COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 46 (2022) e13141

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13141

Verb Metaphoric Extension Under Semantic Strain

Daniel King, Dedre Gentner

Department of Psychology, Northwestern University

Received 15 March 2019; received in revised form 6 April 2022; accepted 8 April 2022

Abstract

This paper explores the processes underlying verb metaphoric extension. Work on metaphor processing has largely focused on noun metaphor, despite evidence that verb metaphor is more common. Across three experiments, we collected paraphrases of simple intransitive sentences varying in semantic strain—for example, *The motor complained* → *The engine made strange noises*—and assessed the degree of meaning change for the noun and the verb. We developed a novel methodology for this assessment using word2vec. In Experiments 1 and 2, we found that (a) under semantic strain, verb meanings were more likely to be adjusted than noun meanings; (b) the degree of verb meaning adjustment—but not noun meaning adjustment—increased with semantic strain; and (c) verb meaning extension is primarily driven by online adjustment, although sense selection also plays a role. In Experiment 3, we replicated the word2vec results with an assessment using human subjects. The results further showed that nouns and verbs change meaning in qualitatively different ways, with verbs more likely to change meaning metaphorically and nouns more likely to change meaning taxonomically or metonymically. These findings bear on the origin and processing of verb metaphors and provide a link between online sentence processing and diachronic change over language evolution.

Keywords: Metaphor; Verb metaphor; Metaphor processing; Verb mutability; Vector space models; Semantic change; word2vec

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.

Correspondence should be sent to Daniel King, Department of Psychology, Northwestern University, 2120 Campus Drive, Suite 162, Evanston, IL 60208–2710, USA. E-mail: king@u.northwestern.edu

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 are *wasting* my time.
This gadget will *save* you hours.
I do not *have* the time to *give* you.
How do you *spend* your time these days?
That flat tire *cost* me an hour.
I have *invested* 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, 2001; Gentner & Wolff, 1997, 2000; Gibbs, 1992; Giora, 1997; Glucksberg & Keysar, 1990; Glucksberg, McGlone, & Manfredi, 1997; Jones & Estes, 2006; A. N. Katz, 1989; Keysar, Shen, Glucksberg, & Horton, 2000; Ortony, 1979; Shen, 1989; Thibodeau & Durgin, 2011; Tourangeau & Rips, 1991; Trick & Katz, 1986; Tourangeau & Sternberg, 1981, 1982; Wolff & Gentner, 2011). Psycholinguistic research on metaphoric uses of verbs is comparatively rare (but see Cardillo, Schmidt, et al., 2010; Cardillo, Watson, & Chatterjee, 2017; Cardillo, Watson, et al., 2012; Gentner & France, 1988; Stamenković, Ichien, & Holyoak, 2019; Torreano, Cacciari, & Glucksberg, 2005).

The dearth of research on verb metaphor is unfortunate, as there is evidence that verb metaphors are more common than noun metaphors (Jamrozik, Sagi, Goldwater, & Gentner, 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 Fig. 1). The nouns and verbs were selected such

		Human		Animal		Artifact		Abstract	
		agree	worship	shiver	limp	soften	cook	succeed	weaken
Human	daughter	*The daughter agreed*		*The daughter shivered*		*The daughter softened*		*The daughter succeeded*	
	politician								
Animal	mule	*The mule agreed*		*The mule shivered*		*The mule softened*		*The mule succeeded*	
	lizard								
Artifact	car	*The car agreed*		*The car shivered*		*The car softened*		*The car succeeded*	
	lantern								
Abstract	responsibility	*The responsibility agreed*		*The responsibility shivered*		*The responsibility softened*		*The responsibility succeeded*	
	courage								

Fig. 1. Grid showing stimuli noun and verbs from Gentner and 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.

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 (H. H. Clark, 1966; Earles & Kersten, 2000, 2017; Earles, Kersten, Turner, & McMullen, 1999; 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 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 E. V. 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., H. H. 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, Watson, Schmidt, Kranjec, & Chatterjee, 2012; Chatterjee, 2010; Dirven, 1985;

Heine, 1997; Hopper & Traugott, 2003; Jamrozik, McQuire, Cardillo, & Chatterjee, 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, the *heart* referred literally to an organ before later gaining metaphoric senses such as *the center of things* (Dirven, 1985). Similarly, a *bridge* originally referred only to a structure linking two physical locations but is now 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 ($ps < .05$), 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 the 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 (one to two senses) and half were high-polysemy (seven to 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 the degree of semantic change. How does one objectively determine the relative degree of change in the noun versus 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. 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 the 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* (WEMs).

1.4. Using vector space WEMs to assess semantic adjustment

In this research, we use word2vec (Mikolov, Chen, Corrado, & Dean, 2013; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), a vector space WEM, to assess the 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, Gershman, Ritter, and Botvinick (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.

1.5. Sense selection versus online adjustment

Using word2vec, we investigated the question of sense selection versus 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 should 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, then 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

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

2.1.2. Materials and 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 eight nouns with eight intransitive verbs for a total of 64 different sentences, which can be visualized as forming a matrix (see Fig. 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 Fig. 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 six nouns and six verbs for a total of 36 of the original 64 sentences (see Fig. 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 toward 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.

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) $\times$ 2 (item strain: strained vs. unstrained, within-subject). Each participant saw each of the six nouns and six verbs exactly once. Two simple unstrained sentences were included as catch trials for checking attention and following directions; the criteria for excluding a subject were repeating a noun and/or verb in both of the catch trials or producing an obviously nonsensical answer in either. As each of the 109 net participants paraphrased six 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: (a) 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 (b) mechanical, word-by-word paraphrases of strained sentences (e.g., paraphrasing *The lantern worshipped* as *The candle honored*). As noted above, participants were 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 a verb score, representing the amount of semantic adjustment the initial noun and verb underwent from the original sentence to the 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 0.47 and 0.22, 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: (a) overall, the change in meaning was greater for

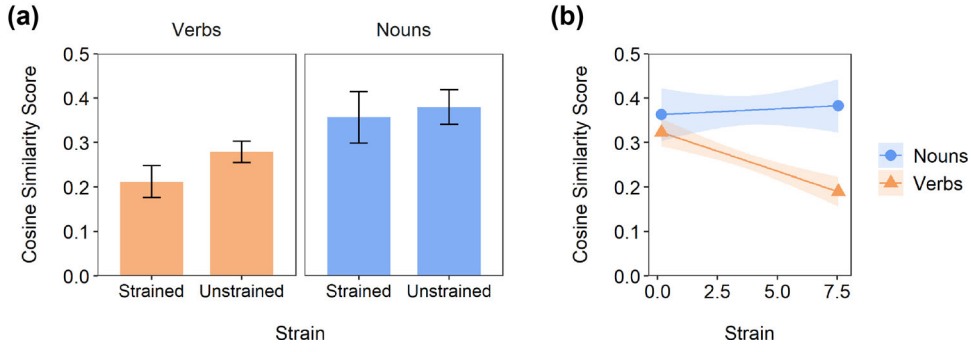


Fig. 2. Noun and verb similarity scores from Experiment 1. Lower scores indicate greater semantic adjustment. Error bars/bands represent 95% confidence intervals. (a) Strain treated as a categorical predictor. (b) Strain as a continuous predictor, derived from the comprehensibility ratings.

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*

verbs than for nouns, and (b) this effect was greater for semantically strained sentences than for unstrained sentences. As Fig. 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.

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 the 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, $\beta = 0.11$, $SE = 0.02$, $t = 5.35$, $p < .001$, indicating that, on average, 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 the degree of meaning change, two additional models were fit: one for nouns and one for verbs. In both models, the 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, $\beta = -0.26$, $SE = 0.08$, $t = 3.09$, $p < .01$, 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, $\beta = 0.07$, $SE = 0.10$, $t = 0.66$, $p = .51$. These results are shown in Fig. 2a.

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 Fig. 1). Although this provides a principled way to classify strained versus 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 four 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, $\beta = -0.38$, $SE = 0.08$, $t = 4.89$, $p < .001$, but not for nouns, $\beta = -0.04$, $SE = 0.11$, $t = 0.39$, $p = .70$ (see Fig. 2b). Notably, the value of the standardized slope coefficient for verbs obtained using the continuous measure of strain (-0.38) was larger than the parameter obtained in the categorical model (-0.26), 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 ($\beta = -0.06$, $SE = 0.04$, $t = 1.47$, $p = .14$) and likewise for verbs ($\beta = -0.004$, $SE = 0.04$, $t = 0.10$, $p = .93$.). 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 (eight 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 0.20 for verbs and 0.32 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 the change of meaning instead of human

			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	+/-	+/+	+/-	+/+	+/-	+/+

Fig. 3. Stimulus matrix for Experiment 2. Shaded cells indicate combinations that result in strained sentences, following Gentner and France's (1988) approach. Pluses and minuses 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*).

judges. The results bear out the key phenomena of the verb mutability hypothesis: (a) 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 paraphrasing.

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.

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 (one to two senses) and half were high polysemy (7+ senses; see Fig. 3). 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 the 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

A total of 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. One 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 six nouns and six 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 Fig. 3). Thus, across the 36 sentences, the four possible combinations of noun and verb polysemy occurred in equal numbers: nine N+/V+ combinations, nine N−/V−-combinations, nine N+/V− combinations, and nine N−/V+ combinations. As in Experiment 1, participants saw each noun and verb exactly once, receiving six target sentences (two strained, four unstrained) comprising an equal number of high- and low-polysemy nouns and verbs (three N−, three V−, three N+, and three 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.

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).

3.1.3. Experimental design

The design was 6 (item grouping, between-subject) × 2 (item strain: strained vs. unstrained, within-subject) × 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 were 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 six 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 who 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 one 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 was 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, $\beta = 0.04$, $SE = 0.02$, $t = 2.62$, $p = .01$, indicating that verbs ($M = 0.24$, $SD = 0.12$) changed significantly more overall than nouns ($M = 0.28$, $SD = 0.13$).

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, the 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 Fig. 4.

For verbs, there was a significant main effect of semantic strain such that the degree of verb meaning change increased as strain increased, $\beta = -0.29$, $SE = 0.08$, $t = 3.51$, $p = .001$. There was also a significant main effect of polysemy, $\beta = -0.22$, $SE = 0.08$, $t = 2.70$, $p = .01$, 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, $\beta = -0.02$, $SE = 0.08$, $t = 0.28$, $p = .78$.

For nouns, a significant main effect of polysemy was found, $\beta = -0.16$, $SE = 0.06$, $t = 2.70$, $p = .01$, with high-polysemy nouns ($M = 0.25$, $SE = 0.01$) changing meaning

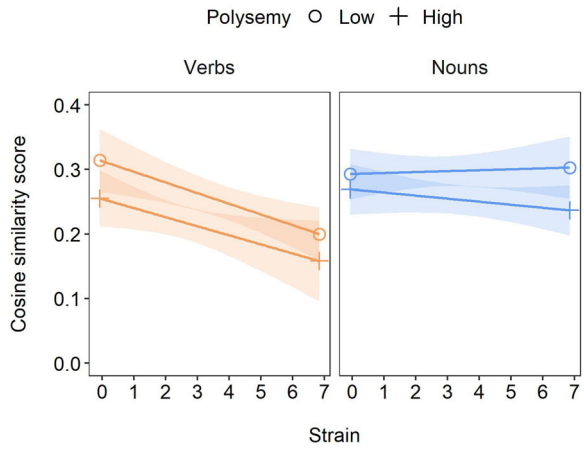


Fig. 4. Fitted model plots showing the effect of strain and polysemy on word2vec scores for verbs and nouns in Experiment 2. Strain increases from left to right. Lower word2vec scores indicate greater meaning change. Shaded ribbons indicate 95% confidence bands.

to a greater extent than low-polysemy nouns ($M = 0.30$, $SE = 0.01$). There was no significant effect of semantic strain, $\beta = -0.03$, $SE = 0.06$, $t = 0.45$, $p = .65$, and the interaction was not significant, $\beta = 0.05$, $SE = 0.06$, $t = 0.85$, $p = .40$.

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, $\beta = -0.03$, $SE = 0.04$, $t = 0.7$, $p = .49$, 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 ($\beta = -0.02$, $SE = 0.03$, $t = 0.58$, $p = .56$). For verbs, a small but significant relationship was found ($\beta = 0.11$, $SE = 0.03$, $t = 4.19$, $p < .001$), 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 toward 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.

Table 2
Example paraphrases from Experiment 2

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

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.

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;* seven noun senses, two verb senses). If sense selection were the primary driver of mutability, we would expect unbalanced sentences like these to be most favorable toward 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, two 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 toward 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.

As in Experiment 1, there were some items with high rates of noncompliant paraphrases, although fewer than previously (five 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 0.19 and 0.17 for low- and high-polysemy verbs and 0.30 and 0.23 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 grant 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—for example, Gerz, Vulić, Hill, Reichart, and Korhonen (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 the 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 (Experiment 2)	**Original paraphrase (Experiment 2)**	**Double paraphrase (Experiment 3)**
The motor complained	The engine did not work well	The motor 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 (although we accepted differences in pluralization or tense) provides an objective scoring procedure. The tradeoff is data loss, since many near-matches are discarded—for example, *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 experiment.

4.1. Method

4.1.1. Participants

Seventy-seven 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. Four participants were excluded for failing the catch trial criteria, and two were excluded due to experimenter error, resulting in a net of 71 participants.

4.1.2. Materials and 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 (which serve as the stimuli/initial sentences in this

experiment) as *singles* and the 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 two catch trials that served as attention checks. The 18 singles were presented in three blocks of six items each, with order randomized within each block. Within each block, each of the original six stimulus nouns and verbs (from which the single paraphrase originated) was represented exactly once (so that each occurred three 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 from 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 six 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.

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

Table 3
Number of resurfacings (hits) versus nonresurfacings (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 Sections 4.1.2 and 4.2.1).

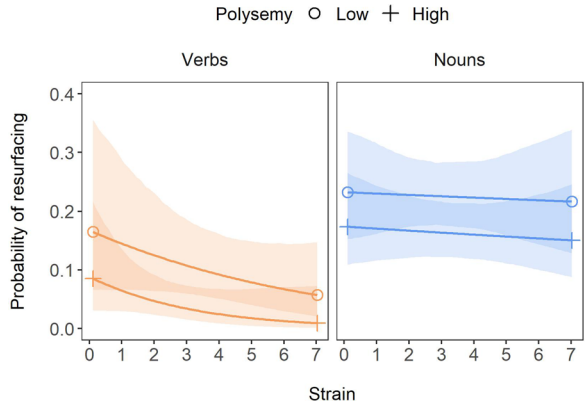


Fig. 5. Fitted models showing the probability of resurfacing for verbs and nouns in Experiment 3. Lower probabilities indicate greater meaning change. Strain increases from left to right. Shaded ribbons indicate 95% confidence bands.

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, $\beta = 1.24$, $SE = 0.33$, 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 Fig. 5.

For verbs, there was a significant main effect of semantic strain, $\beta = -0.25$, $SE = 0.11$, 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, $\beta = -0.63$, $SE = 0.23$, 95% CI [0.21, 1.17], $z = 2.69$, $p < .01$, indicating that low-polysemy verbs (68 resurfacings) were

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

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

4.3. Discussion

A full replication of Experiments 1 and 2 required the following three results: (a) verbs should resurface less often than nouns overall (indicating greater meaning change overall), (b) resurfacings should decrease with semantic strain for verbs but not for nouns, and (c) 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, (a) verbs changed more than nouns overall (they resurfaced less); (b) semantic strain significantly predicted verb—but not noun—change; and (c) 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 serve 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 change

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 by 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 nouns. As discussed earlier, metaphoric uses of verbs appear to be significantly more common in day-to-day language than metaphoric uses of nouns (Jamrozik et al., 2013; Krennmayr, 2011). 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; Qiu, 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—for example, 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* (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, Lenci, Padó, & Zarcone, 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 toward 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, $\kappa = 0.58$, (95% CI, 0.55 to 0.61), $p < .001$; after discussion, consensus was reached on all items.

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*). Fig. 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).

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	

Fig. 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.

Fig. 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

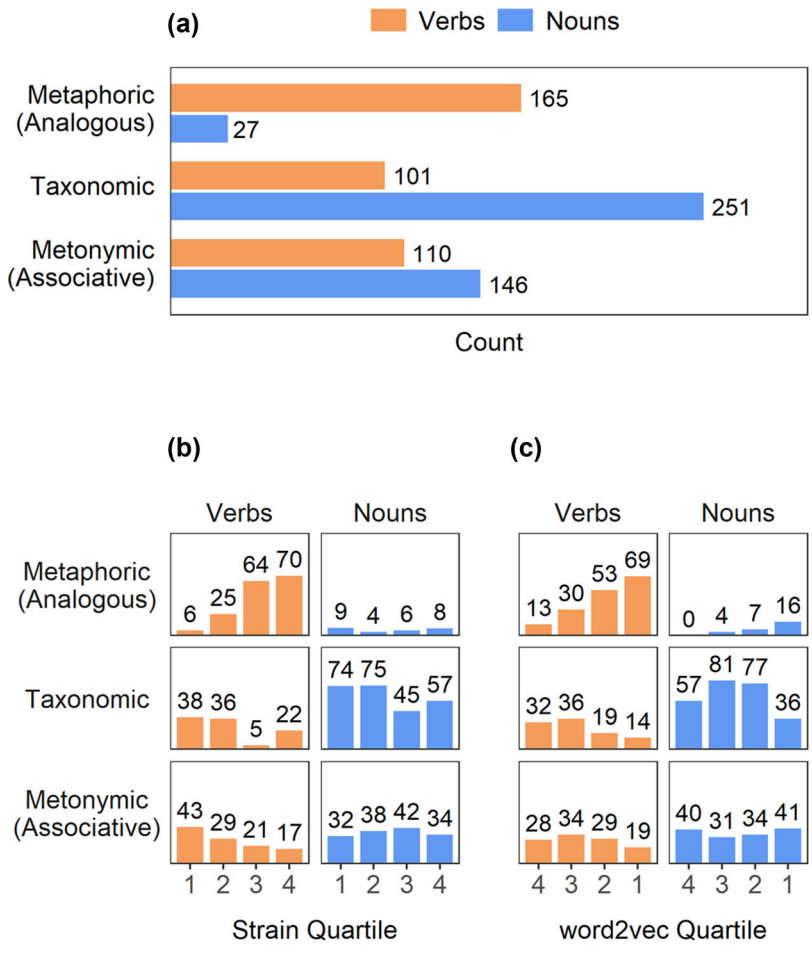


Fig. 6. Tallies for the metaphoric (analogous), associative (metonymic), and taxonomic categories for nouns and verbs from the qualitative analysis. (a) Total counts. (b) Tallies by strain quartile, with strain increasing from left to right. (c) Tallies by word2vec quartile. The x-axes are reversed so 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).

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 Fig. 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 a 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.

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 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: under semantic strain, verb meanings are altered more than 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 (a) that verbs changed more than nouns overall and (b) that the degree of verb meaning change increased with the degree of 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 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 the strain

increased, and both low- and high-polysemy nouns remained equally stable (Fig. 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—for example, 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 from 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 the degree of strain. In contrast, the 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 would not 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 a sound. Sentence (3) is highly strained; boxes are inanimate and also not known for making a 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, Widick, & Chatterjee, 2008; Desai, Binder, Conant, Mano, & Seidenberg, 2011, 2013; Jamrozik et al., 2016; Raposo, Moss, Stamatakis, & Tyler, 2009; Saygin, McCullough, Alac, & Emmorey, 2010; Wallentin, Ostergaard, Lund, Ostergaard, & Roepstorff, 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 their meanings 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., Heine, 1997; Hopper & Traugott, 2003; Joseph et al., 1996; 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. Although 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 a greater semantic change in verbs than in nouns even when the verb was the first word in 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 the 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 (1993) 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 a 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, Vinson, Druks, Barber, & Cappa, 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—for example, *carnivore, barrier* (Asmuth & Gentner, 2017; Gentner & Asmuth, 2017; Gentner & Kurtz, 2005; Goldwater & Markman, 2011; Goldwater, Markman, & Stilwell, 2011; Markman & Stilwell, 2001; Rehder & Ross, 2001).

Emulating Kersten and Earles' (2004) recognition paradigm, Asmuth and Gentner (2017) gave participants phrases consisting of an entity noun and a relational noun—for example, *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 the 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 et al., 2006, 2011; Vigliocco, Vinson, & Siri, 2005). 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 *purr*," the corresponding verb metaphor "the flowers *purred* 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.

5.6.4. 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.

6. Implications and future work

From the theoretical discussion above, one would expect these findings to generalize to transitive sentences. Gentner and France (1988, Experiments 3a and 3b) 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. Furthermore, 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.

7. 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 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 changes under strain, we also found qualitative differences in *how* nouns and verbs change the 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.

Acknowledgments

The research reported here received support from NSF grant SBE-0541957 and from ONR grant N00014-16-1-2613. Research was conducted with the approval of the Institutional Review Board. We thank Eyal Sagi, Sid Horton, Phillip Wolff, and Sandy LaTourrette for their theoretical and technical guidance. We also thank our tireless coders, Kathy Duan, Amelia Emery, Charlie Hansell, Kira Johnson, Anatha Latshaw, Roshaye Poleon, Kyle Rockoff, and Kate Sandberg. The data for all experiments reported here are available at https://osf.io/3swvd/?view_only.

Notes

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

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., J. J. Katz & Fodor, 1963), as it emphasizes the fact that verbs can accommodate arguments that do not match their preferred semantic types, as in the present research.

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, Kintsch, & Landauer, 1998; Landauer & Dumais, 1997; Lenci, 2018). 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; Lenci, 2018; Rimell, Maillard, Polajnar, & Clark, 2016).

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.

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).

6 Noun model: $\beta_{Strain} = 0.02$, $SE = 0.09$, $p = .87$; Verb model: $\beta_{Strain} = -.27$, SE $= .07$, $p < .001$.

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

8 Noun model: $\beta_{Strain} = -.04$, SE $= .06$, $p = .49$; $\beta_{Polysemy} = -0.16$, $SE = 0.05$, $p < .01$; $\beta_{Str*Poly} = -0.05$, $SE = 0.07$, $p = .21$; Verb model: $\beta_{Strain} = -0.23$, $SE = 0.08$, $p < .01$; $\beta_{Polysemy} = -0.22$, $SE = 0.08$, $p < .01$; $\beta_{Str*Poly} = -0.02$, $SE = 0.08$, $p = .84$.

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

10 Metonymy and metaphor are both types of figurative language, but they differ in an important way. Metaphor involves abstract 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, Bowdle, Wolff, & Boronat, 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*).

References

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

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

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

Blacoe, W., & Lapata, M. (2012). A comparison of vector-based representations for semantic composition. In J. Tsujii (Ed.), Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning (pp. 546–556). Stroudsburg, PA: Association for Computational Linguistics (ACL). 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. (2005). The career of metaphor. *Psychological Review*, *112*(1), 193–216. https://doi.org/10.1037/0033-295X.112.1.193

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

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., & 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.

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

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

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. In P. M. Vogel, & B. Comrie (Eds.), *Approaches to the Typology of Word Classes* (pp. 65–102). New York: Mouton de Gruyter. https://doi.org/10.1515/9783110806120.65

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

Croft, W. (1993). The role of domains in the interpretation of metaphors and metonymies. *Cognitive Linguistics*, *4*(4), 335–370.

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, *14*(2), 159–190.

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. In Paprotte, W., & Dirven, R. (Eds.), *The ubiquity of metaphor* (pp. 85–119). Amsterdam: John Benjamin.

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

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.

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

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

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)*, Vancouver, Canada. 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

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

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

Gentner, D., & Asmuth, J. (2017). Metaphoric extension, relational categories, and abstraction. *Language, Cognition and Neuroscience*, *34*(10), 1298–1307. https://doi.org/10.1080/23273798.2017.1410560

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

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). Cambridge, MA: Morgan Kaufmann. https://doi.org/10.1016/B978-0-08-051013-2.50018-5

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). Washington, DC: American Psychological Association. https://doi.org/10.1037/11156-009

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. In E. Dietrich, & A. B. Markman (Eds.), *Cognitive dynamics: Conceptual change in humans and machines* (pp. 295–342). Mahwah, NJ: Lawrence Erlbaum Associates.

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*, Austin, TX (pp. 2173–2182). https://arxiv.org/abs/1608.00869

Gibbs, R. W. (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

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

Glucksberg, S., & Keysar, B. (1990). Understanding metaphorical comparisons: Beyond similarity. *Psychological Review*, *97*(1), 3–18.

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*. Chicago, IL: University of Chicago Press.

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.

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

Heine, B. (1997). *Cognitive foundations of grammar*. Oxford, England: 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

Hopper, P. J., & Traugott, E. C. (2003). *Grammaticalization*. Cambridge, MA: 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.

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*, Atlanta, GA (pp. 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). Berlin: Walter de Gruyter.

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). Verbs are more metaphoric than nouns: Evidence from the lexicon. *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.

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

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

Lakoff, G., & Johnson, M. (2008). *Metaphors we live by*. Chicago, IL: 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, England: Oxford University Press.

Lenat, D. B., & Guha, R. V. (1989). *Building large knowledge-based systems; representation and inference in the Cyc project*. Reading, MA: 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.

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

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

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. (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

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

Narrog, H., & Heine, B. (2021). *Grammaticalization*. Oxford, England: 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

Ortony, A. (1979). Beyond literal similarity. *Psychological Review*, *86*(3), 161.

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 Science Society*. Stanford, CA, (pp. 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)*, Doha, Qatar (pp. 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*. Cambridge, MA: MIT Press.

Qiu, M., Castro, N., & Johns, B. (2021). Structural comparisons of noun and verb networks in the mental lexicon. In Proceedings of the Annual Meeting of the Cognitive Science Society, Vienna, Austria (Vol. 43, No. 43).

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). 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.

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

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

Simmons, S., & Estes, Z. (2006). Using latent semantic analysis to estimate similarity. *Proceedings of the 28th Cognitive Science Society*, Vancouver, Canada (pp. 2169–2173).

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*. Philadelphia, PA: John Benjamins Publishing.

Shen, Y. (1989). Symmetric and asymmetric comparisons. *Poetics*, *18*(6), 517–536. https://doi.org/10.1016/0304-422X(89)90010-7

Sweetser, E. (1990). *From etymology to pragmatics*. Cambridge, MA: 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. 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., & Sternbeg, 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

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*, Potsdam, Germany (pp. 30–39).

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. *Oxford research encyclopedia of linguistics*. Oxford, England: Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.325

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., 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., 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

Wilks, Y. (1975). Preference semantics. In E. L. Keenan (Ed.), *Formal semantics of natural language* (pp. 329–348). Cambridge, MA: Cambridge University Press.

Wisniewski, E. J. (1997). When concepts combine. *Psychonomic Bulletin & Review*, *4*(2), 167–183.

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

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). New York: Routledge. https://doi.org/10.4324/9781315782379-203

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supplementary Material

APPENDIX A

Instructions for Experiments 1 and 2

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 is ok to repeat words like "a" and "the").

Important: Please do not translate mechanically, word by word. Instead, think about what the sentence means. Imagine that you are 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 do not 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 do not 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 on that.

Again, *be sure not to repeat any content words*. Of course, it is ok to repeat words like the, a, an, and so forth, 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 min 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, sorted in descending order of proportion of paraphrases excluded. Only meaningful 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, 1217 out of a total of 1493 paraphrases were included.

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	0.83
The lantern agreed	6	9	0	2	17	0.65
The lizard worshipped	7	11	0	1	19	0.63
The mule worshipped	7	11	0	0	18	0.61
The car agreed	8	10	0	0	18	0.56
The lantern worshipped	8	10	0	0	18	0.56
The car worshipped	8	9	0	0	17	0.53
The mule agreed	11	8	0	0	19	0.42
The lantern shivered	14	2	2	0	18	0.22
The lantern limped	15	3	0	1	19	0.21
The car limped	15	1	2	0	18	0.17
The mule cooked	15	0	1	2	18	0.17
The lantern cooked	16	0	1	2	19	0.16
The daughter cooked	15	0	0	2	17	0.12
The car cooked	16	0	1	1	18	0.11
The daughter limped	16	0	2	0	18	0.11
The lantern softened	16	0	1	1	18	0.11
The mule shivered	16	0	2	0	18	0.11
The politician shivered	16	0	2	0	18	0.11
The car shivered	17	2	0	0	19	0.11
The lizard cooked	17	0	2	0	19	0.11
The daughter worshipped	17	0	1	0	18	0.06
The lizard softened	17	0	1	0	18	0.06
The politician agreed	17	0	1	0	18	0.06
The politician cooked	17	0	1	0	18	0.06
The car softened	18	0	1	0	19	0.05
The daughter agreed	18	0	1	0	19	0.05
The daughter shivered	18	0	1	0	19	0.05
The mule softened	18	0	1	0	19	0.05
The politician worshipped	18	0	0	1	19	0.05
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	0.20

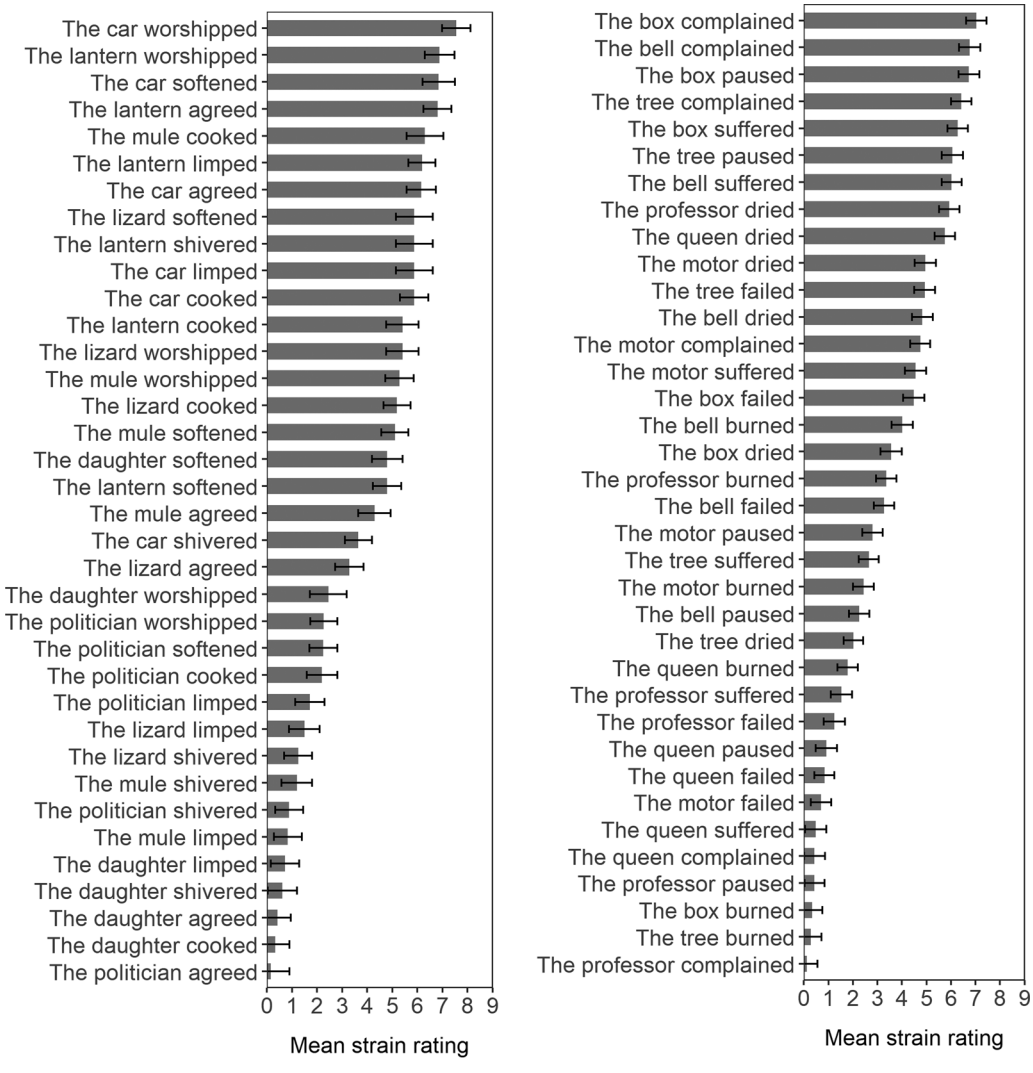
Experiment 2

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	0.18

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.

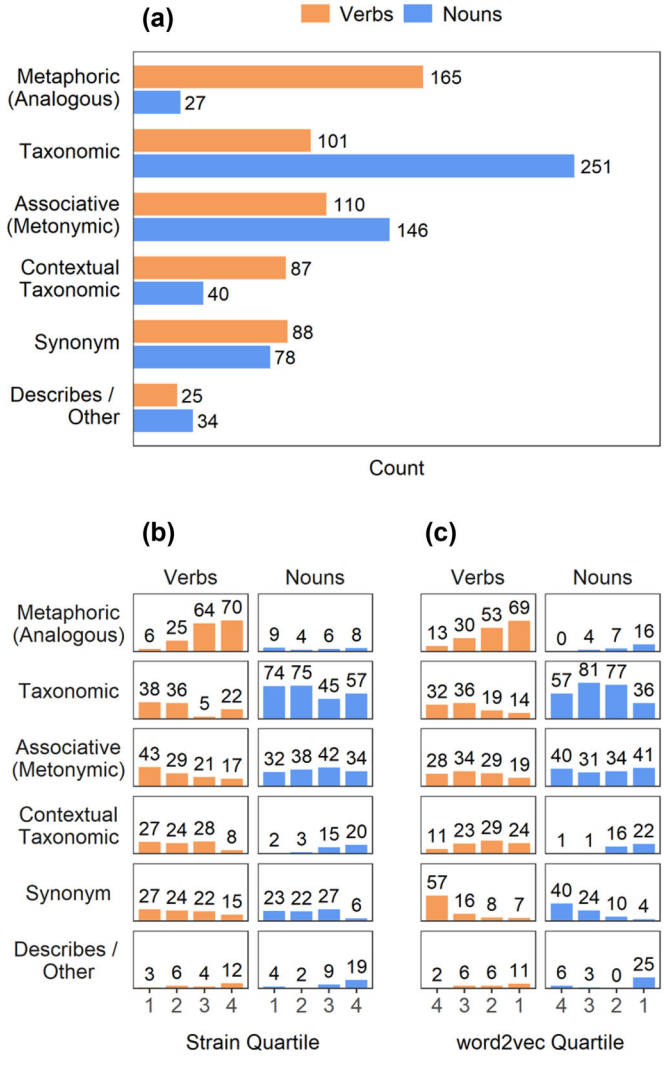
APPENDIX C

Strain rating by items for the 36 items used in Experiment 1 (a) and Experiment 2 (b). 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.



APPENDIX D

Code tallies for all categories from the qualitative analysis in Experiment 3.



(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.