King, D., & Gentner, D. (2019). Polysemy and verb mutability: Differing processes of semantic adjustment for verbs and nouns. In A. K. Goel, C. M. Seifert, & C. Freksa (Eds.). *Proceedings of the 41st Annual Conference of the Cognitive Science Society (pp. 2011–2017)*. Montreal, QB: Cognitive Science Society.

Polysemy and Verb Mutability: Differing Processes of Semantic Adjustment for Verbs and Nouns

Daniel King (dking@u.northwestern.edu)
Department of Psychology
Northwestern University

Dedre Gentner (gentner@u.northwestern.edu)
Department of Psychology
Northwestern University

Abstract

Previous research has found that verbs are more likely to adapt their meaning to the semantic context provided by a noun than the reverse (verb mutability). One possible explanation for this effect is that verbs are more polysemous than nouns, allowing for more sense-selection. We investigated this possibility by testing polysemy as a predictor of semantic adjustment. Our results replicated the verb mutability effect. However, we found no evidence that polysemy predicts meaning adjustment in verbs. Instead, polysemy was found to predict meaning adjustment in nouns, while semantic strain was found to predict meaning adjustment in verbs (but not nouns). This suggests that processes of meaning adjustment may be different for nouns vs verbs.

Keywords: polysemy, mutability, computational linguistics, word2vec, semantics

Introduction

A remarkable aspect of language is that it is both stable enough to reliably convey meaning and flexible enough to accommodate unusual or semantically-strained utterances. For example, the sentence "The hostess galloped to the door" is a bit odd, but we can readily understand it as meaning "The hostess moved rapidly and somewhat gracelessly." While overt metaphorical language has been studied extensively, there is much less work on how people resolve semantically-strained utterances of this type, which may be far more prevalent than traditional "X-is-a-Y" metaphors.

Gentner and France (1988) found that, in paraphrases of simple intransitive sentences of the form *The noun verbed*, participants tended to adjust verb meaning to a greater extent than noun meaning – a phenomenon termed the *verb mutability effect*. *Mutability* can be defined as the degree to which a word's semantic interpretation differs across contexts. The verb mutability effect was found to be strongest when the stimulus sentence was semantically strained – that is, when the noun was incompatible with the paired verb's expected argument, resulting in a nonliteral sentence. For example, one participant paraphrased *The lizard worshipped* as *The reptile stared unblinkingly at the sun*, largely preserving the meaning of the noun *lizard* while shifting the meaning of the verb *worshipped* dramatically.

Little research has examined the processes that drive mutability – that is, how semantic structures are altered during these types of adjustments. In an initial investigation, Gentner and France (1988, Experiment 3b) proposed that verbs are adjusted in a graduated manner, by altering the domain-specific aspects of meaning just as far as is necessary to render a meaningful interpretation – a process they called *minimal subtraction*. We refer to accounts like these as *online adjustment* accounts, as they involve the adjustment of meaning *in situ,* constrained by the context provided by the noun.

Another possibility, however, is that mutability is simply a matter of selecting an appropriate alternate meaning from a word's extant senses. There is evidence suggesting that verbs are more polysemous than nouns across all frequency levels (Gentner, 1981). Thus, higher verb mutability may simply be due to there being more available senses to choose from. We refer to this account as the *sense-selection* view.[1]

Indeed, Gentner and France did not control for polysemy in their original study. We evaluated the polysemy of their stimuli by counting the number of synsets (i.e., senses or meanings) for each word in WordNet 2.1 (Miller, 1995).[2] Our analysis found that the verbs from their study were significantly more polysemous than the nouns, $M_V = 4.13$, $SD_V = 2.17$, $M_N = 2.25$, $SD_N = 1.39$, $t(14) = 2.06$, $p = .03$ – leaving open the possibility that the greater verb mutability they observed was due to the relatively higher polysemy of the stimulus verbs.

Thus, a more precise characterization of the processes underlying these types of semantic adjustments is needed – specifically, the extent to which sense-selection and/or online adjustment drive mutability needs to be better understood.

To investigate this question, we tested polysemy as a predictor of mutability. If polysemy is found to predict mutability, it would provide evidence for the sense-selection account of meaning adjustment. If no such relationship is found, this would instead favor the online adjustment view.

[1] Our descriptions of sense-selection and online adjustment are similar (but not identical) to sense-selection and sense-creation as discussed by Gerrig (1989).

[2] We chose WordNet 2.1 over newer versions due to concerns of a proliferation of synsets in later iterations.

In addition, we seek to understand whether the processes employed vary by word class.

This study follows the paradigm established by Gentner and France. Participants were asked to paraphrase intransitive sentences of varying levels of semantic strain. These sentences were generated by combining 6 nouns and 6 verbs for a total of 36 stimulus sentences (see Figure 1).

For verbs, two expected a human argument (*complain, suffer*), two expected a dynamic artifact object artifact (i.e., a man-made object that functions in some way) as an argument (*pause, fail*), and two expected a static inanimate object as an argument (*dry, burn*). For nouns, two were human (*professor, queen*), two were a dynamic artifact object (*motor, bell*), and two were static inanimate objects (*tree, box*). Combinations in which the noun was incompatible with the verb's expected argument resulted in semantically-strained sentences (e.g., *The box suffered*), while those that were compatible resulted in unstrained sentences (e.g., *The professor complained*).

Half the nouns and verbs used were highly polysemous (at least 10 senses) and half were low in polysemy (1-2 senses; see Figure 1). This resulted in both "balanced" combinations, where the noun/verb polysemy matched—both high (N+/V+) or both low (N-/V-)—and "unbalanced" combinations, where the noun/verb polysemy differed greatly (N+/V- or N-/V+). Thus, across the 36 stimulus sentences, every possible combination of high- and low- polysemy nouns and verbs was realized.

Assessing Meaning Adjustment

A thorny issue in this research is how to quantify meaning adjustment. Gentner and France, using human raters, approached this from three different angles. Across these techniques, they obtained converging evidence for the verb mutability effect; however, each method had drawbacks.

In their *divide and rate* method, raters were asked to divide each paraphrase into *the part that came from the noun* (in the stimulus sentence) and *the part that came from the verb*. They then rated the similarity of each part to the original word. This was problematic for several reasons. It was time-consuming and labor-intensive, and judges often had difficulty deciding how to properly divide the sentence, resulting in a high amount of data loss. Worse, in some cases, some words in the paraphrase were clearly affected by both the original verb and noun, making division impossible. For example, consider the following paraphrase of *The motor complained*: *The badly-functioning engine let out a strange noise from its exhaust*. Here, *badly-functioning* modifies the noun in a context-specific manner (i.e., a motor can function badly but a rock cannot), but it also seems to owe its presence in the paraphrase to the original verb *complained*. The same case can be made for the phrase *from its exhaust.*

Therefore, a way to assess semantic change without dividing paraphrases into noun- and verb-originating components is necessary. Gentner and France employed two such methods: a *double paraphrase* and *retrace* task.[3] While both these methods mirrored the results of the divide-and-rate approach in finding verb mutability, they were similarly labor intensive.

In an attempt to address these issues, we used word2vec (Mikolov et al., 2013) to assess meaning adjustment. This allowed us to quantify semantic change by comparing each paraphrase as a whole to the initial noun and initial verb, without having to divide the paraphrases into components. This provided a hands-off approach that the reduced the time and labor costs of using human judges, as well as data-loss due to low inter-rater agreement. In addition, we hoped to obtain a finer-grained quantification of adjustment than was possible with Gentner and France's methods.

Against these advantages, however, we must ask whether vector-space word embedding models (WEMs) like word2vec can adequately capture human similarity judgments. We next describe these models and discuss issues in using them to assess similarity.

Vector Space Word Embedding Models

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 co-occurrence patterns in the corpus. Thus, each word's meaning is represented as a point in an n-dimensional vector space. The relatedness between any two words is calculated by taking the cosine of the angle between their two associated vectors, resulting in a score between -1 and 1. Scores closer to 1 indicate high levels of relatedness, and scores closer to 0 indicate low levels of relatedness.

While promising in some areas, the evidence regarding WEMs' ability to approximate human similarity judgments is mixed. Latent Semantic Analysis (LSA; Landauer & Dumais, 1997) has been shown in a number of studies to match human judgments of similarity fairly well in certain contexts (Günther et al., 2016; Landauer & Dumais, 1997; Landauer et al., 1998). In addition, previous work has used it as a measure of semantic change over time (Sagi et al., 2011). Other studies, however, suggest that it fails to approximate human intuition, both in literal similarity judgments (cf., Simmons & Estes, 2006), and in relational similarity tasks (Chen et al., 2017) . That the vectors used in WEMs lack explicit relational structure calls into question whether these problems are fully surmountable.

Perhaps the deepest problem lies in the fact that WEMs do not provide a pure measure of similarity, as associations can

[3] In the double paraphrase task, a new group of participants paraphrased the original paraphrases, and any reoccurrences of the original nouns and verbs were scored. There were higher rates of reoccurrence for nouns, indicating greater meaning preservation in the paraphrase. In the retrace task, a new group of participants were given a set of paraphrases, as well as a list of the original eight nouns or verbs used, and asked to guess which noun or verb they thought had occurred in the stimulus sentence. They showed higher accuracy for nouns, indicating greater meaning preservation.

Figure 1: Stimulus nouns and verbs. Shaded cells indicate combinations that result in strained sentences. Pluses and minuses indicate high or low polysemy, respectively. For example, - / + indicates a low-polysemy noun and high-polysemy verb combination, while + / - indicates a high-polysemy noun and low-polysemy verb combination.

			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	10	+ / -	+ / +	+ / -	+ / +	+ / -	+ / +
Static Inanimate	tree	2	- / -	- / +	- / -	- / +	- / -	- / +
	box	10	+ / -	+ / +	+ / -	+ / +	+ / -	+ / +

also influence their scores. For example, the words *cow* and *milk* cooccur frequently, resulting in a high cosine similarity score, despite the obvious fact that a cow is not at all similar to milk.

Thus, we consider the present research to be in part an exploration of WEMs' efficacy in this domain. Future work will involve comparing our word2vec results with human judgments. For now, we will provisionally assume that they can be used as an *approximate* assessment of similarity. We chose word2vec based on evidence that it outperforms other WEMs in approximating human similarity judgments in humans (Pereira, et al., 2016).[4,5]

Method

Participants

112 undergraduates completed the study in person at the lab. One participant was excluded for not being a native speaker of English, one was excluded for providing nonsensical answers to all questions, and two were excluded for failing the catch-trial criteria of repeating a noun and/or verb on both catch trials, for a net of 108 participants.

Materials

6 nouns and 6 verbs were used to generate a total of 36 intransitive sentences. Half of the nouns and half of the verbs used were highly polysemous, and half were low-polysemy. Polysemy was evaluated by counting synsets in WordNet 2.1, excluding any that referred to actual people, places, or events.

The shaded cells in Figure 1 indicate the combinations in which the noun does not satisfy the verb's expected argument, resulting in a semantically-strained sentence (e.g., *The bell suffered*). The unshaded cells indicate those combinations where the noun is compatible with the verb's selectional restrictions, resulting in an unstrained sentence that is literally interpretable (e.g., *The professor complained*).

Procedure

Participants were university students who completed the study on a computer. They saw sentences one at a time and were told to paraphrase each sentence without repeating any of the original content words. They were asked to aim for a plausible interpretation of what the speaker meant, rather than a mechanical, word-by-word translation—e.g., to paraphrase *The slimy senator* as something like *The corrupt politician* rather than *The gooey politician*.

So that each participant saw each noun and verb exactly once, the 36 total stimuli sentences were divided into 6 different assignment factors of 6 sentences. Each assignment factor contained two strained and four unstrained sentences. Sentences were presented in randomized order. In addition, two catch trials were included. The catch trials were simple unstrained sentences designed to test for attention and following directions; the criteria for excluding a subject was repeating a content word in both of the catch trial paraphrases or any obviously nonsensical answers in either.

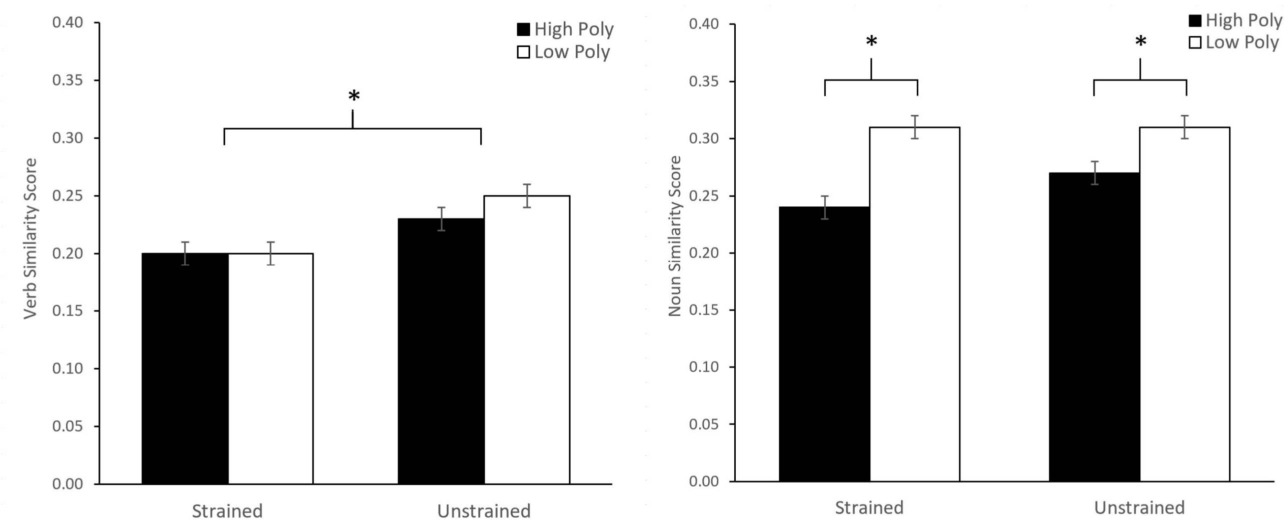
Assessing Semantic Adjustment

For each paraphrase, word2vec was used to obtain two similarity scores, representing the amount of semantic adjustment the initial noun or verb underwent, respectively. The following procedure was used. First, separate normalized vectors were derived for each initial noun and verb. Next, a

[4] We used pretrained vectors available from Google, which were trained using the CBOW method on part of the Google News corpus (about 100 billion words), available at https://code.google.com/archive/p/word2vec/.

[5] We have also begun analyzing our results following the method outlined by Sagi (in press) for using LSA and other WEMs.

Figure 2. Noun and verb similarity scores. Note that lower scores indicate greater semantic adjustment.



A. Verb similarity scores

B. Noun similarity scores

vector for each paraphrase was generated by averaging its normalized component word vectors.[6] Then the similarity of each paraphrase to the initial noun and to the initial verb was computed by taking the cosine of the angle between the vector representing the initial noun/verb and the entire paraphrase vector.

Any words not found in the corpus were skipped (along with any stop words like *the, a*, etc.). If none of the words in the paraphrase were present in the word2vec dictionary, a null vector was generated. Any paraphrases generating null vectors were discarded (this only occurred twice).

Coding

Certain types of responses were excluded from analysis. First, blatantly noncompliant responses (e.g., paraphrasing *The box dried* as just *fruit*) were excluded. Second, responses that did not constitute a meaningful interpretation of the stimulus noun and verb were excluded as well. This included responses that described the context suggested by the stimulus sentence (e.g., paraphrasing *The tree shivered* as *It is cold outside*), as well as any mechanical, word-by-word paraphrases of strained sentences (e.g., paraphrasing *The box complained* as *The object was frustrated*). For unstrained sentences (which are literally interpretable), a word-by-word paraphrase *is* a meaningful paraphrase (e.g., paraphrasing *The professor paused* as *The teacher stopped*) and therefore were not discarded in these cases.

Two human coders were used. Each coder was presented with the original sentence and its paraphrase and was asked to code each paraphrase as described above, resulting in the exclusion of 137 paraphrases. Cohen's κ was run to determine interrater reliability. There was moderate agreement between the two judges, $\kappa = 0.60$, (95% CI, 0.52 to 0.68), $p < .0001$.

Analysis and Results

The 108 participants generated a total of 648 paraphrases. 137 paraphrases were discarded after coding, leaving a net of 511 paraphrases. All analyses were conducted in R (R Development Core Team, 2008) using the *lmer* package (Bates, Mächler, et al., 2015). Fixed-effect hypothesis tests were conducted using a Satterthwaite approximation for degrees of freedom (Luke, 2017).

First, in order to test Gentner and France's initial finding – that verbs are more mutable than nouns overall – a difference score for each paraphrase was calculated by subtracting verb score from noun score. Next, a linear mixed-effect model was fit, with the difference score as the dependent measure, the intercept (mean) as the fixed effect, and random intercepts for subject and item. The mean of the difference scores was found to differ significantly from 0, $t = 2.99$, $p = .01$, indicating that, on average, verbs (M = 0.23, SD = 0.11) changed their meaning significantly more overall than nouns did (M = 0.28, SD = 0.13; lower similarity scores correspond to greater amounts of semantic adjustment).

Next, to test for effects of semantic strain and polysemy, two additional linear mixed models were fit: one for nouns and one for verbs. In both models, similarity score was the dependent measure, and polysemy (high vs. low), strain (strained vs. unstrained), and the interaction term were included as fixed effects. Both models were initially fit with random slopes and intercepts for subjects and random

[6] These sentence vectors can be viewed as representing the "average meaning" of all the words they contain (Landauer, et al., 1998).

intercepts for items. The random effects structure was then simplified as far as necessary as described in Bates, Kliegl, et al. (2015).

For verbs, the effect of semantic strain was significant, β = -0.18, SE = .08, F = 4.90, p = .03, with verb meaning being adjusted to a greater extent in the strained condition (M = 0.20, SD = 0.09) than in the unstrained condition (M = 0.24, SD = 0.11). There was no significant effect of polysemy, F = 0.98, p = .33, and the interaction was not significant, F = 0.62, p = .43). These results are shown in Figure 2a.

For nouns, there was no significant effect of semantic strain, F = 0.11, p = .74. A significant main effect of polysemy was found, β = -0.19, SE = .06, F = 8.95, p = .01, with high-polysemy nouns (M = 0.25, SD = 0.15) being adjusted to a greater extent than low-polysemy nouns (M = 0.30, SD = 0.11). The interaction was not significant, F = 0.60, p = .44. These results are shown in Figure 2b.

Discussion

There were three objectives in the present research: (1) to replicate Gentner and France's finding that verbs are more mutable than nouns under conditions of semantic strain, using new materials and a new method of assessment; (2) to better understand the processes that drive semantic adjustment; and, (3) to investigate possible noun-verb differences in these processes.

The results regarding objective (1) were as predicted: on average, across conditions, participants adjusted verb meaning significantly more than noun meaning. In addition, verbs (but not nouns) were adjusted to a greater extent in strained contexts than in unstrained contexts (see Figure 2a). Both these results replicate Gentner and France's findings and provide additional evidence for the verb mutability effect: during sentence interpretation, the verb's default meaning is more likely to be adjusted to fit the context provided by the noun, rather than the reverse – especially under semantic strain.

More surprising were the results regarding objectives (2) and (3). Polysemy significantly predicted meaning adjustment in nouns, but not verbs; and semantic strain predicted adjustment in verbs, but not nouns.

This leads to the intriguing conclusion that the processes driving semantic adjustment vary by word class. That polysemy predicted noun adjustment favors the sense-selection view. That it did *not* predict verb adjustment is evidence that their increased mutability is not a matter of having more senses to choose from; rather, online adjustment is taking place. Indeed, a qualitative examination of the paraphrases supports this explanation. For example, nouns were frequently paraphrased as close synonyms (e.g., *tree* as *plant* or *oak*; *box* as *container*), while verbs were frequently adjusted to express meanings that were outside the word's existing set of senses (e.g., paraphrasing *The box complained* as *The container couldn't hold all of its contents*).

What explains the noun polysemy effect?

That semantic strain predicted meaning adjustment in verbs but not nouns is consistent with our prediction that verbs are the locus of change in resolving strained utterances. What is more surprising is the effect of polysemy in driving meaning adjustment for nouns. Why did participants consistently adjust high-polysemy nouns to a greater extent than low-polysemy ones—even in unstrained contexts, where no significant adjustment was necessary? We propose three possible explanations.

1. Higher polysemy allows for more creativity. The first possibility is that higher polysemy granted participants more freedom of interpretation, allowing them to choose a more distant meaning than was available with low-polysemy nouns. We believe this to be unlikely. Examining the paraphrases suggested that, regardless of polysemy (or strain), participants usually attempted to choose a meaning as close to the original noun as possible (unlike with verbs, whose meaning was often changed dramatically). For example, it's not clear that, in substituting *container* for *box* (a high-polysemy noun), one has attempted to adjust meaning further than when one substitutes *oak* for *tree* (a low-polysemy noun). The similarity scores for each pair, however, are 0.12 and 0.80 respectively – a relatively large difference.

2. The results reflect a problem with word2vec. A second possibility is that the observed effects of polysemy are simply an artifact of word2vec and don't reflect actual human intuitions. In all WEMs, the meaning of a word derives from the contexts it appears in. A more polysemous word is likely to appear in a wider variety of contexts than a less polysemous word, rendering it less similar, on average, to any one of those meanings (cf., Beekhuizen et al., 2018).

3. High-polysemy words are less similar to their synonyms than low-polysemy words are. A third possibility is that the relationship between higher polysemy and lower similarity scores reflects a psychologically real pattern: namely, that the more polysemous a word is, the less similar it is, on average, to any one of its synonyms. In this account, polysemy significantly predicted adjustment in nouns because, for a high-polysemy word, any synonym one replaces it with will, on average, be less similar to the original word than when the same is done for a low-polysemy word, despite an equal intention to preserve meaning. That is, the subjective similarity between synonyms of a given word is lower for high-polysemy words than for low-polysemy words. If so, the difference in word2vec scores between *box-container* and *tree-oak* reflects a real psychological difference.

To decide between these latter two possibilities, we conducted a preliminary study with human raters. The results suggest that our WEM results do match human intuitions. We asked raters (N=18) to rate the similarity of eight nouns and verbs (drawn from Gentner & France, 1988) to their closest synonyms, as determined by a thesaurus (Lewis, 1978). Each base word was paired with three synonyms as well as one antonym as an attention check. Participants rated the similarity between each base word/synonym pair on a scale of 1 to 10, resulting in 865 target ratings. A linear mixed-effects model analysis was conducted, with human similarity rating as the dependent measure, polysemy of the base word as the fixed effect, and random intercepts and slopes for subject and random intercepts for item. A significant negative correlation between polysemy and similarity rating was found, β = -0.20, SE = 0.09, F = 5.63, p = .02.

Thus, we found evidence in favor of our third explanation: on average, the more polysemous a word was, the less similar it was considered to be to its synonyms. In this way, the human findings paralleled our results with word2vec. If this pattern generalizes across other materials, it will be important to understand the reasons for this converging result in humans and in WEMs.

Do Noun and Verb Change Mean the Same Thing?

There are several outstanding issues to acknowledge before concluding. First, an important question is whether semantic distance means the same thing for nouns as it does for verbs. In other words, are the two scales commensurable? WEMs like word2vec are blind to syntactic category and thus employ the same method of generating and comparing vectors for both nouns and verbs (and all word classes). But whether humans judge similarity between nouns on the same dimensions that they do for verbs is unclear.

Similarly, whether polysemy means the same thing for nouns and verbs is also uncertain. It is possible that verb meanings are extended differently than noun meanings, resulting in qualitatively different patterns of relatedness among senses. At present, little work has examined this issue.

Lastly, one might question whether there is a circularity in assessing mutability using word2vec. As with polysemy, if verbs are more mutable than nouns overall, they likely appear in a wider variety of contexts than nouns. Thus, the vectors for any two verbs may, on average, be further apart than is the case for any two noun vectors.

These objections are important and demand further investigation. At the same time, there are striking qualitative differences in the manner of adjustment for nouns versus verbs. As noted earlier, nouns are often paraphrased as close synonyms, whereas verbs are often extended in quite novel ways. This suggests that the verb mutability phenomenon is not simply a difference in similarity scales, but reflects a qualitative difference in processing.

Conclusion

There are three main findings. First, we replicated Gentner and France's (1988) results: verbs changed their meaning more than nouns overall, and did so to a greater extent in a strained context. Second, we found evidence that both sense selection and online adjustment processes drive mutability. Third, we found that these processes differ between nouns and verbs. Semantic adaptation to context appears to be driven by sense-selection for nouns, but by online adjustment for verbs.

We also presented initial evidence that the relationship between polysemy and meaning adjustment may reflect a property of polysemous words; namely, that higher-polysemy words are, on average, perceived as less similar to their synonyms than low-polysemy words are.

Our results invite a number of future research directions. First, the number of items used in this study is small. We are currently testing new word sets. Future work will also involve more systematic testing of specific verb classes to examine how well the results observed here generalize. Second, we plan to compare the WEM results with human judgments of similarity. Our ultimate goal is to provide a clearer characterization of the processes underlying semantic adjustment in nouns and verbs.

Acknowledgements. We thank Sid Horton, Eyal Sagi, and Phil Wolff for their help and advice.

References

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Beekhuizen, B., Milic, S., Armstrong, B. C., & Stevenson, S. (2018). What Company Do Semantically Ambiguous Words Keep? Insights from Distributional Word Vectors. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, 6.

Chen, D., Peterson, J. C., & Griffiths, T. L. (2017). Evaluating vector-space models of analogy. *ArXiv:1705.04416 [Cs]*. Retrieved from http://arxiv.org/abs/1705.04416

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

Gentner, D., & France, I. M. (1988). Chapter 14 - 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). https://doi.org/10.1016/B978-0-08-051013-2.50018-5

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

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

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. 30.

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

Lewis, N. (1978). *The new Roget's thesaurus*. GP Putnam's Sons.

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

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ArXiv:1301.3781 [Cs]*. Retrieved from http://arxiv.org/abs/1301.3781

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

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

R Development Core Team. (2008). *R: A Language and Environment for Statistical Computing*. Retrieved from http://www.R-project.org

Sagi, E. (in press). Taming big data: Applying the experimental method to naturalistic data sets. *Behavior Research Methods*. https://doi.org/10.3758/s13428-018-1185-6

Sagi, E., Kaufmann, S., & Clark, B. (2011). Tracing semantic change with Latent Semantic Analysis. In K. Allan & J. A. Robinson (Eds.), *Current Methods in Historical Semantics*. https://doi.org/10.1515/9783110252903.161

Simmons, S., & Estes, Z. (2006). Using Latent Semantic Analysis to Estimate Similarity. *Proceedings of the Cognitive Science Society*, 5.