$ < .01; $\beta_{Str*Poly}$ = -.02, SE = .08, $p$ = .84.

verbs and .30 and .23 for low- and high-polysemy nouns. Thus, the pattern of results for items with high rates of discarded paraphrases matched the overall pattern of results in the data.

### *3.3.1 Comparing the word2vec results with human judgments*

Experiments 1 and 2 paint a consistent picture of greater mutability for verbs compared to nouns. But does this effect match human cognition? Our analyses have assumed that word2vec cosine similarity scores capture the degree of meaning adjustment that the noun and verb underwent when paraphrased. That our findings in Experiment 1 replicated Gentner and France's original results grants us some confidence in this assumption. Still, given the novelty of our method, it is important to compare these results with human assessments of the degree of meaning change.

This replication would have the further benefit of addressing possible shortcomings of word2vec (and WEMs in general) that have been identified in the literature. For example, although word2vec and other WEMs have been shown to match human similarity judgments well in some tasks (e.g., Günther, Dudschig, & Kaup, 2016; Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998; Pereira et al., 2016), there are concerns as to their ability to distinguish similarity from association (Hill, Reichart, & Korhonen, 2015; Lenci, 2018; Pereira et al., 2016; Simmons & Estes, 2006). There are also concerns related to polysemy—e.g., Gerz et al. (2016) found that WEM correlations with human similarity judgments were lower for high-polysemy verbs than low-polysemy verbs. Although they did not test nouns, it is plausible that the same pattern applies.

Therefore, to address these concerns, in Experiment 3 we sought to replicate the results of Experiment 2 using a behavioral assessment of meaning change: the double paraphrase task developed by Gentner and France (1988).

## 4 Experiment 3

As described in the introduction, Gentner and France (1988) used three different behavioral approaches to assess the degree to which nouns and verbs changed meaning under paraphrase: *divide and rate*, *retrace*, and *double paraphrase*. All three provided converging evidence for the verb mutability effect, but they were also labor-intensive and prone to high amounts of data loss. Of the three, the double paraphrase task is most appealing for our present purpose because it is the most hands-off approach. No judges are needed to divide the paraphrase into component pieces (as in the divide and rate method), nor is it necessary to ask raters to match each paraphrase with a fixed list of the initial nouns or verbs (as in the retrace task). Further, the strict criterion of requiring an exact match between the initial noun or verb and its appearance in the paraphrase eliminates subjective judgments about degree of change.

In the double paraphrase task, the original paraphrases are given to a new set of participants for them to paraphrase—that is, to produce a "double" paraphrase. The double paraphrase is then scored for noun and verb resurfacings. A resurfacing occurs when the original stimulus noun or verb reappears in the double paraphrase. The assumption is that words whose meaning has been preserved in the original paraphrase will be most likely to resurface in the double paraphrase, as in the following example:

| Stimulus Sentence | Original Paraphrase | Double Paraphrase |
| --- | --- | --- |

| **(Experiment 2)** | | **(Experiment 2)** | | **(Experiment 3)** |
|---|---|---|---|---|
| *The motor complained* | → | *The engine didn't work well* | → | *The* <u>*motor*</u> *functioned badly* |

Here the stimulus noun *motor* from Experiment 2 has resurfaced in the double paraphrase, while the verb *complained* has not. This matches intuition: *engine* is very similar to *motor*, while *functioned badly* represents a much greater adjustment to the meaning of *complained*. The strict criterion of an exact match (though we accepted differences in pluralization or tense) provides an objective scoring procedure. The tradeoff is data loss, since many near-matches are discarded – e.g., *The oak was on fire* would not count as a resurfacing for *The tree burned* for either the noun or the verb. For our present purposes, however, we wished to use unambiguous criteria to serve as a benchmark for the word2vec results from the previous experiments.

## 4.1 Method

### *4.1.1 Participants*

77 participants completed the study online via Mechanical Turk. The task took approximately 15 minutes and they were paid at a rate equivalent to Illinois' minimum wage at the time of the study. 4 participants were excluded for failing the catch trial criteria and 2 were excluded due to experimenter error, resulting in a net of 71 participants.

### *4.1.2 Materials & Design*

The 1216 participant paraphrases from Experiment 2 served as the stimuli for Experiment 3. Participants in Experiment 3 received the same instructions as participants in Experiments 1 and 2, with the addition of a sentence instructing them to use their best guess as to the meaning of any misspelled words in the sentences and to ignore any typos to the best of their ability (see

Appendix A). For brevity and clarity, in what follows we refer to the first set of paraphrases obtained in Experiment 2 (that serve as the stimuli/initial sentences in this experiment) as *singles*, and responses generated in the present experiment (the paraphrases of those singles) as *doubles*.

Singles were grouped into two between-subject item groupings based on their initial stimulus sentence in Experiment 2. These item groupings were organized so that each participant paraphrased 18 singles, as well as 2 catch trials that served as attention checks. The 18 singles were presented in 3 blocks of 6 items each, with order randomized within each block. Within each block, each of the original 6 stimulus nouns and verbs (from which the single paraphrase originated) was represented exactly once (so that each occurred 3 times total for each participant). This blocked design ensured that participants did not paraphrase singles coming from the same original noun or verb consecutively. In addition, because removing mechanical and noncompliant paraphrases in Experiment 2 resulted in an uneven number of singles per original stimulus item, "dummy" singles were included to ensure a uniform experience across participants within each assignment condition. The goal was to obtain doubles of as many of the 1216 singles from Experiment 2 as possible, while also ensuring that each participant was matched on the criteria described above. This resulted in 1385 items in total: 1158 target items and 227 "dummy" items that were paraphrased by participants but excluded in the analysis.

### *4.1.3 Procedure*

The procedure matched that of Experiments 1 and 2, except that participants paraphrased 18 sentences instead of 6. All of the stimulus items were paraphrases obtained from Experiment 2.

## 4.2 Results

### *4.2.1 Scoring*

Of the original 1158 doubles, 101 were excluded due to dropping 6 participants for failing the catch trials. Due to experimenter error, an additional 45 doubles were excluded, for a net of 1012 included in the analysis. Among the 1012 doubles included in the analysis, the number of doubles obtained per original stimulus item from Experiment 2 (e.g., *The motor complained*) ranged from 15 to 34, with a mean of 28.11 and a median of 29.5. Paraphrases were then scored for noun and verb resurfacings. A strict criterion was used: only identical resurfacings counted, except for changes in tense or pluralization.

### *4.2.2 Analysis*

Resurfacing counts by class and polysemy are given in Table 3. As expected, overall data loss (paraphrases where neither the verb nor the noun resurfaced) was high: out of a possible 1012 paraphrases, nouns resurfaced a total of 214 times and verbs resurfaced a total of 104 times.

Table 3

*Number of resurfacings (hits) vs non-resurfacings (misses) for nouns and verbs from Experiment 3.*[a]

| | Verbs | | | | Nouns | | | |
|---|---|---|---|---|---|---|---|---|
| Polysemy | Hits | Misses | Total | % Hits[b] | Hits | Misses | Total | % Hits[b] |
| Low | 68 | 422 | 490 | 13.88 | 124 | 389 | 513 | 24.17 |
| High | 36 | 486 | 522 | 6.90 | 90 | 409 | 499 | 18.04 |
| Total | 104 | 908 | 1012 | 10.28 | 214 | 798 | 1012 | 21.15 |

*Note*. [a] These numbers include 27 instances in which both the noun and verb resurfaced. [b] Percentages do not sum to the number in the *Total* row due to uneven cell counts (see Materials & Design above).

To test whether the overall difference in noun and verb resurfacings was significant, a difference score for each paraphrase was calculated in the following way: if the noun resurfaced but not the verb, it was scored as a 1. If the verb resurfaced but not the noun, it was scored as a 0. If neither or both resurfaced, it was considered a tie, and that response was excluded. There were 27 instances where both the noun and verb resurfaced.

Next, a mixed effect logistic regression model was fit, with difference score as the dependent measure, the intercept as the only fixed effect, and subjects and items as random effects. The intercept differed significantly from 0, $B = 1.24$, SE = 0.33, 95% CI [0.67, 1.97], $z = 3.79$, $p < .001$, indicating that noun-only resurfacings (187 occurrences) were 78% more likely to occur overall than verb-only resurfacings (77 occurrences).[9]

Next, to test the effect of semantic strain and polysemy on verb and noun resurfacings, two additional mixed effect logistic regression models were fit: one for nouns and one for verbs. Noun / verb resurfacings were the dependent measures in their respective models, with polysemy (high vs. low), strain, and the interaction term included as fixed effects and subjects and items as random effects. The fitted model results are plotted in Figure 5.

For verbs, there was a significant main effect of semantic strain, $B = -0.25$, SE = 0.11, 95% CI [-0.50, -0.04], $z = 2.31$, $p = .02$, indicating that verbs resurfaced less often as strain increased. There was also a significant main effect of polysemy, $B = -0.63$, SE = 0.23, 95% CI [0.21, 1.17], $z = 2.69$, $p < .01$, indicating that low-polysemy verbs (68 resurfacings) were more likely to

[9] All beta parameters reported for logistic models are in logits.

resurface than high-polysemy verbs (36 resurfacings). The interaction was not significant, $B$ = 0.08, SE = 0.11, 95% CI [-0.15, 0.35], $z$ = 0.76, $p$ = .45.

For nouns, there was no significant effect of semantic strain, $B$ = -0.02, SE = 0.05, 95% CI [-0.12, 0.07], $z$ = 0.40, $p$ = .69. There was a marginal main effect of polysemy, $B$ = -0.20, SE = 0.10, 95% CI [-.01, 0.41], $z$ = 1.89, $p$ = .06. The interaction was not significant, $B$ = 0.01, SE = 0.05, 95% CI [-0.09, 0.10], $z$ = .12, $p$ = .90.

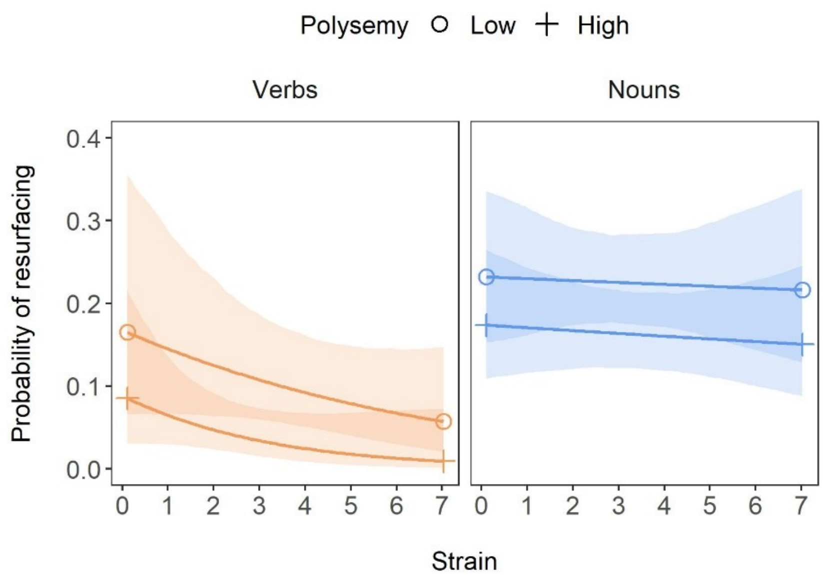



Figure 5. Fitted models showing probability of resurfacing for verbs and nouns in Experiment 3. Lower probabilities indicate greater meaning change. Strain increases from left to right. Shaded ribbons indicate 95% confidence bands.

## 4.3 Discussion

A full replication of Experiments 1 and 2 required the following three results: (1) verbs should resurface less often than nouns overall (indicating greater meaning change overall), (2) resurfacings should decrease with semantic strain for verbs but not for nouns, and (3) high-polysemy nouns and verbs will resurface less often than low-polysemy nouns and verbs, across all levels of strain. The results of the double paraphrase task support all three predictions. As was found in Experiments 1 and 2, (1) verbs changed more than nouns overall (they resurfaced less); (2) semantic strain significantly predicted verb – but not noun – change; and (3) high-polysemy nouns and verbs changed more (resurfaced less often) than low-polysemy nouns and verbs (though the effect was marginal for nouns, $p = .06$).

These results parallel the word2vec results in Experiments 1 and 2, providing support for its use in assessing the degree of meaning change in our paraphrase task. To be clear, we are not suggesting that word2vec's embeddings match human representations of word meaning, nor that calculating cosine similarity scores serves as a model of the human comparison process. Nonetheless, the word2vec scores here appear to capture human patterns in the present task—including the effects of polysemy—rather effectively.

#### *4.3.1 Qualitative differences in noun and verb changes*

Experiments 1 – 3 show that verb change and noun change differ *quantitatively* in the degree of meaning change each is prone to undergo. Another important question is whether verb and noun meaning change differ *qualitatively* as well. That is, in addition to changing *more* than nouns, do verbs also differ in *how* they typically change compared to nouns? Thus far in this paper we have focused mainly on metaphor as the primary way in which verbs extend their meanings. But words can change meaning in many other ways as well, such as through synonymous substitutions (e.g., *motor* → *engine*; *burn* → *combust*), taxonomic substitutions (e.g., *motor* → *machine; burn* → *change*), or metonymic substitutions (e.g., *motor* → *car; burn* → *turned into* *ash*).[10] We ask whether verbs' greater mutability compared to nouns correlates with distinct qualitative patterns of meaning change as well.

We expect that verbs will have a greater propensity for metaphoric/analogical extension than would nouns. As discussed earlier, metaphoric uses of verbs appear to be significantly more common in day-to-day language than metaphoric uses of nouns (Krennmayr, 2011; Jamrozik et al., 2013). A second expectation is that nouns will be more likely than verbs to be paraphrased with a taxonomic substitution (either a more general term (as in *car* → *vehicle*) or a more specific one (as in *car* → *Jeep*). Intuitively, a taxonomic paraphrase is a way to preserve the likely referent of the original noun, while using new content words. Further, taxonomic substitutions may be more available for nouns than for verbs; a number of studies have found that noun concepts are taxonomically structured to a greater extent than verb concepts (e.g., Burnett & Gentner, 2000; Fellbaum, 1999; Graesser, Hopkinson, & Schmid, 1987; Huttenlocher & Lui, 1979; Miller & Fellbaum, 1991; Pavličić & Markman, 1997; Qui, Castro, & Johns, 2021). For example, Graesser et al. (1987) found that participants in a free-sort task consistently categorized nouns—but not verbs—in a way that correlated with the pattern shown in a separate taxonomic organization task. That is, participants spontaneously organized nouns—but not verbs—taxonomically. Further, there is evidence that people sometimes produce "chain reversals" for verbs—e.g., saying both that *drinking* is a kind of *swallowing* and *swallowing* is a kind of *drinking*, or that *thinking* is a type of *reasoning* and *reasoning* is a type of *thinking*

[10] Metonymy and metaphor are both types of figurative language, but they differ in an important way. Metaphor involves abstract, often relational, commonalities between two concepts. These are often relational—for example, *obsession* and *tumor* share the abstract relational schema of "something that grows inside you". Hence, many metaphors can be analyzed as analogies (Gentner et al., 2001). In contrast, *metonymy* involves associations that lack abstract commonalities between concepts, which can be literal (e.g., part-whole relations, as in *engine*→*car*) or figurative (e.g., *flag*→*patriotism*).

(Burnett & Gentner, 2000; Rips & Conrad, 1989). Burnett and Gentner (2000) found that this occurred more often for verbs than for nouns—again suggesting that nouns are organized into stable taxonomies to a greater extent than are verbs.

Finally, a third expectation was that nouns would be more prone to metonymic extensions than would verbs. Metonymy is a well-established aspect of noun usage (e.g., Nunberg 1995; Pustejovsky, 1995), and metonymic relationships are widespread among nouns, both as lexicalized senses (e.g., a *container-contained* relation, as in *I ate the whole box*) and as novel meaning extensions (e.g., saying *the ham sandwich over there* to refer to a customer at a diner; Nunberg, 1979). In contrast, the set of verbs that are frequently used metonymically (e.g., *begin, enjoy*) appears relatively small (Utt et al., 2013). Verb metonymy typically manifests as one part of an event standing for the event as a whole. For example, in *the writer began the novel*, the verb *began* stands for the event *began to write* (Nunberg, 1995).

That nouns and verbs appear to differ in their relative predispositions towards metaphoric, metonymic, and taxonomic organization raises the possibility that these differences might show up at the level of online sentence processing. To investigate this question, we gave a randomly chosen subset of the paraphrases from Experiment 2 (16 paraphrases from each item, for a total of 576 of the original 1216 paraphrases) to two coders who were blind to the hypotheses. The coders were graduate students in linguistics and were paid for their time. For each paraphrase, the judges categorized the type of change the original noun and verb underwent into seven different types: *synonym / highly similar*, *taxonomic, contextual taxonomic*, *associative (metonymic)*, *metaphoric (analogous), describes the situation*, and *other* (see Table 4). Cohen's κ was run to determine interrater reliability. There was moderate initial agreement between the two judges, κ = 0.58, (95% CI, 0.55 to 0.61), $p < .001$; after discussion, consensus was reached on all items.

Table 4

*Codes used in the qualitative analysis. The definitions here are summaries from longer explanations given to the coders; examples are drawn from a larger set that was given to the coders. Coders received an equal number of noun and verb examples for each code.*

| Code | Definition (summarized) | Noun Example | Verb Example |
|---|---|---|---|
| Synonym / Highly Similar | A synonym or highly similar term in a literal sense | *The dad yelled → The father shouted* | *The dog barked → The canine growled* |
| Taxonomic High | A superordinate term | *The car drove → The vehicle moved* | *The car drove → The vehicle moved* |
| Taxonomic low | A subordinate term | *The person walked → The man sauntered* | *The person walked → The man sauntered* |
| Contextual Taxonomic High / Low | A superordinate or subordinate term that is so only in the context established by the sentence | *The barrier melted → The iceberg liquified* | *The radio worked → The receiver received the signal* |
| Associative (Metonymic) | A term that is associated, rather than similar or taxonomically related (e.g., *part-whole*) and does not share an abstract commonality | *The engine functioned → The car worked* | *The dog growled → The canine trembled* |
| Metaphoric (Analogous) | A term involving an analogy or abstract commonality with the original word | *The school was full → The prison was at capacity* | *The car limped → The vehicle drove slowly* |
| Describes the situation | A term that describes the surrounding context instead of providing a paraphrase | *The eggs sizzled → Breakfast is ready* | |
| Other / Uninterpretable | Uninterpretable or not fitting into any of the above categories | No example was provided to the coders | |

The tallies for all code categories are given in Appendix D. In what follows, we focus on our three codes of primary interest: metaphoric/analogous, associative/metonymic, and taxonomic (these were also the most common codes, with the exception of *Synonym / Highly similar*).

Figure 6a shows the overall code tallies for nouns and verbs. As expected, verbs often changed metaphorically (165 occurrences) while nouns did not (27 occurrences) Also as expected, taxonomic substitutions occurred more often for nouns (251 occurrences), than for verbs (101 occurrences), as did associative substitutions (146 for nouns, 110 for verbs).

Figure 6b plots the distribution of codes for nouns and verbs by strain quartile (from participants' ratings in Experiment 2) For verbs, rates of metaphoric responding increased steadily as strain increased, confirming that, as verbs changed meaning in response to strain, they did so mainly via metaphoric extensions. For nouns, however, there were no clear trends across strain associated with most codes, consistent with the idea that verbs were the locus of change. As expected, associative and taxonomic substitutions were more common for nouns, while rates of metaphoric responding were consistently low.

Figure 6c shows the distribution of codes by word2vec quartiles, where quartile 4 represents the paraphrases where the noun or verb changed the least (i.e., had the highest word2vec similarity score), and quartile 1 represents those paraphrases where they changed the most (here the x-axis has been reversed so that the degree of change increases from left to right, matching the direction of increasing semantic strain in Figure 6b).

For verbs, a clear relationship between degree of meaning change (word2vec quartile) and frequency of metaphoric responding can be seen. The further a verb's meaning changed, the more likely that change was to be a metaphoric extension. The pattern was quite different for nouns. For nouns, few metaphoric substitutions were associated with meaning change of any degree. Instead, across all degrees of meaning change (i.e., across all word2vec quartiles), participants mostly made taxonomic substitutions, with associative substitutions next most likely.

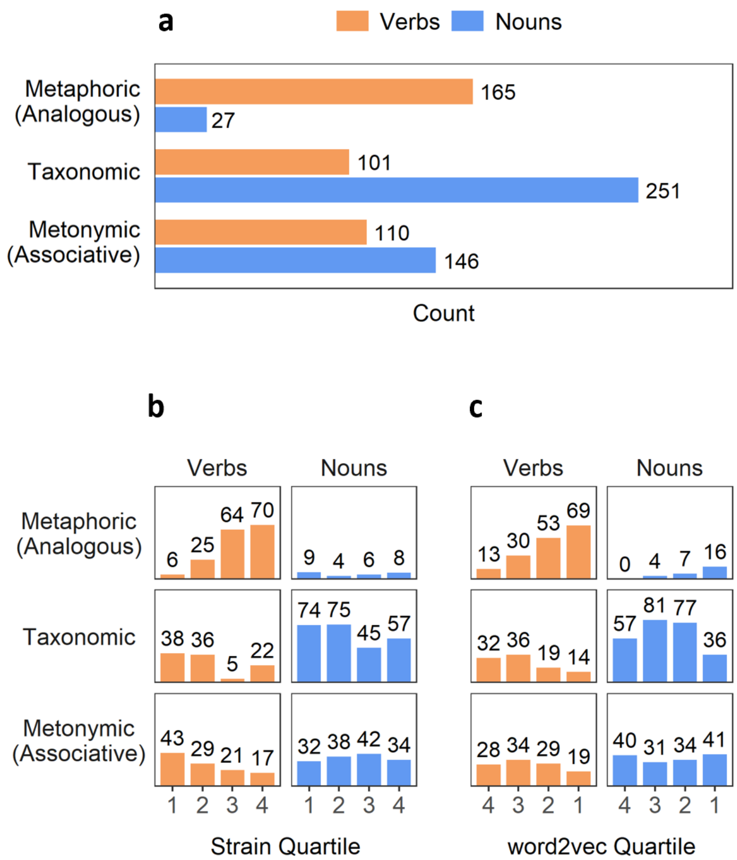



Figure 6. Tallies for the metaphoric (analogous), associative (metonymic), and taxonomic categories for nouns and verbs from the qualitative analysis. (A) Total counts. (B) Tallies by strain quartile, with strain increasing from left to right. (C) Tallies by word2vec quartile. The x-axes are reversed so that change increases from left to right, with quartile 4 representing the least degree of change (highest word2vec scores) and quartile 1 representing the greatest degree of change (lowest word2vec scores).

These results support a novel conclusion: in addition to quantitative differences in meaning change, there are also *qualitative* differences in how nouns and verbs change meaning. When verbs adapt their meanings to context, they mainly do so via metaphor. When nouns adapt their meanings, they do so via taxonomic or associative (metonymic) relations. Thus, in addition to their greater mutability, verbs also appear to be more amenable to metaphoric extensions than do nouns.

# 5 General Discussion

There are three main findings. First, we obtained strong and consistent evidence for the verb mutability effect. Second, we found that online adjustment is the primary driver of verb mutability. Third, we identified qualitative differences in how nouns and verbs change meaning. Also, on a methodological level, we found that word2vec's cosine similarity scores for the original words and their paraphrases aligned well with human judgments of the degree of semantic change. We next review these findings.

## 5.1 Verbs change more than nouns

All three studies provided clear evidence for the verb mutability effect: that under semantic strain, verb meanings are altered more than are noun meanings. In Experiment 1, we replicated Gentner and France's (1988) original verb mutability findings using a subset of their stimuli. We asked people to paraphrase simple *The noun verbed* sentences that varied in semantic strain. The results showed (1) that verbs changed more than nouns overall; and (2) that the degree of verb meaning change increased with the degree of semantic strain. In contrast, noun meanings remained stable across strain. In Experiment 2, we replicated these findings while systematically varying noun and verb polysemy. In Experiment 3, we replicated our word2vec findings from Experiment 2 using a behavioral assessment of meaning change (the double paraphrase task), rather than word2vec scores as in the prior studies. Thus, the verb mutability effect held across different sets of stimuli, different levels of noun and verb polysemy, and different methods of assessing semantic change. When a sentence requires a novel interpretation, it is the verb that alters its meaning.

### 5.2 Online adjustment drives verb mutability

In Experiment 2, we tested whether differential polysemy could explain the greater mutability of verbs. If meaning change occurs largely through selecting an appropriate word sense of the verb (or noun), then more polysemous words should show greater meaning change under strain. To test this, we created a new set of sentences that systematically varied the polysemy of the nouns and verbs, while independently varying semantic strain. Not surprisingly, there was a main effect of polysemy for both nouns and verbs, indicating that some sense selection occurred. Importantly, however, we did not obtain the interaction between polysemy and strain that would be expected if sense selection were the primary driver of mutability. Instead, both low- and high-polysemy verbs showed greater change of meaning as strain increased, and both low- and high-polysemy nouns remained equally stable (Figure 4). Thus, the effect of polysemy was orthogonal to that of semantic strain and cannot explain the asymmetry between nouns and verbs. Further, we observed instances in which people generated novel metaphoric extensions for verbs even when conditions were favorable to greater sense selection in nouns than in verbs—e.g., when a low-polysemy verb was paired with a high-polysemy noun (e.g., *The bell complained* → *The alarm rang annoyingly*). Strikingly, this sometimes happened even when a literal interpretation was available (e.g., *The box dried* → *All of the contents were eaten*).

In sum, selection among existing word senses cannot explain the verb mutability pattern (greater change in verb meaning than in noun meaning, and greater change in verb meaning as strain increases). We are left with the conclusion that online adjustment is the primary driver of verb mutability. In short, verbs appear remarkably willing to extend their meanings in a way that nouns are not. Indeed, it may be that verbs' greater mutability is what leads to their relatively high polysemy.

### 5.3 Qualitative differences in noun and verb change

Our third main finding was that verbs and nouns differ qualitatively in *how* they change meaning. To our knowledge, no prior work has looked at this question. Coding a subset of the paraphrases from Experiment 2, we found that verbs were more likely to extend their meanings metaphorically/analogically than were nouns overall. Noun change was more likely to be via taxonomic substitution or metonymic association; metaphoric extension was rare for nouns. Further, the rate of verb metaphoric extension increased sharply with degree of strain. In contrast, rates of all types of noun substitutions (including taxonomic and metonymic substitutions) were largely flat across strain.

### 5.4 Characterizing verb meaning change

In examining the paraphrases from these studies, we observed another important pattern in meaning change—in this case among the verbs themselves. Across paraphrases, verb meaning change tended to follow two principles. First, verbs typically changed only as far as was required to resolve the semantic strain. Second, verbs changed in such a way that domain-specific meaning components were adjusted before more abstract relational ones. For example, consider the set of paraphrases below for the verb *complained* from Experiment 2.

| | Original Sentence | Paraphrase |
|---|---|---|
| 1. | *The professor complained* | *The adult whined* |
| 2. | *The bell complained* | *The alarm rang annoyingly* |
| 3. | *The box complained* | *The container wouldn't close.* |

In this example, strain increases with the degree of semantic mismatch between noun and verb as one moves from (1) to (3). Sentence (1) is unstrained since the verb receives its preferred (human) subject type. Sentence (2) is moderately strained in that, although bells are inanimate artifacts, they are saliently associated with making sound. Sentence (3) is highly strained; boxes are inanimate and also not known for making sound. As the paraphrases show, the degree of verb change increases progressively with strain. The paraphrase of (1)—which is unstrained and literally interpretable—largely retains the standard meaning of *complain*. In the paraphrase of (2), the domain-specific components of *complain*'s meaning have been adjusted from referring to human verbal communication to a more general meaning involving producing an (annoying) sound. In the paraphrase of (3), the verb is abstracted further so that the meaning components having to do with sound are discarded entirely; only the abstract relational notion that *complaining indicates a bad state of affairs* is retained. Thus, verb meaning change is gradual, rather than radical.

This pattern of progressive meaning change in verbs was first identified by Gentner and France (1988), who termed it *minimal subtraction*. Recent work in cognitive neuroscience looking at verb processing has found activation patterns that are consistent with this pattern. A number of studies have found that cortical activation shifts anteriorly from primary perceptual processing areas when a verb is used literally to adjacent secondary areas when it is used figuratively (Cardillo et al., 2012; Chatterjee, 2008; Chen et al., 2008; Desai et al., 2011, 2013; Jamrozik et al., 2016; Raposo et al., 2009; Saygin et al., 2010; Wallentin et al., 2005). These adjacent anterior areas are associated with the processing of abstract concepts (Cardillo et al., 2012;

Chatterjee, 2008). Thus, our finding that domain-specific meaning components (i.e., sensorimotor components) are retained when a verb is used literally, but are abstracted away when a verb is used metaphorically, parallels imaging studies showing similar shifts from sensorimotor areas to adjacent areas associated with abstract processing.

These findings also bear on the question of personification—an area of debate among linguists. As Dorst (2011) describes, at one level, any instance in which the noun violates the verb's selectional preferences can be considered personification—that is, as an invitation to construe the noun as animate/human. This account appears to stand in contrast to our argument here that the verb, rather than the noun, is what is reconstrued. But Dorst also notes that the interpretation of such violations varies according to the field of study and the purpose of the analysis. Our analysis focused on the semantic-conceptual level—that is, on how people interpreted the words in strained sentences. In this analysis, we found that although there were a few instances in which an inanimate noun was paraphrased as an animate being (e.g.,, *The motor complained* → *The talkative Tracy was on her usual rant*), in the great majority of the paraphrases, the noun largely retained its usual meaning and the verb adapted its meaning to fit the noun's meaning (e.g., *The motor complained* → *The vehicle was noisy and struggling*).

### 5.5 Mutability and meaning change over time

Our findings also connect to work on language evolution. There is evidence that verbs change meaning at a greater rate over time than nouns do (Dubossarsky et al., 2016; Sagi, 2019). For example, Dubossarsky et al. (2016) compared rates of change for nouns, verbs, and adjectives from 1850 to 2000. They found that verbs changed meaning at a higher rate than both nouns and adjectives over the entire period of analysis. Dubossarsky et al. linked their results with the verb mutability effect:

> The verb mutability effect identified by Gentner (1981) may be one kind of synchronic interpretative bias implicated in the diachronic asymmetry observed in the present article: in terms of synchronic processing, verbs are more semantically mutable than nouns; correspondingly, in terms of diachronic change over time, verbs undergo more semantic change than nouns. (p. 20)

An important question is the extent to which these diachronic meaning changes are due to metaphoric extensions of verb meaning. There is widespread agreement among both psychologists (e.g., Bowdle & Gentner, 2005; Gentner & Asmuth, 2017; Gentner & Wolff, 2000; Xu et al., 2017) and linguists (e.g., Joseph et al., 1996; Heine, 1997; Hopper & Traugott, 2003; Narrog & Heine, 2021; Sweetser, 1990; Traugott, 1988) that metaphor is an important vehicle for language change over time. For example, in a computational historical analysis examining 5000 metaphorical mappings spanning 1100 years, Xu et al. (2017) found that new word senses most frequently emerged from metaphorical mappings originating from concrete source domains to more abstract domains. For example, the cognitive sense of *reflect* emerged from a metaphorical mapping from light to thought. Our finding that the verb mutability effect is driven primarily by online adjustment, and that verbs have a higher propensity for metaphoric extensions than nouns, suggests an intriguing link between verb mutability, online metaphoric extensions, and meaning change over time.

## 5.6 Why do verbs change more than nouns?

Our findings here invite an explanation of *why* verbs undergo more online change than do nouns. We next consider factors that may drive verb mutability.

### *5.6.1 Syntactic influences: Word order*

The simplest account is that the SVO word order typical of English (and the SV order of our stimuli) establishes the primacy of the subject noun as the context to which the verb must adapt. Though this is plausible to a certain extent, prior work has shown that word order cannot account for verb mutability on its own. Gentner and France (1988, Experiment 2) found greater semantic change in verbs than in nouns even when the verb was the first word the sentence (e.g., *Worshipped was what the lizard did*). Thus, word order alone is unlikely to be a major driver of verb mutability.

### *5.6.2 Pragmatic influences: Predicate role*

Another possible factor underlying verb mutability lies in the pragmatics of sentence interpretation—specifically, the fact that verbs typically serve in the predicate role in a sentence. As Gentner and France (1988, p. 372) suggested, "… verbs have the job of conveying relations or events that apply to the referents established by the nouns." More generally, Croft (2002) observed that sentence elements that depend on another element for their meaning (like verbs and adjectives) are the ones that typically change meaning in figurative statements, while the autonomous elements they depend on (often nouns) establish the domain to which they must adapt. As support for the claim that occupying the predicate position contributes to mutability, Gentner and France noted that this pattern holds even for within-class constructions, such as noun-noun metaphors. For example, in *That surgeon is a butcher*, the noun in the predicate position (*butcher*) is the one interpreted metaphorically, yielding a sloppy, brutish surgeon. In contrast, the reverse metaphor, *That butcher is a surgeon,* suggests a deft, precise butcher As another example, in noun-noun conceptual combination, the predicate noun typically adapts its meaning to the referent noun (Murphy, 1990; Wisniewski, 1997). Thus, an *acrobat hippopotamus* is an agile hippopotamus, while a *hippopotamus acrobat* is a clumsy acrobat. In both these examples, the meaning of the referent term is held constant, while the predicate term is adapted to provide information about the referent. Thus, we suggest that verb mutability is partly driven by the verb's role as predicate in a sentence.

*5.6.3. Semantic influences: Relationality of meaning*

Another potential contributor to verb mutability is relationality of meaning. It has been argued that relationality is a key feature of verb meaning; that is, while nouns often refer to objects or object concepts, verbs typically express relations among those referents (Baker et al., 1998; Croft, 2000, 2001; Fillmore, 1971; Jackendoff, 1983; Langacker, 1987, 2008; Levin, 1993; Talmy, 1975, 1988, 2000; Vigliocco et al., 2011). We suggest that relationality imposes additional pressure to adjust meaning, over and above the pragmatic function of predication (although, as discussed below, the two factors normally work in tandem). One way to test the importance of relationality of meaning *per se* is to compare the mutability of two words from the same syntactic class that differ in relationality. Asmuth and Gentner (2017) conducted such a test by comparing the mutability of relational nouns and entity nouns. As mentioned earlier, entity nouns are nouns whose referents share common intrinsic properties (as well as common relational structure)—e.g., *tiger, apple*. Relational nouns are nouns whose referents share a common relational pattern, but not common intrinsic properties—e.g., *carnivore, barrier* (Asmuth & Gentner, 2017; Gentner & Asmuth, 2017; Gentner & Kurtz, 2005; Goldwater & Markman, 2011; Goldwater et al., 2011; Markman & Stilwell, 2001; Rehder & Ross, 2001).

Emulating Kersten & Earles' (2004) recognition paradigm, Asmuth and Gentner (2017) gave participants phrases consisting of an entity noun and a relational noun—e.g., *truck limitation*. At a later surprise recognition test, recognition sensitivity was higher for the entity noun (*truck*) than for the relational noun (*limitation*). More tellingly, recognition of relational nouns suffered when they were paired with a new entity noun at test (e.g, *book limitation*)—but this decrement was not found for entity nouns, which were recognized equally well with a new relational noun (e.g., *truck threat*) as with the original relational noun. Thus, the relational nouns had adapted their meaning to the entity nouns, but not the reverse. This pattern mirrors Kersten and Earles' findings for noun-verb sentence memory, discussed above. Asmuth and Gentner showed that this effect held regardless of word order (e.g., for both *tooth opponent* and *opponent tooth*) and also when controlling for the abstractness of the nouns—evidence for the role of relationality of meaning in driving mutability, over and above other influences.

The idea that semantic factors cut across form-class distinctions in influencing sentence processing has recently been gaining currency in cognitive neuroscience. In a review of the cognitive neuroscience literature on differences in noun and verb processing, Vigliocco et al. (2011) showed that the key distinctions in processing at the cortical level are not defined by form-class distinctions between nouns and verbs, but rather by the semantics of the concepts they refer to. Recent fMRI work comparing noun and verb processing has shown that, when these semantic differences are controlled for (e.g., testing only words that refer to events), nouns and verbs generate similar patterns of cortical activation (Cardillo et al., 2012; Vigliocco, Vinson, & Siri, 2005; Vigliocco et al., 2006, 2011). A study by Cardillo et al. (2012) demonstrated that this pattern holds in metaphor processing as well. They conducted an fMRI study that measured cortical activation when people read either noun metaphors or verb metaphors. Crucially, the noun and verb metaphors were matched semantically such that the verbs used were all denominalized verbs (derived from nouns). For example, for the noun metaphor *her smile was a cat's <u>purr</u>*, the corresponding verb metaphor *the flowers <u>purred</u> in the sunlight* was also tested. Cardillo et al. found no differences in cortical activation between the noun and verb metaphors, suggesting that semantics, rather than syntactic class *per se*, was the key factor driving metaphor processing. These findings converge with those of Asmuth and Gentner (2017) in pointing to relationality as a major factor driving verb mutability.

### *5.6.3 Relationality and the predicate role combine to drive verb mutability and online adjustment*

Based on the above discussion, we propose that verb mutability is driven by both semantic factors (that verb meanings tend to be relational) and pragmatic factors (that verbs play the predicate role). These factors compound in driving verbs' greater propensity for online adjustments to their meanings. One specific proposal is that relational concepts like verb meanings have greater *interactive potential* than object concepts (Gentner, 1981). The idea is that verb representations include relations that take external arguments (e.g., CAUSE(Event(X,Y) Event(Y,Z)), where X, Y, and Z are external participants. Entity noun representations, in contrast, have comparatively few external relations. Verbs' higher interactive potential means that they are "relatively more subject than nouns to external contextual influences and less constrained by internal influences" (Gentner, 1981; p. 175). Compounding these semantic pressures is the pragmatic pressure exerted by the predicate role, which requires that the verb meaningfully relate to its external noun argument(s). This will often require adjusting one or more of the verb's typical semantic components, as in our studies.

## 5.7 Implications and Future Work

From the theoretical discussion above, one would expect these findings to generalize to transitive sentences. Gentner and France (1988) (Experiment 3a, b) found evidence that verbs adjust their meanings to those of their direct objects. For example, a sample paraphrase of *Marvin discarded a doctor* was "*Marvin consulted a different practitioner of medicine*. However, the generality of this pattern and its relation to polysemy need further investigation.

Our findings also lead to the intriguing prediction that nouns and verbs should have different characteristic patterns of word senses. First, the patterns found here suggest that verbs' greater polysemy than nouns is the result of their greater propensity for online adjustment. Further, there should be qualitative differences between verbs and nouns in their characteristic word senses. Specifically, verbs should have many word senses that are metaphorically/analogically related to the verb's literal meaning. Nouns should have many metonymic word senses and fewer metaphoric senses. We have found preliminary evidence for this prediction (King, Gentner, & Mo, 2021, 2022). If this pattern holds, it will provide another link between synchronic processes of sentence comprehension and diachronic processes of word-sense formation.

# 6 Conclusion

We have shown that verb meanings are more mutable than noun meanings: under semantic strain, verb meanings are altered to a greater degree than are noun meanings, with the verb's degree of change increasing as strain increases. We further showed that, although sense-selection plays a role for both nouns and verbs, the verb mutability effect is driven chiefly by online adjustment. Further, beyond the difference between nouns and verbs in the *degree* of meaning change under strain, we also found qualitative differences in *how* nouns and verbs change meaning. Whereas nouns were likely to be paraphrased with a taxonomically or associatively related term, verbs were most likely to be paraphrased metaphorically. These findings bear on the nature and processing of *verb* metaphors, an important and underexplored aspect of language use. Finally, these results provide a link between synchronic processing and diachronic change over language evolution.

# Chapter 2

## 1 Introduction

The ultimate goal of the dissertation will be to provide an empirically supported process account of verb metaphor comprehension. In Chapter 1, we laid the foundations for this goal by establishing three key aspects of verb meaning extension: (1) the verb mutability effect, (2) that verb mutability is driven primarily by online adjustment processes rather than sense selection, and (3) that it results in primarily analogical/metaphoric extensions of the verb's meaning.

The objective of this chapter is to build on those findings by more rigorously and specifically delineating the phenomena of verb metaphor. That is, we know that verbs change meaning more readily (and more metaphorically) than nouns under semantic strain, but *how* are they changing? Obtaining a more detailed and systematic understanding of the patterns of meaning change that occur when verbs extend metaphorically will allow for the formulation of a process account that that is consistent with—and explanatory of—those patterns.

### 1.1 Minimal Subtraction

As mentioned in the General Discussion of Chapter 1, Gentner and France's (1988) investigation of the verb mutability effect included a proposal that verb meaning change under strain follows a principle they called *minimal subtraction*. The basic idea is that verbs under strain will typically change the minimal amount necessary in order to accommodate the noun:

> According to the minimal subtraction view, people interpret the [semantically-strained] sentences by performing the minimal necessary adjustments in verb meaning, rather than

> by postulating a general change of state and/or substituting a contextually appropriate verb (Gentner & France, 1988; p. 364)[11]

In Chapter 1, we noted instances where we observed a pattern consistent with view in the paraphrases, e.g.:

| | **Original Sentence** | **Participant Paraphrase** |
|---|---|---|
| (1) | The professor complained | *The adult whined* |
| (2) | The bell complained | *The alarm rang annoyingly* |
| (3) | The box complained | *The container would not close*. |

These paraphrases suggest a close connection between the degree of semantic strain and the degree and nature of verb abstraction. As strain increases from sentence (1) to sentence (3), *complained* becomes increasingly abstracted. Further, this abstraction appears to follow a specific qualitative pattern, dropping domain-specific components (e.g., components related to the biological production of sound) and retaining more abstract, domain-general ones (e.g., that complaining implies a negative state of affairs).  This suggests that the verb is abstracted only to the degree required to be compatible with the noun if is paired with.

Thus, Gentner and France's minimal subtraction  proposal suggests a potential framework for describing how verbs extend metaphorically as a function of the nouns they are paired with. In this chapter, we present a series of experiments that systematically investigate the minimal subtraction hypothesis of verb change in-depth. We begin by formulating a specific set of testable principles that define minimal subtraction. Building on the work of Gentner and France (1988) as well as our findings in Chapter 1, we propose that minimal subtraction comprises the following three principles:

[11] Gentner and France provided initial empirical evidence for their account for eight verbs of possession—see Gentner and France (1988; Experiments 3A & 3B).

1. The degree of meaning change the verb undergoes increases progressively as the degree of strain increases
2. The verb typically changes meaning only as far as necessary to resolve the strain
3. Domain-specific meaning components are adjusted before more abstract, relational components

We tested this account on verbs from three different classes (adapted from Levin, 1993): manner of motion verbs that denote legged motion on land by a human or animal (e.g., *limp, sprint, waddle*), manner of speaking verbs that denote sounds made by the mouth of a human or animal (e.g., *mumble, cackle, bellow*), and bodily process verbs that denote biological processes humans or animals experience (e.g., *sweat, hiccup, exhale*). We will refer to these three classes as motion, sound, and body-process verbs. The verbs were paired with three different types of nouns: humans (e.g., *teacher, doctor, husband*), artifacts (specifically, vehicles: e.g., *plane, boat, scooter*), and abstract concepts (e.g., *melody, wisdom, rumor*). Combining each verb with a human, artifact, and abstract noun allowed for systematically generating sentences with progressively increasing levels of semantic strain (e.g., *The woman limped, The wagon limped, The fantasy limped*). As in Chapter 1, participants were instructed to provide meaningful paraphrases of each sentence (i.e., interpret them without repeating any content words) so that the degree and nature of verb (and noun) meaning change that occurred could be assessed.

The logic of the design is as follows. Selecting verbs from three different classes allowed for testing for general, cross-class patterns, as well as differences between classes and within classes. Across classes, every verb's preferred subject noun was human or animal. This means that shifting the subject noun from human to vehicle to abstract noun should create progressively increasing degrees of semantic strain for all verbs. Assuming that the degree of strain generated by each noun type is roughly for each verb class, then we expect roughly the same degree of abstraction across classes as strain increases.

Within each class, we expect similar, class-specific patterns of abstraction to be shared for verbs of that class. For example, we expect that all motion verbs will retain the notion of physical motion when paired with a vehicle noun (e.g., *The wagon limped*), but that motion will no longer be via legs (the adjustment necessary for a motion verb to accommodate a vehicle noun). When paired with an abstract noun (e.g., *The fantasy limped*), we expect that all motion verbs will discard the notion of physical motion entirely but retain a metaphorical construal of literal motion. Similarly, we expect that all manner of speaking verbs will retain the notion of physical sound production when paired with a vehicle noun (e.g., *The boat bellowed*) but will become abstracted out of the physical domain when paired with an abstraction noun (e.g., *The wisdom bellowed*).

Finally, at the most fine-grained level, while we expect the same general pattern of abstraction for verbs within a class, the interpretation of the metaphor should also vary within each class as a function of the specific verb. For example, *The rumor trudged* should be interpreted differently than *The rumor sprinted*, as should *The van sang* compared to *The van whispered*, or *The scooter slept* vs. *The scooter sneezed*.

The remainder of this chapter is as follows. In Experiments 1A and 1B, we used the same paraphrase paradigm as in Chapter 1 to test the quantitative predictions of minimal subtraction: namely, that the degree of verb change should increase progressively as the verb goes from being paired with a human noun, to an artifact noun, to an abstract noun. Verb mutability also predicts noun meanings to remain stable. In Experiment 2, we go beyond word2vec scores to test the qualitative predictions of minimal subtraction (that verb meaning change proceeds from domain-specific to domain-general meaning components) by coding the paraphrases directly for the nature of the verb abstraction. In Experiments 3A and 3B, we used a retrace task to confirm that our findings are best explained by semantic operations over the verb's representation, rather than by a process that involves discarding the verb's meaning and replacing it with an action saliently associated with the noun.

# 2 Experiment 1A

The purpose of Experiment 1A was threefold. First, it served as an opportunity to replicate the verb mutability effect from Chapter 1. Second, was a test of the quantitative components of the minimal subtraction hypothesis. Third, the paraphrases generated served as stimuli for testing the qualitative predictions of minimal subtraction in Experiment 2.

## 2.1 Method

### *2.1.1 Participants*

80 native English-speaking university undergraduates completed the study in person in the laboratory. They received course credit in an introductory psychology class for their participation.

### 2.1.2 *Materials*

18 nouns and 54 verbs (see Table 5) were combined to generate 162 target stimulus sentences (See Appendix E). 20 literal filler sentences were also included. All items were simple intransitive sentences of the form *The noun verb-ed* (e.g., *The wisdom pranced*).

The 54 verbs selected were divided evenly among the 3 verb classes: 18 manner of motion (motion) verbs, 18 manner of speaking (sound) verbs, and 18 bodily process (body-process) verbs. The 18 nouns were evenly divided among the 3 different types: 6 human nouns, 6 artifact nouns (all vehicles), and 6 abstract nouns. Each verb was combined with a noun of each type to generate sentence triplets that constituted three levels of semantic strain: (1) unstrained (literal), resulting from combining the verb with a human noun (*The doctor pranced*), (2) moderately strained, resulting from combining the verb with an artifact noun (*The boat pranced*), and (3) highly strained, resulting from combining the verb with an abstract noun (*The wisdom pranced*).

Table 5

*Nouns and verbs used in Experiment 1.*

| **Nouns** | | **Verbs** | | |
|---|---|---|---|---|
| **Type** | **Noun** | **Manner of Motion** | **Manner of Speaking** | **Bodily Process** |
| Human | doctor | jogged | babbled | blinked |
| | husband | limped | bellowed | breathed |
| | lawyer | marched | cackled | burped |
| | student | paced | chanted | coughed |
| | teacher | plodded | cried | drooled |
| | woman | pranced | grunted | exhaled |
| Artifact | boat | prowled | moaned | hiccupped |

| | | | | |
|---|---|---|---|---|
| | plane | scampered | mumbled | panted |
| | scooter | scurried | murmured | slept |
| | train | shuffled | muttered | sneezed |
| | van | sprinted | sang | snored |
| | wagon | staggered | screamed | spat |
| Abstract | fantasy | strolled | shouted | swallowed |
| | melody | strutted | stammered | sweated |
| | mood | stumbled | wailed | vomited |
| | religion | trudged | whimpered | wheezed |
| | rumor | waddled | whispered | winked |
| | wisdom | waltzed | yelled | yawned |

### 2.1.3 *Design*

The 162 target stimuli were divided into 9 between-subject item groupings of 18 items each. Each item grouping consisted of 6 literal sentences (containing each of the 6 human nouns), 6 moderately-strained sentences (containing each of the 6 artifact nouns) and 6 highly-strained sentences (containing each of 6 abstract nouns). Across the 18 target items, each noun and verb occurred exactly once. Each participant saw all 18 nouns and 6 motion verbs, 6 sound verbs, and 6 body-process verbs. Each item grouping also included the same 20 literal filler sentences across participants; thus, each participant paraphrased a total of 38 sentences: 18 target items and 20 filler items. Since 6 of the target items were unstrained, participants paraphrased a total of 12 strained and 26 unstrained sentences.

### *2.1.4 Procedure*

Each participant was randomly assigned to one of the nine item groupings and completed the experiment on a computer in the lab. Participants first viewed the same instructions as in Chapter

1, informing them that they would be asked to provide interpretations for a number of different sentences, some of which would be normal sentences, and some of which would be “odd”. They were explicitly instructed not to translate sentences mechanically (word-by-word, e.g., *The slimy orator* → *The gooey speaker*), but rather to think of a *meaningful* interpretation of the sentence that captures what the speaker might have meant (e.g., *The slimy orator* → *The corrupt politician*).

Sentences were presented one at a time, and participants typed their interpretations. Two literal filler sentences were always presented first, and the remaining 36 items (18 target and 18 filler) were presented in random order for each participant. Once a participant had submitted a paraphrase, they could not go back to change it later.

### *2.1.5 Coding*

The 80 participants generated a total of 1440 paraphrases of the target sentences (an average of 8.89 paraphrases per item). Following the same procedure and criteria described in Chapter 1, two human coders, blind to the hypotheses, were used to exclude certain types of paraphrases from the analysis: blatantly noncompliant responses (e.g., *The husband jogged* → *the man*) and responses that did not constitute a meaningful interpretation of the sentence, either by translating the sentence mechanically (e.g., *The wisdom pranced* → *The knowledge leapt*) or describing the surrounding context rather than interpreting (e.g., *The boat stumbled* → *There were rocky waters*). There was substantial initial agreement between the two judges, $\kappa = 0.80$, (95% CI, 0.78 to 0.83), $p < .0001$. A summary of the results of the coding task is shown in Table 6.

Next, any of the 1058 remaining paraphrases containing a pronoun in place of the original subject noun (e.g., *The woman grunted* → *She groaned*) were also omitted; this was done to avoid confounds in our method for assessing meaning change (see below). Only paraphrases where the pronoun clearly referred only to the original noun subject were excluded; any instances where a pronoun was present but did not clearly refer to the original subject (e.g., *The wisdom waddled* → *It became harder to think*) were not excluded. A total of 86 of the remaining 1058 paraphrases were excluded based on this criteria (see Table 6).

Table 6

*Results of the paraphrase coding task from Experiment 1*

| | **Number Excluded** | **% Total** |
|---|---|---|
| Noncompliant Exclusions | | |
| *Describes the Situation* | 193 | 13% |
| *Mechanical* | 162 | 11% |
| *Other / Unsure* | 41 | 3% |
| Pronoun Exclusions | 86 | 6% |
| **Total Excluded** | **482** | **33%** |
| **Net included in analysis** | **958** | **67%** |

*Note.* Percentages calculated out of a total of 1440 paraphrases.

### *2.1.6 Assessing semantic adjustment*

As in Chapter 1, we used word2vec to assess the degree of meaning change each stimulus noun and verb underwent from the original sentence to its paraphrase. Following the same procedure, we obtained a noun cosine similarity score and a verb cosine similarity score for each paraphrase, representing the degree of meaning change each stimulus noun and verb underwent from the original sentence to its paraphrase. Cosine similarity scores range from -1 to 1, with scores close to 1 representing high similarity, and scores close to 0 representing low similarity.

As described above, any paraphrases where the original noun was replaced with a pronoun were excluded from the analysis. The concern was that pronouns would artificially depress noun cosine similarity scores. Extremely high-frequency words like pronouns are problematic for WEMs like word2vec, which rely on differences in the contexts in which words appear to determine differences in their meanings. Words that appear in nearly all contexts (e.g., pronouns, prepositions, articles, etc.) therefore have little meaning in the model and cannot be effectively compared using cosine similarity scores. For our purposes, this means that paraphrases with a pronoun subject can result in vastly lower noun similarity scores than paraphrases with a close noun synonym, despite the fact that pronouns represent an intended total preservation of the original noun referent.[12] Verbs did not appear to have the same problem in our dataset: there were no instances in which the verb was paraphrased in a similar manner (e.g., *The wagon limped* → *The cart did it*).

## 2.2 Results

Our predictions are as follows. First, verbs should change meaning in response to semantic strain, but nouns should not (the verb mutability effect). Second, the degree of verb change should increase progressively as a function of the degree of strain, which itself increases based on the type of subject noun the verb receives. Verbs with a human noun subject constitute an unstrained (literal) sentence and should change the least. Verbs with an artifact subject should result in a moderately-strained sentence and should change more than unstrained verbs, while

[12] For example, following our methodology for generating cosine similarity scores, comparing the noun *daughter* to *the girl limped down the road* generates a fairly high noun score of 0.42, while comparing *daughter* to *she limped down the road* generates a noun score of 0.07. The verb score, however, is relatively unaffected by the presence of the pronoun: it is .48 and .52, respectively.

verbs with an abstract noun subject should result in a highly strained sentence and should change the most of all. Recall that lower word2vec cosine similarity scores correspond to greater degrees of meaning change; therefore, we expect that scores will decrease significantly as strain increases.

To test these predictions, two linear mixed effect models were fit for each of the three verb classes, one testing verb change and one testing noun change. In each model, word2vec score was the dependent measure, noun type (human, artifact, and abstract) was the fixed effect, and subjects and items were included as random effects. The fitted models are plotted in Figure 7. below.

To test the first prediction—that verbs (but not nouns) will change meaning in response to strain—each model was entered into a type III ANOVA test of fixed effects provided an omnibus test for the effect of semantic strain on noun and verb change (see Table 7)[13]. For all three verb models (corresponding to motion, sound, and body-process verbs) there was a significant main effect of strain (all *p*s < .05). In contrast, none of the three noun models showed a significant effect of strain (all *p*s > .05). Thus, as predicted by the verb mutability effect, verbs—but not nouns—changed meaning in response to strain.

To test the second prediction—that the degree of verb meaning change increases progressively as a function of noun type—planned pairwise contrasts of the three levels of the noun type predictor were conducted for each verb model in order to compare the extent to which verbs changed at each level of strain (see Table 8). All *p*-values were adjusted for multiple comparisons using

---

[13] Analyzing our fitted model using a Type III ANOVA allowed for testing for an overall main effect of noun type (i.e., across all three levels).

Tukey's method. Cosine similarity scores for motion verbs were consistent with the minimal subtraction pattern: items with human subjects scored marginally higher than items with artifact subjects ($p = .06$) and significantly higher than items with abstract subjects ($p < .0001$); artifact-noun items scored significantly higher than abstract-noun items, $p = .004$. Sound and body-process verbs, however, showed a different pattern: while items with a human subject noun scored significantly higher than both artifact- and abstract-noun items, there was no significant difference between artifact-noun and abstract-noun items.

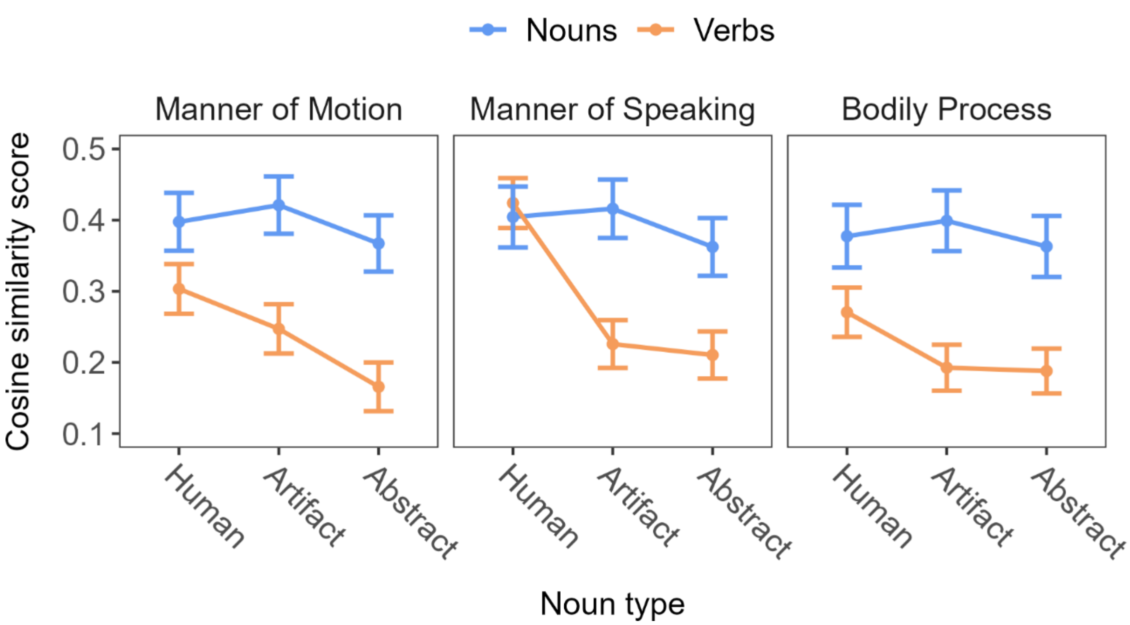



Figure 7. Fitted model plots from Experiment 1A. Each panel represents the results for items from that verb type. Lower word2vec scores correspond to greater degrees of meaning change. Error bars represent 95% CIs.

Table 7

*Results of Type III test of fixed effects for Experiment 1A models.*

| Verb Class | Model | F | *p* |
|---|---|---|---|
| Manner of Motion | Verb | 16.31 | < .0001* |
| | Noun | 2.13 | .13 |
| Manner of Speaking | Verb | 50.23 | < .0001* |
| | Noun | 2.28 | .11 |
| Bodily Process | Verb | 8.39 | .001* |

| | | | |
|---|---|---|---|
| | Noun | .93 | .40 |

Note: For each model, noun type (human, artifact, and abstract) was the only fixed effect. The denominator degrees of freedom were obtained using Satterthwaite's method.

Table 8

*Results of pairwise contrasts of word2vec scores for each noun type within each verb model.*

| Model | Contrast | Estimate | SE | *t* | *p* |
|---|---|---|---|---|---|
| Manner of Motion | *Human – Artifact* | .06 | .02 | 2.31 | .06 |
| | *Human – Abstract* | .14 | .02 | 5.67 | < .0001* |
| | *Artifact - Abstract* | .08 | .02 | 3.73 | .004* |
| Manner of speaking | *Human – Artifact* | .20 | .02 | 8.40 | < .0001* |
| | *Human – Abstract* | .21 | .02 | 9.09 | < .0001* |
| | *Artifact - Abstract* | .02 | .02 | 0.67 | .78 |
| Bodily Process | *Human – Artifact* | .08 | .02 | 3.48 | .001* |
| | *Human – Abstract* | .08 | .02 | 3.64 | < .001* |
| | *Artifact - Abstract* | .004 | .02 | 0.18 | .86 |

Note. *P-values adjusted for multiple comparisons using Tukey's method*.

### 2.3 Discussion

The predictions for this experiment were: (1) we should observe a verb mutability effect, and (2) the degree of verb change should increase as the degree of semantic mismatch between noun and verb increases.

The results matched Prediction 1: across all three verb classes, verbs changed meaning in response to semantic strain, but nouns did not. Our findings regarding Prediction 2 were mixed: motion verbs changed meaning in a manner fully consistent with minimal subtraction, while sound and body-process verb meaning change was partially consistent. For motion verbs, the degree of change increased progressively with the degree of strain. Verbs paired with a human subject changed the least, followed by verbs paired with an artifact subject, and verbs with an abstract subject changed the most. This pattern suggests that, on average, motion verbs changed only as far as was necessary to accommodate the noun. Sound and body-process verbs, however, deviated from this pattern: in both cases, while items with a human subject resulted in the least amount of verb meaning change (as expected), there was no significant difference in the degree of change between items with artifact noun subjects and those with abstract noun subjects (see Figure 7). We will return to this point in detail below.

Lastly, an informal review of the paraphrases provides initial evidence that verbs in all three classes changed meaning in a manner consistent with the *qualitative* predictions of minimal subtraction—that verb change proceeds from domain-specific to domain-general adjustments. Table 9 lists example paraphrases for Experiment 1A. We will delay discussion of this pattern until Experiment 2, where we test it more formally. But before moving on, however, several concerns with the current results bear addressing.

Table 9

*Example paraphrases from Experiment 1A*

| **Verb class** | **Noun type** | **Stimulus sentence** | **Paraphrase** |
| --- | --- | --- | --- |
| Manner of Motion | Human | *The woman limped* | *The girl walked favoring one leg* |
| | Artifact | *The wagon limped* | *The damaged cart creaked along* |
| | Abstract | *The fantasy limped* | *The story moved along slowly* |
| | Human | *The husband paced* | *The wife's significant other walked back and forth* |
| | Artifact | *The scooter paced* | *The vehicle moved steadily* |
| | Abstract | *The rumor paced* | *The gossip went back and forth* |
| | Human | *The doctor pranced* | *The physician danced* |
| | Artifact | *The boat pranced* | *The watercraft glided across the water* |
| | Abstract | *The wisdom pranced* | *Knowledge spread easily* |
| Manner of Speaking | Human | *The doctor bellowed* | *The physician yelled* |
| | Artifact | *The boat bellowed* | *The ship blew its horn* |

| | | | |
|---|---|---|---|
| | Abstract | *The wisdom bellowed* | *The truth spread* |
| | Human | *The student cried* | *The pupil wailed* |
| | Artifact | *The van cried* | *The car broke down* |
| | Abstract | *The mood cried* | *Tension filled the room* |
| | Human | *The student sang* | *The person in class belted out a tune* |
| | Artifact | *The van sang* | *The car engine rumbled* |
| | Abstract | *The mood sang* | *His/her feelings could be read easily* |
| Bodily Process | Human | *The teacher blinked* | *The lecturer's eyes closed and then opened again* |
| | Artifact | *The plane blinked* | *The signals on the plane flashed* |
| | Abstract | *The melody blinked* | *The music sounded rapidly and sporadically* |
| | Human | *The lawyer vomited* | *The lawman threw up* |
| | Artifact | *The train vomited* | *The public transportation let out a ton of people* |
| | Abstract | *The religion vomited* | *The ideology rejected an idea or practice* |
| | Human | *The doctor burped* | *The healthcare professional belched* |
| | Artifact | *The boat burped* | *The vehicle expelled gas* |
| | Abstract | *The wisdom burped* | *A smart idea came out of nowhere* |

### *2.3.1 Paraphrase exclusions*

Our first concern is whether omitting noncompliant paraphrases during the coding section of the analysis distorted the findings (see Table 6). These exclusions resulted in a relatively high proportion of the data being discarded: 396 paraphrases, or 28% of the 1440 total paraphrases were excluded (144 human noun items, 131 artifact noun items, and 121 abstract noun items). This could be problematic if the rate of exclusion is correlated with noun type—that is, with the degree of semantic strain. Fortunately, pairwise *t*-tests indicated no significant differences in rate of exclusion by noun type of strain (all *p*s > .05). This is reassuring in that it suggests that (1) the degree of change observed in verbs in response to strain is not attributable to differential rates of paraphrase exclusions, and (2) participant compliance did not vary significantly as a function of strain.

### *2.3.2 Differing patterns of change by verb class*

Our second concern regards the aforementioned inconsistency in the degree of meaning change between verb classes. As noted above, only motion verbs fully matched the predicted pattern, while sound and body-process verbs deviated from that pattern. For motion verbs, the degree of change increased steadily as a function of noun type, but for sound and body-process verbs, there were no significant differences in the degree of change between artifact and abstract items.

What explains this disparity? One possibility is that the differing patterns of meaning change observed are due primarily to differing patterns of semantic strain between verb classes. We assumed that verbs paired with equivalent noun types would be equivalently strained, regardless of verb class. That is, pairing a sound verb or body-process verb with an artifact noun (e.g., *The boat cackled* or *The boat burped*) will result, on average, in items that are equally strained to those resulting from pairing an motion verb with an artifact noun (e.g., *The boat sprinted*), and likewise for abstract nouns. When items are equally strained, we expect approximately the same degree of verb meaning change during interpretation.

If, however, the strain of items for a particular noun type is *not* equivalent between verb classes, then the minimal subtraction hypothesis would predict differing patterns of meaning change as well. Specifically, if the items containing artifact nouns and abstract nouns are closer to one another in degree of strain for sound and body-process verbs than they are for motion verbs, we would expect them to be closer in terms cosine similarity scores as well, as occurred here. To test this possibility, in Experiment 1B, we obtained direct ratings of the semantic strain of every item tested here, which allowed us to assess the strain of items across verb classes and to model the degree of verb and noun change as a function of the degree of semantic strain directly.

# 3 Experiment 1B

## 3.1 Method

### 3.1.1 *Participants*

90 participants completed the study online via Amazon Mechanical Turk. 20 workers were excluded for failing catch trial criteria for a net of 70 participants total. Participants were paid at a rate equivalent to Illinois' minimum wage at the time of the study. Only participants that responded "Yes" to a question asking if they were native English speakers completed the experiment.

### *3.1.2 Materials*

The materials consisted of the 162 target sentences and 18 literal filler sentences from Experiment 1A, as well as 4 new nonsensical sentences (*Under cloud ran, The of speak, The catch smirking*, and *Quickly above did*). The fillers and nonsensical sentences served as catch trials/attention checks.

### *3.1.3 Design*

The stimulus sentences were divided into the same 9 between-subject item groupings of 18 items each used in Experiment 1A. Each participant saw each of the 18 original nouns and 18 of the original 54 verbs exactly once, consisting of 6 motion verbs, 6 sound verbs, and 6 body-process verbs. The 18 literal filler sentences and 4 nonsense sentences served as catch trials. Thus, each participant rated 40 items total (18 target items and 22 catch trials), comprising 24 literal sentences (6 unstrained target sentences and 18 literal catch trials), 12 strained sentences, and 4 nonsensical sentences.

### *3.1.4 Procedure*

Each participant was randomly assigned to one of the nine item groupings and completed the experiment online. Participants were instructed to rate each sentence for comprehensibility on a scale of 1 to 10 based on how hard the sentence would be for a "typical person" to understand. A rating of 1 corresponded to *very hard for most to understand*, and a rating of 10 corresponded to *very easy for most to understand*.

The first and second trials were always a literal catch trial and nonsense catch trial, respectively, as were the second-to-last and last trials. The remaining 36 trials (comprising 18 target trials, 16 literal catch trials, and 2 nonsense catch trials) were presented in randomized order across participants. Sentences were presented one at time, and participants could not go back to revise previous responses once they submitted a rating.

The criterion for failing a catch trial was a rating of less than 8 out of 10 on a literal catch trial or greater than 3 out of 10 on a nonsense catch trial. Any participant that failed more than 2 catch trials was excluded from the analysis. 20 participants were excluded based on this criteria, for a net of 70 participants. After these exclusions, there was an average of 7.78 ratings per item (SD = 1.32).

## 3.2 Results

If the difference in word2vec scores by verb class is attributable to differences in the strain of the items, this predicts that: (1) the pattern of average strain rating by noun type (human, artifact, abstract) should parallel the pattern of average word2vec scores by noun type for each verb class, and (2) regressing word2vec scores on strain ratings directly should result in a linear relationship. We expect no significant relationship between noun scores and semantic strain.

We first report strain ratings for each item and in aggregate by noun type. Next, we report the results of modeling word2vec scores as a function of both strain ratings and mean strain rating by noun type.

### 3.2.1 *Strain ratings*

As in Chapter 1, to obtain strain ratings for each item, the comprehensibility scale was first inverted by subtracting 10 from each rating such that a low score now corresponds to a low degree of strain and a high score corresponds to a high degree of strain. Next, a linear mixed effect model was fit, with strain rating as the dependent measure, item as the fixed effect, and subjects entered as random effects. The resulting strain ratings for each item are plotted by verb class and noun type in Appendix F.

To test Prediction (1)—that mean strain ratings by noun type will parallel the word2vec scores from Experiment 1A—a second linear mixed effect model was fit, with strain rating as the dependent measure, noun type, verb class, and the interaction included as fixed effects, and subjects and items as random effects. The fitted models are plotted by verb class in Figure 8 below. The figure shows that the pattern of mean strain ratings by noun class closely parallels the pattern of word2vec scores seen in Experiment 1A, with the strain of artifact- and abstract-noun items closer together for sound and body-process verbs than for motion verbs (the scale is inverted compared to the word2vec scores in Figure 7 because strain and cosine similarity scores are inversely related). Planned pairwise contrasts of strain rating by noun type within each verb class confirmed the same pattern held (see Table 10). While items with all three noun types were significantly different from one another in strain for motion verbs (all *p*s < .0001), for sound and body-process verbs there was no significant difference between items with artifact and abstract nouns (*p*s > .24).

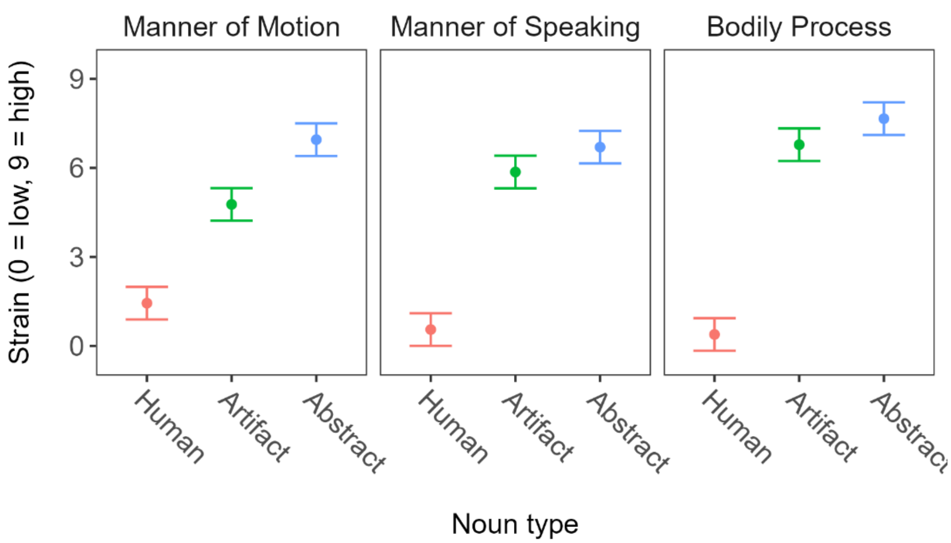



Figure 8. Mean strain ratings by noun type and verb class from Experiment 1B. Higher scores indicate greater semantic strain. Error bars represent 95% confidence intervals.

Table 10

*Pairwise contrasts of strain rating by noun type for each verb class*

| Model | Contrast | Estimate | SE | *t* | *p* |
|---|---|---|---|---|---|
| Manner of Motion | *Human – Artifact* | -3.33 | 0.35 | -9.51 | < .0001* |
| | *Human – Abstract* | -5.51 | 0.35 | -15.74 | < .0001* |
| | *Artifact – Abstract* | -2.18 | 0.35 | -6.23 | < .0001* |
| Manner of speaking | *Human – Artifact* | -5.31 | 0.35 | -15.17 | < .0001* |
| | *Human – Abstract* | -6.15 | 0.35 | -17.56 | < .0001* |
| | *Artifact – Abstract* | -0.84 | 0.35 | -2.39 | .30 |
| Bodily Process | *Human – Artifact* | -6.39 | 0.35 | -18.26 | < .0001* |
| | *Human – Abstract* | -7.27 | 0.35 | -20.77 | < .0001* |

| | | | | |
|---|---|---|---|---|
| *Artifact - Abstract* | -0.88 | 0.35 | -2.51 | .24 |

*Note.* Contrast column indicates noun types of the items for that comparison within that verb class. All *p*-values adjusted using Tukey's method.

### *3.2.2 Materials*

The paraphrases from Experiment 1A served as the stimuli for this experiment. Only the paraphrases that had not been excluded during the coding process were included. The original verbs from Experiment 1A served as the response options for the retrace, with the exception of three body-process verbs (*snored*, *coughed*, and *burped*), which were omitted due to the fact that sound is a salient aspect of these three verbs' meanings, raising concerns of a confound with the sound verbs. In order to ensure that each verb class had the same number of unique distractors in total, three new body-process verbs were selected to replace them as distractors (see below): *choked, inhaled,* and *napped*. After these exclusions, a total of 897 paraphrases from 51 of the original 54 verbs were included in the Experiment.

### *3.2.3 Modeling word2vec score as a function of strain*

To test Prediction 2—that the degree of verb change (but not noun change) is proportional to the degree of semantic strain, we refit new noun and verb models such that the word2vec scores from Experiment 1A were regressed on the new continuous strain ratings rather than in aggregate by noun type. As in Experiment 1A, two linear mixed effect models were fit for each of the three verb classes, one for verbs and one for nouns, for a total of six models. In each model, word2vec score (either noun or verb) was the dependent measure, our new continuous measure of semantic strain was the fixed effect, and subjects and items were entered as random effects. Model summaries are shown in Table 11, and fitted model plots are shown in Figure 9.

As predicted, for all three verb models, the effect of semantic strain was significant (all *p*s < .0001) such that as the degree of semantic strain increased, so did the degree of meaning change. Visual inspection of residual and normal probability plots confirmed a good linear fit for all three models (see

Appendix **G**). Also as expected, for all three noun models there was no significant effect of strain on the degree of noun meaning change.

Table 11.

*Model summaries for each of the three verb classes.*

| Verb class | Model | *β* | SE | *t* | *p* |
|---|---|---|---|---|---|
| Manner of Motion | Verb | -0.38 | 0.09 | -4.29 | < .0001* |
| | Noun | -0.05 | 0.07 | -0.68 | .50 |
| Manner of Speaking | Verb | -0.62 | 0.06 | -9.91 | < .0001* |
| | Noun | -0.09 | 0.08 | -1.24 | .22 |
| Bodily Process | Verb | -0.38 | 0.09 | -4.27 | < .0001* |
| | Noun | -0.01 | 0.07 | -0.14 | .89 |

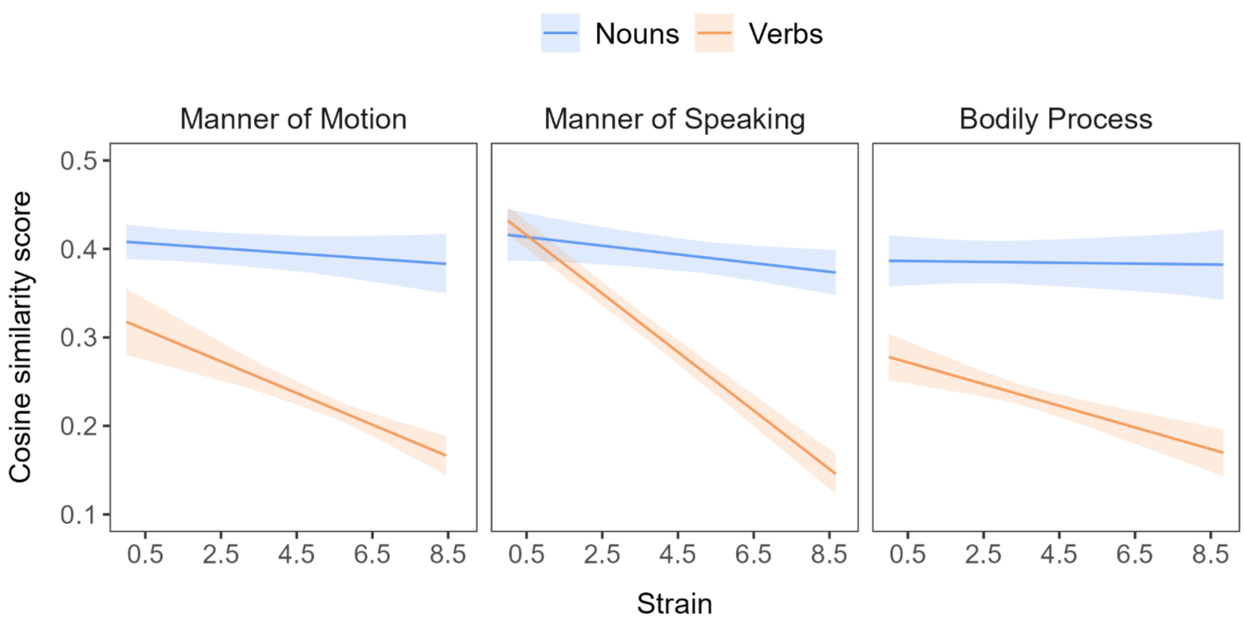



Figure 9. Fitted models using the continuous strain measure. Lower cosine similarity scores indicate greater meaning change. Shaded ribbons indicate 95% confidence bands.

### 3.3 Discussion

We hypothesized that the lack of a significant difference in mean word2vec scores between artifact noun and abstract noun items for sound and body-process verbs was attributable to a corresponding lack of difference in the degree of semantic strain for items of those noun types.

The present results support this hypothesis: for each verb class, mean item strain ratings by noun type matched the pattern of word2vec scores for those items from Experiment 1A. For motion verbs, mean strain ratings followed an approximately linear pattern, while sound and body-process verbs showed no significant difference in ratings for items with artifact and abstract nouns. Further, modeling the word2vec scores from Experiment 1A as a function of semantic strain directly—rather than as a function of noun type—resulted in a significant linear relationship for verbs (but not nouns) in all three verb classes. Thus, as predicted by minimal subtraction, the degree of verb change was proportional to the degree of semantic strain of the original sentence, suggesting that, on average, verbs adapt their meanings just as far as is necessary to resolve the strain.

## 4 Experiment 2

We now turn to our second main question of interest. Experiments 1A and 1B showed that the *degree* of verb change matched the expected pattern under minimal subtraction. In Experiment 2 we investigate the *type* of change that occurs under strain. Minimal subtraction predicts a qualitative pattern of change such that, as strain increases, domain-specific meaning components are adjusted before ore domain-general, abstract components. To test this hypothesis, we gave the paraphrases generated in Experiment 1A to a new group of participants who categorized them based on which components of the original verb's meaning were retained in the paraphrase.

## 4.1 Method

### *4.1.1 Participants*

72 native English-speaking university undergraduates completed the study in person in the laboratory. They received course credit in an introductory psychology class for their participation. A total of 5 participants were dropped for failing catch-trial criteria (described below) for a net of 67 participants included in the analysis.

### *4.1.2 Materials*

The stimuli were 944 of the 958 compliant paraphrases from Experiment 1A (see below). Paraphrases were divided by verb class such that each participant only received paraphrases originating from items from a single verb class (motion, sound, or body-process).

We designed a coding scheme with five categories meant to capture progressively increasing degrees of verb abstraction that applied across all three verb classes in a general sense—but with each code specified class-specific ways within each class (see Table 12). At a general level, Category 1 represented full retention of the verb's typical literal meaning (e.g., *The woman limped* → *The girl walked favoring one leg*). Category 2 was for paraphrases that showed adjustments to domain-specific concrete dimensions, but still retained the verb's meaning as denoting a physical event often occurring in a manner similar to the verb (e.g., *The wagon limped* → *The damaged cart creaked along*). Category 3 was for paraphrases where the verb no longer denotes a physical event, but still retains a metaphorical representation of the verb's original concrete meaning (e.g., *The fantasy limped* → *The dream moved along slowly*).

Categories 4 and 5 did not correspond to a specific degree of change. Category 4 was intended as an "other" category for paraphrases that did not fit into Categories 1 – 3 (e.g., in the case of a motion verb, paraphrases that did not describe physical nor metaphorical motion, such as *The van shuffled* → *The car had trouble*). Category 5 was simply *Unsure*, for ambiguous or nonsensical paraphrases.

Table 12

*Coding scheme used in Experiment 2.*

| **Verb Class** | **Category 1** | **Category 2** | **Category 3** | **Category 4** | **Category 5** |
|---|---|---|---|---|---|
| Manner of motion | *Physical motion involving legs* | *Physical motion not involving legs* | *Metaphorical motion* | *No physical or metaphorical motion* | *Unsure* |
| Manner of speaking | *Physical sound made by a human or other animal* | *Physical sound not made by a human or animal* | *Metaphorical soun d* | *No physical or metaphorical sound* | *Unsure* |
| Bodily Process | *Bodily process occurring in a human or other animal* | *Non-biological process occurring in a nonliving thing* | *Metaphorical process* | *No physical or metaphorical process* | *Unsure* |

*Note*. On each trial, the paraphrase to be coded was displayed at the top of the screen. Underneath was the prompt *Does the above sentence involve….*, followed by the five codes pertaining to that verb class, shown left to right in the order displayed above.

### *4.1.3 Design*

Target items (the paraphrases from Experiment 1A) were separated into nine between-subject item groupings based on the stimulus sentence they originated from. First, they were separated based on the class of the verb in the original stimulus sentence (in Experiment 1, there were 54 stimulus sentences with motion verbs, 54 with sound verbs, and 54 with body-process verbs). This was done so that each participant only had to learn a single coding scheme. Next, within each verb class, the 54 stimulus sentences from Experiment 1 were separated into 9 item groupings of 18 each. This ensured each participant only saw paraphrases originating from each stimulus noun and verb exactly once. That is, if a participant saw a paraphrase originating from the sentence *The wagon limped*, none of the other paraphrases they saw came from sentences containing *wagon* or *limped*. Participants were not shown the original sentence; they only saw the paraphrase.

Within each item grouping, the 18 target paraphrases were presented in 3 blocks of 6 items each. Each block consisted of two paraphrases of items where the original sentence had a human noun, two paraphrases of items with an artifact noun, and two paraphrases of items with an abstract noun. Grouping the target paraphrases into blocks of six ensured that the strain of the underlying stimulus sentences from Experiment 1A and the noun types that composed them were distributed evenly throughout the experiment. Items were presented in randomized order within each block of six.

In addition to 18 target items, each participant received 8 practice items and 8 literal filler items that also served as attention checks/catch trials. Within each verb class, participants saw the same practice and filler items. The eight practice items consisted of two clear examples for each of the

Categories 1 through 4 for that verb class—i.e., for participants in the motion condition, there were two sentences that were clear examples of *physical motion involving legs* (Category 1), two that were clear examples of *physical motion not involving legs* (Category 2), two that were clear examples of *metaphorical motion* (Category 3), and two that were clear examples of *no physical or metaphorical motion* (Category 4). The eight literal filler items were created in the same way and served as catch trials. Any participant that answered incorrectly on more than two catch trials was excluded from the analysis.

Finally, since the exclusion of noncompliant paraphrases during the coding process in Experiment 1A resulted in an uneven number of paraphrases per item, "dummy" paraphrases were also included to allow for coding of the maximum number of paraphrases from Experiment 1A possible, while also ensuring that each participant received a uniform in terms of the criteria outlined above. This resulted in a total of 944 target items and 352 dummy items being tested: for motion, sound, and body-process items, there were 344, 312, and 288 target items, respectively, and 88, 120, and 144 dummy items, respectively. Dummy paraphrases were not included in the analysis; each of the 944 target items received exactly one rating.

### *4.1.4 Procedure*

Participants were randomly assigned to one of the item groupings. The experiment was completed individually, in person, on a computer. Participants were told that they would see a series of sentences and would be asked to rate each sentence on several dimensions related to *physical motion* (for those in the motion group), *sound* (for those in the sound group), or *processes* (for those in the body-process group). Each dimension (corresponding to each of the five categories described above) was explained and several positive examples were listed. The body-process group was also explicitly told that they were not to consider physical changes of location to constitute processes.

Items were presented one at a time, and participants could not go backwards to change earlier responses. Each item (paraphrase) was presented at the top of the screen (participants only saw the paraphrase from Experiment 1A, never the original stimulus sentence). Below was the prompt *Does the above sentence involve…* followed by the five codes for that verb class (e.g., for motion verbs, that as (1) *Physical motion involving legs,* (2) *Physical motion not involving legs,* (3) *Metaphorical motion,* (4) *No physical or metaphorical motion,* and (5) *Unsure*), presented horizontally in the same order as shown in Table 12. Participants could only select one of the five categories before continuing.

Participants first completed 8 practice items with feedback, followed by the remaining target and filler items without feedback. Literal filler items were covertly interspersed between blocks of target items: each block of target items was preceded by two filler items; the final target block was also followed by two filler items. The filler items were presented in the same order for every participant, but the target items were presented in randomized order within each block.

#### 4.1.4.1 *Coding for domain-general metaphors*

After data collection finished, it became clear that our original coding scheme did not account for the full range of verb abstraction possible. Specifically, the criteria for Category 3 were too narrow to account for abstract interpretations that did not explicitly reference the original literal domain. For example, in *The mood marched* → *His attitude didn't change*, the verb *marched* has been abstracted such that it no longer explicitly refers to the domain of motion, yet the verb still retains abstract, domain-general components of *marched*'s meaning, akin to *an event that is persistent*. Compare this to the Category 3 paraphrase *The mood marched* → *The intense feelings moved powerfully through the person*, in which the domain of motion is explicitly referenced, albeit in an abstract, metaphorical way. We will call extensions of this latter type to be *domain-specific* abstract metaphors, and those of the former type to be *domain-general* abstract metaphors.

For items with abstract nouns, both domain-specific and domain-general interpretations are consistent with minimal subtraction, as both involve abstracting away concrete meaning components (as is demanded by an abstract noun). For that same reason, however, domain-general abstractions represent a further degree of change than would typically be expected for items with artifact noun subjects. For example, in *The train stammered* → *The moving vehicle had an interruption in its movement*, the sound verb *stammered* has become highly abstracted and mapped to the domain of motion—i.e., the train's physical movement. Contrary to expectations, no concrete meaning components related the verb's original literal meaning (i.e., physical sound production) are retained. Nevertheless, as in the example of *The mood marched* → *His attitude didn't change*, the paraphrase remains analogically related to the original verb, sharing highly domain-general commonalities related to an interrupted process of some sort.

We found that many of the paraphrases coded as Category 4 (*Neither physical nor metaphorical*) by participants appeared to be these domain-general metaphoric extensions. Thus, we conducted an additional coding task on all 253 paraphrases that were coded by participants as Category 4 (65 motion paraphrases, 103 sound paraphrases, and 85 body-process paraphrases). Coding was conducted using two judges who were blind to the study's hypotheses. The coders received detailed instructions describing the nature of the original paraphrase task (i.e., Experiment 1A) and instructing them to mark any paraphrases that constituted metaphoric extensions of the original verb. A metaphor was defined as involving an analogy or abstract commonality or similarity between the original verb and its interpretation in the paraphrase. Thus, unlike the criteria for Category 3 (which only covered domain-specific metaphoric extensions), this definition allowed for domain-general metaphoric extensions that did not reference the concrete domain of the verb's literal meaning. The judges were given both positive and negative examples of metaphoric verb extensions and completed a practice session before coding the target items.

Paraphrases were presented alongside the original sentence, and the judges indicated whether each paraphrase involved a metaphoric extension of the verb (yes/no). Coding was done in chunks wherein each judge coded a set of paraphrases independently, followed by a reconciliation session where the judges came to an agreement on any disparities. Initial agreement between raters occurred on 89% of items, and the judges were able to reach a final consensus on all items. Cohen's κ was run to determine interrater reliability. There was moderate initial agreement between the two judges, $\kappa = .55$, (95% CI, 0.40 to 0.70), $p < .0001$. Some examples are shown in Table 13 below.

Table 13

*Example domain-general paraphrases from Experiment 2*

| **Stimulus sentence** | **Paraphrase** |
|---|---|
| *The lawyer snored* | *The representative was bored* |
| *The wagon whimpered* | *The wooden object to transport goods started to break down* |
| *The plane trudged* | *The machine struggled* |

| | |
|---|---|
| *The train stammered* | *The moving vehicle had an interruption in its movement* |
| *The wisdom shouted* | *The intelligence was evident* |
| *The melody trudged* | *The song continued* |
| *The religion waltzed* | *The belief was powerful* |

In the results presented below, the original domain-specific Category 3 from Table 12 has been renamed to Category 3A. Any items coded by the judges as being domain-general metaphors are categorized as Category 3B. Finally, any paraphrases not coded as domain-general metaphors by the judges remained in Category 4.

## 4.2 Results

Code tallies are displayed by the verb class and noun type of the original stimulus sentence in Table 14. Within each verb class, the degree of semantic mismatch between noun and verb increases downwards (row-wise), from human to artifact to abstract noun types, and the degree of verb abstraction represented by the categories increases from left to right (column-wise), with the exception of Category 4 and Category 5, which do not correspond to any particular degree of abstraction. Shaded cells indicate the codes predicted to be most frequent for that level of semantic strain (i.e. noun type) under the minimal subtraction hypothesis, and bolded numbers indicate the most frequent response for that noun type (row).

Table 14

*Response tallies and percentages by verb class from Experiment 2*

| **Sentence Type** | | **Category** | | | | | |
|---|---|---|---|---|---|---|---|
| | | (1) | (2) | (3A) | (3B) | (4) | (5) |
| **Verb Class** | **Noun type** | Literal meaning retention | Physical domain | Metaphorical (Domain-specific) | Metaphorical (Domain-general) | Neither physical nor metaphorical | Unsure |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Manner of Motion | Human | **75 (68.81%)** | 8 (7.34%) | 5 (4.59%) | 5 (4.59%) | 2 (1.83%) | 14 (12.84%) |
| | Artifact | 6 (5.41%) | **79 (71.17%)** | 6 (5.41%) | 13 (11.71%) | 1 (0.90%) | 6 (5.41%) |
| | Abstract | 2 (1.74%) | 2 (1.74 %) | **54 (46.96%)** | 37 (32.17%) | 7 (6.09%) | 13 (11.30%) |
| Manner of Speaking | Human | **76 (84.44%)** | 0 (0.00%) | 2 (2.22%) | 2 (2.22%) | 7 (7.78%) | 3 (3.33%) |
| | Artifact | 4 (3.64%) | **67 (60.91%)** | 2 (1.82%) | 29 (26.36%) | 1 (0.91%) | 7 (6.36%) |
| | Abstract | 10 (8.93%) | 10 (8.93%) | 20 (17.86%) | **57 (50.89%)** | 7 (6.25%) | 8 (7.14%) |
| Bodily Process | Human | **43 (55.13%)** | 1 (1.28%) | 8 (10.26%) | 13 (16.67%) | 10 (12.82%) | 3 (3.85%) |
| | Artifact | 1 (1.30%) | **39 (50.65%)** | 6 (7.79%) | 23 (29.87%) | 1 (1.30%) | 7 (9.09%) |
| | Abstract | 1 (1.23%) | 3 (3.70%) | 28 (34.57%) | **38 (46.91%)** | 0 (0.00%) | 11 (13.58%) |

*Note*. *Verb class* and *noun type* refer to the stimulus sentences that were paraphrased in Experiment 1A. Numbers in each cell correspond to the number of times paraphrases of sentences in that cell received the column code. Percentages sum to 100% for each row. Bolded text indicates most frequent response for that row. Shaded cells correspond to responses predicted to be most frequent under minimal subtraction. Columns are labeled generically to capture the degree of abstraction by each code; see Table 12 for the verb-class-specific codes seen by participants.

Significance testing was conducted using the following procedure. The goal was to identify the most frequent code category for each noun type (e.g., each row in Table 14). Since the most frequent category (the bolded text in Table 14) always fell into what was predicted by minimal subtraction (the shaded cells in Table 14), the purpose of each test was to see if the predicted category was the most likely response for that noun type.

To minimize the number of pairwise comparisons necessary, the most frequent category was compared to the next-most frequent category for that noun type / row. For each comparison, the most-frequent category was scored as a 1, the next-most frequent category was scored as a 0, and all other categories were excluded. Next, a logistic mixed effect regression model was fit, with this score as the dependent measure, the intercept as the only fixed effect, and subjects and items as random effects. A significant intercept therefore indicates that the most-frequent category was the most likely of all five for that row. If the first- and second-most frequent codes were not significantly different from each other, the first- and third-most frequent codes were compared using the same procedure. This occurred four times: for motion verbs with abstract noun subjects, and for body-process verbs for all three noun types. The results of all comparisons are shown in Table 15 below.

Table 15

*Pairwise contrasts from Experiment 2*

| Verb Class | Noun Type | Contrast[a] | *b* (logits) | SE | *z* | *p* |
|---|---|---|---|---|---|---|
| Manner of motion | Human | 1 – 5 | 1.77 | 0.41 | 4.38 | < .0001* |
| | Artifact | 2 – 3B | 2.71 | 0.92 | 2.94 | < .01* |
| | Abstract[b] | 3A – 3B | 0.52 | 0.48 | 1.1 | .27 |
| | | 3A – 5 | 1.86 | 0.60 | 3.08 | < .01* |
| Manner of speaking | Human | 1 – 4 | 2.39 | 0.40 | 6.04 | < .0001* |
| | Artifact | 2 – 3B | 0.96 | 0.38 | 2.52 | .01* |
| | Abstract | 3B – 3A | 1.17 | 0.39 | 2.96 | < .01* |
| Bodily Process | Human[b] | 1 – 3B | 2.22 | 1.85 | 1.20 | .23 |
| | | 1 – 4 | 1.68 | 0.54 | 3.09 | < .01* |
| | Artifact[b] | 2 – 3B | 0.53 | 0.28 | 1.94 | .053 |
| | | 2 – 5 | 1.84 | 0.66 | 2.80 | < .01* |
| | Abstract[b] | 3B – 3A | 0.34 | 0.38 | 0.90 | .37 |
| | | 3B – 5 | 1.30 | 0.43 | 3.06 | < .01* |

*Note*. [a] Refers to which two categories were compared (see Table 14). The first category listed is the more frequent of the two. [b] When first comparison was nonsignificant, most frequent category was compared again with third-most frequent category

As Table 15 shows, for motion and sound verbs, the predicted categories were significantly more likely to occur than any other category. For motion verbs, Category 3A (domain-specific metaphors) was not significantly more likely than Category 3B (domain-specific metaphors); however, as discussed earlier, both categories are consistent with the predictions of minimal subtraction for abstract nouns. Category 3A was significantly more likely than the third-most frequent category, Category 5 (Unsure).

Body-process verbs followed the same numeric pattern as motion and sound verbs; however, the most frequent code was not significantly more likely than the second-most frequent category for any of the three noun types (but was always significantly more likely than the third-most frequent category).[14] This pattern may be partly due to the fact that paraphrases from body-process verbs were excluded at higher rates for being noncompliant (see section 2.1.5 from Experiment 1A) compared to motion and sound verbs; thus, in this experiment, 234 body-process paraphrases were tested compared to 335 motion paraphrases and 312 sound paraphrases, possibly reducing power.

Taking all three verb classes together, however, the overall pattern of verb abstraction is strongly consistent with the predictions of minimal subtraction (see Table 16 below). We discuss possible reasons for the observed by-class variation in abstraction patterns in the next section.

Table 16

*Overall response tallies and percentages from Experiment 2*

| | **Category** | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3A) | (3B) | (4) | (5) |
| **Noun type** | Literal meaning retention | Physical domain | Metaphorical (Domain-specific) | Metaphorical (Domain-general) | Neither physical nor metaphorical | Unsure |
| Human | **194 (70.55%)** | 9 (3.27%) | 15 (5.45%) | 20 (7.27%) | 19 (6.91%) | 18 (6.55%) |
| Artifact | 11 | **185** | 14 | 65 | 3 | 20 |

[14] As with motion and sound verbs, the first- and second-most frequent codes for abstract noun items were as predicted under minimal subtraction.

| | | | | | | |
|---|---|---|---|---|---|---|
| | (3.69%) | **(62.08%)** | (4.70%) | (21.81%) | (1.01%) | (6.71%) |
| Abstract | 13<br>(4.22%) | 15<br>(4.87%) | **102<br>(33.12%)** | **132<br>(42.86%)** | 14<br>(4.55%) | 32<br>(10.39%) |

*Note*. Shaded cells indicated predicted response under minimal subtraction. Bolded cells indicate most frequent response for that row. Percentages sum to 100 for each row.

### 4.3 Discussion

Minimal subtraction makes the qualitative prediction that verb abstraction should proceed from domain-specific meaning components to domain-general components, with the extent and nature of these alterations depending on the noun the verb receives as an argument. Overall, the results matched these predictions (see Table 16). When the verb received a human subject, it most often retained its full literal meaning (Category 1). When it received an artifact subject, it was most often adjusted such that it still referred to a concrete event, but with the dimensions of that event adapted to accommodate the noun (Category 2). When it received an abstract subject, it was most often abstracted further such that it no longer referred to a physical event, but still retained relevant abstract relational components of its meaning (Categories 3A and 3B).

For example, in the case of the motion verb *limped*, when paired with the human noun *woman*, the verb retains its typical literal meaning (*The girl walked favoring one leg*); when paired with the artifact noun *wagon*, the manner of motion is adjusted such that the motion no longer involves legs but still retains the notion of slow, physical movement (*The damaged cart creaked along*). When paired with the abstract noun *fantasy*, *limped* is abstracted further such that the motion event is entirely metaphorical, rather than concrete (*The story moved along slowly*). Importantly, in all three cases, *limped* retains the general, highly abstract meaning of *impaired function/operation*.

As another example, the same pattern manifests for the body-process verb *vomited*: In *The lawyer vomited* → *The lawman threw up*, the verb's typical literal meaning is retained; in *The train vomited* → *The public transportation let out a ton of people*, concrete components of the verb's meaning are adjusted such the train ejects passengers rather than bodily fluids; in *The religion vomited* → *The ideology rejected an idea or practice*, the verb retains only highly abstract meaning components representing the general rejection of an unpalatable input.

While the overall pattern was as predicted, there was some variation in the results between verb classes (see Table 14). First, within all three classes, the most frequent codes matched predictions, but the contrasts were more consistently reliable (significant) for motion and sound verbs than body-process verbs—perhaps due to differences in power. Second, there was an interesting pattern regarding between-class differences in domain-specific (Category 3A) vs domain-general abstractions (Category 3B). Domain-general abstractions were more common for sound and body-process verbs than motion verbs, especially when combined with artifact and abstract nouns.

We suspect that there may be two reasons for this. First, the artifact nouns were all vehicles. Since a vehicle's primary purpose saliently involves movement, this may have promoted a focus on the motion domain, perhaps prompting participants to abstract sound and body-process verbs further to accommodate the noun (e.g., *The plane babbled* → *The jet shook*). Second, for many of the abstract nouns used, motion and spatial words are commonly used to describe their behavior. It is common in English for *rumors*, *wisdom*, and *religion* to *spread* or *circulate*, a *melody* or *fantasy* can be *fast* or *slow*, and a *mood* can be *up* or *down*. As with artifact nouns, this may have led participants to further abstract the verbs from their original domains to accommodate these motion events, discarding all domain-specific (i.e., sound or bodily) meaning components (e.g., *The rumor yelled* → *The lie spread fast and everyone knew it*; *The wisdom burped* → *A smart idea came out of nowhere*).

*4.3.1.1 The toss-and-replace possibility*

Thus far we have assumed that the changes to verbs' meanings reflected in the paraphrases result from operations over the verbs' representations—i.e., verb meaning components are adjusted or discarded in response to strain. But there is another possibility that must be addressed. Perhaps when encountering a semantically-strained sentence, participants simply discard most of the verb's meaning and replace it with an event that is saliently associated with the noun. For example, consider the below set of paraphrases from three different participants in Experiment 1A:

| **Original sentence** | **Paraphrase** |
|---|---|
| *The rumor jogged* | *The news spread* |
| *The rumor muttered* | *The gossip spread* |
| *The rumor drooled* | *The gossip spread* |

The most salient thing that rumors are known to do is spread, and all three of the above paraphrases describe spreading events, regardless of the verb involved. By simply discarding the verb and replacing it with an event saliently associated with the noun, the problem of adapting the verb to the noun is solved without needing to operate over the verb's meaning components at all. This account, which we will call *toss-and-replace*, could result in a similar pattern of results to our findings thus far, with the verb only *appearing* to have abstracted its meaning in a graduated manner in order to accommodate the noun (e.g., *The scooter jogged* → *The vehicle drove; The rumor jogged* → *The rumor spread*).

Of course, for some paraphrases, it would seem to be highly implausible that interpretation took place without taking the verb's meaning into account (e.g., *The plane blinked* → *The signals on the plane flashed*). Nevertheless, due to the similarity of outcomes that toss-and-replace and minimal subtraction produce—and because toss-and-replace poses a potential challenge to our finding in Chapter 1 that verb mutability operates primarily through online adjustments to the verb's representation—it is important to assess the extent to which the paraphrases in the present study could be produced with minimal reference to the verbs' meaning. In Experiments 3A and 3B, we conduct two such tests.

## 5 Experiment 3A

To test the toss-and-replace possibility, we used a retrace task (adapted from Gentner & France, 1988) in which participants tried to identify (i.e., retrace) the original verb for each paraphrase from a list of three options: the original verb and two distractors. In this experiment, each of the three options was from a different verb class (motion, sound, or body-process). In Experiment 3B, each of the three options was from the same verb class, making identifying the correct verb a more difficult task (and therefore constituting a finer-grained measure of verb meaning retention).

The logic is that if the toss-and-replace account is correct (that is, if comprehension mainly involves replacing the verb with an event associated with the noun), then the meaning of the paraphrases depends solely on the noun, and participants will be at chance in guessing the true original verb from among the distractors. Alternatively, if toss-and-replace is wrong and participants are indeed taking the verb's meaning into account during comprehension, participants should be above chance in identifying the original verb.

## 5.1 Method

### *5.1.1 Participants*

27 native English-speaking university undergraduates completed the study online via the Qualtrics platform. They received course credit in an introductory psychology class for their participation.

### *5.1.2 Design*

Target items (the paraphrases from Experiment 1A) were divided into three different item groupings of 51 items each based on the original stimulus sentences that the paraphrases originated from. Thus, within each item grouping, each underlying original verb occurred exactly once. Due to the exclusion of the three body-process verbs, each original noun occurred either two or three times. As in Experiment 2, while the item groupings were based on the original stimulus sentences in Experiment 1A, participants in this experiment never saw the original stimulus sentences, only their paraphrases.

Each paraphrase was presented to participants with three verbs listed horizontally underneath: the target verb, representing the true original verb, and two distractors, one from each of the two non-target verb classes. Thus, if the target was a motion verb, there was one body-process distractor and one sound distractor. Distractors were randomly assigned for each item, as was the position of the target verb in the list (left, middle, or right).

In addition to 51 target items, each participant received 5 practice items and 5 literal filler/catch trials that served as attention checks. The practice items consisted of three literal sentences and two involving a highly conventional verb metaphor, while the filler items were all literal sentences. Both practice and filler items were designed so that the correct answer was obvious (e.g., *The dog made a growling noise* was presented with the response options *barked*, *boiled*, and *swam*).

As in Experiment 2, a total of 480 dummy items were included to account for the fact that the paraphrase exclusions in Experiment 1A resulted in an uneven number of paraphrases per item. The goal was to test the maximum number of items from Experiment 1A possible while also providing a uniform experience across participants in terms of the criteria described above. Each of the 897 paraphrases tested received exactly one response.

#### *5.1.2.1 Procedure*

Participants were randomly assigned to one of the three item groupings. They first read instructions informing them that they would be seeing paraphrases of both literal and non-literal sentences that were written by people in another experiment; their job was to determine which verb was in the original sentence from a list of three possible options.

Items were presented one at a time, and participants could not go back to change earlier responses once submitted. On each trial, the paraphrase was shown at the top of the screen and the three verb options (the target verb and two distractors) were listed from left to right directly below. Participants first completed the five practice items with feedback indicating the correct answer. Participants then completed the remaining target and filler items without feedback. Fillers were covertly interspersed throughout the target items.

### 5.2 Results

Minimal subtraction and toss-and-replace predict different patterns of retrace accuracy. Toss and replace predicts that interpreting strained sentences is noun-centric, taking little of the verb's meaning into account. In this view, retrace accuracy should be at chance levels (33%), regardless of the original verb or noun type. Conversely, minimal subtraction predicts that as much of the original verb's meaning is preserved during comprehension as possible. In this view, retrace accuracy should be above chance for all verbs and noun types—however, since the results of the previous three experiments show that the degree of verb meaning change/abstraction increases with the degree of strain, it is also likely that retrace accuracy will decrease as a function a strain.

To test these two accounts, we modeled the probability that the correct original verb was identified for each item as a function of verb class (motion, sound, and body-process) and noun type (human, artifact, and abstract) using mixed effect logistic regression. Retrace success on the original verb (correct vs. incorrect) was the dependent measure, verb class and noun type of the original item and their interaction were entered as fixed effects, and subjects and items (in this case, the original item from Experiment 1) were entered as random effects.

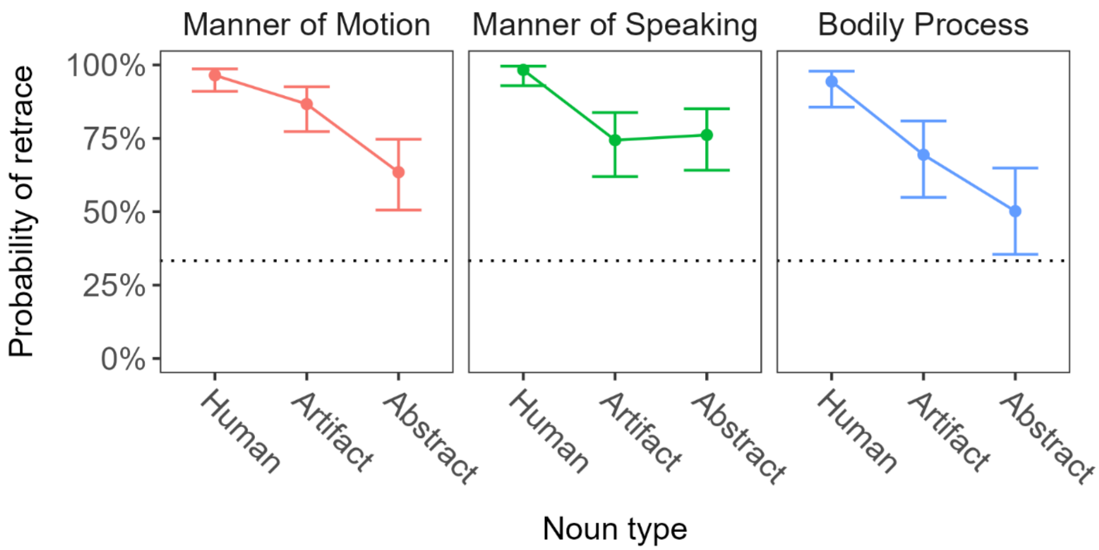



Figure 10 plots the fitted model results by verb class, and Table 17 lists the probability of retrace for each noun type within each verb class. Planned pairwise contrasts by verb class and noun type are shown in Table 18. As predicted, retrace accuracy was significantly above chance for all verb classes and noun types. Also as expected, accuracy varied as a function of strain: accuracy for paraphrases from strained sentences (items with artifact and abstract nouns) was lower than accuracy for paraphrases from unstrained sentences (items with human nouns) for all three verb classes.

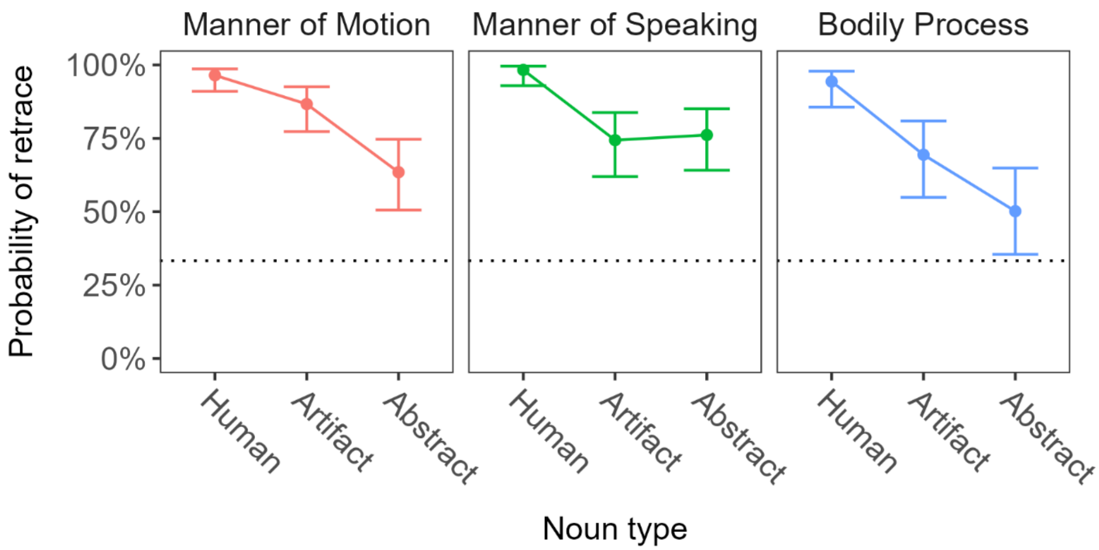



Figure 10. Retrace probabilities by verb class for Experiment 3A. The dotted line represents chance (33%). Error bars represent 95% CIs.

Table 17

*Retrace probabilities from Experiment 3A*

| Verb Class | Noun type | Prob. of guessing correct verb | 95% CI |
|---|---|---|---|
| Manner of Motion | Human | 96% | [91%, 99%] |
| | Artifact | 87% | [77%, 93%] |
| | Abstract | 63% | [51%, 75%] |
| Manner of Speaking | Human | 98% | [93%, 100%] |
| | Artifact | 74% | [62%, 84%] |
| | Abstract | 76% | [64%, 85%] |
| Bodily Process | Human | 94% | [86%, 98%] |
| | Artifact | 69% | [55%, 81%] |
| | Abstract | 50% | [35%, 65%] |

Table 18

*Contrasts by verb class from Experiment 3A*

| Model | Contrast[a] | *b* (logits) | SE | *z* | *p* |
|---|---|---|---|---|---|
| Manner of Motion | *Human – Artifact* | 1.44 | 0.59 | 2.46 | .25 |
| | *Human – Abstract* | 2.76 | 0.56 | 4.92 | < .0001* |
| | *Artifact – Abstract* | 1.32 | 0.41 | 3.25 | .03* |
| Manner of speaking | *Human – Artifact* | 3.00 | 0.80 | 3.76 | < .01* |
| | *Human – Abstract* | 2.90 | 0.80 | 3.65 | < .01* |
| | *Artifact – Abstract* | -0.09 | 0.39 | -0.24 | .99 |
| Bodily Process | *Human – Artifact* | 2.00 | 0.60 | 3.33 | .02* |
| | *Human – Abstract* | 2.81 | 0.60 | 4.68 | < .0001* |
| | *Artifact - Abstract* | 0.81 | 0.43 | 1.90 | .61 |

*Note.* [a] Compares items with that noun subject type. All *p*-values adjusted using Tukey's method.

## 5.3 Discussion

The results of Experiment 3A are consistent with the minimal subtraction hypothesis and inconsistent with the toss-and-replace account. For all three verb classes and all three noun types, participants were significantly above chance in identifying the correct original verb from among the two distractors. In addition, retrace accuracy for strained sentences was lower than accuracy for unstrained sentences for all three verb classes.

That participants were above chance in identifying the original verb for all three verb classes and noun types is strong evidence that at least some components of the verb's meaning are taken into account during comprehension, even for highly strained items. This argues against toss-and-replace in its strongest form. However, it is still possible that a weaker version of toss-and-replace could explain our results. In the weak version, the verb is mostly discarded in favor of an event associated with the noun, but the general domain of the original verb is retained and used to modulate which noun event is selected. For example, pairing the noun *van* with a motion verb like *marched* may cue an associated motion event (e.g., driving), while pairing *van* with a sound verb like *sang* may cue an associated sound event (e.g., honking the horn). In this account, since little beyond the general domain is retained, there should be little *within*-class variation across items.

The below participant paraphrases from Experiment 1A exemplify the types of interpretations that could result from the weak version of toss-and-replace:

| **Original sentence** | **Paraphrase** |
|---|---|
| *The van marched* | *The car drove* |
| *The van shuffled* | *The car moved* |
| *The van sang* | *The car made noise* |
| *The van cried* | *The car was loud* |

Since the two distractors in this experiment were always from different verb classes than the target verb, these results cannot rule out the weak version of toss-and-replace. Thus, in Experiment 3B, we conducted a new retrace task, identical to that of Experiment 3A except that all three verb response options for each trial were always from the same verb class—e.g., if the target verb was a motion verb, then the two distractors were as well.

If participants are above chance in identifying the original verb even when the two distractors are from the same domain, it would support the notion that comprehension involves preserving verb-specific meaning components beyond general domain cues and taking domain- and verb-specific meaning components into account.

# 6 Experiment 3B

The approach of Experiment 3B was identical to Experiment 3A, except that the two distractor verbs were always from the same class as the target verb.

## 6.1 Method

### *6.1.1 Participants*

27 native English-speaking university undergraduates completed the study online via the Qualtrics platform. They received course credit in an introductory psychology class for their participation.

### *6.1.2 Materials*

The materials used were identical to those used in Experiment 3A, with the exception of the distractor verbs used in each trial. For each of the three verb classes (motion, sound, and body-process), the 18 stimulus verbs from that class were divided by the authors into 6 lists of 3 verbs each. Within each list, the verbs were chosen to be semantically distinct. Thus, highly similar verbs (e.g., *trudged* and *plodded*) were placed on different lists. This meant that each verb was distinguishable from the others during the retrace task (e.g., motion: *limped, pranced, sprinted*; sound: *screamed, muttered, sang*; bodily process: *hiccupped, exhaled, drooled*). All 18 lists are shown in Appendix H.

### *6.1.3 Design*

The design was identical to that used in Experiment 3A. For each item, the target verb (i.e., the verb from the original sentence) was presented simultaneously with the other two verbs from that list (e.g., if the target verb was *limped*, the participant saw the choices *pranced, sprinted,* and *limped*). The position of each verb (left, middle, and right) was randomized across trials.

### 6.1.4 *Procedure*

The procedure was identical to that of Experiment 3A.

## 6.2 Results

Using the same approach outlined in Experiment 3A, we modeled the probability of identifying the correct verb for each item as a function of verb class and noun type (human, artifact, and abstract) using a mixed effect logistic regression model. Retrace success on the original verb (correct vs. incorrect) was the dependent measure, verb class and noun type of the original item and their interaction were entered as fixed effects, and subjects and items (in this case, the original item from Experiment 1A) were entered as random effects.

Figure 11 plots the fitted model results by verb class, and Table 19 lists mean accuracy and 95% confidence intervals for each noun type within each verb class. Pairwise contrasts by verb class and noun type are shown in Table 19.

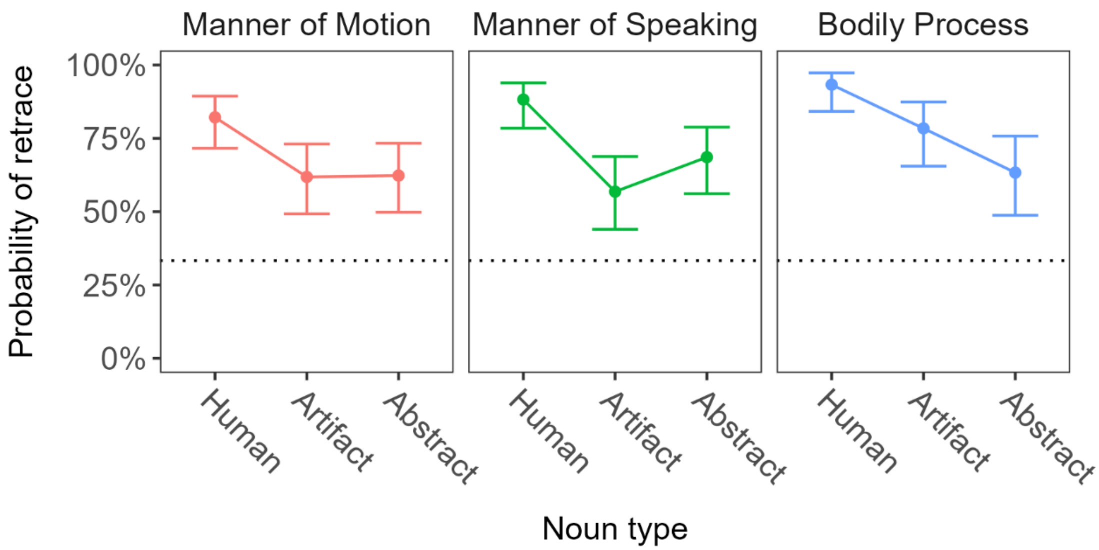



Figure 11. Retrace probabilities by verb class for Experiment 3B. The dotted line represents chance (33%). Error bars represent 95% CIs.

Table 19

*Retrace Probabilities from Experiment 3B*

| Verb Class | Noun type | Prob. of guessing correct verb | 95% CI |
|---|---|---|---|
| Manner of Motion | Human | 0.82 | [.72, .89] |
| | Artifact | 0.62 | [.49, .73] |
| | Abstract | 0.62 | [.50, .73] |
| Manner of Speaking | Human | 0.88 | [.78, .94] |
| | Artifact | 0.57 | [.44, .69] |
| | Abstract | 0.69 | [.56, .79] |
| Bodily Process | Human | 0.93 | [.84, .97] |
| | Artifact | 0.78 | [.65, .87] |
| | Abstract | 0.63 | [.76, .76] |

Table 20

*Model results from Experiment 3B*

| Model | Contrast | *b* (logits) | SE | *z* | *p* |
|---|---|---|---|---|---|
| Manner of Motion | *Human – Artifact* | 1.05 | 0.40 | 2.60 | .18 |
| | *Human – Abstract* | 1.03 | 0.40 | 2.56 | .20 |
| | *Artifact – Abstract* | -0.02 | 0.37 | -0.05 | .99 |
| Manner of speaking | *Human – Artifact* | 1.74 | 0.45 | 3.85 | < .01* |
| | *Human – Abstract* | 1.23 | 0.46 | 2.71 | .15 |
| | *Artifact – Abstract* | -0.51 | 0.38 | -1.34 | .92 |
| Bodily Process | *Human – Artifact* | 1.34 | 0.59 | 2.29 | .35 |
| | *Human – Abstract* | 2.09 | 0.57 | 3.64 | .01* |
| | *Artifact - Abstract* | 0.74 | 0.45 | 1.66 | .77 |

*Note.* [a] Compares items with that noun subject type. All *p*-values adjusted using Tukey's method.

These results replicate Experiment 3A: across all verb classes and noun types, participants were significantly above chance in identifying the original verb. In this case, they were able to do so even when the verbs were from the same semantic class, consistent with our original assumption that people generally include within-domain, verb-specific information when interpreting the semantically-strained sentences.

### 6.3 Discussion

The results of Experiment 3B are consistent with minimal subtraction and inconsistent with the weak version of toss-and-replace. Replicating our findings from Experiment 3A, for all three verb classes and all three noun types, participants were significantly above chance in identifying the original verb from among the two distractors, even though the distractors were all from the same verb class.

These results are consistent with our findings in Chapter 1 that the comprehension of semantically strained sentences involves online adjustment to the verb's representation. In addition, that participants could distinguish between verbs of the same semantic class further demonstrates that this online process can produce remarkably fine-grained meaning adjustments. Even when the original sentence were highly strained (requiring the verb to undergo significant abstraction in order to accommodate the noun) participants were still able to distinguish the original verb from others in the same domain. This argues strongly against a toss-and-replace process in which the verb's meaning is simply replaced by one more compatible with the noun.

Importantly, however, this is not to say that comprehension does not involve accessing salient events associated with the nouns—only that it is not *solely* a matter of doing so. That noun-associated event information is used during comprehension is self-evident in many of the paraphrases—e.g., paraphrases of sentences with *van* as the noun often referenced *driving* events, and paraphrases of sentences with *rumor* tended to describe spreading events—e.g., *The rumor paced* → *The gossip went back and forth*. Thus, it seems clear that noun-related events are critical to the interpretation. But the present results also indicate that verb-specific information surfaces in the paraphrases as well. This invites the important question of how these two information sources are integrated. In the next chapter, we propose a process model with the goal of explaining how this occurs.

# 7 General Discussion

The goal of this chapter was to characterize the phenomena of verb metaphoric extensions. When verbs change meaning under strain, how are they changing? In the introduction, we proposed one possible account, minimal subtraction, for which we laid out three key principles: (1) the degree of verb meaning change increases progressively as strain increases, (2) the verb changes meaning only as far as necessary to accommodate the noun, and (3) domain-specific meaning components of the verb are adjusted before more abstract relational ones.

Our findings from all five experiments support these predictions. In Experiments 1A and 1B, we used word2vec to assess the degree of meaning change that nouns and verbs underwent under paraphrase. We found that verbs—but not nouns—changed meaning in response to strain (replicating the verb mutability effect from Chapter 1). Further, as expected under minimal subtraction, we found that the degree of change increased proportionally with the degree of strain: the greater the strain, the further the verb's meaning changed. Thus, verbs appear to typically change only as far as required by the paired subject noun.

In Experiment 2, we tested the qualitative predictions of minimal subtraction using a coding task that assessed the nature and degree of verb abstraction that occurred under strain. As predicted, the degree of verb abstraction increased as a function of the noun it was paired with. Verbs with a human noun subject retained their typical literal meaning (e.g., *The doctor burped* → *The healthcare professional belched*), verbs with an artifact noun subject were partly abstracted via adjustments to domain-specific, concrete dimensions (e.g., *The boat burped* → *The boat let out a large burst of steam*), and verbs with an abstract noun subject were abstracted further to denote a non-physical, abstract event (e.g., *The wisdom burped* → *A smart idea came out of nowhere*). Regardless of the degree of semantic mismatch, however, higher-order, domain-general relational meaning components were typically preserved. This pattern also converges with the word2vec results in supporting the claim that verbs typically change only as far as demanded by the noun: the degree of verb abstraction represented by the code categories (as shown in Table 14) depended strongly on the paired noun type, with the codes representing greater abstraction becoming more frequent (and those representing less abstraction becoming less frequent) as the degree of semantic mismatch between noun and verb increased.

Finally, in Experiments 3A and 3B, we used a retrace task to show that the minimal subtraction pattern identified in the previous experiments cannot be explained by a *toss-and-replace* process model, wherein the verb is discarded and replaced with an event saliently associated with the noun. At all levels of semantic strain, participants were able to identify the original verb at above-chance levels based on its paraphrase alone, demonstrating that components of the original verb's meaning were preserved regardless of the type of noun subject it received.

Thus, we find that verbs have a quite general pattern whereby they adapt their meanings to their noun arguments in a fine-grained manner, tailoring the degree and kind of abstraction they undergo to match the semantic content of the noun while also preserving the maximal amount of the verb's semantic structure possible.

## 7.1 Implications for verb mutability

In the General Discussion of Chapter 1, we asked *why* verbs change meaning more readily than nouns. We proposed that verb mutability derives from both pragmatic and semantic factors working in tandem. Pragmatically, the verbs' role as predicate in a sentence exerts pressure on the verb to adjust its meaning so that it may meaningfully connect its noun arguments. Semantically, verbs' greater relationality of meaning means that verb meaning components have greater connectivity to external concepts, allowing verbs greater flexibility to adapt to novel contexts.

Our findings here suggest an additional possible factor driving verb mutability: verbs' propensity for minimal subtraction—that is, their ability to abstract in a partial, graduated manner—may widen the range of semantic contexts to which a verb can meaningfully adapt while minimizing the amount of meaning change necessary. Many accounts of verb semantics have characterized verb meanings as composed of interconnected components of meaning (Croft, 2001; Gentner,1981; McCawley, 1972; Jackendoff, 1983; Langacker, 1987; Slobin, 1996; Talmy, 1975). Verbs' ability to undergo minimal subtraction—that is, to partially adapt their meanings—may derive in part from the componential nature of verb representations. In contrast, many noun meanings appear less componential—it is less clear that they can be decomposed into constituent predicates—which may render them less able to partially adapt their meanings to a variety of contexts without changing what they refer to entirely (Gentner, 1981). As such, changing the noun to fit the verb necessarily results in a more radical shift in the meaning of the utterance compared to partially adjusting the verb's meaning to fit the noun. Verb meanings may offer the path of least resistance towards resolving the strain, increasing their mutability.

# 8 Conclusion

We have shown that verb metaphoric extensions follow a minimal subtraction pattern. Under strain, verbs extend metaphorically such that their meanings are adapted to the noun in a fine-grained manner, abstracting only as far as necessary to accommodate the noun, and adjusting or abstracting their meaning components in a domain-specific to domain-general direction as semantic strain increases.

Having now delineated how verbs change their meanings under strain, in Chapter 3 we turn to the underlying process. We will propose a process model that aims to explain the mechanism that ultimately results in the patterns of meaning change reported both here and in Chapter 1.

In our investigation of the possibility of a toss-and-replace of mechanism underlying minimal subtraction, we discussed the apparent salience of noun-related events in the participant paraphrases. That is, participants frequently described prototypical actions associated with the subject noun of the sentence when interpreting the verb metaphors (e.g., *vans* → *drive*, *planes* → *fly*; *rumors* → *spread*, etc.). Experiments 3A and 3B showed, however, that some degree of information from the verb is still retained in the interpretation. How are these two sources of information (the noun-related event and the verb) integrated? The model proposed in the next and final chapter aims to answer this question.

# Chapter 3

## 1 Introduction

In this chapter, we propose and test a novel processing account for verb metaphor. Our proposal is that verb metaphors are comprehended as analogies are—as comparisons processed via structure-mapping. We first lay out the theory and review current supporting evidence. We will show that predictions drawn from the analogical reasoning literature apply well to the phenomena of verb metaphor, as described in Chapters 1 and 2. We then lay out new predictions from the theory and report two additional experiments that provide converging evidence for this account. These latter experiments go beyond the analysis of processing outcomes (i.e., the patterns of the paraphrases and their relation to semantic strain) by testing online processing predictions that follow from the structure-mapping account of verb metaphor comprehension. We conclude with a discussion of the larger implications this account has for theories of metaphor processing and language evolution.

### 1.1 Verb metaphor processing

Our claim is that novel verb metaphors are understood via analogical processing—that is, as comparisons. This claim might at first seem absurd. In analogies such as *Misinformation is like a virus*, it seems clear that a comparison is drawn between *misinformation* (the target of the analogy) and *virus* (the base of the analogy). But in a verb metaphor like *The wagon limped*, what is the verb compared to? Clearly, it would be nonsensical to compare an event like *limped* to an entity like *wagon*. So how can an analogical, comparison-based account apply here?

Our proposal is that the verb is compared not to the entity the noun denotes, but rather to an *event schema* that is activated by the noun. There is substantial evidence, reviewed in the next section, that encountering a noun like *wagon* calls forth not only information about the entity that the noun denotes, but also event schemas that capture our knowledge of what that entity typically *does*—that is, the events it typically participates in. For example, a wagon typically rolls—it moves on wheels across the ground. Thus, our claim is that interpreting a verb metaphor like *The wagon limped* is a matter of comparing two event schemas: an event schema activated by the noun (e.g., *rolling*) and the event schema denoted by the verb (*limping*).[15] The key idea is that, just as with any other analogy, this comparison is a process structural alignment.

### *1.1.1 Structure-mapping in analogy*

As there is substantial evidence in the literature supporting the structure-mapping model of analogy (Gentner, 1983, 1989; Markman & Gentner, 1993; Wolff & Gentner, 2011), we use the structure-mapping framework here. Structure-mapping assumes that analogies involve hierarchical, relationally-structured conceptual representations, which are compared via a two-phase process of structural alignment and inference.[16] Using the example analogy *Misinformation is like a virus,* in the first phase, the base (e.g., *virus*) and target (e.g., *misinformation*) representations are placed into correspondence based on aligning common semantic relations such that structural consistency between the representations is maintained (see

---

[15] In fact, explicit comparisons between nouns and verbs of the form *VERB like a NOUN* are not uncommon in English. We frequently say things like *cry like a baby*, *shake like a leaf*, *bark like a dog*. Each of these can be thought of as eliding a [does] (e.g., *cry like a baby [does]*), emphasizing how the noun stands in for a saliently associated event.

[16] The two phases described here are a simplification for clarity. In the more detailed process model, Phase 1 (structural alignment) is made up of three stages: (i) forming all local matches; (ii) collecting the local matches into structurally consistent partial mappings (kernels), and (iii) combining kernels into a larger mapping (Forbus et al., 2017; Gentner, 2010; Sagi, Gentner & Lovett, 2012).

Gentner & Markman, 1997, Forbus et al., 2017; Sagi, Gentner & Lovett, 2012). Structural alignment renders commonalities between the base and target more salient—especially relational commonalities. Thus, for *Misinformation is like a virus*, the alignment process identifies the shared relational commonalities that both misinformation and viruses spread rapidly among people.

Once the mapping is generated, *inference projection* occurs, wherein any relational structure that is connected to the mapped structure and is present in the base but not in the target is projected from the base to the corresponding location in the target. In this way, spontaneous inferences about the target may be generated based on the identified commonalities with the base. For example, since the mapping process identifies that both *misinformation* and *viruses* spread rapidly among people, if one's representation of *virus* also includes the information that vaccination can reduce the rate of viral spread among people, this may lead to the spontaneous inference that perhaps it is also possible to "vaccinate" against misinformation in order to reduce its spread.

Thus, through structural alignment and inference projection, structure-mapping captures the two hallmarks of analogy and metaphor (1) they highlight abstract commonalities between unlike concepts, and (2) they are informative—they can lead to new knowledge about the target based on projecting inferences from the base.

### *1.1.2 Structure-mapping in verb metaphor*

Our claim is that the same process applies in the case of verb metaphor comprehension. For a verb metaphor like *The wagon limped*, the target of the analogy is an event schema activated by the noun (*rolling*), and the base of the analogy is the event denoted by the verb (*limping*). As in any analogical comparison, processing takes place by first aligning the two relational structures—in this case, the noun and verb event schemas—and identifying their common structure, followed by projecting inferences about the target (the noun event) from the base (the verb event). Thus, just as analogies serve to both highlight commonalities between concepts and generate inferences from the base about the target, so too do verb metaphors.

In the example of *The wagon limped*, the noun event *rolling* and the verb event *limping* are first aligned, and the mapping process identifies the common structure that both involve forms of physical motion across land. Next, the further information that *limping* denotes physical movement that occurs in an awkward or impaired manner is projected to the noun event, resulting in the inference that the wagon is rolling in an impaired manner.[17] Thus, the metaphor comes to be interpreted to mean something like *The cart bumped and rolled awkwardly along the street*.

Thus, our model makes two novel claims with respect to verb metaphor comprehension: (1) During comprehension, nouns activate representations of the events they typically participate in, and (2) The activated noun event is integrated with the verb event via structure-mapping, such that the two events are aligned, and the verb comes to further elaborate the noun event.

Table 21

*Example paraphrases from Chapter 2*

---

[17] Most of the verbs tested in our experiments were manner verbs; therefore, the projected inferences frequently involved manner information. But this is not a requirement of the model: other types of information may be projected depending on the verb and/or the nature of the alignment with the noun event.

| | Original sentence | Paraphrase |
| --- | --- | --- |
| (1) | The wagon limped | *The cart bumped and rolled awkwardly along the street* |
| (2) | The van prowled | *The car drove forward slowly and sneakily* |
| (3) | The plane paced | *The jet flew back and forth in the sky waiting for something* |
| (4) | The boat scampered | *The boat darted across the water* |
| (5) | The rumor paced | *The gossip went back and forth* |
| (6) | The religion swallowed | *The beliefs really brainwashed people and entirely changed them* |
| (7) | The train cackled | *The locomotive made noise when the wheels went over the tracks* |
| (8) | The scooter jogged | *The moped cut in and out of traffic* |

The paraphrases from Chapter 2 shown in Table 21 illustrate these two key claims. First, participants are clearly drawing on their knowledge of what the noun subjects typically *do*—that is, the events that they typically participate in—during comprehension Wagons roll, often down a street, jets fly through the air, trains roll down tracks, and rumors spread among people. While none of these noun-related events are explicitly mentioned in the original sentences, they consistently surface in participants' paraphrases, underscoring that they are an important component of generating a meaningful interpretation of the metaphor.

The paraphrases also demonstrate the role of the verb in  projecting information that serves to further characterize the noun event.  For example, in (1), we infer that the rolling event associated with wagons occurs in an impaired manner, just as *limping* denotes a walking event occurring in an impaired manner. In (2), the driving event associated with vans is specified to occur a slow, sneaky manner, just as *prowling* denotes walking in a slow, sneaky manner. In (3), the flying event associated with planes is specified to be a back-and-forth type of flying, just as *pacing* denotes a back-and-forth walking event.

Thus, the structure-mapping model of verb metaphor provides a process model that can account for the remarkably consistent patterns of interpretation that have emerged throughout the experiments reported thus far in this thesis. This account rests on two key claims: (1) nouns activate event schemas representing the events they typically participate in, and (2) those event schemas are aligned with the verb using structure-mapping, resulting in interpretations that highlight commonalities between the noun event and the verb event and further modify the noun's typical event based on projecting inferences from the verb. In what follows, we first review empirical evidence for claim (1). We then describe new research that tests claim (2).

### 1.2 Noun event schemas in literal sentence processing

There is an extensive literature demonstrating that noun meanings often include information about the events that the noun's referent frequently participates in. This event knowledge appears to be a critical aspect of literal sentence comprehension (for reviews, see Elman, 2011 and Altmann & Mirković, 2009). These noun events are often referred to as *generalized event knowledge* and are assumed to be acquired by abstracting over multiple experiences of the event. Some nouns may have multiple associated event schemas (e.g., a car likely has event schemas for *driving*, *purchasing, being stuck in traffic*, etc.) that may be accessed differentially depending on context. However, often there are one or a few primary event schemas with which the noun is most strongly associated (e.g., for a car, *drive*).

There is evidence that noun event schemas are often activated automatically, even when the nouns are presented in isolation. For example, nouns prime verbs that denote the events that they frequently participate in (Altmann & Kamide, 1999; Ferretti et al., 2001; Hare, Elman, et al., 2009; McRae et al., 2005; McRae & Matsuki, 2009; Moss et al., 1995). There is evidence that these events are activated rapidly upon encountering the noun, on timescales associated with lexical access (~100 ms) (Elman, 2011; Lupyan & Lewis, 2019; McRae & Matsuki, 2009; Zarcone et al., 2014).

Importantly, during sentence processing, the noun's event information is rapidly integrated with the event denoted by the verb such that it immediately constrains expectations about subsequent input (e.g., Altmann, 1999; Altmann & Mirković, 2009; Altmann & Kamide, 1999; Bicknell et al. 2010; Kamide, Altmann, & Haywood, 2003; Matsuki, et al., 2011; McRae, Ferretti, and Amyote, 1997).

For example, Kamide, Altmann, and Haywood (2003) used an eye-tracking paradigm in which participants listened to sentences while looking at a visual array of pictures of various objects—e.g., a man, a girl, a motorcycle, a carousel, a beer, and candy. Kamide et al. found that when participants heard *The man will ride*, they made anticipatory looks to the motorcycle, but when they heard *The girl will ride*, they made anticipatory looks to the carousel. Alternatively, when they heard *The man will taste*, they looked to the image of the beer, but when they heard *The girl will taste*, they looked to the image of the *candy*. Thus, participants rapidly accessed event knowledge associated with the agent nouns (man, girl) and integrated it with the verb such that it immediately constrained their predictions about the likely patient. These predictions depended on the noun and the verb jointly—i.e., people only looked to the motorcycle when the noun was *man* and the verb was *ride*; similarly, they only looked at the beer when the noun was *man* and the verb was *taste*.

Further evidence comes from Zarcone, Pado, and Lenci (2014), who used a speeded rejection paradigm to demonstrate covert activation of noun events in logical metonymies—sentences where an event-expecting verb receives a noun patient instead (e.g., *The writer began the novel*; Nunberg, 1978). Zarcone et al. designed pairs of logical metonymies that were identical except for the agent, and where each agent should call forth a different activity based on its typically-associated actions. For example, the sentence *The baker finished the icing* should lead participants to assume that the baker *spread* the icing, while the sentence *The child finished the icing* should lead them to assume that the child *ate* the icing.

Once participants had finished reading the sentence, the sentence disappeared and one of the two probe words (e.g., SPREAD or ATE) appeared 100 ms later. Thus, each sentence was followed by a probe event that was either typical for the agent (*baker-spread* or *child-ate*) or atypical (*baker-ate*, *child-spread*). Participants were instructed to indicate as quickly as possible whether or not the probe word had appeared in the previous sentence. They found that participants were slower to reject the probe when it matched the agent's typical event than when it didn't match—i.e., participants were slower to reject SPREAD after the sentence *The baker finished the icing* than they were to reject ATE, and vice versa for *The child finished the icing.* Since the verb *finished* was identical in all conditions, this showed that participants activated their knowledge of the typical events associated with the agent and patient nouns during processing. This pattern of rapid activation and integration has been found to occur across a number of different verb role relations, including agent (e.g., Kamide et al., 2003), patient (e.g., Hare et al., 2009), and instrument relations (Matsuki et al., 2011).

In sum, there is ample evidence from research on literal sentence processing that noun representations often include event schemas that are activated upon encountering the noun (even in isolation), and that these are fluently integrated with the verb during comprehension. Our contention is that this integration also occurs when processing verb metaphors.

## 1.3 Predictions of the structure-mapping model of verb metaphor

The proposal is that verb metaphors are processed by structure-mapping between the event denoted by the verb and an event activated by the noun. Viewing verb metaphor as a species of analogy leads to five key empirical predictions. We will discuss each prediction by first introducing the relevant theory from the analogical reasoning literature, and then describing how it applies to the verb metaphor case. We will argue that Predictions 1 through 3 are supported by the results already obtained in Chapters 1 and 2 above. Predictions 4 and 5 will be supported with novel evidence from two new experiments, reported below.

### *1.3.1 Prediction 1: The verb mutability effect*

In an analogy, the target of the analogy is taken as the referent and construed literally (i.e., it retains its full typical meaning), while the base of the analogy applies abstractly—it serves to elaborate the target (Bowdle & Gentner, 2005; see Gleitman et al., 1996 for a similar point concerning similarity statements). As discussed earlier, analogical comparison reveals commonalities (especially relational commonalities) between the two things being compared. This common system may be quite abstract. For example, in the analogy *Misinformation is like a virus,* the common system would be something like "spreads rapidly among people and causes harm." This common system becomes more salient in both concepts relative to concrete, domain-specific features of the two analogs (such as the biological details about viruses) (Christie & Gentner, 2010; Gentner & Namy, 1999).

But the outcome of the comparison differs for base and target. Because the target (misinformation), is construed as the literal referent, it ultimately retains its domain-specific features. The effect of the comparison is to elaborate this literal meaning, often by projecting inferences from the base with the appropriate bindings to the target concept (e.g., that misinformation spreads between people via language, causes cognitive harm, etc.) In contrast, domain-specific features of the base (e.g.,that viruses spread between people via transmission of bodily fluids, that they cause bodily harm, etc.) are discarded in the resulting interpretation. Thus, the comparison results in abstraction of the base while the target remains relatively stable. There is evidence that carrying out such metaphoric comparisons can lead, over time, to this abstraction eventually becoming a new conventional meaning of the base word (Bowdle & Gentner, 2005; Cardillo et al., 2012; Zharikoff & Gentner, 2002).[18]

If verb metaphors are understood in the same way as analogies, then a similar pattern should apply—that is, the target (the noun event) should remain relatively stable as the literal referent, while the base (the verb event) should apply abstractly, by providing information relevant to the target (the noun event). In this account, as with analogies in general, comprehension begins with first aligning the two events. For a metaphor like *The wagon limped*, the mapping process highlights the commonalities that both events involve motion across land, while the nonmatching concrete properties—e.g., that limping usually involves legs are discarded.

Our findings regarding verb mutability in Chapter 1 and 2 support this account (e.g., see Table 21 above). We found a remarkably consistent pattern in which people strongly prefer to interpret

[18] For example, the word *sanctuary* originally referred concretely to a place of worship like a church or temple; it later gained an additional, more abstract sense meaning "a safe place" (Zharikov & Gentner, 2002).

sentences like *The wagon limped* by abstracting the meaning of the verb while preserving the meaning of the noun, as demonstrated by the paraphrase *The wagon limped* → *The cart bumped and rolled awkwardly along the street*.[19] Here, the noun and its event remain stable in the paraphrase (which describes a literal wagon that is literally rolling), while the verb event has been abstracted such that the limping event relates analogically—rather than literally—to the awkward motion of the wagon.

Further evidence for the analogical account comes from the qualitative judgements reported in Chapter 1 (Experiment 3). When given the paraphrases of the verb metaphors, raters judged the events described in the paraphrases to be analogically related to the events denoted by the original stimulus verbs (while the entities in the paraphrases were typically judged to be close synonyms of the original stimulus nouns), suggestive of an underlying analogical mechanism.

### *1.3.2 Prediction 2: Minimal subtraction*

A general assumption of structure-mapping theory is that the mapping process seeks to identify the maximal structurally-consistent mapping possible between two concepts. (Forbus et al.,

[19] While the original verb has become abstracted as a result of comparison with the noun event, the verbs that surface in participants' paraphrases are usually quite concrete/specific (as shown in *The car bumped and rolled awkwardly*). This is expected under our model, which predicts that the event activated by the noun remains stable as the literal referent. It is this event, integrated with the abstracted original verb (which serves to elaborate on the noun event) that surfaces in participants' paraphrases.

2017; Gentner, 2010). This means that the mapping process will retain as much common structure as possible (e.g., Gentner & Kurtz, 2006).[20] For example, compare the two analogies:

*An isthmus is like a bridge*

*An education is like a bridge*

In the former, an isthmus and a bridge share both concrete, domain-specific commonalities (e.g., they are both physical structures that span a physical gap of some sort) and domain-general relational commonalities (they allow objects to cross a physical space that would otherwise be impassable). Thus, the comparison conveys both concrete physical similarities and abstract relational commonalities. There is still a modest degree of abstraction, however, in that the common structure (allowing a crossing) becomes more salient in one's view of the two analogs, eclipsing concrete properties like "manufactured" versus "part of earth's surface."

In *An education is like a bridge*, however, the degree of abstraction is far greater. An education and a bridge share only highly abstract, domain-general relational commonalities (e.g., they provide a means to attain a goal that might otherwise be out of reach). Thus, the common structure is entirely at the abstract relational level. This results in emphasizing a particular abstract aspect of the target, *education,* and strengthening this abstraction in the base term, *bridge*. The construal of *bridge* that results from this analogy is therefore likely to be more

[20] This is certainly not to say that *any* two representations will result in a meaningful alignment. There must be some useful degree of commonality between concepts for an alignment to be successful. For example, metaphors like *water is like a shoe*, or *an education is like a banana* are not likely to be amenable to meaningful alignment and interpretation. We expect the same situation extends to the verb metaphor case. In verb metaphors like, say, *The van raked* or *The rock pondered*, the event schemas may be too dissimilar to align. In such cases, two outcomes are possible: (1) the interpretation attempt is abandoned, either by refusing to comply or by simply providing a synonym for each content word; or (2) the noun is changed to supply a more compatible event schema (e.g., interpreting *rock* as *muscular man* in the second example). We observed both outcomes occasionally in the paraphrases in Chapters 1 and 2.

abstract than that for *isthmus*. Importantly, however, as discussed in Prediction 1, the targets in both comparisons (isthmus and bridge) remain stable and literally construed.

Applying this framework to the verb metaphor case predicts that (1) the degree of verb abstraction that occurs during comprehension depends on the degree to which the noun event and verb event are similar; and (2) this abstraction proceeds in a domain-specific to domain-general direction as a function of the degree to which the two events share domain-specific and domain-general structure. To the extent that the noun and verb events share domain-specific structure in the mapping, it will be retained, but those meaning components will be abstracted away as similarity between events decreases.

This prediction is consistent with the principle of minimal subtraction investigated in Chapter 2. In Experiment 1, we found that (1) as semantic strain increased, the degree of verb meaning change increased, and (2) the abstraction proceeded in a domain-specific to domain general direction, with the verb generally retaining as much of its original meaning (as much domain-specific structure) as allowed by the noun. When *limped* is paired with *wagon*, it retains some domain-specific features related to physical motion over land, but when paired with *rumor*, only domain-general structure is retained (e.g., see Table 22 below).

Table 22

*Example paraphrases illustrating the minimal subtraction pattern*

| | Original sentence | Paraphrase |
|---|---|---|
| (1) | The woman limped | *The female walked unevenly* |
| (2) | The wagon limped | *The cart bumped and rolled awkwardly along the street* |

| | | |
|---|---|---|
| (3) | The rumor limped | *The gossip did not spread easily* |

*1.3.2.1 A new perspective on semantic strain*

This prediction also invites a more process-oriented account of *semantic strain*. In Chapters 1 and 2, we used the traditional account, defining semantic strain as the extent to which the verb's selectional preferences were violated—i.e., the extent to which the features expected by a particular verb's argument (e.g., *human, artifact*, etc.) were satisfied by the paired noun. In Chapter 2, we manipulated strain using a relatively coarse-grained method that involved pairing each verb with three, different noun types (*humans, vehicles,* and *abstracts*) intended to increasingly violate the verb's selectional preferences.

Our current model suggests a more precise definition of strain: it is the similarity between the verb event and the event activated by the paired noun. For example, consider the below two metaphoric extensions of the verb *stammered*:

| | Original sentence | Participant paraphrase |
|---|---|---|
| (1) | The violin stammered | *The violin played intermittently* |
| (2) | The boat stammered | *The boat moved in small, short intervals* |

In both sentences, the human-expecting verb *stammered* receives an artifact noun subject and becomes strained, but the verb has clearly changed its meaning more in (2) than in (1): while it still denotes a sound event when paired with *violin*, it no longer does when paired with *boat* (i.e., it retains more domain-specific components in the former case than in the latter). It seems unlikely that this is due to featural differences in the subject noun entities *violin* and *boat* themselves (e.g., differences in size, shape, or materials) or to the fact that they are different types of artifacts *per se*. Rather, what seems more important is that the sentences differ in terms of how similar the events the nouns activate are to *stammered*.[21] In Experiment 1 below, we formally test the relationship between noun and verb event similarity on perceptions of semantic strain and its effect on metaphor comprehension.

In sum, just as the similarity between concepts predicts the degree of base (but not target) abstraction in analogies in general, so does event similarity (semantic strain) in the verb metaphor case. In both cases, abstraction is seen proceeding in a domain-specific to domain-general direction.

### *1.3.3 Prediction 3: The verb event generates inferences about the nature of the noun event*

As discussed above, one of the hallmarks of analogies is that they can lead to the spontaneous generation of inferences about the analogical target as a result of its mapping with the base. The analogy *Misinformation is like a virus* highlights that both misinformation and viruses spread rapidly among people, inviting the inference that, as is the case with viruses, it may be possible to inoculate against misinformation to reduce that spread.

In the verb metaphor case, this predicts that comprehension should result in inferences about the nature of the noun event (the target) based on the structure of the verb event (the base), as it pertains to the mapping between the two. Since the target of the analogy is an event, the result is

---

[21] One can imagine scenarios where two highly-similar artifacts of the same category generate different degrees of strain when paired with the same verb. For example, the verb *cooed* intuitively seems more strained and difficult to interpret when paired with the musical instrument noun *drums* (e.g., *The drums cooed*) than when paired with the instrument *flute* (e.g., *The flute cooed*). While both nouns are musical instruments that activate sound events, the nature of the sound event activated by *drums* (loud, staccato) is less similar to the event denoted by *cooed* than is the sound event activated by *flute* (quiet, extended in duration, legato). We can see this effect for nouns that are even more similar—e.g., *The flute bellowed* vs. *The oboe bellowed*. Thus, viewing semantic strain in terms of event similarity may capture fine-grained variations that are difficult to capture with featural-based representations of selectional preferences (e.g., Wilks, 1975), but that are likely to be reflected in processing.

that the projected information from the verb serves to further specify or elaborate on that event, characterizing how it unfolds. In addition to the verb mutability effect, this is a remarkably consistent pattern evidenced by participant paraphrases of verb metaphors shown throughout this dissertation. As Table 21 and Table 22 above illustrate, interpretation does not stop at the identification of commonalities between the noun event and verb event—that is, people typically do not paraphrase a metaphor like *The wagon limped* as simply *The wagon moved* or *The wagon rolled* (i.e., relying only on the identified commonality of physical movement). Rather, people typically go further by integrating additional information from the verb (e.g., that limping is physical movement occurring *in an impaired manner*) in order to elaborate more specifically on how the noun event unfolds.

This prediction is further supported by the results of the retrace tasks in Chapter 2, Experiments 3A and 3B. When given paraphrases of verb metaphors and asked to guess what the original verb was, participants were able to correctly identify the original verb from among a list of three choices at all levels of strain, showing that information from the verb was indeed incorporated during comprehension. Since our findings regarding the verb mutability effect demonstrate that the noun event is typically preserved as the referent of the metaphor, this suggests that the information used to identify the verb was projected information, rather than the entire verb event itself (for many of our verbs, this meant manner information).

### *1.3.4 Prediction 4: Event similarity predicts processing time*

Structure-mapping theory predicts that the more similar two conceptual structures are, the easier—and therefore faster—they are to align (Gentner & Kurtz, 2006; Gentner & Wolff, 1997;

Sagi et al., 2012; Wolff & Gentner, 2000).[22] In fact, this effect has been found in the case of verb events specifically: Gentner and Kurtz (2006) showed that people are faster to judge the relatedness of two similarly-structured verbs (e.g., *observe* and *hear*) than two dissimilar verbs (e.g., *observe* and *plan*).

This prediction extends to verb metaphor in a straightforward way: similarity between the noun event and the verb event should predict ease of alignment and thus comprehension times. Put another way, the more semantically-strained the verb metaphor is (see Prediction 2), the longer it should take to interpret. We test this prediction in Experiment 1 below.

### *1.3.5 Prediction 5: Prior activation of relevant event structure facilitates comprehension*

During analogical comparison, structure-mapping highlights relational similarities between concepts, resulting in greater activation of common conceptual structure. In the case of verb metaphor, that shared structure is between the noun event and verb event. More specifically, it is the common structure identified by the mapping process that constitutes basis for the ultimate interpretation or meaning of the metaphor. For example, for a metaphor like *The boat strutted*, we expect that meaning components related to *physical motion* in both the noun event and verb event will be mapped, and therefore will receive greater activation in both representations than, say, meaning components related to the fact that strutting involves legs, or that boats often have cabins for people to sit in.

It follows that prior activation or priming of event structure of that is relevant to the ultimate mapping/interpretation of the metaphor should speed comprehension more than activating

[22] This has also been found for the alignment of perceptual structures (Matlen et al., 2020; Zheng et al., 2022)

irrelevant conceptual structure will. In addition, if noun events are indeed accessed and compared during verb metaphor processing, we expect that priming relevant *noun* event structure specifically should speed interpretation times, just as we would expect priming relevant verb event structure would do. For example, for a metaphor like *The firetruck yelled*, we might expect that priming event structure activated by the word *siren* will speed processing more than priming event structure activated by the word *hose*, since sirens are saliently associated with noise-making events, while hoses are not. We test these predictions in Experiment 2 below.

### 1.4 Summary

To summarize: our proposal is that verb metaphors are processed as analogical comparisons between the event denoted by the verb and an event activated by the noun. As with analogies in general, comprehension involves a two-phase structure-mapping process of alignment followed by inference projection. Alignment serves to identify shared structure between noun event and verb event, and based on that mapping, the verb next projects relevant inferences that further characterize the nature of the referent noun event.

We laid out five key empirical predictions of this account: (1) the verb mutability effect, (2) minimal subtraction, (3) inference projection from verb to noun, (4) event similarity predicts comprehension speed, and (5) priming relevant event structure speeds comprehension. Predictions 1 to 3 are well-supported by the data already presented in Chapters 1 and 2. In the remainder of this paper, we turn to Predictions 4 and 5. Unlike the first three predictions, Predictions 4 and 5 test the role of the comprehension process specifically as it unfolds in real time. We aim to provide converging evidence, using online methodologies, for the role of structure-mapping in verb metaphor processing.

## 2 Experiment 1

As described by Prediction 4 in the introduction, under structure-mapping, similar concepts are faster to align than dissimilar concepts (Gentner & Kurtz, 2006; Gentner & Wolff, 1997; Wolff & Gentner, 2000). The objective of Experiment 1 was to test this prediction in the context of verb metaphor: if verb metaphors are processed as analogical comparisons between an event activated by the noun and the event denoted by the verb, then metaphors comprising noun and verb events that are similar to one another should easier to align—and therefore faster to comprehend—than those comprising dissimilar noun and verb events.

In this experiment, we compared how long it took participants to interpret verb metaphors where the noun event and verb event were similar compared to metaphors where they were dissimilar. Event similarity was operationalized and assessed using the following method. First, we selected two general domains—events primarily about motion and events primarily about sound—under the assumption that events from the same domain (e.g., two motion events) would typically be represented more similarly to one another than events from different domains (e.g., a motion and sound event). Next, we selected nouns and verbs that should activate events in one of the two domains (vehicle nouns and manner of motion verbs for the motion domain, and musical instrument nouns and manner of speaking verbs for the sound domain).[23] For example, a noun like *sailboat* should activate an event schema related to motion over water and is thus categorized as a motion noun, while a musical instrument noun like *violin* should activate an

[23] Here we use the term "vehicle" as a shorthand for artifact nouns whose prototypical event schema primarily involves motion. While some of the nouns were true vehicles in the sense that they are used to transport people (e.g., *raft, sailboat*), others were not (e.g., *kite*). The purpose of the manipulation, however, was to select nouns that saliently engage in motion events; whether or not those nouns typically transport people was not relevant in this experiment.

event schema related to playing music and is categorized as a sound noun (the method for determining the prototypical event activated by the noun is described in the next section). Of course, manner of motion verbs like *strut* inherently denote motion events, and manner of speaking verbs like *whimper* denote sound events.

Next, the noun domains (motion vs. sound) and verb domains (motion vs. sound) were crossed in a 2 X 2 method to create *within-domain* and *cross-domain* verb metaphors, respectively (see Figure 12). Within-domain items were those in which the noun event and verb event were from the same domain (motion-motion or sound-sound; e.g, *The raft strutted* or *The violin whimpered*) and were thus assumed to be represented relatively more similarly to each other. Cross-domain metaphors were those in which the noun event and verb event were from different domains (motion-sound or sound-motion; e.g., *The raft whimpered*, *The violin strutted*) and were assumed to be represented relatively more differently from each other. The idea was that the event schemas constituting within-domain metaphors (those comprising similar events) should be easier and faster to align than those in cross-domain (those comprising dissimilar events) verb metaphors.

| | | Verb Event Domain: Motion | Verb Event Domain: Sound |
|---|---|---|---|
| Noun Event Domain | Motion | *The sailboat strutted* | *The sailboat stammered* |
| | Sound | *The violin strutted* | *The violin stammered* |

Figure 12. Crossing noun event and verb event domains to create within- and cross-domain verb metaphors for Experiment 1. Shaded cells indicate cross-domain items (predicted to be slow to comprehend); unshaded cells indicate within-domain items (predicted to be fast to comprehend).

We followed the same procedure used in Chapters 1 and 2 to collect semantic strain ratings for each stimulus item generated. This was done to account for the possibility that the similarity between noun and verb events may not be equivalent in the cross-domain condition for all types of metaphors. Specifically, we expected that the noun and verb events in cross-domain items with musical instrument noun subjects (e.g., *The violin strutted*) would be less similar to one another than those in cross-domain metaphors with vehicle noun subjects (e.g., *The sailboat stammered*). This would predict an interaction between noun type and condition (within- vs. cross) such that effect of domain similarity would be larger for musical instrument items than vehicle items.

This possibility was raised by the results of a previously-run pilot study (see Appendix I for details and results). The pilot used the same paradigm just described, crossing motion and sound nouns (a different set of vehicles and partially-different set of musical instruments) with motion and sound verbs (manner of motion and manner of speaking) to create within- and cross-domain verb metaphors, and participants' comprehension times for each type of metaphor were recorded. The results indicated that cross-domain metaphors were significantly slower to interpret than within-domain metaphors for musical instrument items, but not for vehicle noun items (see Figure 13) Thus, while comprehension times were slower for items like *The violin sprinted* (cross-domain) than for *The violin stammered* (within-domain), there was no significant difference in RTs for items like *The boat stammered* (cross-domain) vs. *The boat sprinted* (within-domain).

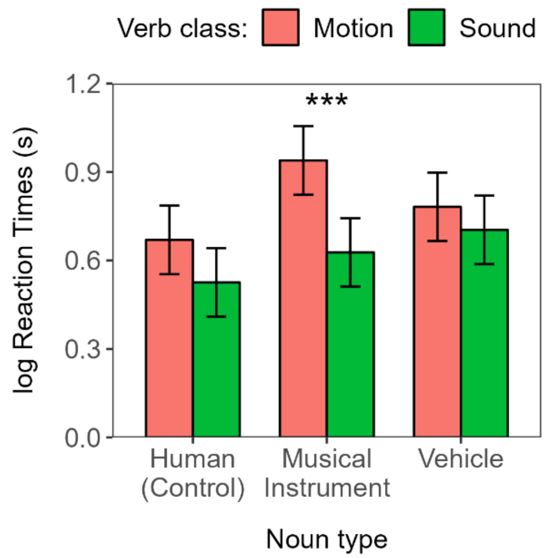



Figure 13. Results from the pilot study. RTs are log-transformed and represent comprehension times as measured from stimulus onset to spacebar press. Error bars represent 95% CIs. *** $p < .001$.

We hypothesized that this asymmetry might be due to the fact that the vehicles tested in the pilot (*plane, boat, jet, van, train, scooter*) saliently make a lot of noise as part of their prototypical motion events; thus, sound may be an important component of their prototypical motion events. In contrast, for the  musical instruments tested (*bell, drums, guitar, horn, piano, violin*), the prototypical events (generating sound) do not involve self-directed motion.[24] Thus, domains of sound and motion could have overlapped more for items with vehicle noun subjects than for those with musical instruments subjects. In the present experiment, we selected vehicles that are not as saliently noisy (e.g., *blimp, raft, sailboat*, etc.), in an attempt to decrease this overlap between motion and sound domains.

To address this issue, we obtained strain ratings of all items in the present experiment. Under the assumption that semantic strain indexes the degree of similarity between noun event and verb event (see discussion in Prediction 2 of our model), obtaining strain ratings of each item on a continuous scale should provide a finer-grained assessment of event similarity than is allowed by a categorical within-domain vs. cross-domain factorial analysis.

Participants were shown a mix of high-similarity and low-similarity verb metaphors one at a time and pressed the space bar as soon as they had thought of a meaningful interpretation. Under the structure-mapping model of verb metaphor, there are three main predictions. First, we predict a verb mutability effect such that verbs should change meaning overall more than nouns (meaning that the noun event should remain stable as the typical referent of the metaphor). Second the degree of verb change should increase with the degree of strain, as the verb must abstract its meaning further to accommodate the increasingly dissimilar noun event. Third, the similarity between the noun event and verb event (as indexed by the degree of semantic strain) should

---

[24] Some of the musical instrument nouns tested (like *drums* and *violin*) involve motion from the musician in order to play them (hitting the drums, bowing the strings). Interestingly participants occasionally adapted motion verbs to describe this motion (e.g., *The violin scampered* → *The player of the violin moved the bow quickly back and forth*. However, it was much more common for motion verbs to be abstracted to apply to the sound produced by the instrument (e.g., *The violin scampered* → *The violin was played such that the notes were fast and frantic*).

predict comprehension times such that relatively lower-strain items (e.g., *The violin stammered*) are faster to interpret than higher-strain items (e.g., *The violin sprinted*).

## 2.1 Method

### *2.1.1 Participants*

A total of 137 native English-speaking Northwestern University undergraduate students from the psychology participant pool completed the study in person in the lab.

While the design and planned analyses for this experiment differ from the pilot study (here we did not include items with human subject nouns, and modeled RTs as a function of our continuous measure of strain, rather than a factorial analysis that crossed noun and verb domains), we obtained a rough estimate of power by conducting an observed power analysis based on the results of the pilot study (N = 139, $\alpha$ = .05). To better approximate the current design, we refit the model from the pilot, excluding items with human subject nouns.[25] Power simulations were conducted using the *simr* package in R (Version 1.0.6; Green & MacLeod, 2016), which indicated an observed power in the pilot of 84.80% (95% CI: [81.35, 87.83%]) for the interaction term of noun type X verb class, $\beta$ = -.08, SE = .03, *p* = .004, and suggesting a power of 84.00% (95% CI: [80.49, 87.10%]) for the present sample size of N = 137 and $\alpha = .05$.

### *2.1.2 Materials*

8 vehicle nouns and 8 musical instrument nouns were selected to activate events in the motion and sound domains, respectively, as were 16 manner of motion and 16 manner of speaking verbs

[25] The pattern of results did not change for the remaining items: those with musical instrument nouns and those with vehicle nouns. The effect of verb class remained significant for the former (*p* < .0001) and nonsignificant for the former (*p* = .49)—see Appendix I.

(see Table 23). In order to reduce the overlap between the motion and sound domains for the vehicle items, we selected vehicles that are not saliently noisy (e.g., *raft, sailboat, glider*, etc.) Vehicle nouns and musical instrument nouns were matched in terms of frequency (determined using the Corpus of Contemporary American English (COCA; Davies, 2008)) and word length ($p$s > .44); as were manner of motion and manner of speaking verbs ($p$s > .23).

Noun domain (motion vs. sound) was crossed with verb domain (motion vs. sound) to create a total of 128 verb metaphors, comprising 64 high-similarity (32 motion-motion and 32 motion-sound) and 64 low-similarity (32 motion-sound and 32 sound-motion) metaphors. A sample of items is shown in Table 24; see Appendix J for the full list.

28 literal filler items and 4 practice items were also included in the study. The literal fillers ensured that each participant saw more normal literal sentences than metaphoric sentences, in order to avoid inducing atypical patterns of thinking. The fillers were disguised by matching the pattern of the target items such that half the fillers had vehicle noun subjects and half had instrument noun subjects. Of the 14 fillers with vehicle nouns, 10 were paired with motion verbs, 2 with sound verbs, and 2 with other verbs (*exploded, smelled*). Conversely, of the 14 fillers with instrument nouns, 10 were paired with sound verbs, 2 with motion verbs, and 2 with other verbs (*rotted, sparkled*). Across all items (literal, filler, and practice), no noun or verb was repeated. Filler and practice items are listed in Appendix J.

Table 23

*Noun and verb stimuli for Experiment 1*

| Target Nouns | |
|---|---|
| Quiet Vehicles | Musical Instruments |
| blimp | banjo |
| glider | violin |
| raft | harp |
| sailboat | flute |
| kayak | trumpet |
| kite | gong |
| rowboat | cello |
| canoe | accordion |

| Target Verbs | |
|---|---|
| Manner of Motion | Manner of Speaking |
| trudged | yelled |
| waddled | muttered |
| staggered | barked |
| sprinted | cackled |
| pranced | grunted |
| strutted | stammered |
| jogged | whimpered |
| prowled | chanted |
| paced | babbled |
| marched | bellowed |
| waltzed | wailed |
| scurried | hollered |
| limped | grumbled |
| strolled | chattered |
| hobbled | screamed |
| danced | chirped |

Table 24

*Sample of metaphoric items from Experiment 1.*

| Noun Type | Verb Type | Item |
|---|---|---|
| Motion | Motion | *The blimp trudged* |
| | | *The glider waddled* |
| | | *The kayak jogged* |
| | | *The raft limped* |
| | Sound | *The blimp babbled* |
| | | *The glider stammered* |
| | | *The kayak barked* |
| | | *The raft grumbled* |
| Sound | Motion | *The banjo trudged* |
| | | *The violin waddled* |
| | | *The trumpet jogged* |
| | | *The harp limped* |
| | Sound | *The banjo babbled* |
| | | *The violin stammered* |
| | | *The trumpet barked* |
| | | *The harp grumbled* |

*Note*. See Appendix J for all 128 metaphoric items.

### *2.1.2.1 Confirming the domains of noun event schemas*

While verb domains are self-evident (i.e., motion verbs like *limped* inherently denote motion events), the domain of the prototypical events activated by entity nouns is less obvious for two main reasons. First, the noun itself refers explicitly only to the entity (e.g., *van*) and not the event. Second, as mentioned earlier, many nouns may have multiple associated event schemas. We assume that in many or most cases, a noun will have a primary (characteristic) event schema that will be called upon by default—that is, in the absence of further or conflicting context or a verb that invites an alternative, less-common schema (e.g., *The van drank deeply* likely activates an event schema relating to filling up the van's gas tank).

As an objective method for identifying a noun's characteristic schema, we used collocation data to select the eight motion and eight sound nouns used in the study. Under the assumption that people's experiences of events in the world (and thus their event representations) are reflected to in patterns of occurrence in the language (Altmann & Mirković, 2009), we used COCA's iWeb Corpus—a 14-billion word corpus drawn from 22 million web pages (Davies, 2018)—to identify the stimulus nouns' most frequent event schemas. Specifically, we used the top verb collocate for a given noun to approximate the characteristic schema for that noun. For example, for the noun *raft*, the top five most frequent verb collocates in the corpus are *float, build, swim, ride,* and *pull*. Since the top verb collocate is *float*, we judge *raft's* characteristic schema to involve motion and therefore consider it appropriate as a noun in our motion domain condition. As another example, the musical instrument *violin*'s top five most frequent verb collocates are *play*, *sound*, *study*, *teach*, and *perform*. Since the top verb collocate for violin is *play*, its characteristic schema can be considered to involve the production of sound, so it is therefore appropriate as a noun in our sound domain condition. Following this method, we confirmed the proper domain for each stimulus subject noun used in the metaphoric items.

*2.1.2.2 Semantic strain ratings*

As discussed above, semantic strain ratings were collected for each item in order to more precisely quantify the degree of similarity between noun event and verb event for that item. Following the same procedure used in Chapters 1 and 2, 81 participants were recruited via Amazon's Mechanical Turk platform to provide ratings and were paid at a rate equivalent to Illinois' minimum wage at the time of the study. Only participants that responded "Yes" to a question asking if they were native English speakers completed the experiment.

To reduce load on each participant, the 128 verb metaphors were divided into 8 item groupings of 16 metaphors each. Each item grouping also included the same 28 literal filler sentences described and 4 nonsensical items (*In sprang then, The of speak, Quickly above did*, and *The catch smirking*) that served as catch trials. Thus, each participant rated a total of 48 items.

Each participant was randomly assigned to one of the eight item groupings and completed the experiment online. Participants rated each sentence for comprehensibility on a scale of 1 to 10 based on how hard the sentence would be for a "typical person" to understand. A rating of 1 corresponded to *very hard for most to understand*, and a rating of 10 corresponded to *very easy for most to understand*.

The first and second trials were always a literal filler sentence and nonsense catch trial, respectively, as were the second-to-last and last trials. The remaining 44 trials (comprising 16 target trials, 26 literal filler trials, and 2 nonsense catch trials) were presented in randomized order across participants. Sentences were presented one at time, and participants could not go back to revise previous responses once they submitted a rating. The criterion for failing a catch trial (a nonsense sentence) was a rating of greater than 2 out of 10, and any participant that failed more than 1 catch trial was excluded from the analysis. 14 participants were excluded based on this criterion, for a net of 67 participants included in the analysis. After these exclusions, there was an average of 8.34 ratings per item (SD = 1.22).

Finally, as was done in Chapters 1 and 2, to convert the comprehensibility scale into a semantic strain scale, the comprehensibility ratings were inverted by subtracting 10 from each score such that a low score now indicated a low degree of strain (i.e., high similarity between noun and verb event) and a high score indicated a high degree of strain (i.e., low similarity between noun and verb event). The strain ratings for all 128 metaphors are shown in Appendix K.

### *2.1.3 Design*

Returning to the main experiment, the 128 target metaphors were divided into 8 between-subject item groupings of 16 metaphors each. Within each item grouping, each participant saw metaphors comprising all 16 stimulus nouns (8 vehicles and 8 musical instruments), 8 unique motion verbs and 8 unique sound verbs. Each participant saw each noun and verb exactly once. All participants received the same 28 filler items and 4 practice items.

### *2.1.4 Procedure*

Participants completed the experiment in the lab, on a computer. Each participant was randomly assigned to an item grouping and began by reading instructions telling them that they would be viewing a mix of normal and "unusual" sentences, and that their job was to interpret those sentences meaningfully as quickly as possible.

Items were presented one at a time, and participants pressed the space bar as soon as they had thought of an interpretation. After pressing the space bar, the metaphor disappeared and they typed their interpretation of the sentence. As in all previous experiments described thus far, they were explicitly instructed to provide a *meaningful* interpretation (e.g., *The anger melted* → *The rage slowly went away*), rather than a simple word-by-word mechanical translation (e.g., *The anger melted* → *The rage unfroze*). Unlike previous experiments, participants were not told to avoid repeating the original noun or verb in their paraphrase, but only to provide a meaningful interpretation. This was done to obtain as pure a measure of natural comprehension times as possible; we were concerned that the need to search for synonyms might artificially inflate reaction times beyond the amount of time it took to understand the sentence.

After reading instructions, participants completed the four practice trials before completing the main experiment. Each participant began the main experiment with the same two filler trials. The remaining target and filler items were presented in randomized order across each participant.

Each trial proceeded as follows (see Figure 14). The words *Get Ready!* appeared on a blank screen for 1000 ms, followed by a blank screen for 1000 ms, a fixation cross for 1250 ms, a blank screen for 100 ms, and followed finally by the target sentence. The target sentence was displayed until the participant pressed the space bar, indicating that they had thought of a meaningful interpretation. The participant then had unlimited time to type their interpretation. Participants then pressed the enter key to proceed to the next trial.

| **Description** | Trial Start | Blank | Fixation cross | Blank | Metaphor | Paraphrase |
|---|---|---|---|---|---|---|
| **Example Text** | *Get Ready!* | | + | | *The raft limped* | *The raft struggled to float* |
| **Duration (ms)** | 1000 | 1000 | 1250 | 100 | Displayed until [SPACE] pressed | Displayed until [ENTER] pressed |

Figure 14. Trial structure for Experiment 1.

On each trial, three key measures were collected: (1) the amount of time from metaphor presentation to space-bar press (indexing comprehension time), (2) the amount of time from space-bar press until the participant began typing their response, and (3) the paraphrase / interpretation of the item.

## 2.2 Results

There are three main predictions. First, we predict a verb mutability effect such that verbs should change meaning overall more than nouns. Second, the degree of verb (but not noun) meaning change should increase with the degree of strain. Third, we predict that the similarity between noun event and verb event (as indexed by semantic strain) should predict comprehension times, such that participant RTs will increase with strain.

The 137 participants generated a total of 2192 responses to the 128 metaphoric items (an average of 17.13 responses per item). To test the first prediction (the verb mutability effect), the same procedure described in the previous Chapters for assessing noun and verb change using word2vec was used to obtain a noun and verb cosine similarity score for each paraphrase. Three paraphrases were excluded for generating null vectors (the paraphrase contained no words that were present in the word2vec model's vocabulary), resulting in a net of 2189 paraphrases included in the similarity analysis. Example paraphrases are shown in Table 25 below.

Next, for each paraphrase, a difference score was calculated by subtracting the verb score from the noun score for that paraphrase. Thus, a positive difference score indicates that the verb changed more than the noun, and vice versa for a negative difference score. The difference scores were fit to a linear mixed effect model, with difference score as the dependent measure, the intercept as the only fixed effect, and subjects and items as random effects. The intercept was significantly greater than 0, $b = .27$, SE = .01, $t = 30.98$, $p < .0001$, indicating that, as expected, verbs changed significantly more under paraphrase (M = 0.24, SE = .02) than did nouns (M = .52, SE = .02).

To test the second prediction—that semantic strain predicts the degree of verb change, but not noun change—two additional linear mixed effect models were fit, one for verbs and one for nouns. In both models word2vec score (noun or verb) was the dependent measure, semantic strain was the fixed effect, and subjects and items were entered as random effects. There was a marginal effect of semantic strain on the degree of verb change, $\beta = -.09$, SE = .04, $t = -1.95$, $p = .053$, indicating that as strain increased, so did the degree of verb change.[26] Also as expected, there was no effect of semantic strain on the degree of noun change, $\beta = -.03$, SE = .03, $t = -0.97$, $p = .33$.

[26] One possible explanation for why the effect of strain on the degree of verb change was smaller here than in previous experiments is because the range of strain across items was relatively restricted here compared to previous experiments, since no literal (low strain items) were included in the present experiment—all items were metaphoric.

To test the third prediction—that event similarity (i.e., semantic strain) predicts comprehension times, RTs were modeled as a function of semantic strain using the following procedure. First, examination of the data revealed a heavy rightward skew in the distribution of RTs; fitting an initial linear model indicated severe departures from the assumptions of residual normality and homoskedasticity. Thus, a log transformation of RTs was used, resulting in an approximately normal distribution of transformed RTs. Next, outlier logRTs were trimmed using the model criticism approach described by Baayen & Milin (2010). First, we fit an initial linear mixed effect model to the full dataset, with RT (comprehension time) as the dependent measure, semantic strain as the fixed effect, and subjects and items as random effects. Next, any observations in the dataset corresponding to residuals that were more greater than +/- 2.5 SDs away from zero were excluded; this led to the omission of 2.19% of the data. The final model (with the same parameters as the initial model) was then fit to this trimmed dataset. Diagnostic plots confirmed a satisfactory linear fit for the final model without significant departures from the assumptions of residual normality and homoskedasticity (see Appendix L).

As predicted, semantic strain significantly predicted comprehension times such that as strain increased, so did the time it took participants to interpret the metaphors, $\beta = 0.14$, SE = 0.02, $t = 6.91$, $p < .0001$ (see Figure 15). Thus, as the similarity between the noun event and verb event decreased, comprehension time increased.[27]

[27] Participants were instructed to press the spacebar as soon as they had thought of a meaningful interpretation. Thus, our measure of comprehension time was RT from stimulus onset to spacebar press. However, some participants did not begin typing immediately after pressing the spacebar, raising the possibility that they had not finished processing the metaphor. To see whether this affected the results, we fit three additional models, excluding trials where participants took longer than 2 seconds, 1.5 seconds, or 1 second (5.47%, 9.22% and 19.62% of the data) to begin typing. In each case, excluding these participants did not change the results (parameter estimates were nearly identical to the full model, all *p*s < .0001).

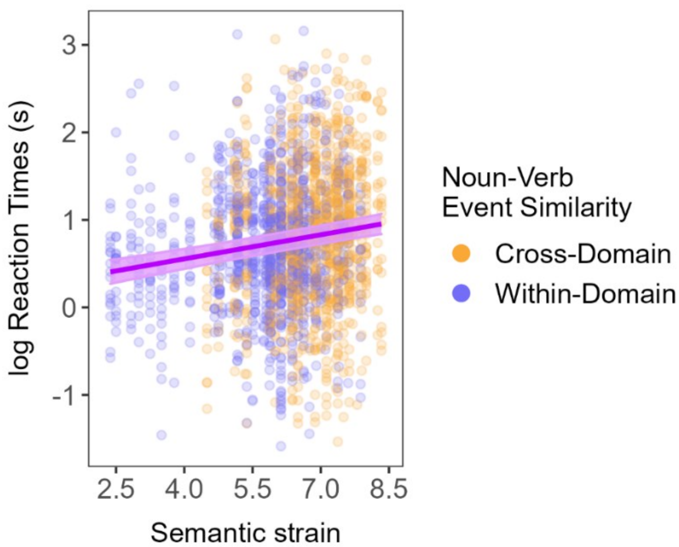



Figure 15. Reaction time results from Experiment 1 showing a significant effect of semantic strain on comprehension times. Shaded ribbon represents 95% confidence bands.

## 2.3 Discussion

The results of Experiment 1 are consistent with the structure-mapping model of verb metaphor comprehension. There were three main findings. First, as predicted, verbs changed meaning significantly more than nouns overall (the verb mutability effect). Second, the degree of verb (but not noun) change increased with the degree of strain (though the effect was marginal for verbs in this instance, $p = .053$). This pattern replicates findings from Chapters 1 and 2 and is consistent with the claim that the events activated by the subject nouns remain as relatively stable referents to which the verb event is adapted during comprehension; and, further, that the degree of verb abstraction increases with semantic strain (i.e., Predictions 1 & 2 from the introduction).

Table 25

*Example paraphrases from Experiment 2*

| Verb Domain | Domain Similarity[a] | Original Sentence | Participant Paraphrase |
|---|---|---|---|
| Motion | Within | The sailboat strutted | The sailboat glided through the water with ease |
| | Cross | The flute strutted | The flute played staggered notes at a constant pace |
| | Within | The kayak hobbled | The small boat wobbled in the water |
| | Cross | The cello hobbled | The cello played a wavering pitch |
| | Within | The kite danced | The kite whirled in the sky, billowing on gusts of air |
| | Cross | The gong danced | The gong moved around after being hit |
| | Within | The raft trudged | The raft moved slowly due to the rapids impeding its progress |
| | Cross | The banjo trudged | The banjo sounded like it was struggling |
| | Within | The sailboat waddled | The sailboat swayed side to side in the wind as it moved along slowly |
| | Cross | The violin waddled | The violin player played clumsily |
| | Within | The canoe scurried | Somebody frantically rode the canoe to the shore |
| | Cross | The accordion scurried | The accordion player opened and closed the instrument very quickly |
| Sound | Within | The violin stammered | The violin played quick, repeated notes |
| | Cross | The glider stammered | The tiny aircraft jerked around in the air |
| | Within | The cello barked | The instrument made a loud cacophonic noise |
| | Cross | The kayak barked | The kayak jumped and spun over the rough waters |
| | Within | The trumpet whimpered | The trumpet let out a sad, tinny sound |
| | Cross | The rowboat whimpered | The rowboat made a concerning noise while being piloted. |
| | Within | The harp grumbled | The large sting instrument played its bass notes. |
| | Cross | The raft grumbled | The raft that had been floating across the ocean made a horrible noise against the sand bar |
| | Within | The accordion chanted | The accordion made a loud, repeated sound |
| | Cross | The kite chanted | The kite made a repetitive sound in the wind |
| | Within | The flute chattered | The notes from the flute were fast and short |
| | Cross | The sailboat chattered | The sails on the boat made loud rapping noises in the large gales |

*Note*. [a] Refers to whether noun event and verb event were from the same domain (motion-motion or sound-sound) or different domains (motion-sound or sound-motion).

These claims are further supported by the example paraphrases shown in Table 25, which show how verb metaphors are consistently interpreted to describe the events that derive from literal interpretations of the nouns. Moreover, verbs' strong propensity to abstract their meanings to accommodate the noun event—even when the two events are quite dissimilar—is once again made clear. For example, in *The cello hobbled* → *The cello played a wavering pitch*, the verb has become highly abstracted by the mapping between the motion event denoted by *hobbled* and the sound event activated by *cello,* such that the cello's pitch is interpreted as wavering in a manner analogous with the wavering physical motion denoted by *hobbled*. For *The glider stammered* → *The tiny aircraft jerked around in the air*, the physical motion of the glider unfolds in a manner analogous to the halting sound production associated with *stammering*.

For both types of nouns, cross-domain pairings sometimes led to interpretations that preserved the domain of the stimulus verb, rather than the domain of the subject noun. As discussed earlier, for vehicles paired with sound verbs, this was expected based on the results of the pilot (e.g., *The sailboat chattered* → *The sails on the boat made loud rapping noises in the large gales*), since even quiet vehicles can make noise. More interesting is what sometimes occurred for musical instrument nouns paired with motion verbs: for those instruments that saliently involve motion in order to produce sound (e.g., *violin, accordion, gong*), participants sometimes drew an analogy between the motion of the verb and the motion of the instrument itself (e.g., *The accordion scurried* → *The accordion player opened and closed the instrument very quickly*; *The gong danced* → *The gong moved around after being hit*). Thus, in the same way that vehicle events appear to include information about sound, some musical instrument events include information about movement, which participants sometimes accessed during comprehension. However, the more typical pattern was that of abstracting the verb more significantly to accommodate the sound domain (e.g., *The accordion chanted* → *The accordion made a loud, repeated sound*; *The gong scurried* → *The gong made small scratching noises*).

The third main finding here provides new evidence for the proposed model: consistent with Prediction 4 of our model, semantic strain predicted comprehension times such that as strain increased, so did the amount of time it took participants to interpret the metaphor. Thus, just as prior work has found that conceptual similarity predicts alignment time (e.g., Gentner & Kurtz, 2006; Sagi et al., 2012), the results here show that the same pattern applies to similarity between event representations in verb metaphors.

The pattern of the strain ratings themselves lends support to our use of semantic strain as an index of event similarity. Recall that event similarity (low vs. high) was initially operationalized as a 2 X 2 grid by crossing nouns and verbs that call forth events from motion and sound domains, under the assumption that within-domain pairings would result in more similarly-structured events than would cross-domain pairings. The strain ratings supported this assumption, indicating that overall, cross-domain items (M = 6.80, SE = 0.22) were significantly more strained than within-domain items (M = 5.77, SE = 0.22), $\beta = 0.22$. SE = .03, $t = 6.54$, $p < .0001$.[28] Thus, the degree of strain appears to vary as a function of the similarity between the event activated by the noun and the verb event.

We further expected that the effect of mismatching domains on semantic strain might be larger for items with musical instrument nouns (*The accordion scurried*) than for those with vehicle

---

[28] Tested by fitting a linear mixed effect model with strain rating as the dependent measure, event similarity (cross-domain vs. within-domain) as the fixed effect, and subjects and items as random effects.

nouns (e.g., *The raft chirped*), since musical instruments cannot move on their own but even quiet vehicles can still make noise. The pattern of strain ratings indicated that the effect of domain similarity on strain was indeed larger for items with musical instrument nouns than for those with vehicle nouns (see Figure 16A). Counter to expectations, however, that difference was due to a difference between musical instruments and vehicles for the within-domain items, rather than the cross-domain items. The degree of strain for cross-domain items (e.g., *The violin sprinted, The sailboat stammered*) did not differ significantly between musical instrument items (M = 6.78, SE = .24) and vehicle items (M = 6.82, SE = .24), $p = .99$; rather, within-domain items (e.g., *The violin stammered*, *The sailboat sprinted*), were significantly more strained for within-domain vehicle items (M = 6.06, SE = 0.24) than within-domain similarity musical instrument items (M = 5.48, SE = .24), $p = .048$.

We are unsure of the reason underlying this unexpected finding. It appears that the motion verbs chosen were less compatible on average with the vehicle nouns than the sound verbs were with the musical instrument nouns. One possible explanation is that our objective of testing vehicles that are relatively quiet led to the selection of nouns whose motion is relatively passive compared to those used in previous experiments (e.g., *blimp, glider, kite, sailboat*). The motion events activated by these nouns may have been harder to align with many of the motion verbs used, which typically suggest active, agentive movement (e.g., *marched, strutted, sprinted, strolled, danced,* etc.). For example, *The glider marched* was rated as relatively high-strain for a within-domain item (M = 6.89, SE = 0.74), perhaps due to the passive motion of glider (i.e., being carried by the wind) conflicting with the purposeful motion denoted by *marched*. Other dimensions of the motion denoted by some of the verbs may have also interacted with the passive motion of the vehicles to increase strain. The two highest-strained within-domain vehicle items were *The rowboat waltzed* (M = 7.87, SE = 0.78 ) and *The blimp pranced* (M = 7.60, SE = 0.70). In both cases, it may have been difficult for raters to imagine how the motion denoted by *waltzing* and *prancing* could apply to the motion that *rowboats* and *blimps* typically engage in.

Although the effect of domain similarity on strain was not as expected for vehicles, the pattern of strain ratings when dichotomized into within- vs. cross-domain categories was nevertheless generally paralleled by both RTs (Figure 16B) and verb word2vec scores (Figure 16C). This affirms the expectation that both comprehension times and degree of verb change depend on the degree of similarity between noun and verb event schemas.[29]

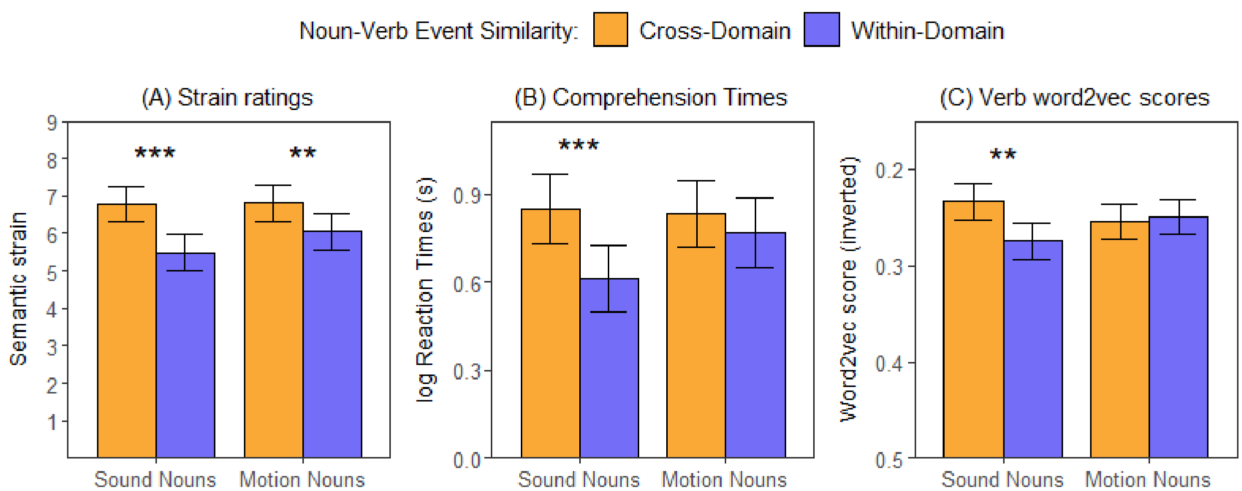



Figure 16. Strain, RT, and word2vec results broken down by noun-verb event similarity. For (C), y-axis direction has been inverted to better depict parallel with (A) and (B). ** p < .01; *** p < .001.

[29] In the factorial models plotted in Figure 16, all three models were fit using the same fixed and random effect structure, with noun type (motion vs. sound), noun-verb event similarity (low vs. high), and the interaction terms as fixed effect, and subjects and items as random effects. The only difference was the dependent measure (strain ratings in A, comprehension times in B, and verb cosine similarity scores in C).

In sum, the results of Experiment 1 constitute novel evidence for the role of structure-mapping in verb metaphor comprehension. Supporting Prediction 4 of our model, the more dissimilar the noun event and verb event that comprise a verb metaphor are, the longer the metaphor takes to comprehend. Thus, just as research in analogy has found that similar conceptual structures are rapidly aligned (e.g., Gentner & Kurtz, 2006; Gentner & Wolff, 1997; Wolff & Gentner, 2000), we find here that the ease of alignment of event structures also depends on their similarity.

## 3 Experiment 2

The objective of Experiment 2 was to build on the results of Experiment 1 by providing further converging online evidence of the involvement of a structure-mapping process in verb metaphor comprehension). Here we turn to the second online prediction of the structure-mapping model (Prediction 5): that structure-mapping results in greater activation of mapped conceptual structure than unmapped conceptual structure. In the case of verb metaphor, that means that structure shared between the noun event and verb event will become activated to a greater extent than non-shared structure. For example, in a metaphor like *The wagon limped*, which is known to be typically interpreted as being about physical motion over land, we expect that structure in the noun event (*roll*) related to physical motion will become activated during mapping, as will structure in the verb event *limped* related to physical motion.

It follows that—assuming one can predict the likely interpretation of a verb metaphor—advance priming of the common structure should speed comprehension. To test this prediction, participants were shown a series of verb metaphors and were instructed to interpret each one as quickly as possible. Each metaphor was preceded by one of two primes (counterbalanced across subjects): (1) a *relevant prime* that activated an event related to the event structure involved in ultimate mapping (and therefore interpretation) of the subsequent metaphor, and (2) an *irrelevant prime* that activated event structure unrelated to the metaphor's interpretation. The idea was that the event structure activated by the relevant prime would overlap with the subsequent metaphoric mapping to a greater extent than the event structure activated by the irrelevant prime. Under structure-mapping, this should result in faster comprehension times.

For example, for the verb metaphor *The bicycle sprinted*, the relevant prime was *pedals* and the irrelevant prime was *handlebar*. Pilot work has shown that people typically interpret *The bicycle sprinted* as being primarily about the bicycle moving rapidly (e.g., *The bike sped down the block*). Our assumption is that this interpretation results from aligning the legged motion event denoted by the verb *sprinted* with the wheeled motion event activated by the noun *bicycle*. Since a bicycle moves as a result of the rider driving its pedals, *pedals* should be a relevant prime—that is, it should activate a noun event similar to that activated in the subsequent metaphor (e.g., bicycle motion).[30] The handlebars of a bicycle, however, do not impart motion and therefore do not activate event structure involved in the subsequent mapping (rather, we expect *handlebar* to activate events related to steering the bicycle). Thus, we expect *The bicycle sprinted* will faster to comprehend when primed by *pedals* than when primed by *handlebar*.[31]

---

[30] The relevant prime does not need to activate an event identical to the one that is activated by the subject noun of the metaphor when it is compared with the verb; rather it should activate one that is more similar to that event than the one activated by the incompatible prime—thus priming it more strongly.

[31] The purpose of the primes was not to influence the interpretation of the metaphor, as the primes were selected to activate events that were either compatible or incompatible with the interpretations known to be typical of the metaphors tested. Thus, we expected interpretations to remain similar across prime types (but for RTs to differ); as will be described in the Results section, however, some influence from the primes on interpretations did occur. As will be discussed, this is not inconsistent with our framework.

For each item, the two primes were always physical components of the entity denoted by the subject noun of the metaphor. While the role-neutral (symmetric) nature of the initial mapping phase of structure-mapping (e.g., Gentner & Wolff, 1997, Wolff & Gentner, 2001, 2011) predicts a benefit from priming components of either the noun or the verb, focusing on primes that are components of the subject noun but that differ in terms of the event schemas they activate acts as a further test of the role of noun events in verb metaphor comprehension.

For each verb metaphor selected, the typical interpretations generated by participants were known based on previous paraphrase studies. In addition, the typical event schemas activated by each noun prime (e.g., *pedals*, *handlebar*) were assessed in a separate norming study (described in the next section). After all metaphor interpretations were collected, the normed prime events were matched to the paraphrases to confirm that the relevant—but not irrelevant—primes activated similar events to those elicited by the subject noun in the metaphors.

A total of 16 verb metaphors and 32 primes were tested (one relevant and one irrelevant prime per metaphor). Each metaphor was presented after either a relevant or irrelevant prime (between subject), and comprehension RTs were collected. Our model makes two key predictions. First, we expect a verb mutability effect: the verb should change meaning significantly more than the noun, which will remain more stable in meaning—thus, the noun event will serve as the referent event to which the verb event is adapted. Second, we expect that relevant primes will facilitate processing more than irrelevant primes, resulting in faster comprehension times.

## 3.1 Method

### *3.1.1 Participants*

A total of 60 native English-speaking Northwestern University undergraduate students from the psychology participant pool completed the study in person in the lab.

An observed power analysis was conducted based on the results of a previously-run pilot study (N = 49, $\alpha = .05$). The pilot study matched the design described above (but tested 14 target items instead of the present 16). Power simulations were conducted using the *simr* package in R (Version 1.0.6; Green & MacLeod, 2016), which indicated an observed power in the pilot of 88.00% (95% CI: [79.98%, 93.64%]) for the effect of prime type (relevant vs. irrelevant; $\beta$ = .08, SE = .02, $p$ = .001), and suggesting a power of 95.20% (95% CI: [93.69%, 96.44%]) for the present sample size of N = 60 and $\alpha = .05$.

### *3.1.2 Materials*

Target stimuli consisted of 16 intransitive verb metaphors. Each verb metaphor comprised an artifact subject noun paired with a human- or animal-expecting verb (e.g., *The bicycle sprinted, The rifle barked, The car limped*, etc.). The metaphors' typical interpretations were known based on previous paraphrase studies. The metaphors, the events expected in their interpretations, and the relevant and irrelevant primes are shown in Table 26.

Based on these anticipated interpretations, a relevant and an irrelevant prime were selected for each metaphor. As described above, the relevant primes were intended to activate event schemas that would overlap with the anticipated mapping of the subsequent verb metaphor; conversely, the irrelevant primes were selected to activate event schemas that overlapped little with the subsequent mapping. Each prime was a physical component of the entity denoted by the metaphoric subject noun (e.g., *pedals* and *handlebar* for *The bicycle sprinted*; *trigger* and *crosshairs* for *The rifle barked*; *tire* and *windshield* for *The car limped*, etc.). Importantly, relevant and irrelevant primes were matched for *a priori* relatedness with both the subject noun and verb of the corresponding verb metaphors using word2vec (all *ps* > .53). This ensured that neither prime type was more strongly associated with either the subject noun or verb, and therefore that any effect of prime type was due to the nature of the event schemas activated by the primes. Target items, the events expected in their interpretations, and primes are shown in Table 26 below.

Table 26

*Target items and primes from Experiment 2*

| | | **Prime** | |
|---|---|---|---|
| **Item** | **Expected event**[a] | **Relevant** | **Irrelevant** |
| *The bicycle sprinted* | Motion over land | *pedals* | *handlebar* |
| *The blender attacked* | Blending food | *blade* | *button* |
| *The boat strutted* | Motion over water | *propellor* | *deck* |
| *The car limped* | Motion over land | *tire* | *windshield* |
| *The firetruck yelled* | Sound generation | *siren* | *hose* |
| *The guitar stammered* | Sound generation | *string* | *knob* |
| *The kettle drooled* | Dripping water from spout | *spout* | *lid* |
| *The mattress shrieked* | Sound generation | *spring* | *tag* |
| *The plane waddled* | Motion through the air | *wing* | *seat* |
| *The rifle barked* | Rifle firing | *trigger* | *crosshairs* |
| *The truck howled* | Sound generation | *horn* | *seatbelt* |
| *The typewriter babbled* | Typing | *keys* | *paper* |
| *The flashlight stuttered* | Illumination | *bulb* | *battery* |

| | | | |
|---|---|---|---|
| *The lantern waltzed* | Illumination | *wick* | *pole* |
| *The vacuum cackled* | Sound generation | *motor* | *handle* |
| *The sprinkler coughed* | Dispensing water | *nozzle* | *valve* |

*Note*. [a]Refers to the general category of event that interpretations are expected to describe based on previous studies.

In addition to the 16 verb metaphors, 24 literal filler items were included. The fillers ensured that participants received a larger number of normal literal sentences than odd metaphoric sentences in order to avoid driving participants into abnormal patterns of thinking. The fillers matched the pattern of the target items such that half the filler items were preceded by a prime that was relevant to the target sentence meaning (e.g., *leg* → *The horse galloped*) and half were preceded by an irrelevant prime (e.g., *fur* → *The dog barked*). In addition, half the fillers had artifact noun subjects, and half had animal noun subjects.

Finally, four practice items were included in order to familiarize the participants with the experimental procedure before beginning the main experiment. Two practice items were literal, one preceded by an irrelevant prime and the other with a relevant prime, and likewise for the remaining two practice items, which were metaphoric. All filler and practice items are shown in Appendix **M**.

*3.1.2.1 Prime event norming*

To confirm that the primes activated the expected event schemas, a norming study was conducted to determine the typical events people associate with each prime. The 32 primes were given to a group of 40 participants who completed the study online via Amazon's Mechanical Turk (MTurk) platform. Participants were paid for their participation the equivalent of Illinois' minimum wage at the time of the study.

Each prime was presented along with the noun subject it was a component of in parentheses (e.g., *pedals (of a bike), blade (of a blender), tire (of a car)*). This was done to ensure that participants reported the typical events for the appropriate sense of any ambiguous words.[32] Each participant was given 16 of the 32 total primes, half of which were relevant primes and half of which were irrelevant. If a participant saw the relevant prime for a given metaphor, they did not see the irrelevant prime for that metaphor. Each participant also completed the same six catch trials that served as attention checks. The criteria for violating a catch trial was providing an anomalous response to the catch trial (e.g., saying the typical action of an *eraser* is to *write*). The stimuli for the norming study are shown in

[32] This approach addressed two potential sources of ambiguity with the primes: semantic and syntactic homonyms (e.g., *truck horn vs. animal horn* or *car tire* vs. *tire out*). As will be described below, the primes were presented without qualification in the main experiment (as single words)—however, there is evidence that presenting words without context elicits initial activation of all senses (i.e., exhaustive access) of ambiguous words (Burgess & Simpson, 1988; Frost & Bentin, 1992; Lin & Chen, 2015; Meade & Coch, 2017); further, even non-dominant senses of ambiguous words have been shown to exert priming effects (Meade & Coch, 2017). Thus, we assume here that the events elicited by participants in the norming study were still activated when the primes were presented without further qualification in the main experiment.

Appendix **N**.

Participants were told that we were interested in their knowledge of the things that everyday objects typically do. They were instructed to provide words or short phrases describing the typical actions that come to mind for each object they saw in the order that they thought of them. Using a paradigm adapted from a noun event norming study by McRae et al. (2005), each participant was given five spaces to list events for each word. They were not required to provide more than one response for each word; rather, they were told to list multiple actions if they came to mind naturally.

5 participants were excluded for failing the catch trials, resulting in a net of 35 participants who generated a total of 1316 responses for the 32 primes (Mean responses per prime = 36.41, SD = 7.91). Next, any responses that clearly referred to the same or similar event were condensed into single categories by the experimenters and given a descriptive label. For example, “directs wheel,” “pivoting the front wheel,” “steer the bike,” “steering,” “steers,” and “directs” were condensed into the single category “steers the bike” for the prime *handlebars (of a bike)*, as they all refer to the same action. Responses that did not clearly refer to the same action were not placed in the same category. For example, for the prime *pedals (of a bike*), the responses “allows the rider to provide power to move,” “causes speed,” and “moves bike forward” were categorized as “propels bike/rider,” while the responses “moves in a circular motion,” “spin around,” and “spins around” were categorized separately as “moves in a circular motion.”

Condensing the 1316 net responses resulted in 132 unique event categories (an average of 4.13 categories per prime.).[33]

Next, following the procedure used by McRae et al. (2005), a weighted score for each event category was calculated by multiplying the frequency with which a given event category occurred by the order it was listed in by each participant (first, second, third, fourth, or fifth). In other words, the number of times that event was listed first by participants was multiplied by five, the number of times it was listed second was multiplied by four, third by three, fourth by two, and fifth by one. Thus, events that were listed first by participants (indicating they were more top-of-mind) were weighted more heavily than those that were listed in later positions. Finally, the scores for each event were summed and sorted in descending order to produce the final rankings. The top-ranked event category for each prime is shown in Table 27 below. The full rankings are shown in Appendix O.

If the primes we selected were appropriate for their intended purpose—that is, if the relevant primes activate events similar to the event activated by the subject noun of the metaphor that followed—and if the irrelevant primes do not—then we expect that (1) each relevant prime will have at least one event listed that is similar to the expected metaphor event, and (2) that event should generally be among the top-ranked events for that prime. For example, *The rifle barked* was expected to be interpreted about a rifle firing. As Table 27 shows, the top event for the relevant prime *trigger* is "causes gun to fire." As another example, *The blender attacked* was

---

[33] Participants occasionally listed properties of the primes instead of events (e.g., "contains the brakes" for *handlebar (of a bike)*). These were treated the same as the event responses in the analysis—that is, similar property responses were condensed into categories and provided to the coders to match with the paraphrases should they appear in the participant's interpretation.

expected to be about blending objects in the blender, and the top event for the relevant prime *blade* is “cuts / blends food.”

An informal initial review of the results suggested that for all relevant primes, at least one event was listed that appeared similar to the expected event in the metaphor (as shown in Table 26 above). For 13 out of the 16 primes (81.2%), a similar event was the top-ranked event. For 2 out of 16 primes (12.5%), a similar event was second or third (*The sprinkler coughed, The vacuum cackled*, respectively), and for 1 prime it was eighth (*The mattress shrieked*)—see Appendix O. Conversely, none of the irrelevant primes appeared to have events listed that matched the expected metaphoric event. As will be described below, independent coders were used to match these prime events to the actual paraphrases resulting from the main experiment.

Table 27

*Top-ranked events from the event norming study*

| *Item* | Prime Type | Prime | Event |
|---|---|---|---|
| *The bicycle sprinted* | Relevant | pedals | propels bike/rider |
| | Irrelevant | handlebar | steers the bike |
| *The blender attacked* | Relevant | blade | cuts / blends food |
| | Irrelevant | button | turns blender on/off |
| *The boat strutted* | Relevant | propellor | propels boat |
| | Irrelevant | deck | provides place for people to stand / sit / lay |
| *The car limped* | Relevant | tire | allows for motion / rolls over street |
| | Irrelevant | windshield | Protects occupants from wind / rain / other objects |
| *The firetruck yelled* | Relevant | siren | makes a loud noise / alerts other people |
| | Irrelevant | hose | dispenses water |
| *The flashlight stuttered* | Relevant | bulb | illuminates objects |
| | Irrelevant | battery | provides energy / power |
| | Relevant | string | produces sound / allows for playing |

| | | | |
|---|---|---|---|
| *The guitar stammered* | Irrelevant | knob | adjusts the sound / volume / tuning / etc. |
| *The kettle drooled* | Relevant | spout | pours liquid / directs flow of liquid |
| | Irrelevant | lid | prevents overflowing / spilling |
| *The lantern waltzed* | Relevant | wick | burns / creates light / illumination |
| | Irrelevant | pole | holds / carries / hangs the lantern |
| *The mattress shrieked* | Relevant | spring | absorbs / supports weight / body |
| | Irrelevant | tag | gives information about the mattress |
| *The plane waddled* | Relevant | wing | allows plane to fly / keeps plane in air |
| | Irrelevant | seat | allows people to sit / provides comfort during flight |
| *The rifle barked* | Relevant | trigger | causes gun to fire |
| | Irrelevant | crosshairs | provides aim / increases accuracy |
| *The sprinkler coughed* | Relevant | nozzle | directs /controls water stream |
| | Irrelevant | valve | allows / prevents water flow / releases water |
| *The truck howled* | Relevant | horn | makes a loud sound / alerts / warns others |
| | Irrelevant | seatbelt | keeps person safe |
| *The typewriter babbled* | Relevant | keys | allows for typing / producing text |
| | Irrelevant | paper | receives text / marks / ink |
| *The vacuum cackled* | Relevant | motor | creates suction / air flow / air pressure / makes vacuum work |
| | Irrelevant | handle | allows vacuum to be gripped / moved / carried / manipulated |

### *3.1.3 Design*

Returning to the main experiment, each participant received all 16 verb metaphors, 24 literal fillers, and 4 practice items. For each participant, half the metaphors were preceded by relevant primes and half were preceded by irrelevant primes (counterbalanced across subjects). Thus, there were two between-subject item groupings based on prime type. If a metaphor was preceded by a relevant prime in the first grouping, it was preceded by an irrelevant prime in the second grouping.

### *3.1.4 Procedure*

Participants were randomly assigned to one of the two item groupings. Participants first read instructions telling them that they would see a series of sentences and that they should interpret each sentence as quickly as possible. They were instructed to press the space bar as soon as they had thought of a *meaningful* interpretation (but not before) and were shown an example of what the structure of each trial would look like. The example indicated that each sentence would be preceded by a word (the prime), but the purpose of the word was not explained.

Participants first completed the four practice trials to familiarize themselves with the experimental procedure before moving on to the main experiment. The first two trials of the main experiment were the same two literal fillers across all participants. The remaining target and filler trials were presented in randomized order across participants.

The structure of each trial is shown in Figure 17 below.

| **Description** | Ready message | | Fixation cross | | Prime mask | Prime | ISI | Metaphor | Paraphrase |
|---|---|---|---|---|---|---|---|---|---|
| **Example Text** | Get Ready! | [Blank] | + | [Blank] | **** | pedals | [Blank] | *The bicycle sprinted* | *The bike went fast* |
| **Duration (ms)** | 1000 | 1000 | 1000 | 100 | 500 | 1500 | 1000 | Displayed until user presses [SPACE] | Participant types paraphrase |

Figure 17. Trial structure for Experiment 2.

On each trial, three key measures were collected: (1) the amount of time from metaphor presentation to space-bar press (indicating comprehension time), (2) the amount of time from space-bar press until the participant began typing their response, and (3) the paraphrase / interpretation of the item.

### 3.2 Results

The 60 participants each paraphrased all 16 items, resulting in a total of 960 responses (60 per item). Half the responses corresponded to trials where the metaphor was preceded by a relevant prime, and half corresponded to trials where it was preceded by an irrelevant prime (counterbalanced across subjects). Two responses where the participant did not provide an interpretation of the metaphor were excluded, for a net of 958 responses included in the analysis. Example paraphrases are shown in Table 28 below.

Table 28

*Example paraphrases from Experiment 2*

| Original Sentence | Paraphrase |
|---|---|
| The bicycle sprinted | The bike went really fast. |
| The blender attacked | The blender blended what was in it ferociously and loudly |
| The boat strutted | The boat moved smoothly on the water. |
| The car limped | The vehicle slowly drove on after popping its tire |
| The firetruck yelled | The firetruck siren made a loud noise |
| The flashlight stuttered | The flashlight started flickering on and off |
| The guitar stammered | The guitarist’s fingers fumbled while they played the guitar |
| The kettle drooled | The kettle had a bit of tea drip out of it at the end |
| The lantern waltzed | The light in the lantern appeared like it was dancing |
| The mattress shrieked | The mattress squeaked when sitting on it |
| The plane waddled | The wings on the plane made it tip side to side while it flew. |
| The rifle barked | The rifle made a short, harsh sound as it went off |
| The sprinkler coughed | The water from the sprinkler came out in bursts |
| The truck howled | The truck sounded its horn loudly for a long period of time |

| | |
|---|---|
| The typewriter babbled | The keys of the typewriter clicked incessantly |
| The vacuum cackled | The vacuum sucked up something that made a loud crackling noise |

There are two main predictions. First, we expect that verbs—but not nouns—will change meaning in response to strain. Thus, the noun event should typically serve as the referent to which the verb's meaning is adapted via metaphoric extension. Second, we expect that metaphors should be faster to understand when preceded by a relevant prime than when preceded by an irrelevant prime.

To test the first prediction, the same procedure used in the previous experiments for testing the average overall difference between noun and verb cosine similarity scores was used. As predicted, verbs changed significantly more under paraphrase (M = 0.23, SE = .02) than did nouns (M = .52, SE = .02), $b = .28$, SE = .02, $t = 12.03$, $p < .0001$.

Next, to test the second prediction—that people will be faster to interpret verb metaphors when preceded by relevant primes than when preceded by irrelevant primes—the following analysis was performed. First, we followed the same procedure described in Experiment 1 for log transforming RTs and excluding outliers, which resulted in the exclusion of 1.77% of the data. Next, a linear mixed effect model was fit on this trimmed dataset, with logRT (comprehension time) as the dependent measure, prime type (relevant vs. irrelevant) as the fixed effect, and subjects and items as random effects (as in Experiment 1, diagnostic residual plots confirmed a good linear fit on the log-transformed data).

As predicted, the effect of prime type was significant: participants were faster to interpret the verb metaphors when they were preceded by the relevant prime compared to the irrelevant prime, $\beta = 0.08$, SE = .02, $t = 4.53$, $p < .0001$. The results are plotted in Figure 18, and mean RTs by condition are shown in Table 29.[34]

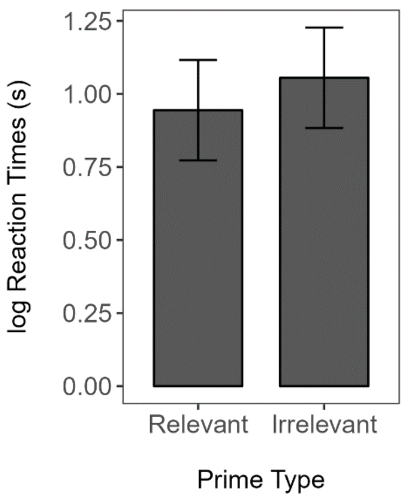



Figure 18. Mean comprehension times from Experiment 2 by prime type. RTs were log-transformed prior to model fitting. Error bars correspond to 95% CIs.

Table 29

*Mean comprehension times by prime type for Experiment 2*

| | Relevant | Irrelevant |
|---|---|---|
| log RTs (s) | 0.87 (0.08) | 0.99 (0.08) |
| RTs (s) | 3.01 (0.28) | 3.33 (0.28) |

Note. Standard errors in parentheses. Non-transformed (raw) RTs are provided for reference.

This result is consistent with our hypothesis that priming noun event knowledge relevant to the meaning of the subsequent verb metaphor speeds comprehension more than priming noun event

[34] As in Experiment 1, participants were instructed to press the spacebar as soon as they had thought of a meaningful interpretation. We ran the same analyses described in Experiment 1 (Footnote 27) to see if participants who did not begin typing immediately after pressing the spacebar (suggesting that they were still processing) affected the results. As before, we fit three additional models, excluding trials where participants took longer than 2 seconds, 1.5 seconds, or 1 second (5.43%, 9.82% and 23.07% of the data) to begin typing. Also as before, in each case, excluding these participants did not change the results (parameter estimates were nearly identical to the full model, $p < .001$).

knowledge that is irrelevant to the meaning of the metaphor. Since both primes were physical components of the metaphoric subject noun, and both were matched in terms of *a priori* relatedness to the target noun (and verb) that followed, the significant effect of prime type suggests that the difference in comprehension times by condition is driven by the different event schemas that the primes activated. Specifically, relevant primes activated event schemas that were more compatible with the noun event schemas activated by the subsequent metaphor than were those activated by the irrelevant primes.

### *3.2.1 Prime and paraphrase event coding*

To further bolster this claim, we conducted an additional coding task on the paraphrases using the norming data for the primes that were obtained via the study described earlier in the Method section. The idea was to assess the extent to which the events listed for the primes by the participants in the norming study matched the events described by metaphor paraphrases in the present RT study. The expectation was that the events listed for the relevant primes from the norming study should match the events in the paraphrases more often than those listed for the irrelevant primes, as only the relevant primes should be activating event schemas that are compatible with the corresponding verb metaphor (i.e., the presumed mechanism behind the observed priming effect). Since the goal of the design was that each metaphor would be interpreted similarly regardless of the prime that preceded it, the events listed for the relevant primes should match the paraphrases most often, regardless of which prime preceded the metaphor that generated the paraphrase (recall that each metaphor was preceded by a relevant prime half the time and an irrelevant prime half the time).

To test this, two RAs who were blind to the study's hypotheses coded a subset of the paraphrases. 30 of the 60 paraphrases for each of the 16 metaphors were randomly selected, half of which corresponded to trials where the item was preceded by a relevant prime, and half of which corresponded to trials where the item was preceded by an irrelevant prime. Thus, a total of 480 paraphrases, or about 50% of the net 958 paraphrases collected, were coded.

The coding process is illustrated for the example of *The blender attacked* in Figure 19. The judges matched the typical events/actions listed for each prime from the norming study (Figure 19A) to the events described by the paraphrases for that item (Figure 19B). For example, the paraphrase *The blender blended up roughly* matches the event "cuts / blends food" from the list of typical events for the relevant prime *blade*; the paraphrase does not match any events listed for the irrelevant prime. The paraphrase *The blender blended what was in it ferociously and loudly* matches two events from the relevant prime event list from the norming study: "cuts / blends food" and "is noisy when used," while the paraphrase *The button provided the command for the appliance to turn on and off* matches the event "turns blender on/off" from the irrelevant prime norming data and no events from the relevant prime norming data. *The machine stopped working* did not match any events from the relevant or irrelevant prime.

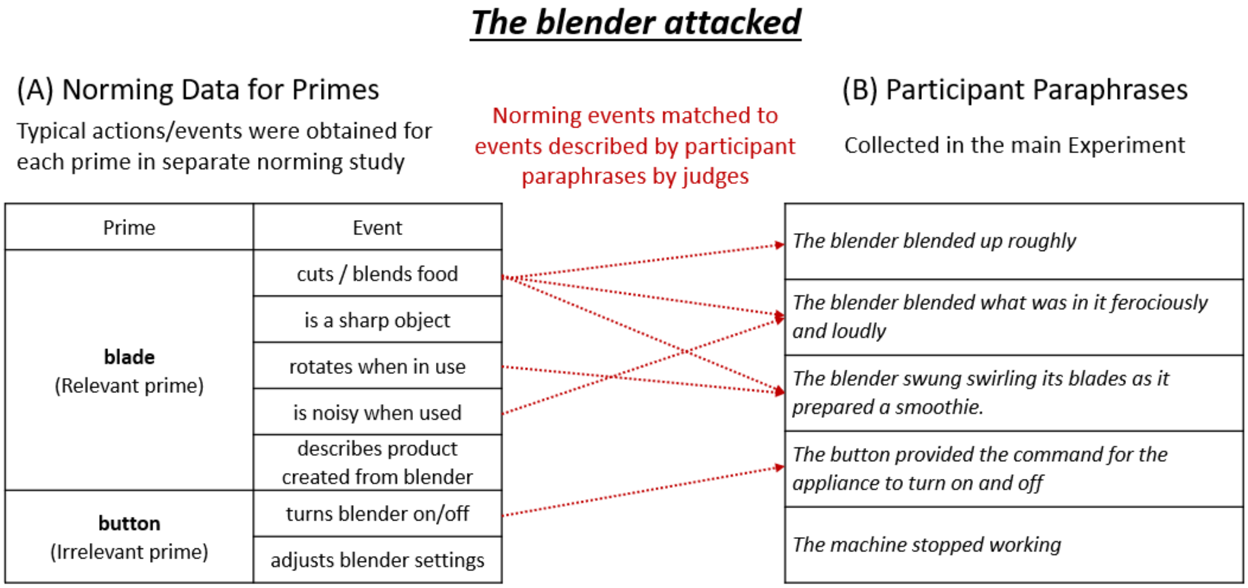



Figure 19. Illustrating the event coding process from Experiment 2. For every item, judges indicated any matches between the events from the norming study listed for both primes for that item (A) and its paraphrases (B). Each paraphrase could have multiple matches or no matches from the prime lists. The arrows indicate the actual matches made by the judges for those paraphrases. A total of 30 paraphrases were coded for each item.

The paraphrases for each item were coded twice: first, judges indicated all matches from the list of events for the relevant prime; they then did the same for the list of events for the irrelevant prime. All paraphrases were coded using both lists, regardless of which prime preceded the corresponding metaphor in the actual experiment (i.e., each item was coded for relevant events 30 times and irrelevant events 30 times, half of which corresponded to trials where the metaphor was preceded by the relevant prime, and half of which it was preceded by the irrelevant prime). If no events from the prime event lists matched the paraphrases, that was recorded as well. Following the same coding procedure outlined in previous chapters, the judges coded in chunks, making initial judgements independently before coming together to reconcile any disagreements. The judges were able to reach a final consensus on all items. There was strong initial agreement between the two judges, $\kappa = 0.81$, (95% CI, 0.78 to 0.84), $p < .0001$.

The results confirmed that the events described by the metaphor paraphrases were highly likely to match the events listed for the relevant primes from the norming data, but not the events listed for the irrelevant primes. Overall, 75.8% of the paraphrases matched at least one event from the relevant prime event list, while only 7.71% matched an event from the irrelevant prime event list.[35] It was most common for the matching event from the relevant prime list to be the top-ranked event in the list for that prime (69.10% of matches), and 86.96% of relevant event matches corresponded at least one of the top three ranked events. This suggests that the events most saliently associated with the relevant primes—that is, the ones most frequently listed and listed first by participants in the norming study—were the ones that best matched the events that typically surfaced in the metaphor interpretations. Thus, as intended, the relevant primes appear to have activated saliently-associated event schemas that were relevant to the meaning of the subsequent metaphors, while the irrelevant primes—whose events rarely matched with the paraphrases—did not.

An important question is whether the primes themselves influenced the interpretations of the metaphors that followed them. The intent of the study was for people to interpret the metaphors in roughly the same way regardless of the type of prime that preceded it, but the ease with which that interpretation was produced should vary based on the prime (as only the relevant prime should activate compatible event structure). However, for participants to sometimes shift to a different interpretation when given the incompatible prime would not be surprising. Under our

[35] A mixed effect logistic regression model confirmed the difference was significant, $b = 3.37$, SE = .50, $z = 6.74$, $p < .0001$.

framework, we assume that in these cases a different noun event from the intended one was primed and successfully aligned with the verb event.

Analyzing the event matches by the prime type that preceded the metaphor (relevant vs. irrelevant) for that response indicated that prime itself did have some effect on participants' interpretations (see Figure 20). When a metaphor was preceded by a relevant prime, its interpretation was significantly more likely to match an event listed for that prime from the norming data (e.g., the list shown in Figure 19A; 83.33% of paraphrases) than when it was preceded by an irrelevant prime (68.33% of paraphrases), $p = .04$. Similarly, when a metaphor was preceded by an irrelevant prime, it was significantly more likely to match an event listed for that prime from the norming data (12.5% of paraphrases) than when it was preceded by a relevant prime (2.92% of paraphrases), $p = .002$.

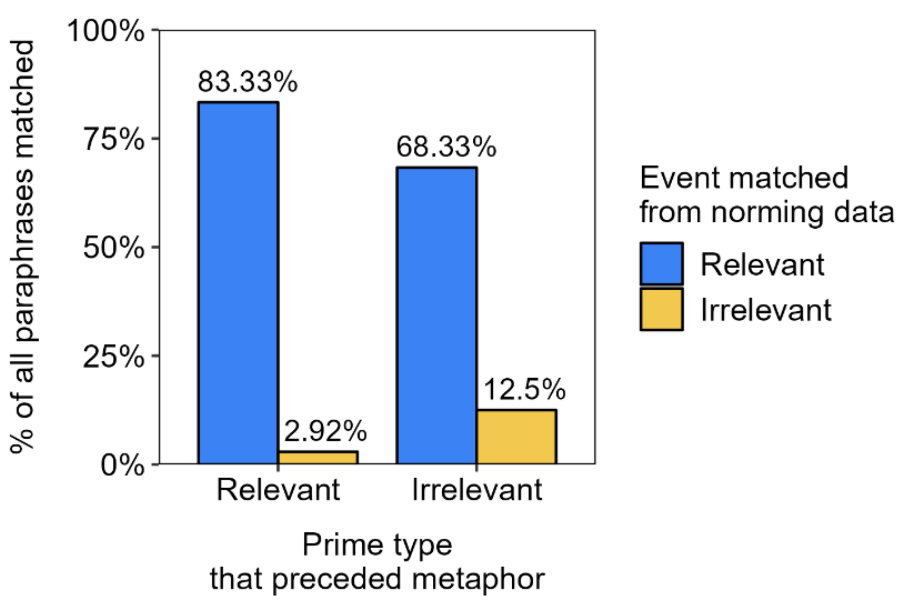



Figure 20. Results of event coding study from Experiment 2. The x-axis indicates the type of prime that preceded the metaphor that was paraphrased. The y-axis indicates the percent of all 480 coded paraphrases that matched for that prime type. Blue bars represent cases where the paraphrase matched an event from the relevant primes event list (obtained from the norming data; e.g., Figure 19A). Yellow bars represent cases where the paraphrase matched an

event from the irrelevant primes event list. For each prime type on the x-axis, percentages do not sum to 100% because some items matched both relevant and irrelevant events, and some matched neither.

Thus, both relevant and irrelevant primes had an effect on the interpretations of the metaphors that followed them: events consistent with those activated by the given prime were significantly more likely to surface in the corresponding paraphrase for both prime types. This effect was relatively small, however; the events in the paraphrases matched the events activated by the relevant primes a strong majority of the time, even when the metaphor was preceded by the irrelevant prime—that is, even when participants had never seen the relevant prime beforehand (and when the prime they did see suggested an alternative event). While seeing the irrelevant primes did increase the frequency with which the corresponding irrelevant events surfaced in the paraphrases, those interpretations remained a small minority of the paraphrases. Therefore, as intended by our design, participants typically interpreted the metaphors in roughly the same way, regardless of the preceding prime, and in a manner consistent with the events activated by the relevant prime.

This pattern therefore supports two of our assumptions in the design of the experiment: (1) the metaphors were generally interpreted in line with our expectations (i.e., matching the expected events shown in Table 26), regardless of the preceding prime, and (2) the relevant (but not irrelevant) primes independently activated event schemas that were consistent with the events in those interpretations, as evidenced by the high rate of matches between events listed for the relevant prime from the norming data and metaphor paraphrases, regardless of the preceding prime. Together, these two findings bolster our posited mechanism underlying the significant effect of priming on comprehension times: RTs were speeded by prior activation of compatible (but not incompatible) noun event structure.

### 3.3 Discussion

The objective of Experiment 2 was to test the involvement of the structure-mapping process in verb metaphor comprehension. If verb metaphors are processed as structure-mappings between noun and verb event structures, then prior activation of event structure that will be involved in the mapping should speed processing. We chose to prime noun event structure specifically to provide further evidence for the claim that verb metaphors involve the activation and integration of prototypical events associated with the subject noun.

There were two main findings. First, as expected, verbs changed meaning significantly more than nouns in the paraphrases. This (as well as the results of the event coding task described above) indicated that the event activated by the subject noun of the metaphor served as the target event to which the verb event was adapted (see Table 28).

Second, we found the predicted effect of prime type on comprehension times: participants were significantly faster to interpret metaphors preceded by relevant than irrelevant primes, suggesting that, on average, the event schemas activated by the relevant primes overlapped with (and therefore activated) the event schemas activated by the metaphor itself, while the events activated by the irrelevant prime did not. As all the primes were components of the subject noun of the metaphor, it follows that the observed effect was due specifically to activation of noun-related event schemas. This supports the idea that noun event schemas are an important part of verb metaphor processing.

The event norming and coding tasks further supported this assertion. Together, the tasks showed that the events most saliently associated with the relevant (but not irrelevant) primes matched those that surfaced most often in the subsequent metaphor interpretations. Thus, for the metaphor *The blender attacked*, the *cutting/blending* event that was ranked highest for the relevant prime *blade* matched the event described most often by the paraphrases (e.g., *The blender eviscerated fruit chunks, The blender cut up the stuff inside, The blade pierced through the fruit in the blender*). For the metaphor *The flashlight stuttered*, the *illuminates objects* event that was ranked highest for the relevant prime *bulb* matched the event described most often by the paraphrases (e.g., *The flashlight was flickering, The flashlight was blinking, The flashlight's bulb kept turning on and off*).

This pattern of results makes clear that (1) both the relevant primes and metaphors activated similar event schemas independently of one another, and (2) the observed effect of prime type on comprehension times is likely due to the priming of common noun event structure that occurred with the relevant (but not irrelevant) prime. Thus, these results support a key claim of our model: that noun event schemas are an important part of verb metaphor comprehension.

### *3.3.1 Ambiguity of the primes*

One issue worth discussing before moving on has to do with the primes themselves. Some of the primes used were ambiguous when presented in isolation (e.g., *horn* can be interpreted as a vehicle horn, a musical instrument, or the horn of an animal; *tire* can be interpreted as a verb or a noun, etc.). For this reason, in the norming study (but not the RT study), the primes were disambiguated by including the relevant subject noun (e.g., *horn (of a truck)*) to ensure that participants listed events for the right sense of the word.

Ambiguous words were used due to the difficulty of finding appropriate primes that satisfied all the necessary criteria for inclusion in the study. However, there is evidence that when ambiguous words are presented in isolation, multiple meanings are activated initially (e.g., Burgess & Simpson, 1988; Frost & Bentin, 1992; Lin & Chen, 2015; Meade & Coch, 2017); in addition, there is evidence that both dominant and non-dominant senses of homonyms exert priming effects (though the effect from non-dominant senses may be weaker in some instances, Meade & Coch, 2017). Thus, our assumption here was that priming would occur even if the relevant meaning of a given prime was not the dominant sense. Nevertheless, the difference between how the primes were presented in the norming study (disambiguated) vs. in the main experiment (ambiguous) raises the question of whether the strength of event activation and/or priming may have differed depending on the degree of prime ambiguity or the dominance of the relevant meaning. Primes where the relevant event schema pertained to the dominant sense of the word may have exerted stronger priming effects than those where the relevant schema pertained to a subordinate sense. Further work restricted to only unambiguous primes could shed light on this issue.

### *3.3.2 Noun event schema selection*

These results bear on an important question that has been discussed little up to this point: how is an appropriate noun event schema selected during verb metaphor comprehension? As mentioned earlier, it seems likely that many nouns have multiple associated event schemas. For example, vans likely have schemas related to *driving, being loaded with things, being filled with gas*, *being purchased* or *sold*, etc. We assume that one schema is typically dominant by virtue of how frequently the event is encountered compared to other events (e.g., in the case of a van, *driving* is likely a more commonly encountered event than *being purchased*), and these are the schemas focused on in this thesis. But it is clear that which event is selected depends also on the paired verb. For most people, the event activated by *The van limped* likely differs from that described by *The van guzzled*.

In this experiment, both the relevant and irrelevant primes were physical components of the subject noun, and both activated event schemas directly related to that noun. For example, for *The firetruck yelled*, the sound event activated by the relevant prime *siren* and the water-spraying event activated by the irrelevant prime *hose* are both common events that firetrucks participate in. One key difference between those events, however, may be their compatibility with (or similarity to) the verb event. The sound event activated by *siren* is more compatible with/similar to the event denoted by *yelled* than is the water-dispensing event activated by *hose*. We speculate that this compatibility/similarity between noun event and verb event is a crucial factor in determining which noun event is selected during comprehension. That the relevant prime resulted in faster comprehension times than the incompatible prime is consistent with this idea, as only the relevant prime activated compatible event structure. We discuss this point further in the General Discussion.

### *3.3.3 Summary*

The results of this experiment support two key claims of our model: (1) that structure-mapping underlies verb metaphor comprehension, and (2) these mappings involve events activated by the noun. First, priming relevant noun-related event structure resulted in faster comprehension times than did priming irrelevant structure. Thus, consistent with Prediction 5, prior activation of mapped relational structure facilitated alignment.

Second, using the normed prime event data and a coding task, we confirmed the importance of these noun events by showing that the events activated by the relevant primes matched the events that surfaced in the metaphor paraphrases much more frequently than did those activated by the irrelevant primes. Consistent with this, there was also a verb mutability effect such that verbs changed significantly more than nouns under paraphrase, indicating that the noun (and its activated event) remained relatively stable as the referent; the paraphrases shown in Table 28 further underscore this point.

# 4 General Discussion

In this chapter, we proposed and tested a novel process account for how verb metaphors are comprehended: they are understood in the same way that analogies are—as analogical comparisons processed via structure-mapping. Metaphors like *The boat pranced* are understood by comparing the event schema denoted by the verb (*prancing*)—the base of the analogy—to an event schema activated by the noun (*motion over water*)—the target of the analogy. As with any other structure-mapping, processing unfolds in two phases. First, the two representations are aligned and common structure is mapped—in this case identifying that both events involve physical motion. Next, information connected to that mapping that is present in the verb event but not the noun event (e.g., *prancing* is physical motion that occurs in a springy, up-and-down manner) is then projected over to the noun event, and an interpretation like *The boat went up and down over the waves* is obtained. Thus, verb metaphors manifest two of the hallmarks of analogies: they identify commonalities between unlike things, and they lead to the spontaneous generation of inferences about the target from the base.

The objective of this chapter was to demonstrate that verb metaphor comprehension fits well into the framework of analogy. We laid out five key empirical predictions that arise by treating verb metaphors as a type of analogy: (1) the verb mutability effect, (2) minimal subtraction, (3) inference projection, (4) event similarity predicts comprehension time, and (5) prior activation of relevant event structure speeds comprehension. In discussing these predictions, we have pointed out the consistent parallels between the behavior of comparisons that are commonly accepted as analogies (e.g., *Misinformation is like a virus*) and the behavior verb metaphors, and have argued that all five predictions are well-supported by the experimental evidence reported in this thesis. In what follows, we first review each prediction and the evidence supporting it. We then conclude with the theoretical implications of the structure-mapping model of verb metaphor comprehension.

## 4.1 Evidence for the structure-mapping model of verb metaphor

### *4.1.1 Prediction 1: The verb mutability effect*

In analogies like *Misinformation is like a virus*, comparison results in the abstraction of the analogical base, while preserving the literal meaning of the target. The base serves to convey relational information that asserted to also hold for the target; irrelevant objects and attributes in the base representation are discarded (abstracted). The target, however, remains literally construed.

If verb metaphors are analogies, a parallel pattern should hold such that the verb (the base of the analogy) becomes abstracted, while the noun remains relatively stable as the literal referent—i.e., the verb mutability effect. The results reported in this dissertation are unequivocal in supporting this prediction: in six different experiments across Chapters 1 and 2, and in both experiments reported in this chapter, there was a consistent pattern of verb mutability and noun stability. Further we found in Chapter 1, Experiment 3 that these verb meaning changes were primarily analogical abstractions of the verb's literal meaning, as would be expected with an underlying analogical mechanism. Moreover, in Experiment 1 of this chapter, this pattern held even when the noun and verb events were highly dissimilar—e.g., participants readily mapped a motion event from the verb to a sound event from the noun, as in *The violin marched* → *The stringed instrument played rhythmically to the beat*. Thus, as predicted, verb metaphor comprehension consistently results in abstraction of the verb as it is adapted to accommodate the literal noun referent.

#### *4.1.2 Prediction 2: Minimal subtraction*

In analogies, the degree of abstraction the base undergoes as a result of comparison depends on how similar it is to the target. The comparison *An isthmus is like a bridge* requires less abstraction of the base (*bridge*) than does *An education is like a bridge*, because isthmuses and bridges share both domain-specific and domain-general commonalities, while educations and bridges share only abstract, domain-general commonalities. Under structure-mapping, mapping between concepts retains as much common structure as possible (Forbus et al., 2017; Gentner, 2010); thus, abstraction proceeds in a domain-specific to domain-general direction as a function of the degree of shared structure between base and target: as shown by the above examples, to the extent that they share domain-specific commonalities relevant to the mapping, those will be retained.

Our model applies this general analogical framework to the more specific verb metaphor case by viewing the type and degree of verb abstraction as a function of the similarity between two *event* concepts: the noun and verb event schemas. We further proposed that semantic strain (previously viewed in terms of the extent to which a verb's selectional preferences are violated) indexes event similarity such that semantic strain increases as event similarity decreases. The less similar two events are, the further the verb must change its meaning to accommodate the noun. Thus, as strain increases, so should the degree of verb abstraction, proceeding from domain-specific to domain-general meaning components.

The minimal subtraction pattern of verb meaning change reported in Chapter 2 supports this prediction. Experiments 1A and 1B showed that as strain increased, so did the degree of verb meaning change, while Experiment 2 demonstrated the expected qualitative pattern of domain-specific to domain-general abstraction as strain increased. In addition, in Experiment 1 of this chapter, we observed a similar pattern in the paraphrases. For example, in *The violin stammered* → *The violin played quick, repeated notes*, the verb *stammered* retains more domain-specific components than it does in *The glider stammered* → *The tiny aircraft jerked around in the air*. Similarly, in *The sailboat strutted* → *The sailboat glided through the water with ease*, the verb *strutted* retains more of its literal meaning than it does in *The flute strutted* → *The flute played staggered notes at a constant pace*. In both examples, event similarity predicts verb abstraction; metaphors where the noun and verb events are from the same event domain (e.g., sound or motion) require less abstraction than those where they are from different domains.

Thus, in addition to verb mutability, we found consistent evidence for a minimal subtraction pattern of verb meaning change, supporting the prediction that event similarity determines the degree of verb (but not noun) abstraction—just as conceptual similarity determines base (but not target) abstraction in analogies more generally.

#### *4.1.3 Prediction 3: The verb event generates inferences about the nature of the noun event*

Analogies are often informative: they lead to the spontaneous generation of inferences about the target based on its mapping with the base. This is made possible by the two phases of structure-mapping: after the initial mapping (commonality-finding) phase, inference projection occurs based on relational structure that is connected to the mapping in the base but is not present in the target. Thus, the analogy *Misinformation is like a virus* might lead to the insight that one can inoculate against the spread of information, just as one can with viruses.

We argued that the same occurs in the case of verb metaphors. The paraphrases throughout this thesis show that interpretation depends on both the identification of commonalities between noun event and verb event, as well as the projection of further information to the noun event from the verb event. Participants rarely paraphrased items solely in terms of the common event structure between noun and verb event (e.g., *The van prowled* → *The large car drove*); instead, the noun event was almost always further specified by additional information from the verb (e.g., *The van prowled* → *The large car drove slowly along the street*). The retrace tasks in Chapter 2, Experiments 3A and 3B, provide further evidence by showing that, when given only the paraphrases, participants were able to identify what the original verb was from among a list of three choices at significantly above-chance levels (and even for high-strain items), indicating that information from the verb was indeed incorporated into the noun event.

*4.1.4 Prediction 4: Event similarity predicts processing time*

In analogies, the more similar two conceptual structures are, the faster they are to align. In the verb metaphor case, this means that the similarity between the noun event and the verb event in the metaphor should predict comprehension times. The results of Experiment 1 of this chapter support this prediction. We created metaphors where the similarity between the constituent noun and verb events was expected to be either high (within-domain items, e.g., *The violin stammered* or *The sailboat sprinted*) or low (cross-domain items, e.g., *The violin sprinted* or *The sailboat stammered*). Next, we obtained semantic strain ratings of each item using human raters in order to confirm this expectation and more directly and precisely quantity event similarity for each item.

As predicted, we found a significant relationship between strain and metaphor comprehension times such that as strain increased, so did the amount of time it took to interpret the metaphors (see Figure 15). Analyzing the strain ratings themselves further supported the hypothesized relationship between domain similarity and strain: overall, within-domain items (motion-sound and sound-motion) were significantly more strained than cross-domain items (motion-motion and sound-sound). The ratings also matched our expectations that this effect would be larger for items with musical instrument nouns than for items with vehicle nouns, but for an unexpected reason: within-domain vehicle items were significantly more strained than within-domain musical instrument items, while cross-domain instrument and vehicle items were equally strained (see Figure 16A).

In sum, the results of Experiment 1 provide additional converging, online evidence for the role of structure-mapping in verb metaphor processing. As expected under our model, event similarity predicted comprehension times, with metaphors comprising dissimilar events taking longer to interpret than those comprising similar events. This finding is consistent with previous work showing the same pattern for the alignment of other types of conceptual representations.

#### *4.1.5 Prediction 5: Prior activation of relevant event structure facilitates comprehension*

In analogies, the result of the structure-mapping process is that the mapped structure (i.e., the structure shared between the concepts being aligned) receives greater activation than unmapped structure. Applying this to our model of verb metaphor means that shared structure between the noun event and verb event representations should become activated by the mapping process to a greater extent than structure that is not shared between them. In other words, meaning components relevant to the meaning of the metaphor should become more activated than those irrelevant to the meaning of the metaphor.

In Experiment 2 of this chapter, we tested this prediction using a priming paradigm. If verb metaphor comprehension involves a mapping between event schemas that highlight structure relevant to the meaning of the metaphor, then prior activation of that structure should speed alignment and therefore comprehension times. We compared interpretation RTs for verb metaphors that were preceded by primes that activated event schemas that were either relevant (similar to) or irrelevant to the events expected to be activated by the metaphor itself. The typical events associated with each prime were collected in a separate norming study.

As predicted, metaphors preceded by relevant primes were significantly faster to interpret overall than those preceded by irrelevant primes. In addition, a coding task that matched the typical events associated with each prime (as obtained in the norming study) with the paraphrases showed that, as expected, the events associated with the relevant primes matched the events described by the paraphrases highly frequently (75.8% of the paraphrases), while events from the irrelevant primes rarely did (7.71% of the time). This pattern held (though to a slightly lesser extent) even when the metaphors were not preceded by the relevant prime, indicating that the relevant primes and the metaphors themselves were activating similar event schemas independently of one another (and that the primes were not exerting substantial influence on the subsequent metaphor's interpretation).

Thus, the RT results and coding results together support our hypothesized mechanism underlying the effect of prime type: consistent with a structure-mapping model, the relevant primes speeded comprehension times by virtue of activating event structure shared with mapping in the subsequent verb metaphor. Moreover, since the primes were all physical components of the subject noun of the metaphor (and were shown by the norming study to activate event structure related to the subject noun's event schema specifically), these findings provide further evidence for our claim that noun event schemas are an important component of verb metaphor processing.

#### *4.1.5.1 Mechanisms for noun event schema selection*

A key question for verb metaphor comprehension is how a noun event schema is selected during processing. Assuming that many nouns have multiple associated event schemas, what determines which one is selected for processing? An obvious candidate for priority is frequency of use. In Experiment 1, we used collocation tables to gauge which noun events were most dominant and likely to come to mind when encountering the noun.

An additional important factor that is predicted by our framework is the similarity between the noun event and verb event. In Experiment 2, we suggested that the key difference between the relevant and the irrelevant primes was that the relevant primes activated event schemas that were more similar to the subsequent verb than were those activated by the irrelevant primes. For example, in *The firetruck yelled*, the sound event activated by the relevant prime *siren* and the water-spraying event activated by the irrelevant prime *hose* are both common events that firetrucks participate in. The siren event, however, is more similar to (and therefore more relevant with) the event denoted by *yelled* than is the water-dispensing event activated by *hose*. If the target metaphor were instead *The firetruck vomited*, it seems likely that *hose* would become the relevant prime and *siren* the irrelevant one, because the event denoted by *vomited* is more similar to what a hose does than is the event denoted by *yelled*.[36] This account further predicts that the noun events selected for *The firetruck yelled* and *The firetruck vomited* will differ even when the metaphor is encountered in isolation (without primes), driven by the differing similarity of the noun events to the two verbs.[37]

This proposal therefore entails that (1) noun event schema selection depends jointly on the noun and the verb, and (2) which event is selected is influenced by the similarity between the noun event and verb event. The RT and coding findings from Experiments 1 and 2 support both these assertions. In Experiment 2, the irrelevant primes did not exert a strong influence on which event surfaced in the metaphoric paraphrases, indicating that foregrounding an alternative noun event—even one that is associated with the noun of the metaphor, like *handlebar* → *steers the*

[36] Of course, changing the verb should only change the noun event if an appropriate alternative is available.
[37] Assuming that this is the case, a further question would be whether noun events are selected serially or in parallel.

*bike*—is generally not sufficient to alter the interpretation. Only those events that are compatible with (i.e., similar to) the verb are typically integrated.

Further, as predicted, the relevant primes speeded comprehension times more than the irrelevant primes. In other words, consistent with the results of Experiment 1, the primes that activated events that were more similar to the verb (the relevant primes) were aligned faster than those that activated less-similar events (the irrelevant primes). This is consistent with the involvement of structure-mapping specifically in the comparison process.

#### *4.1.6 Summary of predictions*

The results presented in across the three chapters of this thesis point to structure-mapping as the process underlying verb metaphor comprehension. Viewing verb metaphors as a type of analogy that is understood by aligning noun and verb event structures fits remarkably well within the framework drawn from the broader literature on analogy. Our argument is that the experiments presented here provide converging evidence for this claim by supporting each of the five empirical predictions put forth. This includes evidence centering on the outcome of processing (i.e, the paraphrases), as well as the process itself as it unfolds (i.e., the online results). Our contention is that the analogy model of verb metaphor provides a single parsimonious framework under which all the phenomena described above are accounted for.

The structure-mapping model of verb metaphor has implications that go beyond the domain of verb metaphors. We conclude with a discussion of the theoretical implications for theories of metaphor processing and language evolution more broadly.

## 4.2 Theoretical implications

### *4.2.1 A unified framework for different types of metaphors*

Our claim that verb metaphors are analogies builds upon the work of Gentner and colleagues, who previously made the same argument for noun metaphors (e.g., *jealousy is like a tumor*; Gentner & Bowdle, 2008; Gentner et al., 2001; Gentner & Wolff, 1997; Wolff & Gentner, 2000, 2011). Across several studies, Gentner et al. provided empirical evidence for the role of structure-mapping in noun metaphor processing, showing that comprehension takes place via the two-phase model described here of structural alignment followed by inference projection (Gentner & Wolff, 1997; Gentner et al., 2000; Wolff & Gentner, 2000, 2011). That our findings here parallel this account invites the possibility that the same process may underly comprehension of both kinds of metaphors.

Having a single process model for both noun and verb metaphor is appealing, in that it readily allows metaphorical processing to extend beyond single local metaphors in isolation to extended metaphorical passages. Such passages often mix metaphoric uses of nouns and verbs in a single utterance—e.g., *The rockets came like* <u>*locusts,*</u> <u>*swarming*</u> *and settling in* <u>*blooms*</u> *of rosy smoke*.[38] Structure-mappings can be updated incrementally as new information is encountered (Forbus et al., 1994)—meaning that it is not necessary to know in advance whether the current input is part of a local or extended metaphoric mapping. Since language input is always encountered incrementally, this is a crucial feature for any model of figurative language processing. Thus,

[38] From Ray Bradbury's *The Martian Chronicles*.

structure-mapping provides a natural mechanism for the comprehension of extended metaphors (e.g., Gentner & Boronat, 1991; Keysar et al., 2000; Thibodeau & Durgin, 2008).

Understanding the role of events in verb metaphor comprehension also raises interesting questions about their possible involvement in metaphors comprising words from other syntactic classes. A natural starting point is nouns: if structure-mapping underlies both noun and verb metaphor processing, might event schemas be involved in noun metaphors as well? The obvious case involves nouns that denote events, which can clearly be used metaphorically (e.g., *The hearing was a circus; This dinner is a party in my mouth*). But events seem intuitively to be involved in some noun metaphors involving concrete nouns as well. For example, in *My heart is a chainsaw* (the title of a book by Stephen Graham Jones), *chainsaw* seems to call forth a "chain-sawing" event, which is mapped to the target *heart* to imply some sort of romantic destruction.[39] In the earlier example that began *The rockets came like locusts*, the noun *locusts* likely activates a *swarming* event, which is aligned with the *flying* event likely activated by *rockets*.[40]

Certainly, not all metaphors or analogies involve aligning event schemas. First, attributive metaphors clearly describe states (properties), rather than events—e.g., *The sun is like an orange; Pancakes are like nickels*. Second, not all relational metaphors seem to involve events. It seems wrong to classify the relational knowledge activated by analogies like *An atom is like the solar system*, *Misinformation is like a virus,* and *Analogy is the pinnacle of cognition* as

[39] Indeed, one phenomenon that suggests permeability between syntactic boundaries when it comes to event representations is the case of metaphoric novel denominal verbs—e.g., *You really Trojan-horsed that idea into his head* or *The man was tommy-gunning his screen with spittle, etc*.—where the events activated by the noun become verb metaphors as a result of being placed in the ditransitive syntactic frame. In the case of *Trojan horse*, the relevant meaning is a conventionalized metaphoric sense.

[40] Note also that in this example, *swarming* denotes flying that is occurring in a particular manner, and the metaphor suggests that the rockets were flying in a similar manner—that is, the interpretation involves elaborating on the target event based on information in the base event, just as we observed for verb metaphors in this thesis.

constituting events. Regardless, it may be that events play a role in metaphors beyond those involving verbs.

### *4.2.2 Language evolution and the Career of Metaphor*

As discussed in previous chapters, metaphor is widely considered to be an important vehicle for language change over time. Bowdle and Gentner's (2005) Career of Metaphor hypothesis outlines a single framework linking analogical reasoning, noun metaphor processing, and language evolution over time, in order to explain how novel metaphors encountered in day-to-day language may eventually lead to long-term changes in word meaning. The basic idea is that, just as work in analogy has shown that repeated comparisons of an analogical base lead to the induction of more domain-general abstractions of that base (Gentner & Hoyos, 2017; Gick & Holyoak, 1980; Kotovsky & Gentner, 1996), so too do the comparisons involved in noun metaphor processing. If the base in a noun metaphor is repeatedly used in a novel way that suggests a useful metaphoric abstraction (e.g., if the word *butterfly* is repeatedly used to signify something graceful, as in *A ballerina is like a butterfly*, *A figure skater is like a butterfly*, etc.), it may over time enter common circulation and become lexicalized as a new conventional metaphoric sense of the base—i.e., stored as is any other word sense in LTM. Thus, the Career of Metaphor points to the abstractions that result from repeated comparisons as one driver of language change and metaphoric polysemy.

Just as with our proposed model for verb metaphor comprehension, the Career of Metaphor therefore constitutes another application of the general principles of analogy to metaphoric language processing. If verb metaphors are also processed as analogical comparisons, it invites the possibility that the Career of Metaphor applies to verbs as well, perhaps through a very similar process. For example, if the verb *paced* is repeatedly used as the base in different verb metaphors in a way that suggests a common abstract event category (e.g., *The joke paced*, *The idea paced*, *The rumor paced*, etc.), *paced* may eventually gain a new conventional metaphoric sense, perhaps meaning something akin to the back-and-forth transmission of information.

Applying the Career of Metaphor to verbs provides a mechanism for linking a number of different findings regarding verbs and language evolution. As was found consistently in this thesis, verbs change meaning more than nouns in online sentence processing and do so primarily through analogical abstractions of the verb's literal meaning. Verbs' greater mutability and metaphoricity in sentence processing provides a possible explanation for findings that verbs also change meaning at a substantially higher rate than nouns (and adjectives) over historical time periods (Dubossarsky et al., 2016). Viewing these two phenomena from the opposite direction, this faster rate of meaning change over time raises the possibility that verbs are not only more likely to change meaning than nouns under strain, but also that these meaning extensions are happening more often in day-to-day language than they are for nouns. That is, verbs may have an accelerated Career of Metaphor: being more frequent participants in metaphoric comparisons may lead to greater meaning change over time compared to nouns.

There are several further reasons to believe this might be the case: verbs are more polysemous than nouns, controlling for frequency (Gentner, 1981; Miller & Fellbaum, 1991), and verb senses (as reflected in the dictionary) tend to be more metaphoric than noun senses (King, Gentner, & Mo, 2021). If, as is proposed by the Career of Metaphor, words can gain new senses as the result of repeated analogical comparisons, and if those comparisons occur more frequently for verbs than they do for other classes of words, one might expect these patterns.

In sum, just as structure-mapping provides the appealing possibility of unifying both noun and verb metaphor processing in a single framework, so too does extending Genter and Bowdle's Career of Metaphor from nouns to verbs. This suggests a broad analogical framework for language evolution for words of multiple syntactic classes. Indeed, an interesting further question is how the role of event representations specifically in metaphors and the Career of Metaphor might also interact. There are many conventional metaphoric expressions that are short event descriptions, such as *step up to the plate, circle the wagons, cast a wide net, go off the rails, spit in the wind, take a shot across the bow, kick the hornets' nest,* etc. If, as we have argued here, many metaphors involve analogical comparisons between literal event schemas, perhaps these events have a Career of their own.

## 5 Conclusion

In this chapter, we proposed a novel process account for verb metaphor comprehension: they are understood as analogical comparisons between an event activated by the noun and the event denoted by the verb. The results from all three chapters in this thesis support the predictions that arise from applying the analogical framework to verb metaphor. In addition, this model has important implications for our understanding of how metaphors of other types and syntactic classes are understood, as well as language evolution over time.

# Summary of Chapters

The goal of this dissertation was to characterize the nature of verb metaphor. To do so, we conducted a two-pronged investigation, examining the patterns of meaning change that result from processing, as well as the nature of the underlying process itself. Across three chapters, we found that verb metaphor comprehension results in a consistent and predictable pattern of verb meaning change. Our proposed process model—that verb metaphors are understood as analogical comparisons—provides a mechanism that accounts well for this behavior, while also fitting naturally into the well-attested structure-mapping framework from the broader literature on analogy. We conclude with a summary of the findings of this thesis, followed some brief parting thoughts on the topic.

## Chapter 1: The verb mutability effect

In Chapter 1, we laid the foundation for this project by investigating the verb mutability effect. In Experiments 1 and 2, we found that verbs are more likely to change their meanings under semantic strain than nouns, and that this effect increases with the degree of strain. In Experiment 2, we showed that verb mutability derives primarily from online adjustments to verb representations, rather than from sense selection. In the qualitative analysis in Experiment 3, we found that verb and noun meaning change follow qualitatively distinct patterns. Verb change was primarily analogical (metaphorical)—e.g., *The motor complained* → *The engine was making suspicious noises*. In contrast, noun meaning change occurred predominantly through the substitution of synonymous or taxonomically-related terms—e.g., *motor* → *engine; motor* → *machine*.

This chapter also introduced a novel methodology that used off-the-shelf word2vec vectors to assess the degree of meaning change that words undergo under paraphrase. In the double-paraphrase task in Experiment 3, we showed that this approach was well-calibrated to human judgments of similarity in this context by replicating previous pattern of word2vec scores from Experiment 2. This paradigm therefore provides a hands-off manner for tackling the potentially thorny issues inherent in using human raters to make these judgments—while also vastly reducing the time and labor required to assess the thousands of paraphrases analyzed here.

**Chapter 2: Minimal subtraction**

Chapter 2 built on the results of Chapter 1 by investigating more deeply the patterns of verb change that result from verb mutability. The goal was to provide a specific delineation of how verbs change meaning under strain. Building on Gentner & France's (1988) minimal subtraction hypothesis, we proposed and tested three principles of verb meaning change: (1) the degree of meaning change the verb undergoes increases progressively as the degree of strain increases, (2) the verb typically changes meaning only as far as necessary to resolve the strain, and (3) domain-specific meaning components are adjusted before more abstract, relational ones. We tested these predictions on a large number of verbs (54) drawn from a diverse set of semantic classes (manner of motion, manner of speaking, and bodily process), systematically straining each verb by pairing it with three classes of nouns (human, artifact, and abstract). In Experiments 1 and 2, we replicated the findings of Chapter 1 by finding that verbs changed meaning more than nouns, and—consistent with Predictions 1 and 2—this effect increased proportionally to the degree of strain.

In Experiment 2, a qualitative coding task showed that, consistent with Prediction 3, verb abstraction occurred in a domain-specific to domain-general direction, with the extent of abstraction increasing as a function of the paired noun type. When paired with a human noun, the verb typically retained its full literal meaning (e.g., *The husband paced* → *The wife's significant other walked back and forth*); when paired with an artifact noun, the verb typically retained some domain-specific meaning components while adjusting or discarding those that were incompatible with the noun (e.g., *The scooter paced* → *The child raced back and forth on their scooter*); when paired with an abstract noun, the verb's meaning became abstracted out of the physical domain entirely (e.g., *The rumor paced* → *The gossip went back and forth*). We further identified some interesting variation between verb classes in the patterns of verb abstraction: manner of speaking and bodily process verbs tended to abstract their meanings further under strain than did manner of motion verbs. We speculated that this may owe partly to the fact that the artifact nouns selected were all vehicles, which may have promoted a focus on the domain of motion, and to a general pattern in English of discussing (and possibly representing) abstract nouns in terms of spatial metaphors (e.g., rumors, wisdom, and religions *spread*). Finally, in Experiments 3A and 3B, we confirmed that the prior results involved operations over the verb's representations, rather than being solely driven by the selection of event schemas saliently associated with the noun.

**Chapter 3: Verb metaphors are analogies**

In Chapter 3, we synthesized the results of Chapters 1 and 2 and proposed a novel account of verb metaphor comprehension: that they are understood as analogical comparisons between an event activated by the noun (the target of the analogy) and the event denoted by the verb (the base of the analogy), processed via structure-mapping. In our model, the same two-phase structure-mapping process underlying analogical comparisons in general applies to the verb metaphor case. In the first phase, the noun event and verb event are placed into alignment and common structure between the two events is identified. (e.g., for *The wagon limped*, the *roll* event activated by *wagon* and the legged motion event denoted by *limped* both involve physical motion over land). After mapping is complete, in the second phase inferences are projected from the verb event to the noun event based on the established mapping (e.g., The physical motion event of *rolling* is inferred to occur in an impaired manner, just as the legged motion event denoted by *limped* does. Thus, verb metaphor comprehension parallels two of the key phenomena of analogy in general: they highlight commonalities between unlike concepts, and they can lead to the spontaneous generation of inferences about the target based on those commonalities with the base.

We showed that five key empirical predictions that arise from applying the structure-mapping analogical framework to the verb metaphor are supported by the experiments reported in this thesis: Prediction 1 (the verb mutability effect) was supported by the patterns of meaning change observed in all three chapters; Prediction 2 (minimal subtraction) was supported by the results of Chapters 2 and 3; Prediction 3 (inference projection from verb event to noun event) was also demonstrated by the paraphrases in all three chapters; Predictions 4 and 5 (event similarity predicts processing time, and prior activation of event structure speeds comprehension time) were supported by Experiments 1 and 2 in Chapter 3.

In each case, the experimental findings demonstrate a remarkable parallel between prior work in the analogical reasoning literature and the behavior of verb metaphors. The structure-mapping model provides a natural mechanism that accounts for that behavior, as verb mutability, minimal subtraction, and inference projection arise naturally out of the semantic-matching process of structure-mapping and the structure-mapping engine. The online methodologies of Experiments 1 and 2 from Chapter 3 provide converging evidence for the involvement of that process specifically.

Finally, viewing verb metaphors as analogies raises important connections with other areas of sentence processing and language evolution. Based on the empirical work of Gentner and colleagues supporting the theory that noun metaphors are also processed as analogies, our model here offers the potential of a unified account that explains both verb metaphor and noun metaphor processing (e.g., *Misinformation is like a virus*)—and perhaps that involving other syntactic classes as well. This idea fits well with other work supporting the role of structure-mapping in extended metaphors. Finally, the structure-mapping model of verb metaphor invites the possibility that Bowdle and Gentner's (2005) Career of Metaphor account of language evolution applies to verb as well; an idea that unites the findings of this dissertation (verbs' greater mutability compared nouns, and the role of analogical comparison in verb metaphor comprehension) with other work showing that verbs change meaning over time faster than do nouns, are more polysemous than nouns, and have senses in the dictionary are more metaphoric than nouns.

**Parting words: As verb metaphor goes, so goes cognition?**

The findings of this dissertation show that verbs have a remarkable ability to adapt their meanings to contexts that are radically different than those they typically occupy—something that requires seeing (sometimes quite abstract) similarities between disparate concepts. For this to happen does not require hours of reflection and analysis on the part of the comprehender; it often happens immediately, and even in the minds of stressed and tired undergraduates who are only interpreting these weird sentences in order to fulfill class requirements (we are very grateful for their willingness to nevertheless do so). It is worth pausing for a moment to appreciate that this is kind of amazing!

One cannot help but see verb metaphor as a microcosm of peoples' remarkable ability to identify commonalities between unlike concepts in general, and to realize new insights as a result—that is, their distinctive capacity for analogy. These are the insights drive our lives as humans—through the seemingly-mundane reasoning demanded of us by daily life and the need to communicate with others, to the awesome scientific discoveries that propel the species forward. We believe that further understanding of verb metaphor comprehension (as well as metaphor comprehension in general) will benefit our understanding of both language processing in particular and human cognition more broadly. We therefore hope that researchers will continue to make strides forward in this important area of work.

# References


Altmann, G. T. M. (1999). Thematic Role Assignment in Context. *Journal of Memory and Language*, *41*(1), 124–145. https://doi.org/10.1006/jmla.1999.2640

Altmann, G. T. M. (2017). Abstraction and generalization in statistical learning: Implications for the relationship between semantic types and episodic tokens. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160060. https://doi.org/10.1098/rstb.2016.0060

Altmann, G. T. M., & Ekves, Z. (2019). Events as intersecting object histories: A new theory of event representation. *Psychological Review*, *126*(6), 817–840. https://doi.org/10.1037/rev0000154

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264. https://doi.org/10.1016/S0010-0277(99)00059-1

Altmann, G. T. M., & Mirković, J. (2009). Incrementality and Prediction in Human Sentence Processing. *Cognitive Science*, *33*(4), 583–609. https://doi.org/10.1111/j.1551-6709.2009.01022.x

Asmuth, J., & Gentner, D. (2017). Relational categories are more mutable than entity categories. *Quarterly Journal of Experimental Psychology*, *70*(10), 2007–2025. https://doi.org/10.1080/17470218.2016.1219752

Baayen, R. H., & Milin, P. (2010). Analyzing Reaction Times. *International Journal of P Sychological Research*, *2*, 18.

Baker, C., Fillmore, C., and Lowe, J.  (1998). The Berkeley FrameNet project. *Proceedings of the COLING-ACL*, Montreal, Canada.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language, 68(3). https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *ArXiv:1506.04967 [Stat]*. http://arxiv.org/abs/1506.04967

Bicknell, K., Elman, J. L., Hare, M., McRae, K., & Kutas, M. (2010). Effects of event knowledge in processing verbal arguments. *Journal of Memory and Language*, *63*(4), 489–505. https://doi.org/10.1016/j.jml.2010.08.004

Blacoe, W., & Lapata, M. (2012). A comparison of vector-based representations for semantic composition. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 546–556. https://www.aclweb.org/anthology/D12-1050

Blank, G. D. (1988). Metaphors in the Lexicon. Metaphor and Symbolic Activity, 3(3), 21–36. https://doi.org/10.1207/s15327868ms0301_2

Bowdle, B. F., & Gentner, D. (1997). Informativity and Asymmetry in Comparisons. *Cognitive Psychology*, *34*(3), 244–286. https://doi.org/10.1006/cogp.1997.0670

Bowdle, B. F., & Gentner, D. (2005). The career of metaphor. *Psychological Review*, *112*(1), 193–216. https://doi.org/10.1037/0033-295X.112.1.193

Burgess, C., & Simpson, G. B. (1988). Cerebral hemispheric mechanisms in the retrieval of ambiguous word meanings. Brain and Language, 33(1), 86–103. https://doi.org/10.1016/0093-934X(88)90056-9

Burnett, R. C., & Gentner, D. (2000). *What is strolling a kind of.*

Camblin, C. C., Gordon, P. C., & Swaab, T. Y. (2007). The interplay of discourse congruence and lexical association during sentence processing: Evidence from ERPs and eye tracking. *Journal of Memory and Language*, *56*(1), 103–128. https://doi.org/10.1016/j.jml.2006.07.005

Cardillo, E. R., Schmidt, G. L., Kranjec, A., & Chatterjee, A. (2010). Stimulus design is an obstacle course: 560 matched literal and metaphorical sentences for testing neural hypotheses about metaphor. *Behavior Research Methods*, *42*(3), 651–664. https://doi.org/10.3758/BRM.42.3.651

Cardillo, E. R., Watson, C. E., Schmidt, G. L., Kranjec, A., & Chatterjee, A. (2012). From novel to familiar: Tuning the brain for metaphors. *NeuroImage*, *59*(4), 3212–3221. https://doi.org/10.1016/j.neuroimage.2011.11.079

Cardillo, E. R., Watson, C., & Chatterjee, A. (2017). Stimulus needs are a moving target: 240 additional matched literal and metaphorical sentences for testing neural hypotheses about metaphor. *Behavior Research Methods*, *49*(2), 471–483.

Chambers, N. (n.d.). *Event Schema Induction with a Probabilistic Entity-Driven Model*. 11.

Chatterjee, A. (2008). The neural organization of spatial thought and language. *Seminars in Speech and Language*, *29*(03), 226–238. https://doi.org/10.1055/s-0028-1082886

Chatterjee, A. (2010). Disembodying cognition. *Language and cognition*, *2*(1), 79-116.

Chen, E., Widick, P., & Chatterjee, A. (2008). Functional–anatomical organization of predicate metaphor processing. *Brain and Language*, *107*(3), 194–202. https://doi.org/10.1016/j.bandl.2008.06.007

Chiappe, D. L., & Kennedy, J. M. (2001). Literal bases for metaphor and simile. *Metaphor and Symbol*, *16*(3–4), 249–276. https://doi.org/10.1080/10926488.2001.9678897

Chiappe, D., Kennedy, J. M., & Smykowski, T. (2003). Reversibility, Aptness, and the Conventionality of Metaphors and Similes. *Metaphor and Symbol*, *18*(2), 85–105. https://doi.org/10.1207/S15327868MS1802_2

Gentner, D., & Christie, S. (2010). Mutual bootstrapping between language and analogical processing. Language and Cognition, 2(02), 261–283. https://doi.org/10.1515/langcog.2010.011

Clark, E. V., & Clark, H. H. (1979). When nouns surface as verbs. *Language*, *55*(4), 767–811. https://doi.org/10.2307/412745

Clark, H. H. (1966). The prediction of recall patterns in simple active sentences. *Journal of Verbal Learning and Verbal Behavior*, *5*(2), 99–106. https://doi.org/10.1016/S0022-5371(66)80001-4

Clark, H. H., & Gerrig, R. J. (1983). Understanding old words with new meanings. *Journal of Verbal Learning and Verbal Behavior*, *22*(5), 591–608.

Clausner, T. C., & Croft, W. (1997). Productivity and schematicity in metaphors. *Cognitive Science*, *21*(3), 247–282. https://doi.org/10.1207/s15516709cog2103_1

Croft, W. (2000). Parts of speech as language universals and as language-particular categories. *Empirical Approaches to Language Typology*, 65–102.

Croft, W. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press on Demand.

Croft, W. (2002). The role of domains in the interpretation of metaphors and metonymies. *Metaphor and Metonymy in Comparison and Contrast*, 161–205.

Davies, M. (2018). The iWeb corpus. *Online At< Https://Www. English-Corpora. Org/IWeb/>(Accessed April 1, 2021)*.

Davis, C. P., Paz-Alonso, P. M., Altmann, G., & Yee, E. (2019). *Abstract concepts and the suppression of arbitrary episodic context* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/nfy4s

Desai, R. H., Binder, J. R., Conant, L. L., Mano, Q. R., & Seidenberg, M. S. (2011). The neural career of sensory-motor metaphors. *Journal of Cognitive Neuroscience*, *23*(9), 2376–2386. https://doi.org/10.1162/jocn.2010.21596

Desai, R. H., Conant, L. L., Binder, J. R., Park, H., & Seidenberg, M. S. (2013). A piece of the action: Modulation of sensory-motor regions by action idioms and metaphors. *NeuroImage*, *83*, 862–869. https://doi.org/10.1016/j.neuroimage.2013.07.044

Dirven, R. (1985). Metaphor as a basic means for extending the lexicon. *The Ubiquity of Metaphor*, 85–119.

Dorst, A. G. (2011). Personification in discourse: Linguistic forms, conceptual structures and communicative functions. *Language and Literature*, *20*(2), 113–135. https://doi.org/10.1177/0963947010395522

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, *67*(3), 547–619. https://doi.org/10.1353/lan.1991.0021

Dubossarsky, H., Weinshall, D., & Grossman, E. (2016). Verbs change more than nouns: A bottom-up computational approach to semantic change. *Lingue e Linguaggio*, *15*(1), 7–28.

Earles, J. L., & Kersten, A. W. (2000). Adult age differences in memory for verbs and nouns. *Aging, Neuropsychology, and Cognition*, *7*(2), 130–139. https://doi.org/10.1076/1382-5585(200006)7:2;1-U;FT130

Earles, J. L., & Kersten, A. W. (2017). Why are verbs so hard to remember? Effects of semantic context on memory for verbs and nouns. *Cognitive Science*, *41*, 780–807. https://doi.org/10.1111/cogs.12374

Earles, J. L., Kersten, A. W., Turner, J. M., & McMullen, J. (1999). Influences of age, performance, and item relatedness on verbatim and gist recall of verb-noun pairs. *The Journal of General Psychology*, *126*(1), 97–110.

Elman, J. L. (2009). On the Meaning of Words and Dinosaur Bones: Lexical Knowledge Without a Lexicon. *Cognitive Science*, *33*(4), 547–582. https://doi.org/10.1111/j.1551-6709.2009.01023.x

Elman, J. L. (2011). Lexical knowledge without a lexicon? *The Mental Lexicon*, *6*(1), 1–33. https://doi.org/10.1075/ml.6.1.01elm

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). *The Structure-Mapping Engine: Algorithm and Examples*. 62.

Fauconnier, G., & Turner, M. (1998). Conceptual integration networks. *Cognitive Science, 22*(2), 133-187.

Federmeier, K. D., & Kutas, M. (1999). A Rose by Any Other Name: Long-Term Memory Structure and Sentence Processing. *Journal of Memory and Language*, *41*(4), 469–495. https://doi.org/10.1006/jmla.1999.2660

Fellbaum, C. (1999). English verbs as a semantic net. *International Journal of Lexicography*, *4*(3), 278–301.

Ferretti, T. R., Kutas, M., & McRae, K. (2007). Verb aspect and the activation of event knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(1), 182–196. https://doi.org/10.1037/0278-7393.33.1.182

Ferretti, T. R., McRae, K., & Hatherell, A. (2001). Integrating Verbs, Situation Schemas, and Thematic Role Concepts. *Journal of Memory and Language*, *44*(4), 516–547. https://doi.org/10.1006/jmla.2000.2728

Fillmore, C. J. (1971). Verbs of judging: An exercise in semantic description. In C. J. Fillmore & D. T. Langendoen (Eds.), *Studies in linguistic semantics* (pp. 273-296). New York: Holt, Rinehart, & Winston.

Finley, G., Farmer, S., & Pakhomov, S. (2017). What analogies reveal about word vectors and their compositionality. *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, 1–11. https://doi.org/10.18653/v1/S17-1001

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, *25*(2–3), 285–307. https://doi.org/10.1080/01638539809545029

Forbus, K. D., Ferguson, R. W., & Gentner, D. (1994). *Incremental Structure-Mapping*. 6.

Forbus, K. D., Ferguson, R. W., Lovett, A., & Gentner, D. (2017). Extending SME to Handle Large-Scale Cognitive Modeling. *Cognitive Science*, *41*(5), 1152–1201. https://doi.org/10.1111/cogs.12377

Frisson, S., & Pickering, M. J. (2007). The processing of familiar and novel senses of a word: Why reading Dickens is easy but reading Needham can be hard. *Language and Cognitive Processes*, *22*(4), 595–613. https://doi.org/10.1080/01690960601017013

Frost, R., & Bentin, S. (1992). Processing phonological and semantic ambiguity: Evidence from semantic priming at different SOAs. Journal of Experimental Psychology: Learning, Memory, and Cognition, 18(1), 58–68.

Garrido Rodriguez, G., Norcliffe, E., Brown, P., Huettig, F., & Levinson, S. C. (2023). Anticipatory Processing in a Verb-Initial Mayan Language: Eye-Tracking Evidence During Sentence Comprehension in Tseltal. *Cognitive Science*, *47*(1), e13292. https://doi.org/10.1111/cogs.13219

Gentner, D. (1981). Some interesting differences between nouns and verbs. *Cognition and Brain Theory*, *4*, 161–178.

Gentner, D. (1987). *Mechanisms of analogical learning.* ILLINOIS UNIV AT URBANA DEPT OF COMPUTER SCIENCE.

Gentner, D. (2010). Bootstrapping the Mind: Analogical Processes and Symbol Systems. Cognitive Science, 34(5), 752–775. https://doi.org/10.1111/j.1551-6709.2010.01114.x

Gentner, D., & Asmuth, J. (2017). Metaphoric extension, relational categories, and abstraction. *Language, Cognition and Neuroscience*, *0*(0), 1–10. https://doi.org/10.1080/23273798.2017.1410560

Gentner, D., & Boronat, C. B. (1991). Metaphors are (sometimes) processed as generative domain-mappings. In symposium on Metaphor and Conceptual Change, Meeting of the Cognitive Science Society, Chicago.

Gentner, D., & Christie, S. (2010). Mutual bootstrapping between language and analogical processing. Language and Cognition, 2(02), 261–283. https://doi.org/10.1515/langcog.2010.011

Gentner, D., & France, I. M. (1988). The verb mutability effect: Studies of the combinatorial semantics of nouns and verbs. In S. L. Small, G. W. Cottrell, & M. K. Tanenhaus (Eds.), *Lexical Ambiguity Resolution* (pp. 343–382). Morgan Kaufmann. https://doi.org/10.1016/B978-0-08-051013-2.50018-5

Gentner, D., & France, I. M. (1988). The verb mutability effect: Studies of the combinatorial semantics of nouns and verbs. In S. L. Small, G. W. Cottrell, & M. K. Tanenhaus (Eds.), *Lexical Ambiguity Resolution* (pp. 343–382). Morgan Kaufmann. https://doi.org/10.1016/B978-0-08-051013-2.50018-5

Gentner, D., & Gentner, D. R. (1983). Flowing waters or teeming crowds: Mental models of electricity. *Mental Models*, *99*, 129.

Gentner, D., & Hoyos, C. (2017). Analogy and Abstraction. *Topics in Cognitive Science*, *9*(3), 672–693. https://doi.org/10.1111/tops.12278

Gentner, D., & Kurtz, K. J. (2005). Relational categories. In W. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. Wolff (Eds.), *Categorization inside and outside the laboratory: Essays in honor of Douglas L. Medin.* (pp. 151–175). American Psychological Association. https://doi.org/10.1037/11156-009

Gentner, D., & Kurtz, K. J. (2006). Relations, Objects, and the Composition of Analogies. *Cognitive Science*, *30*(4), 609–642. https://doi.org/10.1207/s15516709cog0000_60

Gentner, D., & Markman, A. B. (1993). *Analogy—Watershed or Waterloo? Structural alignment and the development of connectionist models of analogy*. 8.

Gentner, D., & Markman, A. B. (1994). Structural Alignment in Comparison: No Difference Without Similarity. *Psychological Science*, *5*(3), 152–158. https://doi.org/10.1111/j.1467-9280.1994.tb00652.x

Gentner, D., & Markman, A. B. (1995). *Similarity is like analogy: Structural alignment in comparison*.

Gentner, D., & Markman, A. B. (1997). Structure Mapping in Analogy and Similarity. *American Psychologist*, 12.

Gentner, D., & Namy, L. L. (1999). Comparison in the Development of Categories. Cognitive Development, 14(4), 487–513. https://doi.org/10.1016/S0885-2014(99)00016-7

Gentner, D., & Wolff, P. (1997). Alignment in the processing of metaphor. *Journal of Memory and Language*, *37*(3), 331–355. https://doi.org/10.1006/jmla.1997.2527

Gentner, D., & Wolff, P. (2000). Metaphor and knowledge change. *Cognitive Dynamics: Conceptual Change in Humans and Machines*, 295–342.

Gentner, D., Bowdle, B., Wolff, P., & Boronat, C. (2001). Metaphor is like analogy. In *The analogical mind: Perspectives from cognitive science* (pp. 199–253). MIT Press. https://pdfs.semanticscholar.org/d6f2/945bf8f21be0f463436fea2959e16ac679d0.pdfhttps://pdfs.semanticscholar.org/d6f2/945bf8f21be0f463436fea2959e16ac679d0.pdf

Gentner, D., Jai, A., Gentner, D., & Boronat, C. B. (1991). Metaphors are (sometimes) processed as generative domainmappings. Paper presented at the symposium on. *Metaphor and Conceptual Change, Meeting of the Cognitive Science Society*.

Gentner, D., Loewenstein, J., Thompson, L., & Forbus, K. D. (2009). Reviving Inert Knowledge: Analogical Abstraction Supports Relational Retrieval of Past Events. *Cognitive Science*, *33*(8), 1343–1382. https://doi.org/10.1111/j.1551-6709.2009.01070.x

Gerrig, R. J. (1989). The time course of sense creation. *Memory & Cognition*, *17*(2), 194–207. https://doi.org/10.3758/BF03197069

Gerrig, R. J., & Bortfeld, H. (1999). Sense creation in and out of discourse contexts. *Journal of Memory and Language*, *41*(4), 457–468.

Gerz, D., Vulić, I., Hill, F., Reichart, R., & Korhonen, A. (2016). Simverb-3500: A large-scale evaluation set of verb similarity. *Proceedings of EMNLP*, 2173–2182. http://arxiv.org/abs/1608.00869

Gibbs Jr, R. (1992). Categorization and metaphor understanding. *Psychological Review*, *99*(3), 572–577.

Gibbs, R. W. (2006). Metaphor interpretation as embodied simulation. *Mind & Language*, *21*(3), 434–458. https://doi.org/10.1111/j.1468-0017.2006.00285.x

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, *12*(3), 306–355. https://doi.org/10.1016/0010-0285(80)90013-4

Gildea, P., & Glucksberg, S. (1983). On understanding metaphor: The role of context. *Journal of Verbal Learning and Verbal Behavior*, *22*(5), 577–590. https://doi.org/10.1016/S0022-5371(83)90355-9

Giora, R. (1997). Understanding figurative and literal language: The graded salience hypothesis. *Cognitive Linguistics (Includes Cognitive Linguistic Bibliography)*, *8*(3), 183–206.

Gleitman, L. R., Gleitman, H., Miller, C., & Ostrin, R. (1996). Similar, and similar concepts. Cognition, 58(3), 321–376.

Glucksberg, S., & Haught, C. (2006). On the Relation Between Metaphor and Simile: When Comparison Fails. *Mind & Language*, *21*(3), 360–378. https://doi.org/10.1111/j.1468-0017.2006.00282.x

Glucksberg, S., & Keysar, B. (1990). Understanding metaphorical comparisons: Beyond similarity. *Psychological Review*, *97*(1), 3.

Glucksberg, S., & Keysar, B. (1990). Understanding metaphorical comparisons: Beyond similarity. *Psychological Review*, *97*(1), 3.

Glucksberg, S., Gildea, P., & Bookin, H. B. (1982). On understanding nonliteral speech: Can people ignore metaphors? *Journal of Verbal Learning and Verbal Behavior*, *21*(1), 85–98. https://doi.org/10.1016/S0022-5371(82)90467-4

Glucksberg, S., McGlone, M. S., & Manfredi, D. (1997). Property attribution in metaphor comprehension. *Journal of Memory and Language*, *36*(1), 50–67.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Goldstone, R. L. (1994). Similarity, interactive activation, and mapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(1), 3.

Goldstone, R. L., Medin, D. L., & Gentner, D. (1991). Relational similarity and the nonindependence of features in similarity judgments. *Cognitive Psychology*, *23*(2), 222–262. https://doi.org/10.1016/0010-0285(91)90010-L

Goldwater, M. B., & Markman, A. B. (2011). Categorizing entities by common role. *Psychonomic Bulletin & Review*, *18*(2), 406–413.

Goldwater, M.B., Markman, A.B., & Stilwell, C.H. (2011). The empirical case for role-governed categories. *Cognition*, 118, 359-376.

Graesser, A. C., Hopkinson, P. L., & Schmid, C. (1987). Differences in interconcept organization between nouns and verbs. *Journal of Memory and Language*, *26*(2), 242–253.

Graesser, A. C., Robertson, S. P., & Anderson, P. A. (1981). Incorporating inferences in narrative representations: A study of how and why. *Cognitive Psychology*, *13*(1), 1–26. https://doi.org/10.1016/0010-0285(81)90002-5

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. Methods in Ecology and Evolution, 7(4), 493–498.

Günther, F., Dudschig, C., & Kaup, B. (2016). Latent semantic analysis cosines as a cognitive similarity measure: Evidence from priming studies. *Quarterly Journal of Experimental Psychology*, *69*(4), 626–653. https://doi.org/10.1080/17470218.2015.1038280

Hare, M., Elman, J. L., Tabaczynski, T., & McRae, K. (2009). The Wind Chilled the Spectators, but the Wine Just Chilled: Sense, Structure, and Sentence Comprehension. *Cognitive Science*, *33*(4), 610–628. https://doi.org/10.1111/j.1551-6709.2009.01027.x

Heine, B. (1997). *Cognitive foundations of grammar*. Oxford University Press.

Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665–695. https://doi.org/10.1162/COLI_a_00237

Holyoak, K. J., & Stamenković, D. (2018). Metaphor comprehension: A critical review of theories and evidence. *Psychological Bulletin*, *144*(6), 641–671. https://doi.org/10.1037/bul0000145

Hopper, P. J., & Traugott, E. C. (2003). *Grammaticalization*. Cambridge University Press.

Horowitz, L. M., & Prytulak, L. S. (1969). Redintegrative memory. *Psychological Review, 76*(6), 519–531. https://doi.org/10.1037/h0028139

Huttenlocher, J., & Lui, F. (1979). The semantic organization of some simple nouns and verbs. *Journal of Verbal Learning and Verbal Behavior*, *18*(2), 141–162. https://doi.org/10.1016/S0022-5371(79)90091-4

Jackendoff, R. (1983). *Semantics and cognition.* Cambridge, MA: MIT Press.

Jameson, J., & Gentner, D. (2003). Mundane comparisons can facilitate relational understanding. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *25*(25).

Jamrozik, A., McQuire, M., Cardillo, E. R., & Chatterjee, A. (2016). Metaphor: Bridging embodiment to abstraction. *Psychonomic Bulletin & Review*, *23*(4), 1080–1089. https://doi.org/10.3758/s13423-015-0861-0

Jamrozik, A., Sagi, E., Goldwater, M., & Gentner, D. (2013). Relational words have high metaphoric potential. *Proceedings of the First Workshop on Metaphor in NLP*, 21–26.

Jones, L. L., & Estes, Z. (2006). Roosters, robins, and alarm clocks: Aptness and conventionality in metaphor comprehension. *Journal of Memory and Language*, *55*(1), 18–32. https://doi.org/10.1016/j.jml.2006.02.004

Joseph, B. D., Hock, H. H., & Joseph, B. D. (1996). *Language history, language change, and language relationship: An introduction to historical and comparative linguistics* (Vol. 93). Walter de Gruyter.

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, *49*(1), 133–156. https://doi.org/10.1016/S0749-596X(03)00023-8

Kang, X., Eerland, A., Joergensen, G. H., Zwaan, R. A., & Altmann, G. T. M. (2020). The influence of state change on object representations in language comprehension. *Memory & Cognition*, *48*(3), 390–399. https://doi.org/10.3758/s13421-019-00977-7

Kaschak, M. P., & Glenberg, A. M. (2000). Constructing meaning: The role of affordances and grammatical constructions in sentence comprehension. *Journal of Memory and Language*, *43*(3), 508–529. https://doi.org/10.1006/jmla.2000.2705

Katz, A. N. (1989). On choosing the vehicles of metaphors: Referential concreteness, semantic distances, and individual differences. *Journal of Memory and Language*, *28*(4), 486–499. https://doi.org/10.1016/0749-596X(89)90023-5

Katz, J. J., & Fodor, J. A. (1963). The structure of a semantic theory. *Language*, *39*(2), 170–210. https://doi.org/10.2307/411200

Kersten, A. W., & Earles, J. L. (2004). Semantic context influences memory for verbs more than memory for nouns. *Memory & Cognition*, *32*(2), 198–211. https://doi.org/10.3758/BF03196852

Keysar, B., Shen, Y., Glucksberg, S., & Horton, W. S. (2000). Conventional language: How metaphorical is it? *Journal of Memory and Language*, *43*(4), 576–593. https://doi.org/10.1006/jmla.2000.2711

King, D., Gentner, D., & Mo, F. (2021, July). *Verbs are More Metaphoric than Nouns: Evidence from the Lexicon* [Poster]. 43rd Annual Conference of the Cognitive Science Society, Vienna, Austria.

King, D., Gentner, D., & Mo, F. (2022). *Qualitative differences between noun and verb senses in the lexicon*. Manuscript in preparation.

Kintsch, W. (2001). *Predication*. https://ac.els-cdn.com/S0364021301000349/1-s2.0-S0364021301000349-main.pdf?_tid=d590e19b-cb20-4e8a-953d-5bacb46888f3&acdnat=1529949499_2529123fec3c8b685d81faacb7eda6b0

Kotovsky, L., & Gentner, D. (1996). *Comparison and Categorization in the Development of Relational Similarity*. 27.

Krennmayr, T. (2011). *Metaphor in newspapers* (Vol. 276). LOT Dissertation Series.

Lakoff, G., & Johnson, M. (1980). Conceptual metaphor in everyday language. *The Journal of Philosophy*, *77*(8), 453–486. JSTOR. https://doi.org/10.2307/2025464

Lakoff, G., & Johnson, M. (2008). *Metaphors we live by*. University of Chicago press.

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284. https://doi.org/10.1080/01638539809545028

Langacker, R. W. (1987). Nouns and verbs. *Language*, *63*(1), 53–94. https://doi.org/10.2307/415384

Langacker, R. W. (2008). *Cognitive grammar: A basic introduction*. Oxford University Press.

Lazic, S. E. (2008). Why we should use simpler models if the data allow this: Relevance for ANOVA designs in experimental biology. BMC Physiology, 8, 16. https://doi.org/10.1186/1472-6793-8-16

Lenat, D. B., & Guha, R. V. (1989). *Building large knowledge-based systems; representation and inference in the Cyc project*. Addison-Wesley Longman Publishing Co., Inc.

Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics*, *4*(1), 151–171. https://doi.org/10.1146/annurev-linguistics-030514-125254

Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago, IL: University of Chicago Press.

Levy, O., & Goldberg, Y. (2014). Dependency-based word embeddings. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 302–308. https://doi.org/10.3115/v1/P14-2050

Lin, C.-J. C., & Chen, Y.-R. (2015a). Exhaustive semantic activation for reading ambiguous verbs in Chinese sentences. Lingua Sinica, 1(1), 7. https://doi.org/10.1186/s40655-015-0008-2

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*(4), 1494–1502. https://doi.org/10.3758/s13428-016-0809-y



Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*(4), 1494–1502. https://doi.org/10.3758/s13428-016-0809-y



Lupyan, G., & Lewis, M. (2019). From words-as-mappings to words-as-cues: The role of language in semantic knowledge. *Language, Cognition and Neuroscience*, *34*(10), 1319–1337. https://doi.org/10.1080/23273798.2017.1404114

MacAvaney, S., & Zeldes, A. (2018). A deeper look into dependency-based word embeddings. *ArXiv:1804.05972 [Cs]*. http://arxiv.org/abs/1804.05972

MacGregor, L. J., Bouwsema, J., & Klepousniotou, E. (2015). Sustained meaning activation for polysemous but not homonymous words: Evidence from EEG. Neuropsychologia, 68, 126–138. https://doi.org/10.1016/j.neuropsychologia.2015.01.008

Markman, A. B., & Gentner, D. (1993). Structural Alignment during Similarity Comparisons. *Cognitive Psychology*, *25*(4), 431–467. https://doi.org/10.1006/cogp.1993.1011

Markman, A. B., & Gentner, D. (1996). Commonalities and differences in similarity comparisons. *Memory & Cognition*, *24*(2), 235–249.

Markman, A. B., & Stilwell, C. H. (2001). Role-governed categories. *Journal of Experimental & Theoretical Artificial Intelligence*, *13*(4), 329–358.

Matlen, B. J., Gentner, D., & Franconeri, S. L. (2020). Spatial alignment facilitates visual comparison. *Journal of Experimental Psychology: Human Perception and Performance*, 46(5), 443.

Matsuki, K., Chow, T., Hare, M., Elman, J. L., Scheepers, C., & McRae, K. (2011). Event-based Plausibility Immediately Influences On-line Language Comprehension. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *37*(4), 913–934. https://doi.org/10.1037/a0022964

McCawley, J. D. (1972). A program for logic. *Semantics of Natural Language*, 498–544.

McRae, K., & Matsuki, K. (2009). People Use their Knowledge of Common Events to Understand Language, and Do So as Quickly as Possible. *Language and Linguistics Compass*, *3*(6), 1417–1429. https://doi.org/10.1111/j.1749-818X.2009.00174.x

McRae, K., Brown, K. S., & Elman, J. L. (2021). Prediction-Based Learning and Processing of Event Knowledge. *Topics in Cognitive Science*, *13*(1), 206–223. https://doi.org/10.1111/tops.12482

McRae, K., Ferretti, T., & Amyote, L. (1997). Thematic Roles as Verb-specific Concepts. *Language and Cognitive Processes*, *12*(2–3), 137–176. https://doi.org/10.1080/016909697386835

McRae, K., Hare, M., Elman, J. L., & Ferretti, T. (2005). A basis for generating expectancies for verbs from nouns. *Memory & Cognition*, *33*(7), 1174–1184. https://doi.org/10.3758/BF03193221

Meade, G., & Coch, D. (2017). Word-pair priming with biased homonyms: N400 and LPC effects. Journal of Neurolinguistics, 41, 24–37. https://doi.org/10.1016/j.jneuroling.2016.09.002

Medin, D. L., Goldstone, R. L., & Gentner, D. (1990). Similarity Involving Attributes and Relations: Judgments of Similarity and Difference Are Not Inverses. *Psychological Science*, *1*(1), 64–69. https://doi.org/10.1111/j.1467-9280.1990.tb00069.x

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, *100*(2), 254.

Metusalem, R., Kutas, M., Urbach, T. P., Hare, M., McRae, K., & Elman, J. L. (2012). Generalized event knowledge activation during online sentence comprehension. *Journal of Memory and Language*, *66*(4), 545–567. https://doi.org/10.1016/j.jml.2012.01.001

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *ArXiv:1301.3781 [Cs]*. http://arxiv.org/abs/1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *ArXiv:1310.4546 [Cs, Stat]*. http://arxiv.org/abs/1310.4546

Miller, G. A. (1993). Images and models, similes and metaphors. Metaphor and Thought, 2, 2–25.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41.

Miller, G. A., & Fellbaum, C. (1991). Semantic networks of english. *Cognition*, *41*(1), 197–229. https://doi.org/10.1016/0010-0277(91)90036-4

Mirković, J., & Altmann, G. T. M. (2019). Unfolding meaning in context: The dynamics of conceptual similarity. *Cognition*, *183*, 19–43. https://doi.org/10.1016/j.cognition.2018.10.018

Moss, H. E., Ostrin, R. K., Tyler, L. K., & Marslen-Wilson, W. D. (1995). Accessing different types of lexical semantic information: Evidence from priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 863.

Murphy, G. L. (1990). Noun phrase interpretation and conceptual combination. *Journal of Memory and Language*, *29*(3), 259–288. https://doi.org/10.1016/0749-596X(90)90001-G

Nakamura, T., Sakamoto, M., & Utsumi, A. (2010). *The Role of Event Knowledge in Comprehending Synesthetic Metaphors*. 7.

Narayanan, S. S. (1997). *Knowledge-based Action Representations for Metaphor and Aspect (KARMA)*.

Narrog, H., & Heine, B. (2021). *Grammaticalization*. Oxford University Press.

Nunberg, G. (1979). The non-uniqueness of semantic solutions: Polysemy. *Linguistics and Philosophy*, *3*(2), 143–184. https://doi.org/10.1007/BF00126509

Nunberg, G. (1995). Transfers of meaning. *Journal of Semantics*, *12*(2), 109–132. https://doi.org/10.1093/jos/12.2.109

Nunberg, G. D. (1978). *The pragmatics of reference.* City University of New York.

Ortony, A. (1979). Beyond literal similarity. *Psychological Review*, *86*(3), 161.

Otten, M., & Van Berkum, J. J. A. (2007). What makes a discourse constraining? Comparing the effects of discourse message and scenario fit on the discourse-dependent N400 effect. *Brain Research*, *1153*, 166–177. https://doi.org/10.1016/j.brainres.2007.03.058

Pavlicic, T., & Markman, A. (1997). The structure of the verb lexicon: Evidence from a structural alignment approach to similarity. *The Proceedings of the 19th Annual Conference of the Cognitive Sciencety. Stanford, CA: Lawrence Eribaum Associates*, 590–595.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Pereira, F., Gershman, S., Ritter, S., & Botvinick, M. (2016). A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive Neuropsychology*, *33*(3–4), 175–190. https://doi.org/10.1080/02643294.2016.1176907

Pritchard, T. (2019). Analogical cognition: An insight into word meaning. *Review of Philosophy and Psychology*, *10*(3), 587–607. https://doi.org/10.1007/s13164-018-0419-y

Pustejovsky, J. (1995). *The generative lexicon*. MIT Press.

Qiu, M., Castro, N., & Johns, B. (2021). Structural comparisons of noun and verb networks in the mental lexicon. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 8.

Raposo, A., Moss, H. E., Stamatakis, E. A., & Tyler, L. K. (2009). Modulation of motor and premotor cortices by actions, action words and action sentences. *Neuropsychologia*, *47*(2), 388–396. https://doi.org/10.1016/j.neuropsychologia.2008.09.017.

Rapp, D. N., & Gerrig, R. J. (1999). Eponymous verb phrases and ambiguity resolution. *Memory & Cognition*, *27*(4), 612–618. https://doi.org/10.3758/BF03211555

Rehder, B., & Ross, B. H. (2001). Abstract coherent categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(5), 1261.

Reyna, V. (1980, November). *When words collide: Interpretation of selectionally opposed nouns and verbs*. Sloan Symposium on Metaphor and Thought, Chicago, IL.

Rimell, L., Maillard, J., Polajnar, T., & Clark, S. (2016). RELPRON: A relative clause evaluation data set for compositional distributional semantics. *Computational Linguistics*, *42*(4), 661–701. https://doi.org/10.1162/COLI_a_00263

Rips, L. J., & Conrad, F. G. (1989). Folk psychology of mental activities. *Psychological Review*, *96*(2), 187.

Rips, L. J., & Hespos, S. J. (2019). Concepts of objects and substances in language. *Psychonomic Bulletin & Review*, *26*(4), 1238–1256. https://doi.org/10.3758/s13423-019-01613-w

Ronderos, C. R., Guerra, E., & Knoeferle, P. (2021). The Role of Literal Features During Processing of Novel Verbal Metaphors. *Frontiers in Psychology*, *11*, 3899. https://doi.org/10.3389/fpsyg.2020.556624

Royston, P., Altman, D. G., & Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: A bad idea. Statistics in Medicine, 25(1), 127–141. https://doi.org/10.1002/sim.2331

Sagi, E. (2019). Taming big data: Applying the experimental method to naturalistic data sets. *Behavior Research Methods*, *51*(4), 1619–1635. https://doi.org/10.3758/s13428-018-1185-6

Sagi, E., Gentner, D., & Lovett, A. (2012). What Difference Reveals About Similarity. Cognitive Science, 36(6), 1019–1050. https://doi.org/10.1111/j.1551-6709.2012.01250.x

Saygin, A. P., McCullough, S., Alac, M., & Emmorey, K. (2010). Modulation of bold response in motion-sensitive lateral temporal cortex by real and fictive motion sentences. *Journal of Cognitive Neuroscience*, *22*(11), 2480–2490. https://doi.org/10.1162/jocn.2009.21388

Shen, Y. (1989). Symmetric and asymmetric comparisons. *Poetics*, *18*(6), 517–536. https://doi.org/10.1016/0304-422X(89)90010-7

Simmons, S., & Estes, Z. (2006). Using latent semantic analysis to estimate similarity. *Proceedings of the 28th Cognitive Science Society*, 2169–2173.

Slobin, D. I. (1996). From "thought and language" to "thinking for speaking." In *Rethinking linguistic relativity* (pp. 70–96). Cambridge University Press.

Stamenković, D., Ichien, N., & Holyoak, K. J. (2019). Metaphor comprehension: An individual-differences approach. *Journal of Memory and Language*, *105*, 108–118. https://doi.org/10.1016/j.jml.2018.12.003

Steen, G. (2007). *Finding metaphor in grammar and usage: A methodological analysis of theory and research*. John Benjamins Publishing.

Sweetser, E. (1990). *From etymology to pragmatics*. Cambridge: Cambridge University.

Talmy, L. (1975). Semantics and syntax of motion. In J. Kimball (Ed.), *Syntax and semantics* (Vol. 4, pp. 181-238). New York: Academic Press.

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science, 12*(1), 49-100.

Talmy, L. (2000). *Toward a cognitive semantics, Vol. 1: Concept structuring systems*. Cambridge, MA: The MIT Press.

Thibodeau, P., & Durgin, F. H. (2008). Productive figurative communication: Conventional metaphors facilitate the comprehension of related novel metaphors. *Journal of Memory and Language, 58*(2), 521–540. https://doi.org/10.1016/j.jml.2007.05.001

Thibodeau, P. H., & Durgin, F. H. (2011). Metaphor Aptness and Conventionality: A Processing Fluency Account. *Metaphor and Symbol, 26*(3), 206–226. https://doi.org/10.1080/10926488.2011.583196

Torreano, L. A., Cacciari, C., & Glucksberg, S. (2005). When dogs can fly: Level of abstraction as a cue to metaphorical use of verbs. *Metaphor and Symbol*, *20*(4), 259–274. https://doi.org/10.1207/s15327868ms2004_2

Tourangeau, R., & Rips, L. (1991). Interpreting and evaluating metaphors. *Journal of Memory and Language, 30*(4), 452–472. https://doi.org/10.1016/0749-596X(91)90016-D

Tourangeau, R., & Sternberg, R. J. (1981). Aptness in metaphor. *Cognitive Psychology*, *13*(1), 27–55. https://doi.org/10.1016/0010-0285(81)90003-7

Tourangeau, R., & Sternberg, R. J. (1982). Understanding and appreciating metaphors. *Cognition*, *11*(3), 203–244.

Traugott, E. C. (1988). Pragmatic strengthening and grammaticalization. *Annual Meeting of the Berkeley Linguistics Society*, *14*, 406–416.

Trick, L., & Katz, A. N. (1986). The Domain Interaction Approach to Metaphor Processing: Relating Individual Differences and Metaphor Characteristics. *Metaphor and Symbolic Activity*, *1*(3), 185–213. https://doi.org/10.1207/s15327868ms0103_3

Ünal, E., Ji, Y., & Papafragou, A. (2021). From Event Representation to Linguistic Meaning. *Topics in Cognitive Science*, *13*(1), 224–242. https://doi.org/10.1111/tops.12475

Utt, J., Lenci, A., Padó, S., & Zarcone, A. (2013). The curious case of metonymic verbs: A distributional characterization. *IWCS 2013 Workshop Towards a Formal Distributional Semantics*, 30–39.

Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating Upcoming Words in Discourse: Evidence From ERPs and Reading Times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 443–467. https://doi.org/10.1037/0278-7393.31.3.443

Vassallo, P., Chersoni, E., Santus, E., Lenci, A., & Blache, P. (2018, May). Event Knowledge in Sentence Processing: A New Dataset for the Evaluation of Argument Typicality. *LREC 2018 Workshop on Linguistic and Neurocognitive Resources (LiNCR)*. https://hal.archives-ouvertes.fr/hal-01724286

Vicente, A. (2018). Polysemy and word meaning: An account of lexical meaning for different kinds of content words. *Philosophical Studies*, *175*(4), 947–968. https://doi.org/10.1007/s11098-017-0900-y

Vicente, A., & Falkum, I. L. (2017). Polysemy. In A. Vicente & I. L. Falkum, *Oxford Research Encyclopedia of Linguistics*. Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.325

Vigliocco, G., Vinson, D. P., & Siri, S. (2005). Semantic similarity and grammatical class in naming actions. *Cognition*, *94*(3), B91–B100. https://doi.org/10.1016/j.cognition.2004.06.004

Vigliocco, G., Vinson, D. P., Druks, J., Barber, H., & Cappa, S. F. (2011). Nouns and verbs in the brain: A review of behavioural, electrophysiological, neuropsychological and imaging studies. *Neuroscience & Biobehavioral Reviews*, *35*(3), 407–426. https://doi.org/10.1016/j.neubiorev.2010.04.007

Vigliocco, G., Warren, J., Siri, S., Arciuli, J., Scott, S., & Wise, R. (2006). The role of semantics and grammatical class in the neural representation of words. *Cerebral Cortex*, *16*(12), 1790–1796.

Wallentin, M., Ostergaard, S., Lund, T., Ostergaard, L., & Roepstorff, A. (2005). Concrete spatial language: See what I mean? *Brain and Language*, *92*(3), 221–233. https://doi.org/10.1016/j.bandl.2004.06.106

Whelan, R. (2008). Effective Analysis of Reaction Time Data. *The Psychological Record*, *58*(3), 475–482. https://doi.org/10.1007/BF03395630

Wilks, Y. (1975). Preference semantics. In E. L. Keenan (Ed.), *Formal Semantics of Natural Language* (pp. 329–348). Cambridge University Press.

Wisniewski, E. J. (1997). When concepts combine. *Psychonomic Bulletin & Review*, *4*(2), 167-183.

Wolff, P., & Gentner, D. (2000). Evidence for role-neutral initial processing of metaphors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(2), 529–541. https://doi.org/10.1037/0278-7393.26.2.529

Wolff, P., & Gentner, D. (2011). Structure-mapping in metaphor comprehension. *Cognitive Science*, *35*(8), 1456–1488. https://doi.org/10.1111/j.1551-6709.2011.01194.x

Xu, Y., Malt, B. C., & Srinivasan, M. (2017). Evolution of word meanings through metaphorical mapping: Systematicity over the past millennium. *Cognitive Psychology*, *96*, 41–53. https://doi.org/10.1016/j.cogpsych.2017.05.005

Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin*, *127*(1), 3.

Zarcone, A., Padó, S., & Lenci, A. (2014). Logical Metonymy Resolution in a Words-as-Cues Framework: Evidence From Self-Paced Reading and Probe Recognition. *Cognitive Science*, *38*(5), 973–996. https://doi.org/10.1111/cogs.12108

Zharikov, S. S., & Gentner, D. (2002). Why do metaphors seem deeper than similes? In W. D. Gray & C. D. Schunn (Eds.), *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society* (1st ed., pp. 976–981). Routledge. https://doi.org/10.4324/9781315782379-203

Zheng, Y., Matlen, B., & Gentner, D. (2022). Spatial alignment facilitates visual comparison in children. *Cognitive Science*, 46(8), e13182.

# Appendices

# Appendix A

Instructions for Experiments 1 and 2 from Chapter 1

In this experiment you will see a number of sentences. Your task is to write a paraphrase -- that is, please write out what you think the sentence means, **without using any of the same content words** (but it's ok to repeat words like "a" and "the").

Important: Please don't translate mechanically, word by word. Instead, think about what the sentence means. Imagine that you're walking by someone in a restaurant and you hear them utter the sentence to a friend. What could they be trying to communicate? Try to capture that possible meaning in your paraphrase.

Some of the sentences you see might seem a little odd, but please try your best to come up with a plausible overall meaning.

(*shown in Experiments 1 and 3 only*)

Example:

If you saw the phrase "the slimy executive," you could translate it:

Mechanical way: "the gooey person" (please don't do this)

Natural way: "the corrupt CEO" (a more plausible meaning)

(*shown in Experiment 2 only*)

Example:

If you saw the phrase "the slimy orator," you could translate it:

Mechanical way: "the gooey speaker" (please don't do this)

Natural way: "the corrupt politician" (a more plausible meaning)

(*shown in Experiment 3 only*)

Some of the sentences you encounter may have typos in them. In those cases, use your intuition to determine what you think was meant, and base your paraphrase off that.

Again, **be sure not to repeat any content words**. Of course, it's ok to repeat words like the, a, an, and, etc., but notice how, in the above example, the words slimy and executive were not repeated.

Some of the sentences you encounter may be odd, but please do your best to provide a **meaningful interpretation**. Ask yourself – if someone else read the sentence I just wrote, would they know what I meant?

This task should take about 10-15 minutes to complete.

Thank you for your time and effort! Good luck.

# Appendix B

Code tallies for each item from the coding task in Experiments 1 and 2 (Chapter 1), sorted in descending order of proportion of paraphrases excluded. Only Meaningful (Mg) paraphrases were included in the analysis; all other codes were excluded. Thus, for Experiment 1, 526 out of a total of 654 paraphrases were included in the analysis. For Experiment 2,

Codes for paraphrases are as follows: Mg = Meaningful (the net number of paraphrases included in the analysis, all other codes were excluded), Mc = Mechanical, D = Describes the situation, N = Noncompliant. Total = total number of paraphrases generated for that item.

**Experiment 1**

| Item | Mg | Mc | D | N | Total | Prop. Excluded |
|---|---|---|---|---|---|---|
| The lizard agreed | 3 | 15 | 0 | 0 | 18 | .83 |
| The lantern agreed | 6 | 9 | 0 | 2 | 17 | .65 |
| The lizard worshipped | 7 | 11 | 0 | 1 | 19 | .63 |
| The mule worshipped | 7 | 11 | 0 | 0 | 18 | .61 |
| The car agreed | 8 | 10 | 0 | 0 | 18 | .56 |
| The lantern worshipped | 8 | 10 | 0 | 0 | 18 | .56 |
| The car worshipped | 8 | 9 | 0 | 0 | 17 | .53 |
| The mule agreed | 11 | 8 | 0 | 0 | 19 | .42 |
| The lantern shivered | 14 | 2 | 2 | 0 | 18 | .22 |
| The lantern limped | 15 | 3 | 0 | 1 | 19 | .21 |
| The car limped | 15 | 1 | 2 | 0 | 18 | .17 |
| The mule cooked | 15 | 0 | 1 | 2 | 18 | .17 |
| The lantern cooked | 16 | 0 | 1 | 2 | 19 | .16 |
| The daughter cooked | 15 | 0 | 0 | 2 | 17 | .12 |
| The car cooked | 16 | 0 | 1 | 1 | 18 | .11 |
| The daughter limped | 16 | 0 | 2 | 0 | 18 | .11 |
| The lantern softened | 16 | 0 | 1 | 1 | 18 | .11 |
| The mule shivered | 16 | 0 | 2 | 0 | 18 | .11 |
| The politician shivered | 16 | 0 | 2 | 0 | 18 | .11 |
| The car shivered | 17 | 2 | 0 | 0 | 19 | .11 |
| The lizard cooked | 17 | 0 | 2 | 0 | 19 | .11 |

| | | | | | | |
|---|---|---|---|---|---|---|
| The daughter worshipped | 17 | 0 | 1 | 0 | 18 | .06 |
| The lizard softened | 17 | 0 | 1 | 0 | 18 | .06 |
| The politician agreed | 17 | 0 | 1 | 0 | 18 | .06 |
| The politician cooked | 17 | 0 | 1 | 0 | 18 | .06 |
| The car softened | 18 | 0 | 1 | 0 | 19 | .05 |
| The daughter agreed | 18 | 0 | 1 | 0 | 19 | .05 |
| The daughter shivered | 18 | 0 | 1 | 0 | 19 | .05 |
| The mule softened | 18 | 0 | 1 | 0 | 19 | .05 |
| The politician worshipped | 18 | 0 | 0 | 1 | 19 | .05 |
| The daughter softened | 18 | 0 | 0 | 0 | 18 | 0 |
| The lizard limped | 18 | 0 | 0 | 0 | 18 | 0 |
| The lizard shivered | 17 | 0 | 0 | 0 | 17 | 0 |
| The mule limped | 17 | 0 | 0 | 0 | 17 | 0 |
| The politician limped | 19 | 0 | 0 | 0 | 19 | 0 |
| The politician softened | 17 | 0 | 0 | 0 | 17 | 0 |
| **Total** | 526 | 91 | 24 | 13 | 654 | .20 |

**Experiment 2**

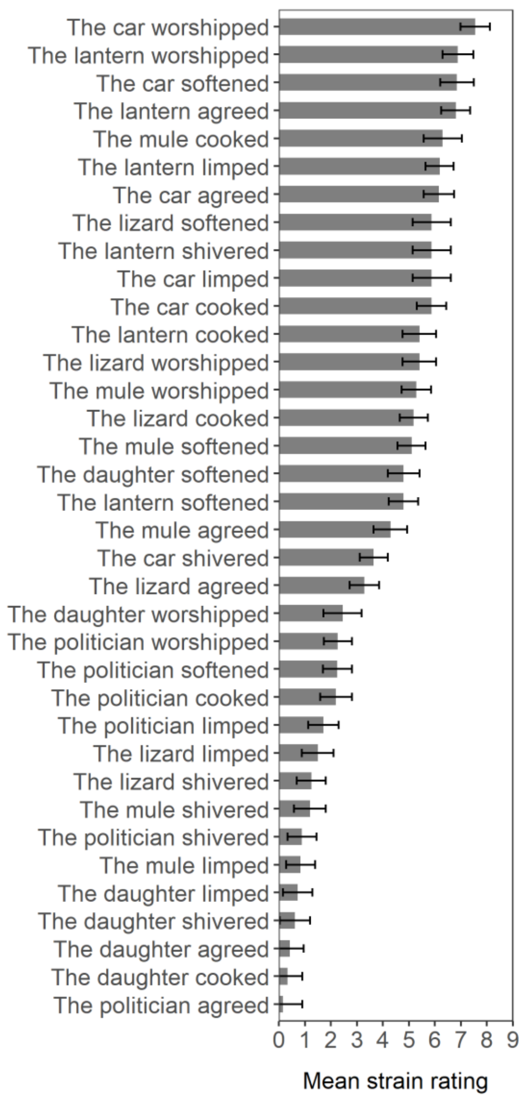
| Item | Mg | Mc | D | N | Total | Prop. Excluded |
|---|---|---|---|---|---|---|
| The box paused | 19 | 18 | 4 | 2 | 43 | 0.56 |
| The box complained | 20 | 19 | 3 | 0 | 42 | 0.52 |
| The tree complained | 20 | 16 | 3 | 1 | 40 | 0.50 |
| The tree failed | 26 | 12 | 5 | 0 | 43 | 0.40 |
| The box suffered | 25 | 9 | 5 | 0 | 39 | 0.36 |
| The bell complained | 29 | 7 | 6 | 0 | 42 | 0.31 |
| The tree burned | 29 | 0 | 10 | 3 | 42 | 0.31 |
| The tree paused | 29 | 5 | 7 | 1 | 42 | 0.31 |
| The box failed | 31 | 8 | 2 | 1 | 42 | 0.26 |
| The tree dried | 30 | 0 | 9 | 1 | 40 | 0.25 |
| The professor failed | 32 | 0 | 10 | 0 | 42 | 0.24 |
| The professor suffered | 33 | 0 | 10 | 0 | 43 | 0.23 |
| The tree suffered | 33 | 4 | 5 | 0 | 42 | 0.21 |
| The motor suffered | 34 | 5 | 2 | 1 | 42 | 0.19 |
| The bell suffered | 33 | 2 | 3 | 2 | 40 | 0.18 |
| The professor paused | 33 | 0 | 6 | 1 | 40 | 0.18 |
| The queen failed | 33 | 0 | 6 | 1 | 40 | 0.18 |
| The box burned | 34 | 0 | 5 | 1 | 40 | 0.15 |
| The queen burned | 34 | 0 | 5 | 1 | 40 | 0.15 |
| The motor complained | 35 | 1 | 4 | 0 | 40 | 0.13 |
| The box dried | 37 | 0 | 2 | 3 | 42 | 0.12 |
| The queen paused | 37 | 0 | 5 | 0 | 42 | 0.12 |

| | | | | | | |
|---|---|---|---|---|---|---|
| The motor burned | 38 | 0 | 5 | 0 | 43 | 0.12 |
| The bell failed | 36 | 0 | 4 | 0 | 40 | 0.10 |
| The queen suffered | 38 | 1 | 3 | 0 | 42 | 0.10 |
| The professor dried | 38 | 0 | 1 | 2 | 41 | 0.07 |
| The bell paused | 39 | 0 | 2 | 1 | 42 | 0.07 |
| The motor dried | 39 | 0 | 2 | 1 | 42 | 0.07 |
| The queen dried | 39 | 0 | 2 | 1 | 42 | 0.07 |
| The queen complained | 40 | 0 | 3 | 0 | 43 | 0.07 |
| The professor complained | 40 | 0 | 2 | 0 | 42 | 0.05 |
| The motor paused | 38 | 0 | 1 | 0 | 39 | 0.03 |
| The bell burned | 41 | 0 | 0 | 1 | 42 | 0.02 |
| The motor failed | 41 | 0 | 1 | 0 | 42 | 0.02 |
| The professor burned | 41 | 0 | 1 | 0 | 42 | 0.02 |
| The bell dried | 43 | 0 | 0 | 0 | 43 | 0 |
| **Total** | 1217 | 107 | 144 | 25 | 1493 | .18 |

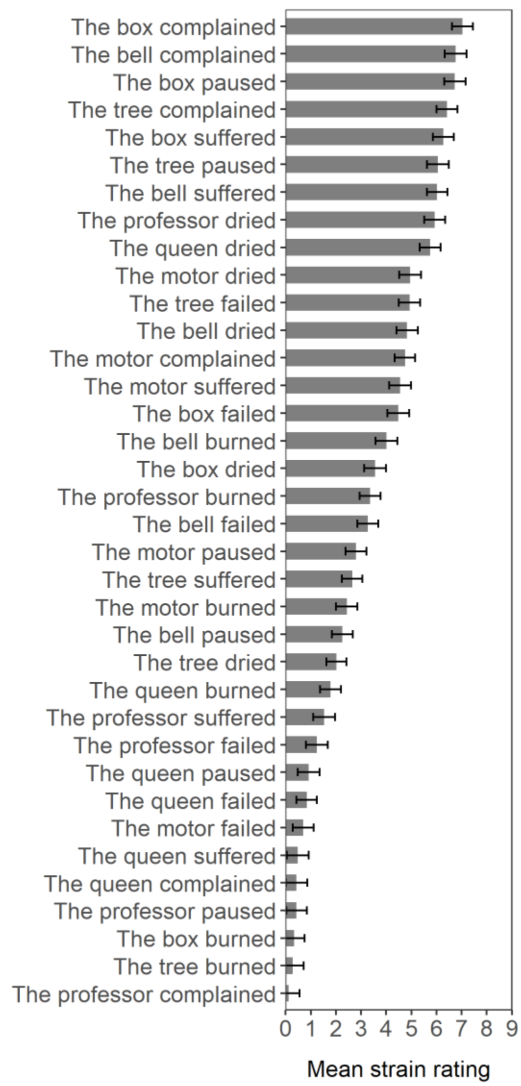
*Note*: The totals for *Meaningful* and *Total* paraphrases here (1217 and 1493) are different from those included in the final analysis in Experiment 2 (1216 and 1491 respectively) due to two paraphrases generating null vectors in word2vec (i.e., containing no words present in word2vec's dictionary). Of the total 1493 paraphrases generated in Experiment 2, two of them generated null vectors, meaning that only 1491 were included in the analysis. Of those two, one of them was excluded during coding, meaning that the 1217 paraphrases coded as meaningful included one paraphrase that generated a null vector. Thus, only 1216 were included in the analysis.

# Appendix C

Strain rating by items for the 36 items used in Experiment 1 (A) and Experiment 2 (B) from Chapter 1. Error bars represent 1 standard error of the mean. Adjusted means were obtained by fitting a linear mixed model with rating as the dependent measure, item as the fixed effect, and subjects as the random effect.



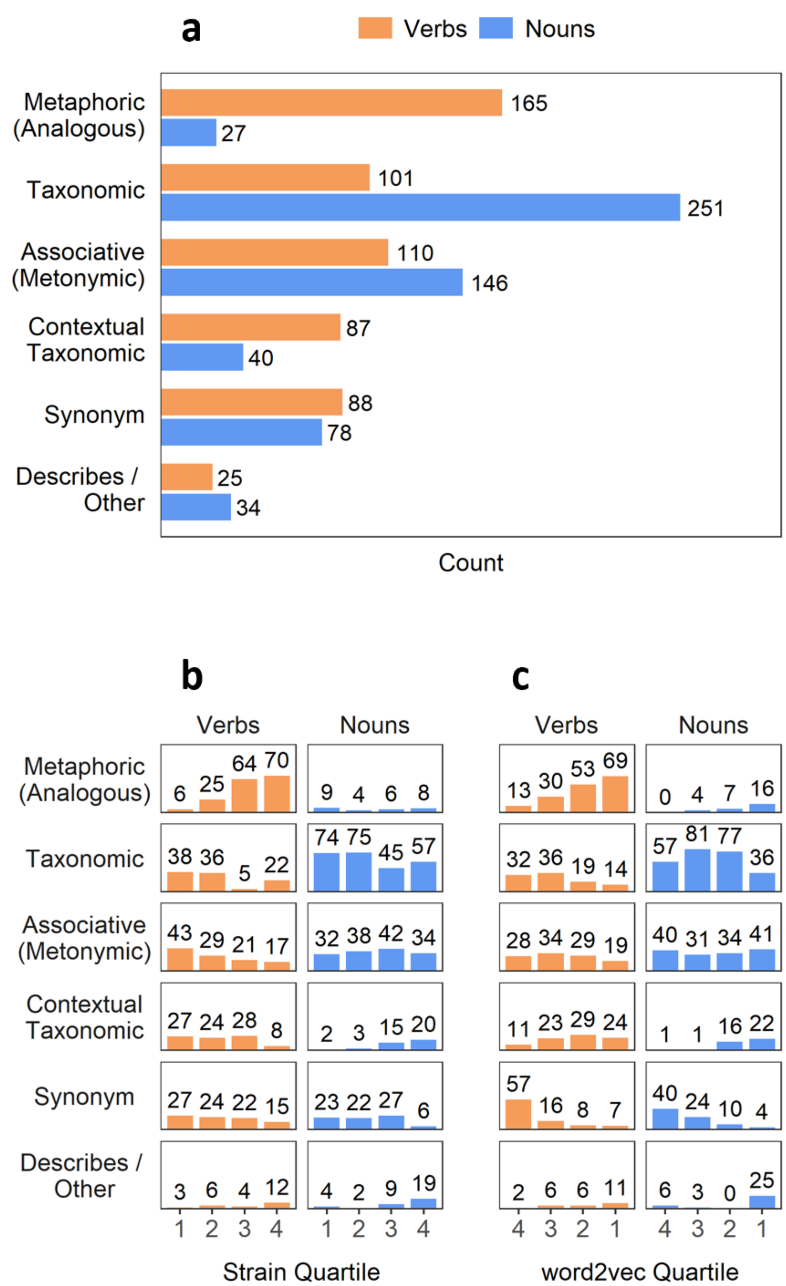

A. Experiment 1




B. Experiment 2

# Appendix D

Code tallies for all categories from the Experiment 3 qualitative analysis from Chapter 1




(A) Total counts. (B) Tallies by strain quartile, with strain increasing from left to right. (C) Tallies by word2vec quartile. In this figure, degree of change increases from left to right, with quartile 4 representing the least degree of change and quartile 1 representing the greatest degree of change

# Appendix E

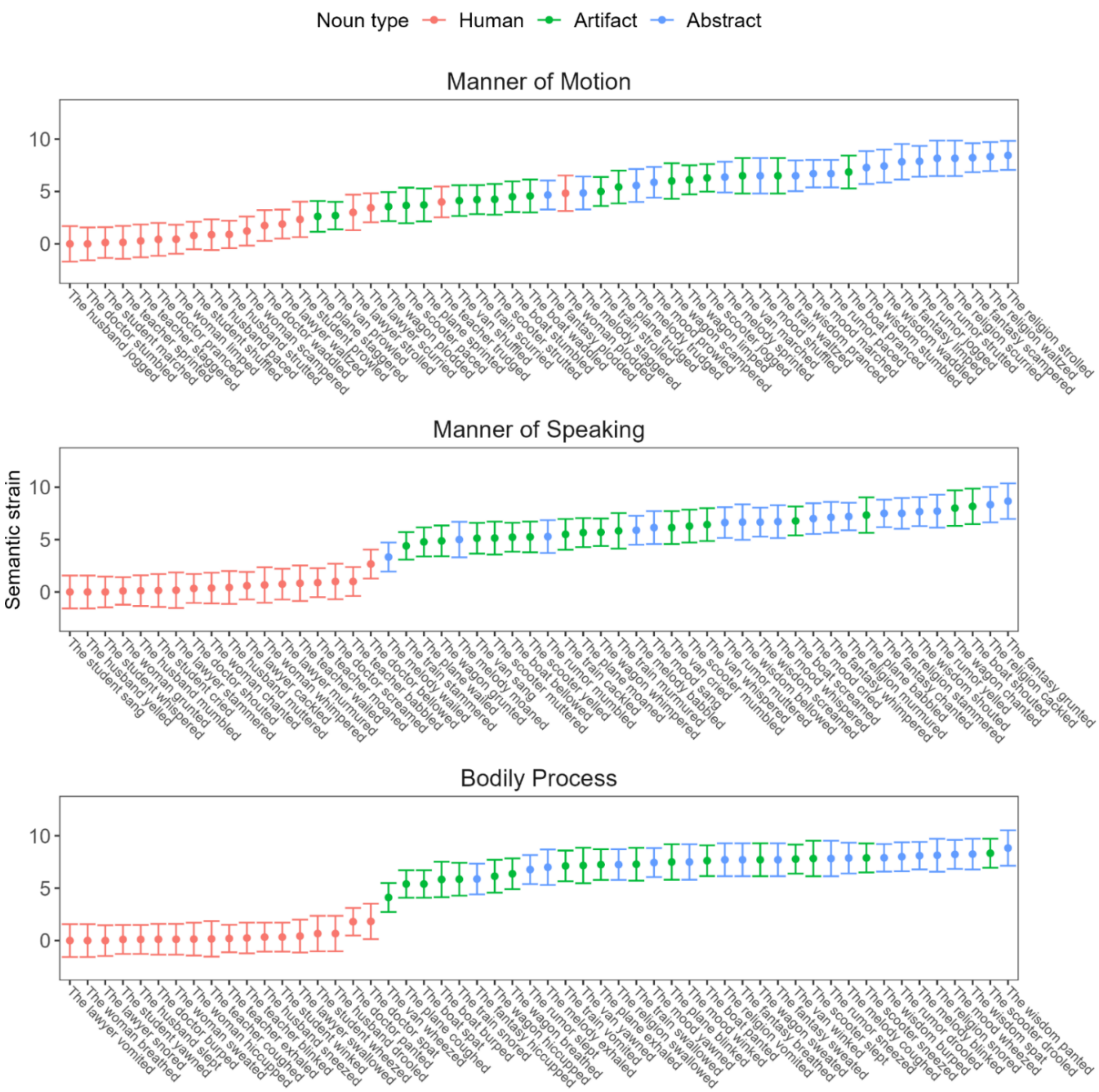
Stimuli used in all experiments in Chapter 2. Each list is sorted in alphabetical order by verb; within each verb items are sorted in ascending order of strain.

| Manner of motion | Manner of speaking | Bodily Process |
|---|---|---|
| The husband jogged | The teacher babbled | The teacher blinked |
| The scooter jogged | The plane babbled | The plane blinked |
| The rumor jogged | The melody babbled | The melody blinked |
| The woman limped | The doctor bellowed | The woman breathed |
| The wagon limped | The boat bellowed | The wagon breathed |
| The fantasy limped | The wisdom bellowed | The fantasy breathed |
| The student marched | The lawyer cackled | The doctor burped |
| The van marched | The train cackled | The boat burped |
| The mood marched | The religion cackled | The wisdom burped |
| The husband paced | The woman chanted | The teacher coughed |
| The scooter paced | The wagon chanted | The plane coughed |
| The rumor paced | The fantasy chanted | The melody coughed |
| The woman plodded | The student cried | The husband drooled |
| The wagon plodded | The van cried | The scooter drooled |
| The fantasy plodded | The mood cried | The rumor drooled |
| The doctor pranced | The woman grunted | The teacher exhaled |
| The boat pranced | The wagon grunted | The plane exhaled |
| The wisdom pranced | The fantasy grunted | The melody exhaled |
| The student prowled | The teacher moaned | The woman hiccupped |
| The van prowled | The plane moaned | The wagon hiccupped |
| The mood prowled | The melody moaned | The fantasy hiccupped |
| The woman scampered | The husband mumbled | The doctor panted |
| The wagon scampered | The scooter mumbled | The boat panted |
| The fantasy scampered | The rumor mumbled | The wisdom panted |
| The lawyer scurried | The lawyer murmured | The husband slept |
| The train scurried | The train murmured | The scooter slept |
| The religion scurried | The religion murmured | The rumor slept |
| The student shuffled | The husband muttered | The husband sneezed |
| The van shuffled | The scooter muttered | The scooter sneezed |
| The mood shuffled | The rumor muttered | The rumor sneezed |
| The teacher sprinted | The student sang | The lawyer snored |
| The plane sprinted | The van sang | The train snored |
| The melody sprinted | The mood sang | The religion snored |
| The teacher staggered | The doctor screamed | The doctor spat |
| The plane staggered | The boat screamed | The boat spat |

| | | |
|---|---|---|
| The melody staggered | The wisdom screamed | The wisdom spat |
| The lawyer strolled | The doctor shouted | The lawyer swallowed |
| The train strolled | The boat shouted | The train swallowed |
| The religion strolled | The wisdom shouted | The religion swallowed |
| The husband strutted | The lawyer stammered | The woman sweated |
| The scooter strutted | The train stammered | The wagon sweated |
| The rumor strutted | The religion stammered | The fantasy sweated |
| The doctor stumbled | The teacher wailed | The lawyer vomited |
| The boat stumbled | The plane wailed | The train vomited |
| The wisdom stumbled | The melody wailed | The religion vomited |
| The teacher trudged | The woman whimpered | The student wheezed |
| The plane trudged | The wagon whimpered | The van wheezed |
| The melody trudged | The fantasy whimpered | The mood wheezed |
| The doctor waddled | The student whispered | The student winked |
| The boat waddled | The van whispered | The van winked |
| The wisdom waddled | The mood whispered | The mood winked |
| The lawyer waltzed | The husband yelled | The student yawned |
| The train waltzed | The scooter yelled | The van yawned |
| The religion waltzed | The rumor yelled | The mood yawned |

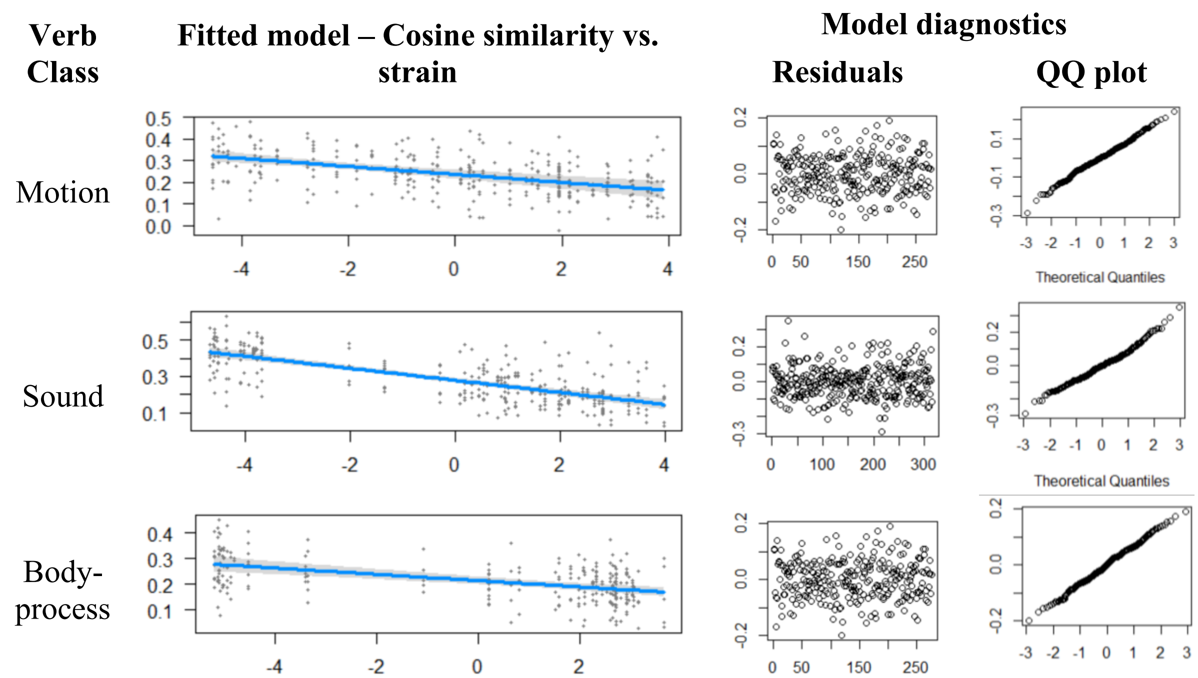
# Appendix F

Strain ratings from Chapter 2, Experiment 1B, sorted in ascending order of strain. Error bars represent 95% CIs.



# Appendix G

Diagnostic plots for models from Chapter 2, Experiment 1B, demonstrating linear fit



# Appendix H

Verb lists used during retrace task in Chapter 2, Experiment 3B

**Manner of Motion**

| **List 1** | **List 2** | **List 3** |
|---|---|---|
| waddled | limped | scurried |
| scampered | pranced | paced |
| plodded | sprinted | stumbled |
| **List 4** | **List 5** | **List 6** |
| prowled | marched | strutted |
| staggered | strolled | trudged |
| jogged | shuffled | waltzed |

**Communication**

| **List 1** | **List 2** | **List 3** |
|---|---|---|
| whimpered | yelled | shouted |
| bellowed | moaned | whispered |
| cackled | mumbled | stammered |
| **List 4** | **List 5** | **List 6** |
| chanted | wailed | screamed |
| cried | murmured | sang |
| grunted | babbled | muttered |

**Bodily Process**

| **List 1** | **List 2** | **List 3** |
|---|---|---|
| snored | burped | sneezed |
| panted | slept | winked |
| vomited | wheezed | yawned |
| **List 4** | **List 5** | **List 6** |
| drooled | breathed | sweated |
| hiccupped | blinked | coughed |
| exhaled | swallowed | spat |

# Appendix I

Reaction time Pilot study from Chapter 3, Experiment 1

Experiment 1 was based on the paradigm established by a previously-run pilot study, which we describe here. In the pilot, we used the same 2 X 2 design described in Experiment 1, crossing noun event domain (motion vs. sound) with verb event domain (motion vs sound) to generate within-domain (high-similarity events: motion-motion or sound-sound) and cross-domain items (low-similarity events: motion-sound or sound-motion). Six vehicle nouns (*plane, boat, jet, van, train,* and *scooter*) and six musical instrument nouns (*bell, violin, horn, piano, drums, guitar*) were used to activate event structures in the motion and sound domains. The verbs were 24 manner of motion verbs (e.g., *sprinted, strutted, pranced,* etc.) and 24 manner of speaking verbs (e.g., *mumbled, cackled, screamed*, etc.), drawn from the same lists used in Chapter 2.[41] A total of 96 unique metaphoric items were generated by combining the nouns and verbs, half of which were within-domain pairings (predicted to be fast to interpret) and half of which were cross-domain pairings (predicted to be slow to interpret; see Figure 21).

[41] There was also a condition with items comprising body-process verbs; since it is not relevant for our present purposes we omit it here.

| | | Verb Event Domain | |
|---|---|---|---|
| | | Motion | Sound |
| Noun Event Domain | Motion | *The boat sprinted* | *The boat stammered* |
| | Sound | *The violin sprinted* | *The violin stammered* |

Figure 21. Experimental design for target items from the pilot and Experiment 1 showing noun event domains crossed with verb event domains with example sentences for each condition in the cells. Shaded cells indicate cross-domain items (low-similarity events, predicted to be slow to comprehend); unshaded cells indicate within-domain items (high-similarity events, predicted to be fast to comprehend).

A literal control condition/manipulation check was also included wherein verbs of both classes were paired with human subject nouns (e.g., *The chef strutted/whimpered*), resulting in 48 unique unstrained (literal) sentences. We expected (1) that as literal sentences, they would be interpreted significantly faster than both within- and cross-domain metaphors; and (2) that there should be no significant difference in RTs by verb class: human nouns paired with motion verbs (*The chef strutted*) should be interpreted equally fast as human nouns paired with sound verbs (*The chef whimpered*). This served to confirm that one verb class was not faster to process overall than another. All three noun types (human, vehicle, and musical instrument) were matched to each other in terms of frequency and word length (all *p*s > .05), as were both verb classes (all *p*s > .05).[42]

Participants were shown the metaphors on a screen one at a time and were instructed to press the space bar as soon as they had thought of a meaningful interpretation. They then typed their

[42] Frequency data was drawn from the Corpus of Contemporary American English (COCA; Davies, 2008).

interpretation before moving on to the next item. The key dependent measure was comprehension time, operationalized as the time from stimulus onset until spacebar press.

The results of the pilot are plotted in Figure 22 below. First, as expected, participants were faster to interpret items in the literal control condition compared to both target conditions (*p*s $< .01$), and interpretation speed for the literal control items did not vary significantly by verb class (though the effect was marginal, $p = .09$). Second for the target metaphoric conditions (where the subject nouns were either vehicles or musical instruments), the results were mixed. Items with musical instrument subject nouns matched predictions: participants were significantly faster to interpret within-domain metaphors (e.g., *The violin stammered*) than cross-domain metaphors (e.g., *The violin sprinted*), $p < .0001$. Items with vehicle subject nouns, however, did not match predictions: there were no significant differences in RTs between within-domain metaphors (e.g., *The boat sprinted*) and cross-domain metaphors (e.g., *The boat stammered*), $p = .70$.

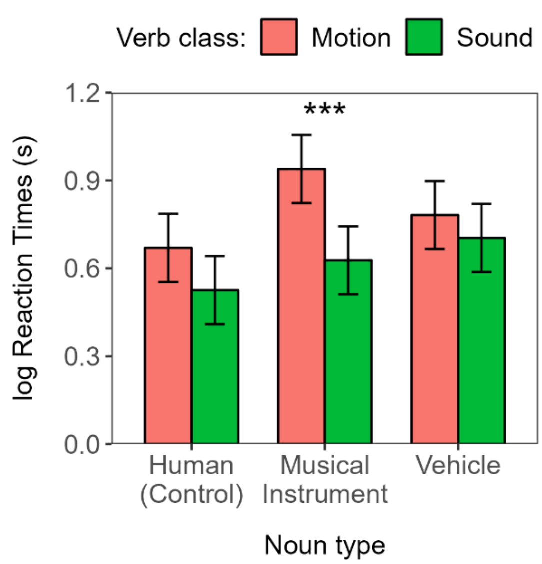



Figure 22. Pilot results. RTs are log-transformed. Error bars indicate 95% confidence intervals. *** $p < .001$.

# Appendix J

Target, filler, and practice items from Chapter 3, Experiment 1

Target items (128)

The accordion cackled
The accordion chanted
The accordion chirped
The accordion danced
The accordion hollered
The accordion prowled
The accordion scurried
The accordion sprinted
The banjo babbled
The banjo grumbled
The banjo grunted
The banjo limped
The banjo paced
The banjo pranced
The banjo trudged
The banjo yelled
The blimp babbled
The blimp grumbled
The blimp grunted
The blimp limped
The blimp paced
The blimp pranced
The blimp trudged
The blimp yelled
The canoe cackled
The canoe chanted
The canoe chirped
The canoe danced
The canoe hollered
The canoe prowled
The canoe scurried
The canoe sprinted
The cello barked
The cello hobbled
The cello jogged
The cello screamed
The cello staggered
The cello wailed
The cello waltzed
The cello whimpered
The flute bellowed
The flute chattered
The flute marched
The flute muttered
The flute stammered
The glider bellowed
The glider chattered
The glider marched
The glider muttered
The glider stammered
The glider strolled
The glider strutted
The glider waddled
The gong cackled
The gong chanted
The gong chirped
The gong danced
The gong hollered
The gong prowled
The gong scurried
The gong sprinted
The harp babbled
The harp grumbled
The harp grunted
The harp limped
The harp paced
The harp pranced
The harp trudged
The harp yelled
The kayak barked
The kayak hobbled
The kayak jogged
The kayak screamed
The kayak staggered
The kayak wailed
The kite cackled
The kite chanted
The kite chirped
The kite danced
The kite hollered
The kite prowled
The kite scurried
The kite sprinted
The raft babbled
The raft grumbled
The raft grunted
The raft limped
The raft paced
The raft pranced
The raft trudged
The rowboat barked
The rowboat hobbled
The rowboat jogged
The rowboat screamed
The rowboat staggered
The rowboat wailed
The rowboat waltzed
The rowboat whimpered
The sailboat bellowed
The sailboat chattered
The sailboat marched
The sailboat muttered
The sailboat stammered
The sailboat strolled
The sailboat strutted
The sailboat waddled
The trumpet barked
The trumpet hobbled
The trumpet jogged
The trumpet screamed
The trumpet staggered
The trumpet wailed
The trumpet waltzed
The trumpet whimpered
The violin bellowed
The violin chattered
The violin marched
The violin muttered
The violin stammered
The violin strolled
The violin strutted
The violin waddled

The flute strolled
The flute strutted
The flute waddled
The raft yelled

Filler items (28)

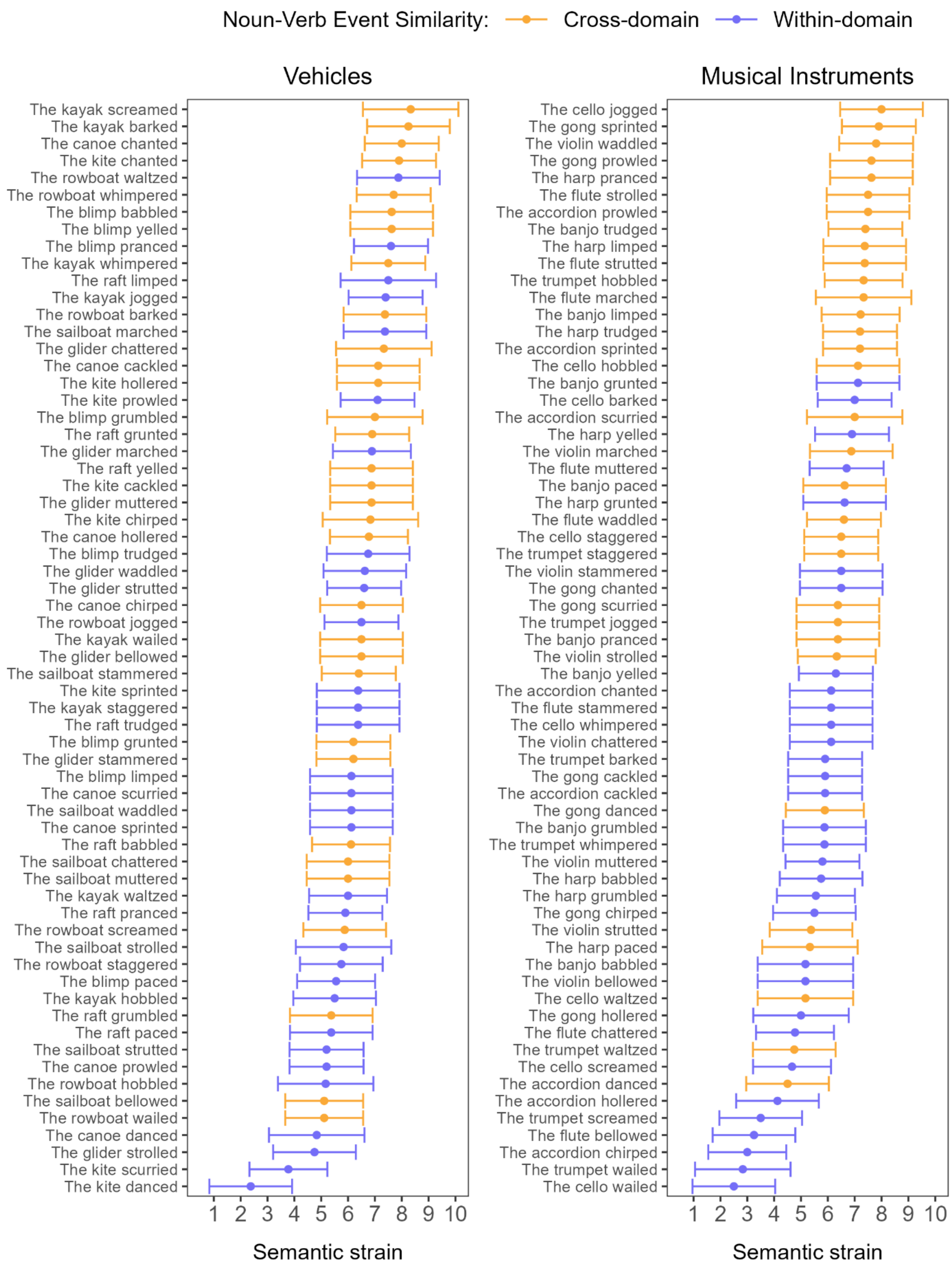
The car drove
The jet flew
The van moved
The boat sailed
The plane landed
The train arrived
The buggy rolled
The carriage departed
The spaceship launched
The scooter skidded
The truck honked
The rocket boomed
The missile exploded
The subway smelled
The bell rang
The trombone sounded
The cymbals crashed
The tambourine rattled
The harmonica blared
The xylophone echoed
The mandolin twanged
The tuba tooted
The guitar played
The drums banged
The saxophone bounced
The oboe fell
The piano rotted
The clarinet sparkled

Practice items (4)

The tiger pounced
The courage collapsed
The submarine surfaced
The speaker beeped

# Appendix K

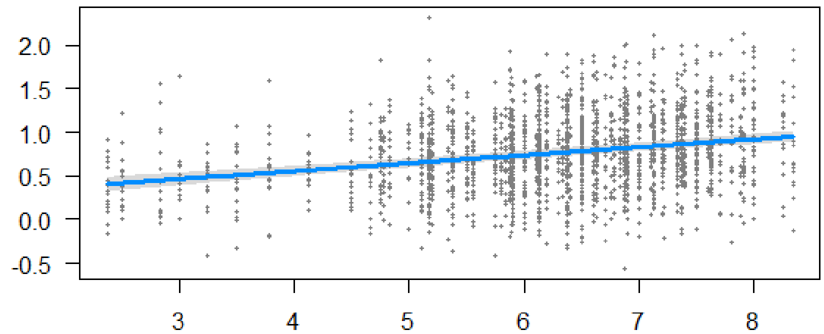
Strain ratings for target items from Chapter 3, Experiment 1, sorted in ascending order of strain



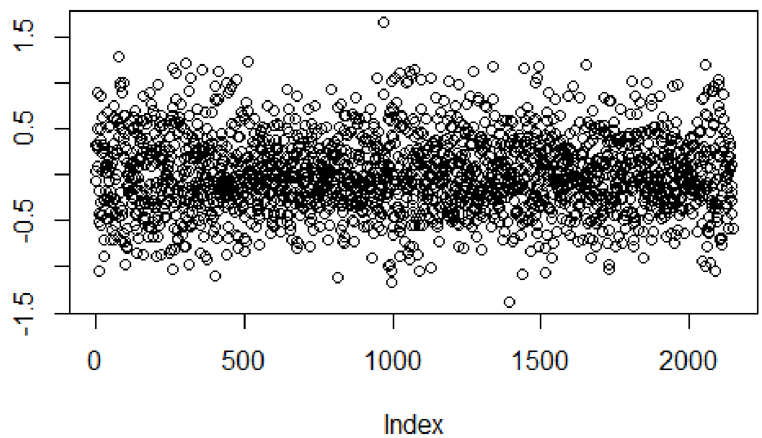
# Appendix L

Diagnostic plots from the linear model from Chapter 3, Experiment 1, demonstrating linear fit

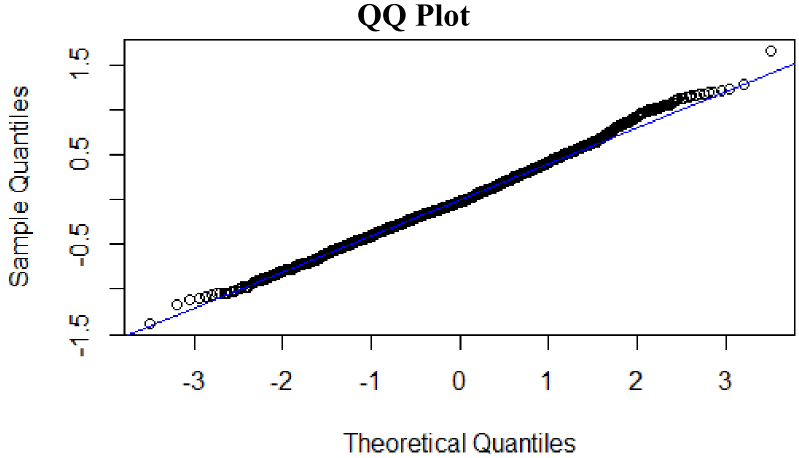
**Fitted model – Log RT (s) vs. Strain**




**Residuals**




**QQ Plot**



# Appendix M

Filler and practice items from Chapter 3, Experiment 2

Filler Items (24)

| Item | Prime | Prime type |
|---|---|---|
| *The dog growled* | fur | Irrelevant |
| *The horse galloped* | leg | Relevant |
| *The hawk flew* | eye | Irrelevant |
| *The bird sang* | beak | Relevant |
| *The cat scratched* | tail | Irrelevant |
| *The fish swam* | fin | Relevant |
| *The snake slithered* | scale | Irrelevant |
| *The donkey drank* | tongue | Relevant |
| *The bear burrowed* | mouth | Irrelevant |
| *The kangaroo jumped* | foot | Relevant |
| *The beaver gnawed* | ear | Irrelevant |
| *The mouse scurried* | paw | Relevant |
| *The sword cut* | hilt | Irrelevant |
| *The alarm rang* | bell | Relevant |
| *The marker drew* | cap | Irrelevant |
| *The pen leaked* | ink | Relevant |
| *The cup shattered* | rim | Irrelevant |
| *The van rusted* | fender | Relevant |
| *The TV glowed* | antenna | Irrelevant |
| *The oven warmed* | coil | Relevant |
| *The fan whirred* | cord | Irrelevant |
| *The book tore* | page | Relevant |
| *The desk burned* | drawer | Irrelevant |
| *The belt tightened* | clasp | Relevant |

Practice Items (4)

| Type | Item | Prime | Prime type |
|---|---|---|---|
| Literal | The violin played | bow | Relevant |
| | The lion roared | mane | Irrelevant |
| Metaphoric | The dam murmured | water | Relevant |
| | The engine complained | gasoline | Irrelevant |

# Appendix N

Stimulus items for event norming study from Chapter 3, Experiment 2

Target Items (32)

| **Relevant Primes** | **Irrelevant Primes** |
|---|---|
| pedals (of a bicycle) | handlebar (of a bicycle) |
| blade (of a blender) | button (of a blender) |
| tire (of a car) | deck (of a boat) |
| propellor (of a boat) | windshield (of a car) |
| siren (of a firetruck) | hose (of a firetruck) |
| string (of a guitar) | knob (of a guitar) |
| spout (of a kettle) | lid (of a kettle) |
| spring (of a mattress) | tag (of a mattress) |
| wing (of a plane) | seat (of a plane) |
| trigger (of a rifle) | crosshairs (of a rifle) |
| horn (of a truck) | seatbelt (of a truck) |
| keys (of a typewriter) | paper (of a typewriter) |
| bulb (of a flashlight) | battery (of a flashlight) |
| wick (of a lantern) | pole (of a lantern) |
| motor (of a vacuum) | handle (of a vacuum) |
| nozzle (of a sprinkler) | valve (of a sprinkler) |

Catch trials (6)

Lock (of a suitcase)
Eraser (of a pencil)
Bristles (of a toothbrush)
Cap (of a pen)
Sail (of a sailboat)
Speaker (of a stereo)

# Appendix O

Results of the event norming study from Chapter 3, Experiment 2. The *Match* column refers to the experimenters' informal judgments of which events matched the events expected to surface in the paraphrases. Top-ranked events are highlighted.

| Item | Prime Type | Prime | Event | Rank | Score | Match |
|---|---|---|---|---|---|---|
| The bicycle sprinted | Relevant | pedals | propels bike/rider | 1 | 97 | X |
| | | | is a place for feet | 2 | 15 | |
| | | | causes braking | 3 | 12 | |
| | | | moves in circular motion | 4 | 11 | |
| | Irrelevant | handlebar | steers the bike | 1 | 78 | |
| | | | is something to hold onto when riding | 2 | 36 | |
| | | | provides balance for rider | 3 | 23 | |
| | | | contains brakes / gears / bells / other things | 4 | 20 | |
| | | | secures rider | 5 | 9 | |
| | | | made of metal | 6 | 6 | |
| The blender attacked | Relevant | blade | cuts / blends food | 1 | 148 | X |
| | | | is a sharp object | 2 | 26 | |
| | | | rotates when in use | 3 | 19 | |
| | | | is noisy when used | 4 | 7 | |
| | | | describes product created from blender | 5 | 6 | |
| | Irrelevant | button | turns blender on/off | 1 | 118 | |
| | | | adjusts blender settings | 2 | 43 | |
| The boat strutted | Relevant | propellor | propels boat | 1 | 83 | X |
| | | | steers boat | 2 | 17 | |
| | | | spins / rotates | 3 | 11 | |
| | | | pushes water | 4 | 10 | |
| | | | adjusts speed | 5 | 7 | |
| | Irrelevant | deck | provides place for people to stand / sit / lay | 1 | 87 | |
| | | | place for storage/holds items | 2 | 37 | |
| | | | covers / protects / roof for lower levels | 3 | 25 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | provides structure / support for boat | 4 | 13 | |
| | | | can be slippery or hazardous | 5 | 7 | |
| | | | has certain physical properties | 6 | 4 | |
| The car limped | Relevant | tire | allows for motion / rolls over street | 1 | 88 | X |
| | | | supports / stabilizes the car | 2 | 34 | |
| | | | provides traction / grip on road | 3 | 24 | |
| | | | goes flat / becomes deflated | 4 | 14 | X |
| | | | rotates / spins | 5 | 13 | |
| | | | holds air / inflates | 6 | 9 | |
| | | | allows for steering | 7 | 7 | |
| | Irrelevant | windshield | Protects occupants from wind / rain / other objects | 1 | 129 | |
| | | | Allows for seeing surroundings | 2 | 21 | |
| | | | Improves aerodynamics | 3 | 7 | |
| | | | Keeps occupant from being ejected | 4 | 6 | |
| The firetruck yelled | Relevant | siren | makes a loud noise / alerts other people | 1 | 156 | X |
| | | | clears traffic | 2 | 12 | |
| | Irrelevant | hose | dispenses water | 1 | 76 | |
| | | | fights fires | 2 | 32 | |
| | | | controls flow / directs water | 3 | 20 | |
| | | | receives water from hydrant or other source | 4 | 12 | |
| | | | is long | 5 | 6 | |
| | | | saves homes / lives | 6 | 5 | |
| The flashlight stuttered | Relevant | bulb | illuminates objects | 1 | 129 | X |
| | | | gets warm with use | 2 | 12 | |
| | | | can break / wear out with use | 3 | 8 | |
| | Irrelevant | battery | provides energy / power | 1 | 70 | |
| | | | allows flashlight to work | 2 | 54 | |
| The guitar stammered | Relevant | string | produces sound / allows for playing | 1 | 114 | X |
| | | | is plucked | 2 | 9 | |
| | | | is tuned / adjusted | 3 | 7 | |
| | Irrelevant | knob | adjusts the sound / volume / tuning / etc. | 1 | 131 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | is used by turning it | 2 | 14 | |
| | | | has various physical properties | 3 | 7 | |
| The kettle drooled | Relevant | spout | pours liquid / directs flow of liquid | 1 | 100 | X |
| | | | vent for steam to escape | 2 | 23 | |
| | | | makes a whistling noise | 3 | 22 | |
| | Irrelevant | lid | prevents overflowing / spilling | 1 | 55 | |
| | | | speeds up boiling / increases heat | 2 | 26 | |
| | | | keeps steam in kettle | 3 | 21 | |
| | | | covers kettle | 4 | 15 | |
| | | | prevents foreign objects from entering kettle | 5 | 9 | |
| | | | helps cook food | 6 | 7 | |
| | | | protects people from burns | 7 | 6 | |
| The lantern waltzed | Relevant | wick | burns / creates light / illumination | 1 | 112 | X |
| | | | absorbs oil / draws fuel | 2 | 17 | |
| | | | changes properties when wet / dry | 3 | 5 | |
| | Irrelevant | pole | holds / carries / hangs the lantern | 1 | 82 | |
| | | | makes it possible to carry lantern | 2 | 18 | |
| | | | keeps heat / flames away from other things | 3 | 8 | |
| | | | mounted in ground or on wall | 4 | 7 | |
| | | | protects lantern | 5 | 6 | |
| | | | has various physical properties | 6 | 3 | |
| The mattress shrieked | Relevant | spring | absorbs / supports weight / body | 1 | 66 | |
| | | | bounces / provides bounce | 2 | 31 | |
| | | | creates firmness | 3 | 13 | |
| | | | provides comfort | 4 | 13 | |
| | | | allows for sleep | 5 | 10 | |
| | | | makes mattress soft | 6 | 9 | |
| | | | gives mattress structure | 7 | 8 | |
| | | | squeaks / makes noise | 8 | 7 | X |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | can become damaged with age | 9 | 5 | |
| | | | is coil / spiral shaped | 10 | 3 | |
| | Irrelevant | tag | gives information about the mattress | 1 | 124 | |
| The plane waddled | Relevant | wing | allows plane to fly / keeps plane in air | 1 | 91 | X |
| | | | keeps plane steady / balanced / stabilized | 2 | 30 | X |
| | | | holds engines / landing gear | 3 | 8 | |
| | | | allows plane to land | 4 | 7 | |
| | | | allows plane to maneuver / steer | 5 | 5 | |
| | Irrelevant | seat | allows people to sit / provides comfort during flight | 1 | 92 | |
| | | | allows for relaxing / sleeping / resting | 2 | 18 | |
| | | | has various components / properties | 3 | 13 | |
| | | | protects passengers | 4 | 11 | |
| | | | can be reclined | 5 | 7 | |
| | | | keeps passengers separated | 6 | 6 | |
| | | | is arranged in rows | 7 | 4 | |
| The rifle barked | Relevant | trigger | causes gun to fire | 1 | 110 | X |
| | | | has various physical properties | 2 | 9 | |
| | | | allows shooter to control gun | 3 | 8 | |
| | Irrelevant | crosshairs | provides aim / increases accuracy | 1 | 109 | |
| The sprinkler coughed | Relevant | nozzle | directs /controls water stream | 1 | 63 | |
| | | | dispenses / sprays / streams water | 2 | 51 | X |
| | | | rotates / oscillates | 3 | 9 | |
| | | | prevents leaks / keeps water in hose | 4 | 8 | |
| | | | waters plants | 5 | 8 | |
| | Irrelevant | valve | allows / prevents water flow / releases water | 1 | 105 | |
| | | | adjusts / controls flow rate of water | 2 | 39 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | is moveable / rotated to operate | 3 | 9 | |
| | | | channels / directs water | 4 | 7 | |
| The truck howled | Relevant | horn | makes a loud sound / alerts / warns others | 1 | 183 | X |
| | | | entertains children | 2 | 7 | |
| | Irrelevant | seatbelt | keeps person safe | 1 | 89 | |
| | | | holds person in place in seat | 2 | 40 | |
| | | | beeps / clicks | 3 | 9 | |
| The typewriter babbled | Relevant | keys | allows for typing / producing text | 1 | 101 | X |
| | | | makes clicking sound | 2 | 18 | X |
| | | | represents letters and numbers | 3 | 17 | |
| | | | acts as a lever | 4 | 14 | |
| | Irrelevant | paper | receives text / marks / ink | 1 | 118 | |
| | | | is rolled through typewriter | 2 | 7 | |
| | | | can be thrown away | 3 | 6 | |
| The vacuum cackled | Relevant | motor | creates suction / air flow / air pressure / makes vacuum work | 1 | 65 | |
| | | | generates power | 2 | 29 | |
| | | | makes noise | 3 | 25 | X |
| | | | rotates / spins | 4 | 22 | |
| | | | can become worn out with use | 5 | 7 | |
| | | | scares others | 6 | 7 | |
| | Irrelevant | handle | allows vacuum to be gripped / moved / carried / manipulated | 1 | 122 | |