Journal of Experimental Psychology:
Learning, Memory, and Cognition
2017, Vol. 43, No. 10, 1537–1556

0278-7393/17/$12.00 http://dx.doi.org/10.1037/xlm0000390

Distinguishing Discrete and Gradient Category Structure in Language: Insights From Verb-Particle Constructions

Laurel Brehm and Matthew Goldrick
Northwestern University

The current work uses memory errors to examine the mental representation of verb-particle constructions (VPCs; e.g., *make up* the story, *cut up the meat*). Some evidence suggests that VPCs are represented by a cline in which the relationship between the VPC and its component elements ranges from highly transparent (*cut up*) to highly idiosyncratic (*make up*). Other evidence supports a multiple class representation, characterizing VPCs as belonging to discretely separated classes differing in semantic and syntactic structure. We outline a novel paradigm to investigate the representation of VPCs in which we elicit illusory conjunctions, or memory errors sensitive to syntactic structure. We then use a novel application of piecewise regression to demonstrate that the resulting error pattern follows a cline rather than discrete classes. A preregistered replication verifies these findings, and a final preregistered study verifies that these errors reflect syntactic structure. This provides evidence for gradient rather than discrete representations across levels of representation in language processing.

Keywords: mental representation, sentence processing, illusory conjunction, break point modeling, gradient symbolic computation

Distinguishing Discrete and Gradient Category Structure in Language

Language's hallmark is its flexibility and productivity: In producing a sentence, words are combined in structures to represent graded shades of meaning, conveying novel messages with potentially novel forms. In discrete symbolic formalisms (e.g., Chomsky, 1965), the flexibility of what can be communicated derives from recombination of the categorical elements that make up language: Words and structures are associated with discrete mental representations such that each element exclusively belongs to one Class X or Y (e.g., for words: classes such as noun, verb, or preposition; for structures, classes such as noun phrase, verb phrase, or transitive sentence).

More recent alternative formalisms have appealed to gradience, in which linguistic elements belong to particular classes to varying degrees. In some frameworks in syntactic and semantic theory, structures are associated with a graded probability distribution over classes (e.g., a structure is in Class X with 70% probability, Class Y with 30% probability; see Bresnan & Hay, 2008; Goodman & Lassiter, 2015). In other frameworks, the structural representations themselves are gradient, simultaneously exhibiting properties of both classes (e.g., a structure is represented as a blend of 0.7 X and 0.3 Y; Aarts, 2007; Dowty, 2003; see Smolensky, Goldrick, & Mathis, 2015, for discussion).

The current experiments focus on discrete versus gradient properties in the mental representation of verb-particle constructions (VPCs; e.g., *make up*; *lock up*; *cut up*). We outline a novel paradigm to elicit memory errors for VPCs. The distribution of these memory errors allows us to determine whether VPCs form clear classes associated with a single structure of Type X or Type Y (assumed by a discrete framework) or whether VPCs are better described as a graded cline with properties of both X and Y (consistent with gradient frameworks). We demonstrate statistical techniques that can be used to distinguish the error-making behavior that would result from underlying discrete versus graded representations. Together, these tools are used to demonstrate that although VPCs vary in semantics and syntactic structure, the underlying representation is more consistent with a cline, consistent with formalisms incorporating graded structure.

Semantic and Structural Variability in VPCs

English VPCs have varied syntactic and semantic properties (for a broader review, including discussion of VPCs in other languages, see Dehé, Jackendoff, McIntyre, & Urban, 2002; Jackendoff, 2002; McIntyre, 2007). One key observation is that VPCs are found in two distinct configurations. In one of these configurations, the verb and particle are adjacent (*cut up the meat*), and in the other, the particle is shifted from the verb and placed after the object (*cut* the meat *up*). The acceptability of the two structures varies by item. Sometimes both are acceptable (e.g., as for *cut up*)

This article was published Online First March 13, 2017.

Laurel Brehm and Matthew Goldrick, Department of Linguistics, Northwestern University.

Laurel Brehm is now at the Center for Language Science, Pennsylvania State University.

This research was supported by a grant from the National Science Foundation (NSF; BCS1344269). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF. The authors thank the members of the Gradient Symbolic Computation research team (especially Pyeong-Whan Cho, Geraldine Legendre, Akira Omaki, Paul Smolensky, and Colin Wilson) and the Northwestern SoundLab for helpful comments.

Correspondence concerning this article should be addressed to Laurel Brehm, Center for Language Science, Moore Building, Pennsylvania State University, University Park, PA 16802. E-mail: laurel.brehm@gmail.com

whereas sometimes there is a clear preference for the adjacent structure over the shifted one (compare *He will make up the story* to *He will make* the story *up*).

Differences across VPCs in the degree to which adjacent versus shifted structures are preferred imply that they may be represented with different underlying syntactic structures. The logic is that if the verb and particle were represented as a single syntactic element, then the particle would be required to appear in the verb-adjacent position. In contrast, if the verb and particle were represented as separate elements, then the particle would be free to shift after the object or could remain adjacent to the verb. Syntactic theory has proposed distinct analyses along these lines (see Figure 1 for a coarse schematic). In one analysis (Figure 1, left), the verb and particle are considered a single constituent, with the verb and particle comprising a "Complex Head" (e.g., Johnson, 1991; Stiebels & Wunderlich, 1994). In an alternative analysis (Figure 1, right), the particle is fully separate from the verbal head, having the option to form a "Small Clause" with the VPC's object (e.g., den Dikken, 1995; Hoekstra, 1988). Current syntactic frameworks propose that both representations exist in English, dividing VPCs into two or more separate classes with different syntactic structures and associated differences in meaning and/or historical origin (Punske, 2013; Wurmbrand, 2000; Zeller, 2002).

The use of the adjacent versus shifted configuration is also associated with semantic properties. Critical for the present work is semantic compositionality, or the degree to which the meaning of the whole VPC reflects the meaning of its component parts. High compositionality supports particle movement, facilitating the availability of the shifted structure. This facilitation has been demonstrated using multiple behavioral techniques (e.g., masked priming and self-paced reading; Gonnerman, 2012; Gonnerman & Hayes, 2005) and has also been observed in corpora (Lohse, Hawkins, & Wasow, 2004).

Under the assumption that there are two clearly separated categories of VPCs associated with distinct semantic and syntactic properties, we would expect that ratings of compositionality would be bimodal. This is not the case; compositionality ratings seem to fall in a cline rather than discrete clusters (see Figure 1 in Gonnerman & Hayes, 2005; Schnoebelen & Kuperman, 2010). Adding a third level to the compositionality predictor improves the ability to account for the frequency of shifted structure use (e.g., Gonnerman & Hayes, 2005), but even here, it remains clear that other factors contribute to the structural preferences of particular VPCs. To uniquely divide different VPCs into Complex Head versus Small Clause structures, as required by a discrete analysis, current formal frameworks still need to appeal to a range of semantic, structural, and pragmatic factors to appropriately describe individual items. Simply appealing to two distinct structural classes does not itself solve the classification problem (see Gonnerman, 2012; Lohse et al., 2004; McIntyre, 2007).

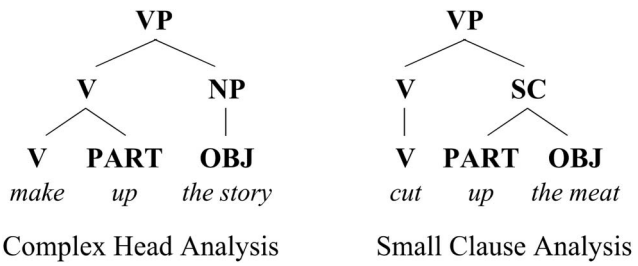

Figure 1. Schematic of VPC structure.

We suggest an alternative approach. Rather than uniquely dividing VPCs into two discrete categories, we propose that VPCs reflect a gradient cline of association with these two syntactic representations (mediated by semantic properties including semantic compositionality). At one end of the cline, highly compositional VPCs (*cut up*) are strongly associated with the Small Clause structure and weakly associated with the Complex Head structure; at the other, noncompositional VPCs (*make up*) are strongly associated with the Complex Head structure and weakly associated with the Small Clause structure. Other VPCs could lie in the middle of the cline and be associated with both structures to varying degrees. Such VPCs would exhibit a *mixture* of properties from each structure—yielding a range of behaviors and not a simple dichotomy.

Illusory Conjunctions and Representational Structure

To assess whether a given VPC can be associated to varying degrees with two representations—the hallmark of gradient structure—we adapted the illusory conjunction paradigm originating in the visual perception literature (e.g., Prinzmetal & Millis-Wright, 1984; Treisman & Gelade, 1980; Treisman & Schmidt, 1982). The basic phenomenon is that when participants view a field of colored letters (e.g., black Ns and gray Ds), they will sometimes erroneously report seeing one letter in the color of another (e.g., a gray N). The implication is that these recombinations can occur only in the case that features have a cognitive identity separate from the whole, providing evidence for the abstract symbolic structure behind the stimuli (see, e.g., Treisman & Gelade, 1980).

Figure 2 schematizes the underlying processes that elicit such errors. Visual stimuli activate mental representations that correspond to the correct combination of stimulus features as well as representations that share only some of the features. Random variation in processing occasionally allows these illusory conjunctions of separable input features to be retrieved.

It is critical to note that the probability of eliciting illusory conjunctions follows from the representational structure of the source stimuli and the coherence of potential illusory outcomes. Evidence for this in the domain of language comes from illusory conjunctions in spoken and written word perception (e.g., Ali & Ingleby, 2010; Kolinsky, Morais, & Cluytens, 1995; Mattys & Samuel, 1997; Prinzmetal, Hoffman, & Vest, 1991; Prinzmetal, Treiman, & Rho, 1986; Rapp, 1992). Illusory conjunctions reflect recombinations of the sources' distinct linguistic components (e.g., vowel of one source, consonants of the other; Mattys & Samuel, 1997). The current work uses this relationship to assess the underlying structure of VPCs. To make the logic and predictions clear, we first walk through the experimental paradigm and then outline the critical statistical diagnostics for differentiating discrete versus gradient structure.

Illusory Conjunctions of Sentences

We presented participants with sets of written sentences, aiming to elicit the percept of elements moving between sentences. Each

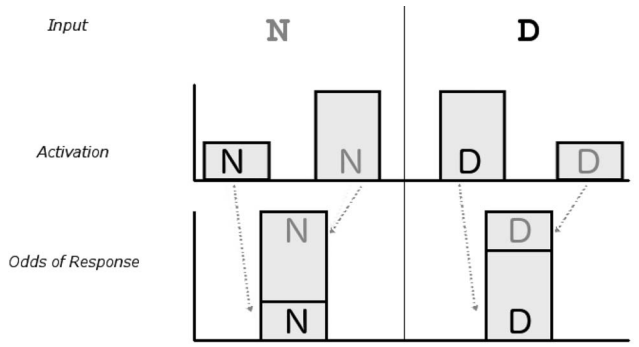

Figure 2. Illusory conjunctions in letter and color features. Inputs most strongly activate representations with the correct conjunction of elements (large bars) but also activate representations that share one element to a smaller degree (small bars). This leads to occasional responses consistent with the wrong analysis.

trial in this paradigm was made of three sentences. The first two were source sentences. One source contained a VPC and would remain a grammatical sentence if the particle were subtracted (e.g., *He will lock up* the bicycle; *lock* is a sensible bare verb) whereas the other source contained a bare verb that would form a grammatical VPC if the particle from the other sentence were added (e.g., *He will cut* the meat; *cut up* is a sensible VPC). After reading the source sentences, participants were presented with a test sentence and were asked to judge whether the test matched one of the two sources. This test sentence might be a repetition of one of the source sentences or, critically, might be a version of a source sentence with the particle subtracted (*He will lock the bicycle*) or with the particle from the other source added (*He will cut up the meat*).

We used the accuracy and speed of the test sentence judgments to delineate the underlying structure of the source stimuli. The critical logic mirrors the findings from visual attention and sound perception: If verbs and particles in VPCs are represented separately in the mind (as in the Small Clause structure in Figure 1), then participants will be susceptible to the illusion that the particle occurred with the bare verb; if verbs and particles are represented as a single entity (as in the Complex Head structure in Figure 1), then the illusion should be infrequent. Rates of illusory conjunctions index the association of a VPC to these different representational structures. By assessing the representation of a range of VPCs, we can establish whether the set of VPCs reflects a single gradient dimension (a cline of variability) or multiple discrete classes.

Figure 3 illustrates the logic of these predictions in three examples, focusing on the sentences' verbs and VPCs. In this figure, a bare verb source sentence is represented by a single verb structure (e.g., <make>, <lock>). We consider two possible representations of a VPC source: a Complex Head (<make_up>) versus a Small Clause consisting of two independent elements (≪make>+<up≫).

In the case of noncompositional VPC sources (e.g., *make up*), the VPC is strongly associated with a Complex Head representation (<make_up>). This implies little activation of the bare verb (<make>; Figure 3, bottom) and the independent particle (<up>; not depicted). The lack of activation on the independent particle entails that the Small Clause representation binding the particle to the other verb will also be inactive (≪lock>+<up≫). As such, noncompositional VPCs will elicit few illusory conjunctions, and responses will largely reflect the original sources ("make up;" "lock").

In contrast, fully compositional VPC sources (e.g., *cut up*) are strongly associated with a Small Clause representation (≪cut>+<up≫). In turn, the activation of the independent elements <cut> and <up> activates the alternative representations where the particle is bound to another verb (e.g., ≪lock>+<up≫; Figure 3, top). Therefore, fully compositional sources will lead to many illusory conjunctions where the VPC appears as a bare verb ("cut") or where the bare verb appears as a VPC ("lock up").

To test between gradient and discrete accounts of VPC structure, the critical cases involve VPCs intermediate in compositionality.

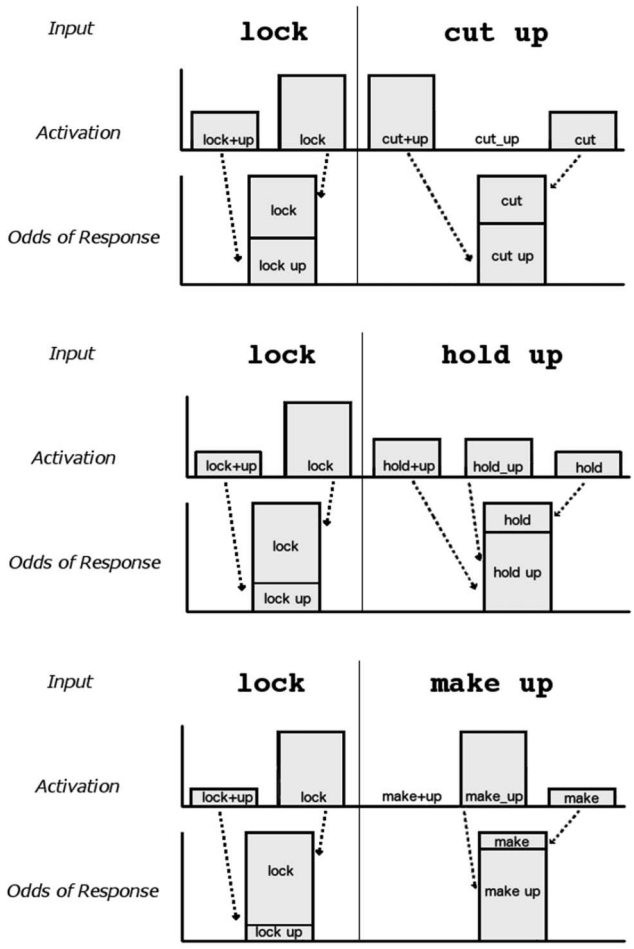

Figure 3. Illusory conjunctions in VPCs. Inputs most strongly activate representations with the correct, appropriately structured conjunction of verb and particle (large bars) but also activate appropriately structured representations that share one element to a smaller degree (small bars), leading to occasional errors (odds of response). Our analysis assumes that VPCs vary in the degree to which they activate single vs. multiple unit VPC representations, with some VPCs activating both structures (middle panel).

Under a gradient account, moderately compositional VPC sources (e.g., *hold up*) might be associated with activation of both the Complex Head and Small Clause representations (see Figure 3, middle). Graded activation of these two representations would lead to equally graded rates of illusory conjunctions, such that as the relative activation of the multiple unit analysis increases, so would the number of illusory conjunction errors. In contrast, under a discrete account, VPCs would be sharply divided such that items lead to either few or many errors with a clear division at some intermediate level of compositionality. This sharp division would reflect the fact that although compositionality might occur along a gradient, the underlying syntactic structure does not. Under such an account, items are associated with either one structure or the other. This would mean that intermediate VPCs such as *hold up* would either lead to many errors, similar to fully compositional VPCs (Figure 3, top), or would lead to few errors, similar to noncompositional VPCs (Figure 3, bottom); no intermediate VPC items should lead to intermediate error rates.

It is important to note that this analysis relies on the assumption that the probability of errors reflects the syntactic structure of VPCs (i.e., the strength of association with the Complex Head vs. Small Clause structure). This assumption is critically examined in our experiments, verifying the independent influence of syntactic structure on the probability of illusory conjunction errors.

Statistically Distinguishing Discrete From Gradient

The success of using the illusory conjunction paradigm to examine discrete versus gradient class structure is also contingent on the ability to statistically distinguish the outcomes of these structures in the resulting data. In statistical terms, this is a problem of establishing the dimensionality of a predictor (e.g., compositionality) to show whether it behaves as a categorical factor with two levels (discrete classes) or as one continuous linear predictor (gradient cline). If there were no noise in the dependent measures, then establishing the predictor's dimensionality would be trivial: Discrete and gradient predictors imply different consequences for behavior. Ideal categorical predictors should show a clear separation in outcomes, leading to easily distinguishable clusters in the data; ideal linear predictors should not. The problem is that in real-world data this pattern is often obscured by noise, requiring a statistical approach.

We propose to use a statistical test to examine whether there is a change in the way the predictor affects the outcome variable at some intermediate value. The premise is that by definition, discrete predictors have a discontinuity in the way the predictor affects the outcome: Discrete categories have a sharp category boundary. The presence of a discontinuity can be quantified using a statistical tool— piecewise regression, also called segmented regression or break point or change point modeling (see Knell, 2009; Muggeo, 2003; and Ruiz, García, Muriel, Andrés, & Ventanas, 2002 for examples of this approach in ecology, epidemiology, and food science, respectively). Such a model fits separate regression lines with break points between them, accurately describing discrete (multiple discontinuous lines) and gradient (one continuous line) data patterns. This statistical technique is robust to random noise and affords similar hypothesis testing and model comparison as standard linear regression, providing a way to assess the meaningfulness of middle levels of semantic compositionality for noisy behavioral outcomes.

Experiment 1

The goal of Experiment 1 was to examine how a continuous measure of semantic compositionality predicted illusory conjunctions involving VPCs and therefore the underlying representational structure of this construction. We used a piecewise regression method to distinguish whether compositionality elicited memory errors in a fashion consistent with two discrete classes or with a single graded cline. To assess the validity of this technique, we first ran Experiment 1a to establish estimates of the relevant effect sizes and then ran Experiment 1b with a larger number of participants drawn from estimates of power to detect null effects observed in Experiment 1a. We walk through the methods and data analysis used for both experiments here.

Method

Construction of the materials and procedure was informed by the results of an in-laboratory pilot experiment that utilized a categorical manipulation of compositionality. A write-up of the pilot study can be found at https://osf.io/z4wya/; anonymized data and analyses can be found at https://osf.io/g6kvc/ and https://osf.io/s6vxh/.

Equipment. The experiment was run on participants' computers in an in-browser script presented with IbexFarm (Drummond, 2013). Participants were instructed to use the F and J keys on their keyboard for "yes" and "no" decisions, respectively. Key-press latencies were collected and sent to the server upon completion of the experiment.

Materials. There were 160 items in the experiment. Items were made up of three sentences: two source sentences and a test sentence that was either identical to one of the two sources or differed by one word (a "foil" sentence). All sentence subjects were pronouns (*he, she, they, we*), and all verbs were presented in the future tense.

In the 32 critical trials, one source contained a VPC and the other contained a bare verb, counterbalanced for position. The VPC source would remain grammatical if the particle were removed (e.g., Source: *He will lock up the bicycle*—> Foil: *He will lock the bicycle*); the bare verb source would remain grammatical if the particle from the other sentence were added (e.g., Source: *He will cut the meat*—> Foil: *He will cut up the meat*). Source VPCs spanned the range of compositionality from Gonnerman and Hayes (2005) and compositionality ratings of the (illusory) foils were all relatively high, above the midpoint of the compositionality scale.[1] The four forms of test sentences (repetitions of the two sources; foils with added/removed particle) were counterbalanced across lists so that each item appeared in each form an equal number of times and each participant saw only one version of each item. See Table 1 for examples; a list of critical trials can be found in the Appendix.

The remaining items were fillers that controlled the predictability of source and test sentences and allowed the general assessment

[1] One of these items ("carry up") was added based upon author judgments—analyses were performed with and without this item and the pattern of results was identical.

Table 1
Example Experimental Stimuli

Item type	Stimulus triad			Test type
	Source 1	Source 2	Test	
Critical	He will **lock up** the bicycle	He will **cut** the meat	He will lock up the bicycle	Matches S1
			He will cut the meat	Matches S2
			He will **lock** the bicycle	**Illusory conjunction**: S1 – particle
			He will **cut up** the meat	**Illusory conjunction**: S2 + particle
Adverb swap	She will finally earn the raise	She will melt the ice	She will finally earn the raise	Matches S1
			She will melt the ice	Matches S2
			She will earn the raise	S1 – adverb
			She will finally melt the ice	S2 + adverb
VPC-noun swap	They will let out the prisoners	They will drive the horses	They will let out the prisoners	Matches S1
			They will drive the horses	Matches S2
			They will let out the horses	S1 + S2 noun
			They will drive the prisoners	S2 + S1 noun
Adverb-noun swap	We will happily advertise the cereal	We will mix the paint	We will happily advertise the cereal	Matches S1
			We will mix the paint	Matches S2
			We will advertise the cereal	S1 + S2 noun
			We will happily mix the paint	S2 + S1 noun
Two-VPC	She will let out the goats	She will phase in the plans	She will let out the goats	Matches S1
			She will phase in the plans	Matches S2
			She will let in the goats	S1 + S2 particle
			She will phase out the plans	S2 + S1 particle
Two-adverb	He will finally wash the socks	He will usually scrub the dishes	He will finally wash the socks	Matches S1
			He will usually scrub the dishes	Matches S2
			He will usually wash the socks	S1 + S2 adverb
			He will finally scrub the dishes	S2 + S1 adverb
Bare noun swap	We will watch the cheetah	We will examine the cannon	We will watch the cheetah	Matches S1
			We will examine the cannon	Matches S2
			We will watch the cannon	S1 + S2 noun
			We will examine the cheetah	S2 + S1 noun

Note. S1 = Source 1; S2 = Source 2. Bold underlined text indicates critical verb-particle phrases and verbs.

of the paradigm's difficulty. Filler source sentences either contained a VPC, a verb with an adverb modifier, or a bare verb; in total, 60% of trials had one or more source sentences with a VPC, 30% had one or more source sentences with an adverb, and 10% of trials had two bare verb sentences. Filler source sentences were paired such that the presence or absence of a particle (or adverb) in the first source sentence had no predictive value for the likelihood of seeing a particle (or adverb) in the second source sentence; both outcomes were equally likely. See Table 1 for example filler items.

Filler test sentences were based upon one of the two source sentences and, similar to the critical test sentences, were either identical to one of the two sources or differed from one of the two sources by one word; these were again counterbalanced across lists so that each item appeared in each test form an equal number of times. Foil test sentences involved changes to particles, adverbs, and nouns and were most likely to involve changes to a particle (40% particle, 25% adverb, and 35% noun changes). Across all sentences, "yes" and "no" responses to test sentences were equally likely.

Procedure. Trials began with a fixation cross presented centrally for 500 msec. Next, the first source sentence appeared centrally for 1,200 msec, followed by a visual mask (a series of hash marks) presented in the same location for 100 msec. This was followed by the second source sentence and a second visual mask, presented for 1,200 and 100 msec, respectively. After these, the test sentence was then displayed with the words "Yes" and "No" below it, with its position offset from the source sentences by half of a line. The test sentence stayed on the screen for 2,500 msec while the participant's response was collected and timed. If no response occurred within 2,500 msec, then the trial was terminated and the feedback "Too slow" was presented to the participant. The intertrial interval was 1,000 msec; a break was offered to participants every 54 trials. Typeface was the participants' browser default in black on a white background.

Design and data analysis. The critical factors in this experiment were source VPC compositionality (ranging from 1.58 to 8.55, with a mean of 4.70) and source particle presence (present/absent). Source sentence position was also entered as a control factor (VPC in first/second source). Source particle presence, source sentence position, and test type were within-item factors, counterbalanced across lists; compositionality was a between-item factor. All of the two-level factors were effects coded (contrasts of −.5, .5), and compositionality was centered. Anonymized data and R analysis scripts are archived on the Open Science Framework: Experiment 1a, https://osf.io/n3dfs/ and https://osf.io/a46bj/; Experiment 1b, https://osf.io/hxgdu/ and https://osf.io/dvkmj/.

Analysis I: Signal detection/linear regression. Analysis was performed in R with the package lme4 (version 1.1–7; Bates, Maechler, Bolker, & Walker, 2014; R Core Team, 2014). Analysis of responses was based upon signal detection theory, building on DeCarlo (1998). Responses were coded for whether they were consistent with the belief that a particle was absent in the test sentence such that correct "yes" responses to tests without particles and incorrect "no" responses to tests with particles were classified as "particle absent" responses. These were analyzed with a logistic mixed-effects regression that separated bias (overall odds of responding particle absent) from discriminability (odds of responding particle absent when appropriate). The interactions of VPC source compositionality with bias and discriminability were examined and source sentence position was entered secondarily as a control predictor. The random effect structure was the maximum justified by the data, assessed using the procedure of Bates, Kliegl, Vasishth, and Baayen (2015).

Analysis of response latencies was based upon a standard linear regression model. Log correct reaction time (RT) was predicted by VPC source compositionality and source particle presence, with source sentence position entered secondarily as a control predictor. Again, the random effect structure was the maximum justified by the data.

Analysis II: Break point analysis. The second stage of data analysis involved searching for a break point in the original response and RT models. The idea is that a piecewise model of an underlyingly discrete predictor will have a markedly higher likelihood as compared with an analogous model done with standard linear or logistic regression, supporting the division of the predictor in multiple classes. In contrast, a gradient predictor would be fit well by a standard linear or logistic regression, with minimal improvement gained from a piecewise link function. Our analyses used models that allow a single break point in one chosen predictor at the level that maximizes model likelihood. If this single break point significantly improves the original model, then there is support for a discrete predictor; if it fails to significantly improve the original model, then a gradient predictor is more parsimonious.

Regardless of model fit, it is also the case that in a piecewise model, idealized discrete and gradient predictors reflect qualitatively different patterns of the piecewise segments' slopes and intercepts. We can capitalize upon this fact in hypothesis testing, as illustrated in Figure 4. For discrete predictors, the two halves of the regression model should have different intercepts but identical slopes that are equal to zero—the pattern apparent in the left side of Figure 4. This would be the case regardless of randomly distributed noise (compare top and bottom left in Figure 4). For gradient predictors, even if we force a model to find a break point, the two "halves" of the regression should have identical intercepts and identical nonzero slopes—apparent in the right side of Figure 4. Again, this would be the case regardless of randomly distributed noise (top and bottom right). This means that by placing a confidence interval (CI) around the differences in intercepts and slopes in the two halves of a piecewise regression model, we can explicitly test support for the contrasting outcomes of discrete and gradient predictors.

To implement the break point analysis, we took the original signal detection/linear regression models from Stage 1, simplifying the random effects structure to random intercepts only to minimize convergence issues due to the increased complexity of these models. A series of piecewise regressions based upon the original models was then performed separately for responses and latencies. This meant running a series of models with a varying break point that dichotomized compositionality into two classes at all possible values between 25% and 75% of its range (3.06–6.09). The break point was coded as a logical statement such a regression model was fit in each of the two segments (left half and right half) on either side of the break point. The model in each of these series with maximum likelihood and full statistical convergence was selected. This procedure corresponds to a grid search for at most one change (AMOC) in a piecewise regression allowing for a discontinuity between the halves (see Crawley, 2007 for a worked example; see http://osf.io/6v3r9 for archived R code).

Following Crawley (2007), we report the regression terms for the right half of the model (intercept, slope estimates for predictors) and the left-half adjustment terms—the difference between the right and left model segments for the intercept and for each slope term. To assess support for discrete versus gradient predictors, Markov chain Monte Carlo profile CIs for the adjustment parameters were calculated, assessing whether slope and intercept differences between the model halves were meaningful. The deviation tolerance for these CIs was set to 1×10^{-6} (vs. the 1×10^{-9} default) because of convergence issues; this is still fairly conservative.

Degrees of freedom were adjusted in the piecewise model by adding an additional penalty for the break because we allowed it to freely vary in the first phase of piecewise model fitting. The adjusted degrees of freedom were then used to perform likelihood ratio tests comparing the break point model to the simpler no-break point version (a nested comparison) to assess whether the more complex model was supported (i.e., Kim, 1994). This is analogous to standard model comparison techniques.[2]

Experiment 1a

Participants

Data were collected from 73 participants recruited through Amazon Mechanical Turk, all of who had IP addresses from the United States and were older than 18 years of age. All participants were compensated $5 for their time, which was approximately 30 min of active participation (excluding breaks). One participant was excluded from the final analysis for reporting learning another language before English and the second run of one participant who chose to do the experiment twice was also excluded. Seven more participants were excluded for an overall accuracy less than 60%, leaving a total of 64 participants contributing data to the final analyses. All experimental procedures and protocols were approved by the Northwestern University Institutional Review Board.

Results

Responses. The mean error rate was 21%. VPC trials elicited the most errors, with the highest error rate in the two-VPC-swap

[2] An alternative technique would be to use an information criterion such as Akaike information criterion (AIC) or Schwartz information criterion (SIC). For these data, patterns are similar with all approaches.

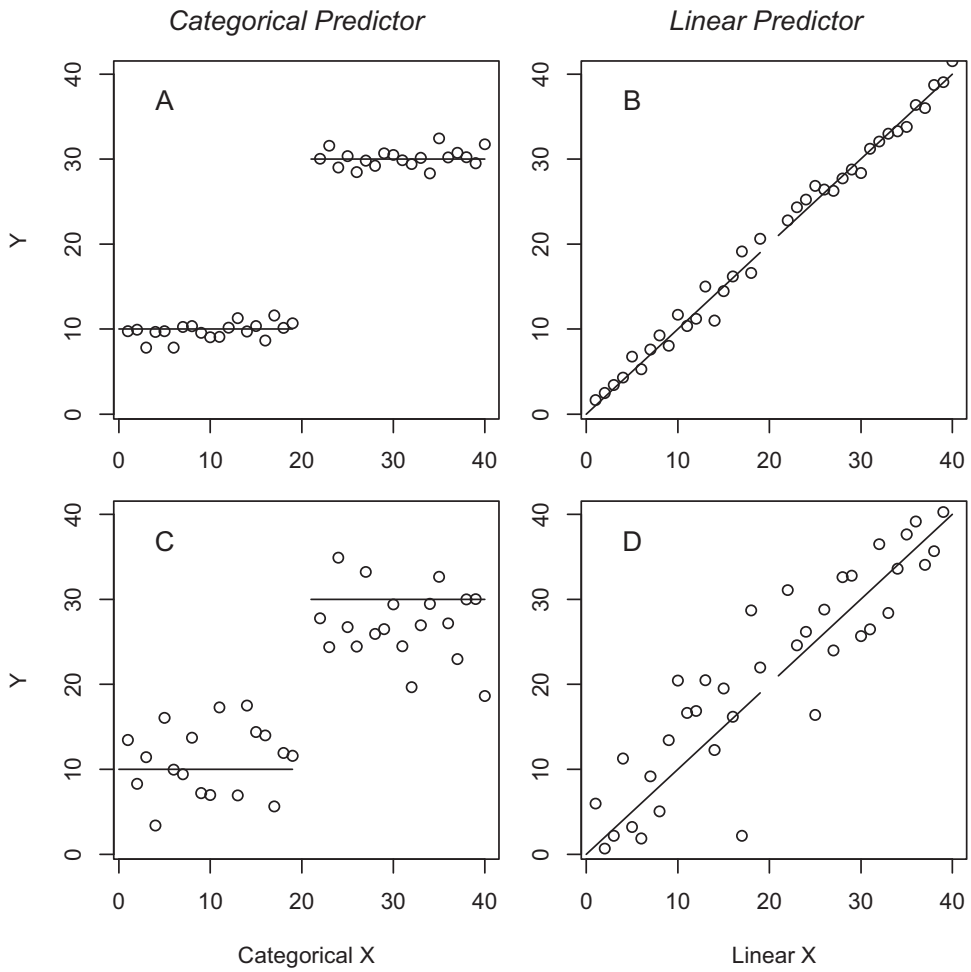

Figure 4. Categorical vs. linear predictors. In both columns, an X value below the mean leads to the same average Y, despite the fact that the link functions are different. The most drastic difference comes at the median, where there is a clear break point in the categorical panels. This holds for low (A and B) and high random noise (C and D).

fillers (29%), followed by the critical trials (26%). This was followed by the adverb-swap fillers (two-adverb = 19%, one-adverb = 19%) and the noun-swap fillers (VPC = 16%, adverb = 18%, bare = 15%).

Signal detection analysis: Illusory conjunctions are more likely with high-compositionality VPCs. In the analysis of critical trial responses (summarized in Table 2), there was an overall bias toward responding "present," with 52% of responses consistent with a "present" belief. Discriminability was also high, meaning that participants experienced illusory conjunctions on a minority of trials. Discriminability was modulated by compositionality of the VPC source such that participants experienced more illusory conjunctions when the particle occurred in a source with higher compositionality. As shown in the top panel of Figure 5, participants were less accurate in reporting that the particle was absent from a bare verb source when it was paired with a high-compositionality VPC versus a low-compositionality VPC. This reflects a higher likelihood of particle intrusion from higher compositionality VPCs. The same occurred for deletions: Particles were likely to migrate away from higher versus lower compositionality VPCs (shown in the bottom panel of Figure 5).

The maximal random effect structure justified by these data was random intercepts by subjects and items only. Adding the control variable of source sentence position and its interactions with source particle presence, compositionality as well as the three-way interaction did not improve model fit,[3] $\chi^2(4) = 1.63$, $p = .80$.

Break point analysis: A gradient, rather than categorical, model best explains the distribution of illusory conjunctions. The maximum log-likelihood piecewise model that met all convergence criteria inserted a break point at a compositionality rating of 4.07. In this model, profile CIs around the adjustment parameters indicated there were no significant differences in intercept or slopes between the two piecewise segments, consistent with a gradient view of compositionality (see Table 2). This was confirmed with model comparison between the break point model (with four adjustment parameters and an additional penalty for the

[3] This model did not converge to tolerance: 0.0019, rather than the 0.001 standard. Removing all nonsignificant effects allows the model to converge to tolerance. This reduced model shows the same overall pattern.

Table 2
Mixed-Effect Model Outputs From Experiment 1a

Model	Estimate	*SE*	*z* value	$p(\chi^2)$
Particle belief analysis				
Zero-break model				
Overall bias	0.14	0.06	2.50	.01
Discriminability	2.64	0.12	23.60	<.001
Source compositionality × Bias	0.00	0.03	0.04	.97
Source compositionality × Discriminability	−0.12	0.06	−2.06	.04
	Estimate (lower, upper CI)	*SE*	*z* value	
Break point model				
Right-half parameters				
Overall bias (intercept)	0.30 (0.08, 0.52)	0.11	2.64	
Discriminability (slope)	2.60 (2.16, 3.04)	0.23	11.53	
Source compositionality × Bias	−0.08 (−0.21, 0.04)	0.06	−1.38	
Source compositionality × Discriminability	−0.12 (−0.36, 0.12)	0.12	−0.99	
Left-half parameter adjustments				
Bias	−0.21 (−0.70, 0.29)	0.25	−0.83	
Discriminability	0.79 (−0.19, 1.79)	0.50	1.57	
Source compositionality × Bias	0.09 (−0.17, 0.35)	0.13	0.67	
Source compositionality × Discriminability	0.38 (−0.13, 0.90)	0.26	1.45	
	Estimate	*SE*	*t* value	$p(\chi^2)$
Log correct RT analysis				
Zero-break model				
Intercept	7.04	0.02	316.2	<.001
Source particle presence	0.02	0.01	2.30	.03
Source compositionality	−0.01	0.00	−2.10	.04
Source particle presence × Source compositionality	0.01	0.01	0.70	.47
	Estimate (lower, upper CI)	*SE*	*t* value	
Break point model				
Right-half parameters				
Intercept	7.07 (6.99, 7.14)	0.04	181.04	
Source particle presence	0.10 (−0.03, 0.22)	0.07	1.48	
Source compositionality	−0.02 (−0.05, 0.01)	0.01	−1.51	
Source particle presence × Source compositionality	−0.01 (−0.07, 0.04)	0.03	−0.48	
Left-half parameter adjustments				
Intercept	−0.01 (−0.08, 0.05)	0.03	−0.43	
Source particle presence	0.10 (−0.24, 0.03)	0.07	−1.51	
Source compositionality	0.02 (−0.01, 0.05)	0.02	1.46	
Source particle presence × Source compositionality	−0.00 (−0.07, 0.05)	0.03	−0.20	

Note. Model comparison was used to calculate *p* values, omitting one factor at a time and comparing models with χ^2 tests.

break) and a model with no break point: adding a break point failed to significantly improve model fit, $\chi^2(5) = 5.70$, $p = .34$

RTs. Across conditions, mean correct responses ranged between 1,100 and 1,200 msec. Correct responses to the two-VPC and two-adverb trials were the slowest (mean for two-VPC trials = 1,197 msec, two-adverb = 1,203 msec), with the other trial types eliciting similar correct latencies (critical trials = 1,147 msec, one-adverb = 1,158 msec, bare noun swap = 1,153 msec, adverb-noun swap = 1,149 msec, VPC-noun swap = 1,146 msec). Error responses tended to be slower but followed a similar pattern (two-VPC $M = 1,239$ msec, two-adverb $M = 1,241$ msec, critical trials $M = 1,232$ msec, one-adverb $M = 1,212$ msec, bare noun $M = 1,198$ msec, adverb-noun swap $M = 1,232$ msec, VPC-noun swap $M = 1,195$ msec).

Linear regression: RTs are lower for high-compositionality VPCs. In the critical trials, correct latencies varied by source particle presence and source compositionality. There was a significant effect of source particle presence such that source particle-present trials were slowest (particle present $M = 1,223$ msec, particle absent $M = 1,197$ msec). There was also a significant effect of compositionality such that responses were faster as the compositionality of the source VPC increased (mean for items below 2.5: 1,214 msec; mean for items above 7: 1,164 msec). This pattern is consistent with a standard speed–accuracy trade-off such that high compositionality led to increased errors and faster responses. There was no interaction between compositionality and particle presence. See Figure 6 for a graphical representation of the data; see Table 2 for mixed-effect analyses.

The maximum random effect structure justified by the data was the full random effect structure (random intercepts by participants and items, random slopes by source particle presence, source compositionality and their interaction for participants, and random slopes by source particle presence for items). Adding the control variable of source sentence position and its interactions did not significantly improve model fit, $\chi^2(4) = 7.34$, $p = .12$.

Break point analysis: A gradient, rather than categorical, model best explains the distribution of RTs. The maximum log-likelihood piecewise regression inserted a break point at

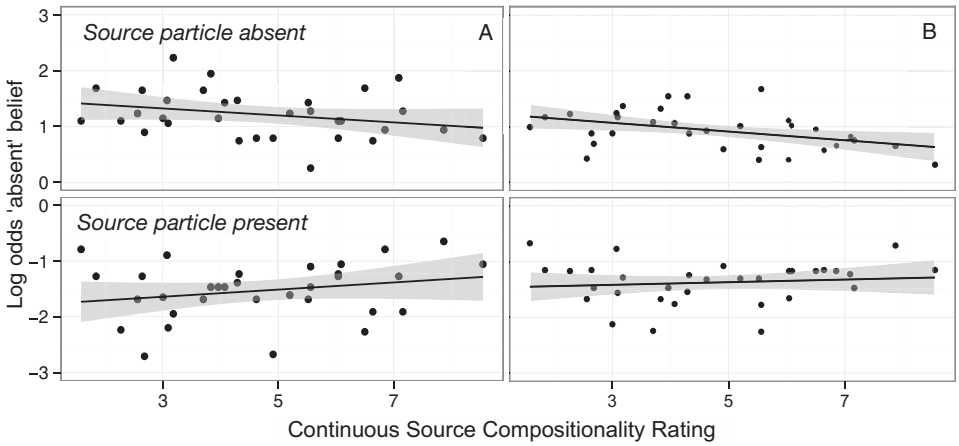

Figure 5. Log odds particle "absent" belief in Experiment 1a (left) and Experiment 1b (right) across items by source particle presence and continuous source compositionality. The solid lines represent the gradient model (with no break points), which was the best-fitting model in Experiment 1a. Confidence bands are for slope estimates.

6.04—a different location from the break point analysis of responses. In this model, profile CIs around the adjustment parameters indicated there were again no significant differences in intercepts or slopes between the left and right halves of the model, consistent with compositionality behaving as a linear predictor (see Table 2). This was confirmed with model comparison between the broken model (with four adjustment parameters and an additional penalty for the break) and a model with no break point. This showed that adding a break point did not significantly improve model fit, $\chi^2(5) = 9.27$, $p = .10$.

Experiment 1b

To verify the applicability of the analysis technique to the present domain, we ran a preregistered replication of Experiment 1a (accessible at https://osf.io/k69cz/). This study was identical to Experiment 1a in every way except an increase in the participant sample size (to $N = 152$) to achieve the best possible statistical power, critical for determining the validity of null effects.

Statistical power was estimated with a series of Monte Carlo simulations. Using the estimated CIs from the zero-break regression models in Experiment 1a, simulations were used to generate sample data at varying participant sample sizes to estimate our ability to recover critical effects. These simulations showed that with 152 participants, the critical interaction in the Experiment 1a response model (between compositionality and source particle presence) could be detected with 80% power. This sample size also provides good power to observe a possible categorical effect. In a second set of simulations, the Experiment 1a break point model was modified to specify a pure categorical effect—in statistical terms, an intercept adjustment for high compositional items only. We then examined the ability of the break point analysis to detect this pure categorical effect with varied effect sizes. These simula-

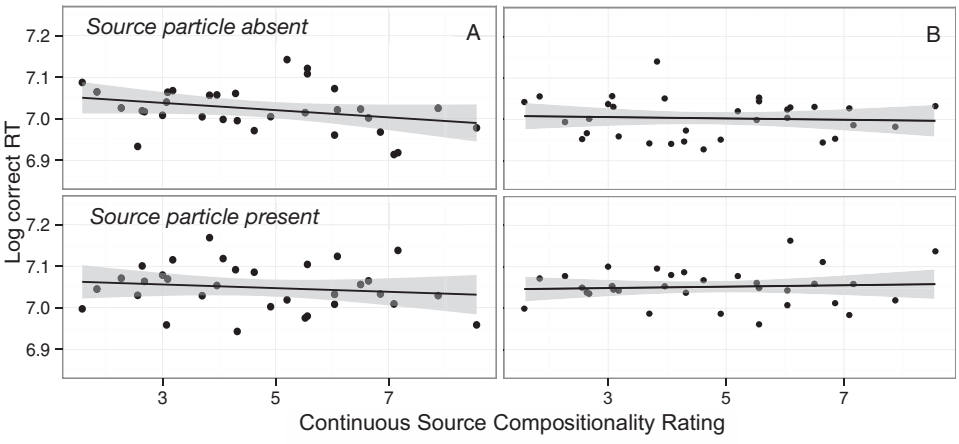

Figure 6. Log correct response latencies in Experiment 1a (left) and Experiment 1b (right) across items by source particle presence and continuous source compositionality. The solid lines represent the best-fitting model in Experiments 1a and 1b, with no break points. Confidence bands are for slope estimates.

tions reveal that with 152 participants, a significant categorical effect roughly equivalent in magnitude to the observed nonsignificant effects in Experiment 1a could be detected with 80% power (Response model: within 1.5 *SD*s; RT model: within 1 *SD*); this projected high power makes it likely that any null results in Experiment 1b will indicate the true absence of an effect.

Participants. Data were collected from 176 participants recruited through Amazon Mechanical Turk, all of who had IP addresses from the United States, were older than 18 years of age, and had not participated in Experiment 1a. As in Experiment 1a, all participants were compensated $5 for their time, approximately 30 min of active participation (excluding breaks). We excluded 23 participants for having an overall accuracy under 60% and excluded the second run of one participant who ran in the experiment twice, leaving a total of 152 participants contributing data to the final analyses. All experimental procedures and protocols were approved by the Northwestern University Institutional Review Board.

Results

Responses. As in the previous experiment, VPC trials elicited the most errors, with the highest error rate in the two-VPC-swap fillers (27%), followed by the critical trials (25%) and the adverb-swap fillers (two-adverb = 15%, one-adverb = 15%). Again, the noun-swap fillers elicited the fewest errors (VPC = 12%, adverb = 14%, bare = 13%). The overall error rate was similar to Experiment 1a (18%, compared with 21%).

Signal detection analysis: Illusory conjunctions are more likely with high-compositionality VPCs. Results of the mixed-effects regression are summarized in Table 3. As in Experiment 1a, there was an overall bias toward responding "present" in the critical trials, with 54% of responses consistent with a "present" belief (compare with 52% in Experiment 1a). Replicating Experiment 1a, discriminability was high and was modulated by compositionality of the source VPC such that participants experienced more particle intrusions and deletions when the particle occurred in a source with higher compositionality (see Figure 5).

The maximum random effect structure justified by the data included only random intercepts for subjects and items. Adding the control variable source sentence position and its interactions improved model fit, but critically, source position did not alter any of the effects of compositionality[4] (see Table 3).

Break point analysis: A gradient, rather than categorical, model best explains the distribution of illusory conjunctions. The maximum log-likelihood piecewise model that met all convergence criteria inserted a break point at a compositionality rating of 3.83 (vs. the response model break point in Experiment 1a: 4.07). Profile CIs around the model's adjustment parameters indicated a significant difference in the source compositionality by discriminability slope term, accounting for the fact that there were more particle deletions (particle "absent" responses when the particle was present) for the lowest compositionality items than items near the midpoint (see Figure 5). Following this pattern, adding a break point improved model fit, $\chi^2(5) = 21.27$, $p < .001$.

Despite this improvement in fit for the break point versus no break point model, the results did not favor a discrete analysis. The hallmark of a clear discrete analysis is an intercept difference on either side of the break point with no slope differences, such that the two line segments are flat and clearly separated. We observed the opposite, a difference in slopes but not intercepts. To interpret this finding, note that a change in slopes without a change in intercepts means that the slope in the left segment of the model is either flatter or in the opposite direction as the slope in the right segment. This is a nonlinearity in the way the predictor affects the outcome. In this case, the nonlinearity comes from compositionality's differential impact on the lowest compositionality items.

Given the diverse factors shown to affect processing of VPCs (see, e.g., Lohse et al., 2004), it is possible that the normed compositionality rating for some specific items, such as the lowest compositionality item *help out* (compositionality value of 1.58), is not appropriate for the particular sentence contexts used here. Consistent with this logic, an anonymous reviewer points out that low compositionality items such as *help out* seem to behave more compositionally when they take pronouns as objects than full noun phrases (e.g., *She will help me out* vs. *She will help the waiter out*), suggesting the plausible importance of accounting for sentence context in compositionality ratings and in the availability of various structures. It is critical to note that this possibility is still more consistent with a graded effect of compositionality than with a sharp division of VPCs into two classes.

Response times. Overall response latencies were similar to the previous experiments, with correct responses typically taking between 1,100 and 1,200 msec. As in Experiment 1a, correct responses to the two-VPC and two-adverb trials were slowest (two-VPC M = 1,207 msec, two-adverb M = 1,237 msec), with the rest of the trial types eliciting similar correct latencies (critical trials M = 1,179 msec, one-adverb M = 1,163 msec, bare noun M = 1,173 msec, adverb-noun swap M = 1,160 msec, VPC-noun swap M = 1,170 msec). Again, error responses tended to be slower than correct ones, following a similar pattern as Experiment 1a (two-VPC M = 1,234 msec, two-adverb M = 1,321 msec, critical trials M = 1,229 msec, one-adverb M = 1,253 msec, bare noun M = 1,204 msec, adverb-noun swap M = 1,168 msec, VPC-noun swap M = 1,249 msec).

Linear regression: RTs are lower for high-compositionality VPCs. As in Experiment 1a, there was a significant effect of source particle presence such that source particle-present trials were slowest (particle present M = 1,201 msec, particle absent M = 1,153 msec). Unlike in Experiment 1a, there was no significant effect of compositionality, neither as a main effect nor as an interaction. However, the numerical pattern was similar to Experiment 1a, with the fastest responses for high-compositional items (mean for items below 2.5: 1,212 msec; mean for items above 7: 1,175 msec). See Figure 6 for a graphical representation of the data, and see Table 3 for mixed-effect analyses. The maximum random effect structure justified by the data was the full random effect structure (random intercepts by participants and items, random slopes by source particle presence, source compositionality and their interaction for participants, and random slopes by source particle presence for items).

Adding the control variable of source sentence position and its interactions improved model fit, $\chi^2(4) = 11.48$, $p = .02$. In this model, there was a significant effect of source sentence position

[4] This model did not converge to tolerance: 0.0015, rather than the 0.001 standard. Removing all nonsignificant effects allows the model to converge to tolerance. This reduced model shows the same overall pattern.

Table 3
Mixed-Effect Model Outputs From Experiment 1b

Model	Estimate	*SE*	*z* value	$p(\chi^2)$
Particle belief analysis				
Zero-break model				
Overall bias	0.21	0.04	5.81	<.001
Discriminability	2.25	0.07	33.21	<.001
Source compositionality × Bias	0.03	0.02	1.52	.13
Source compositionality × Discriminability	−0.09	0.04	−2.57	.01
	Estimate	*SE*	*z* value	$p(\chi^2)$
Zero-break model containing particle source				
Overall bias	0.22	0.04	6.01	<.001
Discriminability	2.27	0.07	33.18	<.001
Source compositionality × Bias	0.03	0.02	1.36	.18
Source sentence position × Bias	−0.22	0.07	−3.23	<.01
Source compositionality × Discriminability	−0.10	0.04	−2.81	<.01
Source sentence position × Discriminability	−0.49	0.14	−3.56	<.001
Source compositionality × Source sentence position × Bias	0.03	0.04	0.93	.35
Source compositionality × Source sentence position × Discriminability	0.10	0.07	1.38	.17
	Estimate (lower, upper CI)	*SE*	*z* value	
Break point model				
Right-half parameters				
Overall bias (intercept)	0.21 (0.10, 0.33)	0.06	3.67	
Discriminability (slope)	2.52 (2.30, 2.74)	0.11	22.57	
Source compositionality × Bias	0.02 (−0.04, 0.09)	0.03	0.70	
Source compositionality × Discriminability	−0.28 (−0.40, −0.15)	0.06	−4.36	
Left-half parameter adjustments				
Bias (intercept)	0.36 (−0.10, 0.81)	0.23	1.57	
Discriminability (slope)	0.85 (0.00, 1.72)	0.44	1.96	
Source compositionality × Bias	0.17 (−0.05, 0.39)	0.11	1.61	
Source compositionality × Discriminability	0.79 (0.38, 1.19)	0.21	3.82	
	Estimate	*SE*	*t* value	$p(\chi^2)$
Log correct RT analysis				
Zero-break model				
Intercept	7.03	0.01	523.7	<.001
Source particle presence	0.04	0.01	4.3	<.001
Source compositionality	−0.00	0.00	−0.2	.81
Source particle presence × Source compositionality	0.00	0.00	0.2	.86
	Estimate	*SE*	*t* value	$p(\chi^2)$
Zero-break model containing particle source				
Intercept	7.03	0.01	524.0	<.001
Source particle presence	0.04	0.01	4.4	<.001
Source compositionality	−0.00	0.00	−0.3	.79
Source sentence position	0.03	0.01	3.2	<.01
Source particle presence × Source compositionality	0.00	0.01	0.2	.86
Source particle presence × Source sentence position	0.00	0.02	0.3	.82
Source compositionality × Source sentence position	0.02	0.00	1.1	.29
Source particle presence × Source compositionality × Source sentence position	0.00	0.01	0.3	.80
	Estimate (lower, upper CI)	*SE*	*t* value	
Break point model				
Right-half parameters				
Intercept	7.01 (6.97, 7.06)	0.02	323.90	
Source particle presence	−0.01 (−0.06, 0.04)	0.03	−0.35	
Source compositionality	0.01 (−0.01, 0.03)	0.01	0.78	
Source particle presence × Source compositionality	0.03 (0.00, 0.05)	0.01	1.86	
Left-half parameter adjustments				
Intercept	0.00 (−0.05, 0.05)	0.02	0.16	
Source particle presence	0.09 (0.02, 0.16)	0.04	2.63	
Source compositionality	−0.02 (−0.04, 0.01)	0.01	−1.14	
Source particle presence × Source compositionality	−0.00 (−0.04, 0.03)	0.02	−0.25	

Note. Model comparison was used to calculate *p* values, omitting one factor at a time and comparing models with χ^2 tests.

such that tests based upon the more recent source were faster; this factor did not modulate any effects of compositionality. The model's random effect structure contained random intercepts by participants and items, random slopes by source particle presence and source compositionality for participants, and random slopes by source particle presence for items. See Table 3 for mixed-effect analyses.

Break point analysis: A gradient, rather than categorical, model best explains the distribution of illusory conjunctions. The maximum log-likelihood piecewise regression inserted a break point at 4.91 (contrast with the response time break point in Experiment 1a: 6.04). As in Experiment 1a, profile CIs around the model's adjustment parameters showed no significant difference in intercepts between the left and right halves of the model. One small but statistically significant difference was observed in the slope adjustment for source particle presence. This was such that there was a slightly negative slope for particle-absent low compositionality trials (β estimate $= -0.05$), and a slightly positive slope for particle-present low compositionality trials (β estimate $=$ 0.04); no other significant differences were observed (see Table 3).

Critically, again, the overall lack of improvement for the broken model over the model with no break point was confirmed with model comparison. Adding a break point did not significantly improve model fit ($\chi^2(5) = 8.42$ $p = .13$).

Discussion

Experiment 1 revealed that semantic compositionality modulated forced-choice memory responses: Participants experienced more illusory conjunctions involving particles from high versus low compositional VPC sources. The tendency toward illusory conjunctions followed a graded pattern across the whole spectrum of compositionality, more consistent with a representational cline than two discrete classes. This was further evidenced by our novel application of piecewise regression in modeling potential break points between classes. This technique also failed to show any evidence in favor of the discrete account in analysis of RTs.

A key assumption of our analysis is that the illusory conjunctions we observe reflect the syntactic structure of the VPCs (e.g., Complex Head vs. Small Clause). Although evidence from other paradigms suggests that syntactic structure independently impacts how VPCs are used (e.g., through syntactic priming; Konopka & Bock, 2009), it is possible that the effects we observe here purely reflect semantic similarity. For example, particles could migrate simply because the source and target sentence have roughly the same meaning, leading to errors in recall based upon gist memory (e.g., Potter & Lombardi, 1990). This motivated Experiment 2, in which syntactic and semantic contributions to illusory conjunctions were crossed to assess the source of these errors.

Experiment 2

To dissociate the role of syntactic and semantic factors in eliciting false memory errors, Experiment 2 capitalized on the availability of two forms of many VPCs. These were an adjacent form, as used in the previous experiments (e.g., *He will cut up the meat*) and a shifted form (e.g., *He will cut* the meat *up*). The notion is that although the adjacent and shifted form are highly similar in meaning, they are associated with different syntactic structures (e.g., as in Figure 1), allowing the separation of syntactic and semantic factors in eliciting illusory conjunction errors.

In Experiment 2, we examined false memories for adjacent VPC sources as induced by adjacent and shifted VPC foil test sentences. The critical aspect of this design is that an adjacent VPC test sentence contains the same [V Part Obj] structure as its source whereas a shifted VPC test appears in a different structure than its source [V Obj Part]. This means that although the same particle intrudes in both conditions, only the adjacent test matches the structure of the particle's source. The prediction is that if structure is the main driver of false memory errors, then rates of illusory conjunctions will be higher for adjacent than for shifted tests.

As a baseline, we also examined error rates to foils that matched the source in meaning but which substituted another bare verb instead of a VPC (e.g., *He will slice the meat*). Error rates for these foil sentences establish a baseline of error responding due to gist meaning. Any errors observed in this condition do not reflect "illusory conjunctions" of the source stimuli because the inserted verb did not appear in either source. This allows us to estimate the base rate of errors due to shared meaning only.

Unlike the previous experiments, the critical comparisons in Experiment 2 focused on false alarms to various types of foil sentences. This change in the critical comparison led us to simulate the effect sizes from the previous experiments to determine an appropriate sample size. As in Experiment 1b, we ran Monte Carlo simulations based upon our previous data to estimate the sample sizes needed to observe the effects of interest with 80% power. We ran these simulations with effect sizes ranging from 100% of the critical trial false alarm rate in Experiments 1a and 1b (pooled, compared with the filler sentence baseline) down to 25% of the original false alarm rate (compared with the filler sentence baseline). These simulations suggest that with a set of 24 items, collecting data from 60 participants should allow us to observe an effect that is 33% of the size originally observed or larger. A plot of simulation results can be found at https://osf.io/k4f9v/, the code used to generate these simulations can be found at https://osf.io/q567r/, and a registration of the study can be found at https://osf.io/sudn4.

Method

Participants. Data were collected from 69 participants recruited through Amazon Mechanical Turk, all of who had IP addresses from the United States, were older than 18 years of age, and had not participated in Experiments 1a or 1b. As in the previous experiments, all participants were compensated $5 for their time, which was approximately 30 min of active participation (excluding breaks). Nine participants were excluded for having an overall accuracy of less than 60%, leaving a total of 60 participants contributing data to the final analyses. All experimental procedures and protocols were approved by the Northwestern University Institutional Review Board.

Equipment. Identical to Experiment 1.

Materials. The items were versions of the same 160 items as in Experiment 1. Of these, 24 were critical trials. These were a subset of those used in the previous experiment, comprised of the items that were determined by both authors to be acceptable in the shifted structure. As in the previous experiments, critical items were made up of triads of sentences that contained two source

sentences and a test sentence. As previously, one of the two source sentences contained a VPC (e.g., *He will lock up the bicycle*) and the other contained a verb that would become a grammatical VPC if the particle from the other sentence were added to it (e.g., *He will cut the meat*).

Unlike the previous experiments, all of the Experiment 2 critical test sentences were what we term *foils*—items that mismatched both of the source sentences. Three types of foils were developed for each critical item. The first matched the source-plus-particle foils in Experiment 1, containing a VPC derived from the non-VPC source sentence where the verb and particle were directly adjacent (e.g., *He will cut up the meat*). The second used a shifted form of the same foil VPC (e.g., *He will cut* the meat *up*), placing the particle in a structural position different from its original source. The third was a synonym of the foil VPC, designed to match the source in semantic content but to contain a verb that did not appear in either original source (e.g., *He will slice the meat).* To equate meaning across these types of foils as best as possible, the semantic similarity of sources and the three types of foils was normed by a separate group of participants. Results of this norming study are outlined here; a list of critical items can be found in the Appendix.

To control for the new types of critical trials, some additional types of filler trials were created by modifying items from Experiment 1. Half of the adjacent VPC-containing filler items from Experiment 1 were used in the shifted VPC form instead; the test sentences associated with these fillers had edits to both particles and nouns. In addition, eight of the adjacent VPC items that had been used as critical items in Experiment 1 were used as fillers in this experiment. These were used in combination with other filler items to balance the ratio of "yes" and "no" responses to VPC trials. As in Experiment 1, the ratio of "yes" to "no" answers across the experiment was equivalent, with 50% "yes" trials overall. As in Experiment 1, 60% of the items contained a VPC in one or both of the sources, 30% of the items contained an adverb in one or both of the sources, and 10% contained neither; again, the presence of a particle in the second source sentence was not predictable from the form of the first source sentence. Similar to Experiment 1, test sentences were more likely to involve changes to a particle than anything else, although the ratio was increased in this experiment (50% particle change trials, 25% adverb change trials, and 25% noun change trials).

Norming. To assess whether the three critical foil sentence types were equivalent in meaning, we normed the semantic similarity of foils and the source sentences they were based upon using a separate group of 60 participants on Amazon Mechanical Turk. These participants were asked to rate the similarity of pairs of sentences using a 1–9 scale, in which 1 indicated *very dissimilar* and 9 indicated *very similar*. Two examples were provided, one on either end of the scale, and the 24 critical items were intermixed with 32 filler item pairs that had varying degrees of semantic similarity. In the norming survey, items were divided across six lists, counterbalancing the three types of foil sentences (Adjacent/Shifted/Synonymous) and which item (Experimental Source/Foil) appeared first.

Results of this norming survey suggest that the critical sources and their foils were rated as much more similar than the various types of filler sentences were (Critical trial $M = 7.13$, 95% CI [6.93, 7.32] vs. fillers $M = 3.89$, 95% CI [3.67, 4.12]). There were also some small differences between types of critical trials. Within the critical trials, the shifted VPC foils and sources were rated as more similar than the adjacent VPC foils and sources, and these were both rated as more similar than the synonymous foils and sources were (Shifted: $M = 7.36$, 95% CI [7.11, 7.60]; Adjacent: $M = 7.12$, 95% CI [6.89, 7.35], Synonymous: $M = 6.93$, 95% CI [6.68 7.18]). There were no systematic differences in similarity rating depending on whether the experimental source or foil sentence was presented first (Shifted foil first $M = 7.41$, 95% CI [7.09, 7.72] vs. source first $M = 7.31$, 95% CI [7.00, 7.62]; Adjacent foil first $M = 7.18$, 95% CI [6.93, 7.48] vs. source first $M = 7.06$, 95% CI [6.71, 7.38]; Synonymous foil first $M = 6.88$, 95% CI [6.55, 7.29] vs. source first $M = 6.97$, 95% CI [6.67, 7.23]). See Appendix for by-item norming results.

Procedure. Identical to Experiment 1.

Design and data analysis. The critical factor in this experiment was foil type (Adjacent/Shifted/Synonymous). As in the previous experiments, source sentence position was entered as a control variable (VPC first/second). The normed similarity rating was also entered as a control variable. Foil type and source sentence position were within-item factors, counterbalanced across six lists such that each item was presented an equal number of times in each form and each participant viewed only one version of each item. Foil type was contrast coded to compare the VPC sentences to the synonymous foil (Adjacent/Shifted to Synonymous, 0.25, 0.25, −0.5) and to each other (Adjacent: 0.5; Shifted: −0.5). Source sentence position was effects coded (.5, −.5), and the normed similarity rating was mean centered.

Unlike the previous experiments, the critical dependent measure for the response analysis was false alarm rate (in contrast to the signal detection approach); the response time analysis was the same as the previous experiments. As in the previous experiments, the random effect structure was the maximum justified by the data. Anonymized data from the experiment and norming task and R scripts are archived on the Open Science Framework: https://osf.io/rkzfv/, https://osf.io/g8ys6/, and https://osf.io/d6z45/.

Results

Responses. The overall error rate was similar to that of Experiment 1 (17%, compared with 21% in Experiment 1a and 18% in Experiment 1b). As in the previous experiments, VPC trials elicited the most errors, with the highest error rate in the two-VPC-swap fillers (27%) followed by the one-VPC-swap trials (18%, pooling critical trials and fillers). The other types of trials elicited fewer errors (Adverb trials: two-adverb = 15%; one-adverb = 13%; Noun swap trials: VPC = 12%; adverb = 11%, bare = 13%).

The critical comparisons in this experiment focused on false alarms, or erroneous "yes" responses to novel foils. The VPC trials elicited the most false alarms, with the highest false alarm rate in the two-VPC-swap fillers (45%), followed by the critical trials (25%). The other types of trials elicited fewer false alarms (Adverb trials: two-adverb = 17%; one-adverb = 16%; Noun swap trials: VPC = 14%; adverb = 13%, bare = 17%).

Critical trial analyses: Illusory conjunctions are sensitive to syntactic structure. The VPC-adjacent foils, where a particle was inserted in the same structural position as it had appeared in its original source, elicited the most false alarms (39%). These were

followed by the VPC-shifted foils (25%), where the particle was inserted in a different structural position. This difference points to the importance of structure in eliciting false alarms. In addition, both types of VPC foils elicited more false alarms than the synonymous foils; at 12% false alarms, these had the lowest error rate in the experiment (see Figure 7). These patterns were confirmed with mixed-effect analyses (see Table 4).

The increased false alarm rate in the adjacent and shifted VPC critical trials versus the synonymous critical trials emphasizes the importance of structure in eliciting errors in this paradigm. False alarms to both types of VPC critical trials largely reflect illusory conjunctions of the verbs and particles from the source sentences driven by shared structure. In contrast, the synonymous trials reflect only the role of meaning; the low error rate here suggests that meaning is not a major source of errors.

Adding the control variable of source sentence position and its interactions improved model fit, $\chi^2(3) = 9.52$, $p = .02$. This was due to the higher rate of false alarms for synonymous trials when the test sentence was in the first position (16%) versus the second (8%; see Table 4). Both types of VPC foils elicited similar false alarm rates regardless of position, providing further evidence that the synonymous trials disclosed a qualitatively and quantitatively different pattern than the VPC foil trials did. Adding the normed similarity rating and its interactions failed to improve model fit, showing that the critical trial differences did not follow from semantic similarity. This was the case for a model adding only VPC type, $\chi^2(3) = 2.00$, $p = .57$, and for a model adding VPC type and item order, $\chi^2(6) = 5.18$, $p = .52$. In the final model, the maximum random effect structure included only random intercepts for subjects and items.

Response times. Response latencies were similar to the previous experiments, with correct responses typically taking between 1,100 and 1,300 msec. As in the previous experiments, correct responses to the two-VPC and two-adverb trials were the slowest (two-VPC $M = 1{,}213$ msec, two-adverb $M = 1{,}238$ msec). Critical trials were similarly slow (1,222 msec); the remaining trial types elicited slightly faster correct latencies (one-VPC controls $M = 1{,}143$ msec, one-adverb $M = 1{,}170$ msec, bare noun swap $M = 1{,}179$ msec, adverb-noun swap $M = 1{,}216$ msec, VPC-noun swap $M = 1{,}179$ msec). Again, error responses tended to be slower than correct ones (two-VPC $M = 1{,}285$ msec, two-adverb $M = 1{,}276$ msec, critical trials $M = 1{,}239$ msec, one-VPC controls $M = 1{,}307$ msec, one-adverb $M = 1{,}239$ msec, bare noun $M = 1{,}183$ msec, adverb-noun swap $M = 1{,}197$ msec, VPC-noun swap $M = 1{,}245$ msec).

Critical trials: RTs are slower for VPCs relative to synonymous non-VPC foils. For correct responses to critical trials (see Figure 8), the pattern was that the VPC trials—whether adjacent or shifted—elicited slower responses than the synonymous foil trials (adjacent $M = 1{,}280$ msec; shifted $M = 1{,}276$ msec; synonymous $M = 1{,}136$ msec), again showing the importance of shared structure in leading to response difficulty. This pattern was confirmed with mixed-effect analyses (see Table 4).

Adding the control variable of source sentence position and its interactions significantly improved model fit. This was due to the main effect of Source 2 tests eliciting faster responses than Source 1 tests regardless of foil type (see Table 4), perhaps reflecting a recency effect. As with the response rate analysis, adding the normed similarity rating failed to improve model fit in a model that contained only VPC type, $\chi^2(3) = 1.85$, $p = .61$, and in a model that contained VPC type and source sentence position, $\chi^2(6) = 4.16$, $p = .65$. In the final model, the maximum random effect structure justified by the data included random slopes for VPC type, source sentence position and their interaction by subjects, and random intercepts only by items.

Discussion

In Experiment 2, we examined whether memory errors for VPCs were driven by meaning similarity or by syntactic structure. This was done by contrasting false alarm rates to three types of foil sentences. Adjacent VPC foils, which contained an intruded particle that appeared in the same position as it did in its original source, elicited the highest rate of "particle-present" false alarms. Shifted VPC foils, which contained the same intruded particle in a position that was different from its original source, elicited fewer

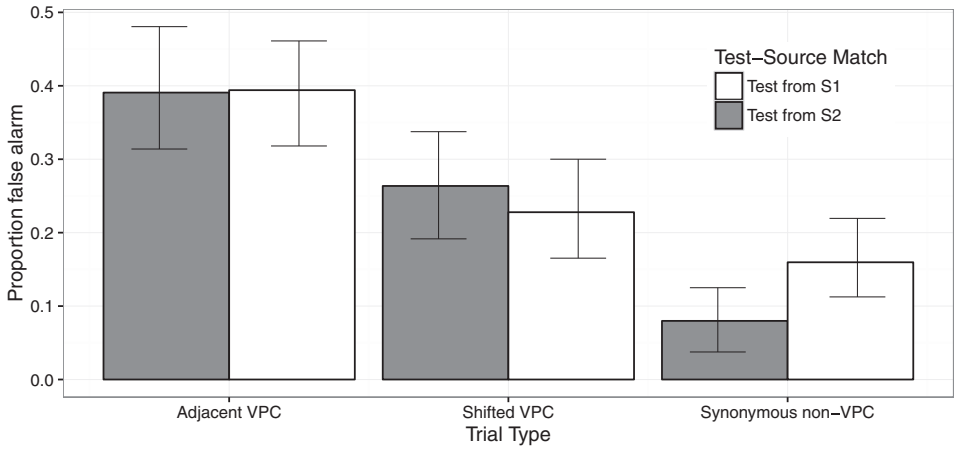

Figure 7. Mean false alarm rate in Experiment 2 by foil type. CIs represent 95% range of mean values from 1,000 bootstrapped samples.

Table 4
Mixed Effect Model Outputs From Experiment 2

Model	Estimate	*SE*	*z* value	$p(\chi^2)$
False alarms by VPC type and particle source				
Intercept	1.60	0.22	7.44	<.001
Adjacent/Shifted VPC vs. Synonymous	−2.13	0.25	−8.70	<.001
Adjacent vs. Shifted VPC	−0.88	0.16	−5.40	<.001
Source sentence position	−0.25	0.15	−1.64	.11
Adjacent/Shifted VPC vs. Synonymous × Source sentence position	1.38	0.48	2.88	<.001
Adjacent vs. Shifted VPC × Source sentence position	−0.23	0.32	−0.70	.49
	Estimate	*SE*	*t* value	$p(\chi^2)$
Log correct RT by VPC type and particle source				
Intercept	7.09	.02	305.23	<.001
Adjacent/Shifted VPC vs. Synonymous	0.17	.02	6.96	<.001
Adjacent vs. Shifted VPC	0.00	0.02	−0.05	.96
Source sentence position	0.06	0.01	4.69	<.001
Adjacent/Shifted VPC vs. Synonymous × Source sentence position	0.02	0.04	0.50	.62
Adjacent vs. Shifted VPC × Source sentence position	0.03	0.03	0.98	.33

Note. VPC = verb-particle constructions. Model comparison was used to calculate *p* values, omitting one factor at a time and comparing models with χ^2 tests.

false alarms. The increased error rate in adjacent versus shifted foils appeared despite the fact that the shifted VPC foils were rated as slightly more similar in meaning to the original target sentences than the adjacent VPCs were. This provides clear evidence in favor of syntactic structure as a primary driver of memory errors in this paradigm.

Critically, both types of VPC foils elicited slower correct responses and more false alarms than did foils where a synonymous bare verb intruded in the sentence instead of a particle. The meaning of the synonymous foil was similar to the meaning of the original source sentence, but in contrast to the VPC foils, the intruded element (here, a bare verb) did not appear in either source sentence. The low false alarm rate and quick responding in this condition implies that meaning similarity induces few errors and the errors that are observed in this paradigm do reflect true illusory conjunctions of elements of the original source sentences (i.e., the particle of one source, verb of the other).

In general, the importance of structure over meaning in eliciting false memories for sentences is consistent with a large body of work on syntactic priming. This work supports the isolable impact of structure from meaning and lexical content (e.g., Bock, 1986; Bock & Loebell, 1990; Konopka & Bock, 2009; Tooley & Bock, 2014). It is important to note that, as with syntactic priming paradigms, errors in the present paradigm come from the presence of a structured representation of the stimuli.

General Discussion

Debates over the structure of VPCs mirror a general tension in theories of cognition: whether mental representations are fundamentally discrete or whether they incorporate graded structure. To examine this issue, we elicited illusory conjunctions of verbs and particles. This allowed us to assess the extent to which VPCs are associated with two distinct structures. In one of these distinct

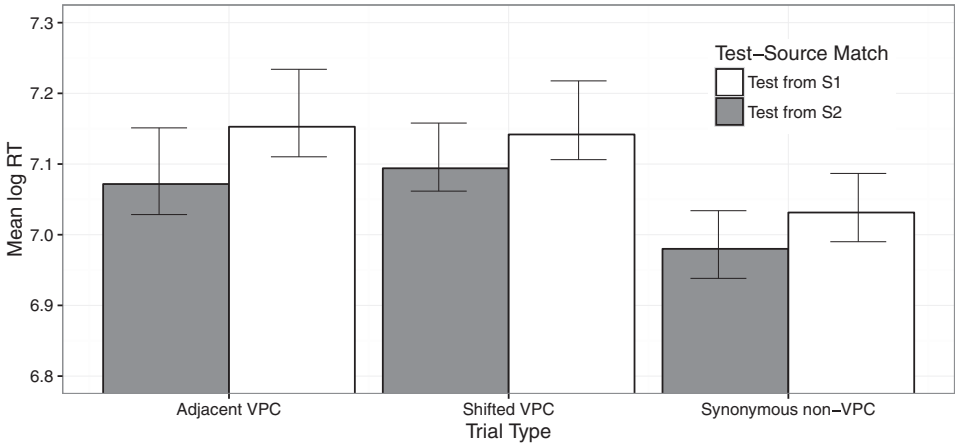

Figure 8. Mean log response time in Experiment 2 by foil type. CIs represent 95% range of mean values from 1,000 bootstrapped samples.

structures, the verb and particle are represented as syntactically distinct structural elements (a Small Clause), facilitating the independent recombination of particle and verb in illusory conjunctions. In the other structure, the verb and particle are represented as a single syntactic element (a Complex Head), suppressing such conjunctions. Under a discrete account, VPCs should be associated with either one or the other structure; under a gradient account, VPCs should be associated, in varying degrees, to both structures.

Our results are consistent with the predictions of the gradient account. We demonstrated that participants make responses in a forced-choice perception paradigm that are consistent with the intrusion and disappearance of particles in VPCs. This tendency was modulated by semantic compositionality, a factor known to be associated with Small Class versus Complex Head syntactic representations of VPCs. The general pattern was that VPCs high in compositionality were more likely to induce illusory conjunctions where the particle intrudes on another verb source. Piecewise regression of the Experiment 1 data showed that illusory conjunctions increased in a gradient fashion with source compositionality. Experiment 2 showed that these effects are not simply due to similarity in meaning, but that they reflect structural overlap between the source and target in illusory conjunction errors. The combination of these experiments suggests that VPCs are gradiently associated with two distinct structural options.

Forms of Gradience in Sentence Processing and Syntactic Structure

Although this work provides novel support for gradient representations in syntax, a key area for future work is determining the structure of gradience in the language processing system. In sentence processing and within probabilistic syntactic models more generally, it is common to incorporate gradience in probability distributions over structures (e.g., Hale, 2001; Jaeger & Snider, 2008; Levy, 2008; inter alia). Applying this perspective to VPCs, a moderately compositional VPC such as *look up* fits a structure in which the VPC is a single unit (*look* + *up*) and a structure in which the verb and particle are part of separate syntactic units (*look*)+(*up*). A reader selects between these two discrete structures in a gradient fashion such that the overall response distribution is probabilistic (e.g., for a VPC, a Small Clause structure is selected 70% of the time, a Complex Head is selected the remaining 30% of the time). A virtue of such models is that they are easy to incorporate with traditional views of linguistic structure.

An alternative approach is to assume that the representations themselves are gradient (Aarts, 2007; Dowty, 2003; Smolensky et al., 2014; see Baayen & del Prado Martin, 2005, for parallel arguments at the semantics/morphology interface). Two structures might be simultaneously available to varying degrees, both partially activated within a single structure. For example, a moderately compositional VPC is associated with a single parse in which the structural elements of a Small Clause structure are active at 0.70 and, simultaneously, the elements of a Complex Head structure are active at 0.30. Such representations form a key part of dynamical computational architectures, in which linguistic representations are realized with a continuous, gradient representational space (e.g., Smolensky et al., 2014; Tabor & Tanenhaus, 1999), proving a transparent mapping to a plausible neural representation (see, e.g., beim Graben, Gerth, & Vasishth, 2008).

It is important to note that both approaches have advantages; distinguishing between them could shine light on the organizing principles of language and of the mind in general. A clear area for future research is examining the relative strength of each account.

Erroneous Recombinations and Gradient Structure in Cognition

Paradigms examining the reassociation of "unbound" stimulus features form a springboard for literature in attention. This is the premise of the extremely influential "feature integration theory" (i.e., Treisman & Gelade, 1980). Supporting work has shown robust evidence for the cognitively distinct identity of source stimulus properties. These independent elements are often called "features" and include properties such as color and size (for shapes) as well as letter position and morphological category (for written words). The data show that features sometimes dislocate and are erroneously rebound in perception. These misbindings of features—illusory conjunctions—reflect properties of the sources and of the similarity structure that underlies isolable features, demonstrating that as a generic property across many domains in cognitive science, what can become mentally unstuck occasionally unsticks. This is robustly supported within the domains of object attention (e.g., Treisman, 1991; Treisman & Schmidt, 1982), word reading (e.g., Esterman, Prinzmetal, & Robertson, 2004; Prinzmetal et al., 1986, 1991; Prinzmetal & Millis-Wright, 1984; Rapp, 1992), and speech processing (Ali & Ingleby, 2010; Kolinsky et al., 1995; Mattys & Samuel, 1997). Here, we have extended such work to investigate the building blocks of sentences. We demonstrate clear evidence of illusory conjunctions of source elements in sentence stimuli, providing a novel measure of linguistic representational structure.

We also note that a gradient account of linguistic structure is consistent with gradience in representational structures across cognitive domains. This is especially apparent for the case of semantic/object categories, which research has shown to be nonlinearly separable and not discretely defined. For example, the category "game" includes the activities chess, football, and Tetris, which have no one unifying property, and the typical use of the category "sandwich" includes hoagies but not hot dogs—both of which comprise meat on a roll. The suggestion is that conceptual categories are defined by a probabilistic collection of features instead of hard and fast rules (e.g., Medin & Schaffer, 1978; Rosch & Mervis, 1975; Wittgenstein, 1953). This underscores the fact that human categories can be gradient: In language and other aspects of cognition, rather than associating an item with cognitive representation X or Y, it may be more fruitful to suppose that item is associated with X and Y to varying degrees.

Statistical Analysis of Categories

Across scientific domains, proposals relying on gradient structure are often contrasted with proposals relying on discrete structure. This means that many statistical techniques have been developed to determine the latent structure of variables. These techniques include taxometric analysis, a technique that determines latent category structure from multiple dependent measures based upon means or covariances (e.g., Meehl, 1999), as well as the related techniques of mixture modeling and latent class anal-

ysis (e.g., Lubke & Muthen, 2005; Vermunt & Magidson, 2003). These tools are powerful and useful for domains in which there is a large amount of data, especially where there are many dependent measures that can provide converging evidence regarding the nature of underlying constructs. In the present work we used piecewise regression to answer a similar question: Does a predictor represent a categorical or linear factor? This type of technique works well for a domain where regression models are most appropriate: One where there are relatively few independent measures, relatively few dependent measures, and an unknown link function.

Novel statistical techniques are most useful when their effects are shown to generalize. Experiment 1b was a registered replication study, designed to confirm the applicability of the piecewise regression technique for describing linguistic structure. Replicating the behavioral results from Experiment 1a and confirming the gradient pattern in responding in a high-powered sample demonstrates the general applicability of this technique: Piecewise regression is capable of determining the underlying structure of predictors in noisy behavioral data.

The replication study underscores a primary virtue of the piecewise regression technique: It is simple and robust to random noise. Even in the present domain, which has a relatively high signal-to-noise ratio, we were able to make consistent inferences about the dimensionality of predictors. This is because the linear predictor in the piecewise technique has an optional split and can fit the data well in either the categorical or continuous case. Furthermore, as a form of regression, this technique preserves all of the desiderata of mixed-effect models (relative to techniques such as ANOVA) including crossed random effects and the ability to appropriately model categorical outcomes (Jaeger, 2008). As a robust, simple method that can be applied to accuracy and latency data, piecewise regression may prove more broadly useful for addressing questions of class structure in other domains within cognitive science.

Conclusion

In three experiments, we have shown that VPCs are gradiently associated with multiple syntactic structures. Using illusory conjunctions to reveal the strength of association between syntactic elements, and applying statistical techniques in a novel way, we have provided new support for the claim that the variable behavioral patterns of VPCs reflect an underlying gradient structure. This provides further support for cognitive architectures that incorporate gradient mental representations across all levels of processing.

References

Aarts, B. (2007). *Syntactic gradience: The nature of grammatical indeterminacy*. Oxford, United Kingdom: Oxford University Press.

Ali, A. N., & Ingleby, M. (2010). Gradience in morphological decomposability: Evidence from the perception of audiovisually incongruent speech. *Laboratory Phonology, 1,* 263–282. http://dx.doi.org/10.1515/labphon.2010.013

Baayen, R. H., & del Prado Martín, F. M. (2005). Semantic density and past-tense formation in three Germanic languages. *Language, 81,* 666–698. http://dx.doi.org/10.1353/lan.2005.0112

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1. 1–7.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models (arXiv:1506.04967). Manuscript submitted for publication.

beim Graben, P., Gerth, S., & Vasishth, S. (2008). Towards dynamical system models of language-related brain potentials. *Cognitive Neurodynamics, 2,* 229–255.

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology, 18,* 355–387. http://dx.doi.org/10.1016/0010-0285(86)90004-6

Bock, K., & Loebell, H. (1990). Framing sentences. *Cognition, 35,* 1–39. http://dx.doi.org/10.1016/0010-0277(90)90035-I

Bresnan, J., & Hay, J. (2008). Gradient grammar: An effect of animacy on the syntax of give in New Zealand and American English. *Lingua, 118,* 245–259. http://dx.doi.org/10.1016/j.lingua.2007.02.007

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Crawley, M. J. (2007). *The R book*. Chichester, United Kingdom: Wiley. http://dx.doi.org/10.1002/9780470515075

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods, 3,* 186–205. http://dx.doi.org/10.1037/1082-989X.3.2.186

Dehé, N., Jackendoff, R., McIntyre, A., & Urban, S. (Eds.) (2002). *Verb-particle explorations* (Vol. 1). Boston, MA: Walter de Gruyter.

den Dikken, M. (1995). *Particles. On the syntax of verb-particle, triadic, and causative constructions*. New York, NY: Oxford University Press.

Dowty, D. (2003). The dual analysis of adjuncts/complements in categorical grammar. In E. Lang, C. Maienborn, & C. Fabricius-Hansen, *Modifying adjuncts* (pp. 33–66). Berlin, Germany: De Gruyter Mouton.

Drummond, A. (2013). *Ibex farm*. Retrieved from http://spellout.net/ibexfarm/

Esterman, M., Prinzmetal, W., & Robertson, L. (2004). Categorization influences illusory conjunctions. *Psychonomic Bulletin & Review, 11,* 681–686. http://dx.doi.org/10.3758/BF03196620

Gonnerman, L. (2012). The roles of efficiency and complexity in the processing of verb particle constructions. *Journal of Speech Sciences, 2,* 3–31.

Gonnerman, L. M., & Hayes, C. R. (2005). The professor chewed the students . . . out: Effects of dependency, length, and adjacency on word order preferences in sentences with verb particle constructions. In *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society* (pp. 785–790). Mahwah, NJ: Lawrence Erlbaum.

Goodman, N. D., & Lassiter, D. (2015). Probabilistic semantics and pragmatics: Uncertainty in language and thought. In S. Lappin & C. Fox (Eds.), *The handbook of contemporary semantic theory* (pp. 655–686). Oxford, United Kingdom: Wiley-Blackwell. http://dx.doi.org/10.1002/9781118882139.ch21

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies* (pp. 1–8). Stroudsburg, PA: Association for Computational Linguistics.

Hoekstra, T. (1988). Small clause results. *Lingua, 74,* 101–139. http://dx.doi.org/10.1016/0024-3841(88)90056-3

Jackendoff, R. (2002). English particle constructions, the lexicon, and the autonomy of syntax. In N. Dehé, A. McIntyre, & R. Jackendoff (Eds.), *Verb-particle explorations* (pp. 67–94). Berlin, Germany: Mouton de Gruyter. http://dx.doi.org/10.1515/9783110902341.67

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59,* 434–446. http://dx.doi.org/10.1016/j.jml.2007.11.007

Jaeger, T. F., & Snider, N. (2008). Implicit learning and syntactic persistence: Surprisal and cumulativity. In *The 30th Annual Meeting of the Cognitive Science Society (CogSci08)* (p. 827). Austin, TX: Cognitive Science Society.

Johnson, K. (1991). Object positions. *Natural Language and Linguistic Theory, 9,* 577–636. http://dx.doi.org/10.1007/BF00134751

Kim, H.-J. (1994). Tests for a change-point in linear regression. *Lecture Notes–Monograph Series, 23,* 170–176. http://dx.doi.org/10.1214/lnms/1215463123

Knell, R. J. (2009). On the analysis of non-linear allometries. *Ecological Entomology, 34,* 1–11. http://dx.doi.org/10.1111/j.1365-2311.2008.01022.x

Kolinsky, R., Morais, J., & Cluytens, M. (1995). Intermediate representations in spoken word recognition: Evidence from word illusions. *Journal of Memory and Language, 34,* 19–40. http://dx.doi.org/10.1006/jmla.1995.1002

Konopka, A. E., & Bock, K. (2009). Lexical or syntactic control of sentence formulation? Structural generalizations from idiom production. *Cognitive Psychology, 58,* 68–101. http://dx.doi.org/10.1016/j.cogpsych.2008.05.002

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106,* 1126–1177. http://dx.doi.org/10.1016/j.cognition.2007.05.006

Lohse, B., Hawkins, J. A., & Wasow, T. (2004). Domain minimization in English verb-particle constructions. *Language, 80,* 238–261. http://dx.doi.org/10.1353/lan.2004.0089

Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods, 10,* 21–39. http://dx.doi.org/10.1037/1082-989X.10.1.21

Mattys, S. L., & Samuel, A. G. (1997). How lexical stress affects speech segmentation and interactivity: Evidence from the migration paradigm. *Journal of Memory and Language, 36,* 87–116. http://dx.doi.org/10.1006/jmla.1996.2472

McIntyre, A. (2007). Particle verbs and argument structure. *Language and Linguistics Compass, 1,* 350–367. http://dx.doi.org/10.1111/j.1749-818X.2007.00013.x

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85,* 207–238. http://dx.doi.org/10.1037/0033-295X.85.3.207

Meehl, P. E. (1999). Clarifications about taxometric method. *Applied & Preventive Psychology, 8,* 165–174. http://dx.doi.org/10.1016/S0962-1849(05)80075-7

Muggeo, V. M. (2003). Estimating regression models with unknown break-points. *Statistics in Medicine, 22,* 3055–3071. http://dx.doi.org/10.1002/sim.1545

Potter, M. C., & Lombardi, L. (1990). Regeneration in the short-term recall of sentences. *Journal of Memory and Language, 29,* 633–654. http://dx.doi.org/10.1016/0749-596X(90)90042-X

Prinzmetal, W., Hoffman, H., & Vest, K. (1991). Automatic processes in word perception: An analysis from illusory conjunctions. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 902–923. http://dx.doi.org/10.1037/0096-1523.17.4.902

Prinzmetal, W., & Millis-Wright, M. (1984). Cognitive and linguistic factors affect visual feature integration. *Cognitive Psychology, 16,* 305–340. http://dx.doi.org/10.1016/0010-0285(84)90012-4

Prinzmetal, W., Treiman, R., & Rho, S. H. (1986). How to see a reading unit. *Journal of Memory and Language, 25,* 461–475. http://dx.doi.org/10.1016/0749-596X(86)90038-0

Punske, J. (2013). Three forms of English verb particle constructions. *Lingua, 135,* 155–170. http://dx.doi.org/10.1016/j.lingua.2012.09.013

Rapp, B. C. (1992). The nature of sublexical orthographic organization: The bigram trough hypothesis examined. *Journal of Memory and Language, 31,* 33–53. http://dx.doi.org/10.1016/0749-596X(92)90004-H

R Core Team. (2014). *R: A language and environment for statistical computing Version 3.1.2*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7,* 573–605. http://dx.doi.org/10.1016/0010-0285(75)90024-9

Ruiz, J., García, C., Muriel, E., Andrés, A. I., & Ventanas, J. (2002). Influence of sensory characteristics on the acceptability of dry-cured ham. *Meat Science, 61,* 347–354. http://dx.doi.org/10.1016/S0309-1740(01)00204-2

Schnoebelen, T., & Kuperman, V. (2010). Using Amazon Mechanical Turk for linguistic research. *Psihologija (Beograd), 43,* 441–464. http://dx.doi.org/10.2298/PSI1004441S

Smolensky, P., Goldrick, M., & Mathis, D. (2015). Optimization and quantization in gradient symbol systems: A framework for integrating the continuous and the discrete in cognition. *Cognitive Science, 38,* 1102–1138. http://dx.doi.org/10.1111/cogs.12047

Stiebels, B., & Wunderlich, D. (1994). Morphology feeds syntax: The case of particle verbs. *Linguistics, 32,* 913–968. http://dx.doi.org/10.1515/ling.1994.32.6.913

Tabor, W., & Tanenhaus, M. K. (1999). Dynamical models of sentence processing. *Cognitive Science, 23,* 491–515. http://dx.doi.org/10.1207/s15516709cog2304_5

Tooley, K. M., & Bock, K. (2014). On the parity of structural persistence in language production and comprehension. *Cognition, 132,* 101–136. http://dx.doi.org/10.1016/j.cognition.2014.04.002

Treisman, A. (1991). Search, similarity, and integration of features between and within dimensions. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 652–676. http://dx.doi.org/10.1037/0096-1523.17.3.652

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12,* 97–136. http://dx.doi.org/10.1016/0010-0285(80)90005-5

Treisman, A., & Schmidt, H. (1982). Illusory conjunctions in the perception of objects. *Cognitive Psychology, 14,* 107–141. http://dx.doi.org/10.1016/0010-0285(82)90006-8

Vermunt, J. K., & Magidson, J. (2003). Latent class models for classification. *Computational Statistics & Data Analysis, 41,* 531–537. http://dx.doi.org/10.1016/S0167-9473(02)00179-2

Wittgenstein, L. (1953). *Philosophical investigations*. Oxford, United Kingdom: Blackwell Publishing.

Wurmbrand, S. (2000). *The structure(s) of particle verbs*. Unpublished manuscript, McGill University, Montreal, Quebec.

Zeller, J. (2002). Particle verbs are heads and phrases. In N. Dehé, A. McIntyre, & R. Jackendoff (Eds.), *Verb-particle explorations* (pp. 233–267). Berlin, Germany: Mouton de Gruyter. http://dx.doi.org/10.1515/9783110902341.233

Appendix

Critical Stimuli From Experiments 1 and 2

Experiment 1

Subject + Aux	Source VPC	Source V	Foil V	Foil VPC	Continuous source VPC compositionality	Continuous foil VPC compositionality
She will	help out the waiter	clear the drawer	help the waiter	clear out the drawer	1.58	6.73
They will	throw over the frisbee	check the essays	throw the frisbee	check over the essays	1.84	5.81
We will	set off the timer	show the ponies	set the timer	show off the ponies	2.27	4.74
He will	bring down the crowd	hold the line	bring the crowd	hold down the line	2.56	5.78
He will	run up the hills	hold the lizard	run the hills	hold up the lizard	2.64	4.32
We will	make over the model	smooth the frosting	make the model	smooth over the frosting	2.68	4.64
He will	run over the trail	pass the dessert	run the trail	pass over the dessert	3.00	5.30
They will	keep up the secret	carry the groceries	keep the secret	carry up the groceries	3.07	-
They will	live down the lie	pull the photo	live the lie	pull down the photo	3.09	4.86
They will	take over the position	look the part	take the position	look over the part	3.18	7.30
They will	boil down the soup	shoot the hawk	boil the soup	shoot down the hawk	3.70	5.64
He will	see through the rumors	think the thoughts	see the rumors	think through the thoughts	3.83	7.04
She will	check out the books	drive the trucks	check the books	drive out the trucks	3.96	4.81
We will	pull off the sticker	mark the pages	pull the sticker	mark off the pages	4.07	5.93
She will	throw out the trash	buy the store	throw the trash	buy out the store	4.29	5.17
We will	dress up the doll	charge the phone	dress the doll	charge up the phone	4.32	5.75
He will	strike down the balance	hold the job	strike the balance	hold down the job	4.62	5.78
They will	catch up the students	mark the lumber	catch the students	mark up the lumber	4.91	5.84
We will	tie up the canoe	buy the supplies	tie the canoe	buy up the supplies	5.20	5.56
She will	break off the stick	pair the geese	break the stick	pair off the geese	5.52	6.17
He will	buy up the land	round the fractions	buy the land	round up the fractions	5.56	4.08
We will	call forth the juror	bring the gifts	call the juror	bring forth the gifts	5.56	7.36
They will	seal off the bag	lead the dance	seal the bag	lead off the dance	6.04	6.13
We will	bring in the dog	throw the towel	bring the dog	throw in the towel	6.04	6.00
She will	close down the slide	shut the computer	close the slide	shut down the computer	6.09	4.16
She will	drive away the candidate	give the prizes	drive the candidate	give away the prizes	6.50	6.14
We will	bust out the moves	bail the prisoner	bust the moves	bail out the prisoner	6.64	6.52
He will	lock up the bicycle	cut the meat	lock the bicycle	cut up the meat	6.85	7.09
He will	eat up the spaghetti	cover the research	eat the spaghetti	cover up the research	7.09	5.75
They will	wring out the rag	find the word	wring the rag	find out the word	7.16	6.26
She will	patch up the tire	set the boundaries	patch the tire	set up the boundaries	7.87	5.83
She will	finish up the game	close the café	finish the game	close up the café	8.55	6.27

(*Appendix continues*)

Experiment 2

	Source		Foil			Semantic similarity rating (mean source-foil [*SD*]/mean foil-source [*SD*])		
	VPC	V	Adjacent	Shifted	Synonymous	Adjacent	Shifted	Synonymous
She will	help out the waiter	clear the drawer	clear out the drawer	clear the drawer out	clean the drawer	7.8 (2.4)/ 7.9 (1.7)	8.1 (1.4)/ 8.7 (0.7)	7.4 (1.4)/ 6.7 (2.3)
They will	throw over the frisbee	check the essays	check over the essays	check the essays over	review the essays	8.6 (0.7)/ 8.9 (0.3)	8.5 (1)/ 8.6 (0.7)	8.3 (0.9)/ 8.5 (0.8)
We will	set off the timer	show the ponies	show off the ponies	show the ponies off	display the ponies	4 (2.8)/ 8.9 (0.3)	5.1 (2.3)/ 7.6 (1.3)	6.8 (2.6)/ 8.7 (0.7)
He will	run up the hills	hold the lizard	hold up the lizard	hold the lizard up	carry the lizard	7.2 (1.9)/ 7.3 (1.9)	7.6 (1.2)/ 8 (1)	7 (1.7)/ 6.1 (3.2)
He will	run over the trail	pass the dessert	pass over the dessert	pass the dessert over	skip the dessert	5 (2.9)/ 5.1 (3)	7.3 (1.8)/ 8.3 (1.1)	4.9 (3.5)/ 6 (2.7)
They will	keep up the secret	carry the groceries	carry up the groceries	carry the groceries up	lift the groceries	8.8 (0.4)/ 7.8 (0.8)	8.1 (1)/ 8.2 (1.1)	7.8 (1.3)/ 7.4 (1.2)
They will	live down the lie	pull the photo	pull down the photo	pull the photo down	remove the photo	6 (2.1)/ 6.2 (2.1)	5.9 (3)/ 7 (2.1)	7.5 (1.1)/ 7.3 (2.5)
They will	take over the position	look the part	look over the part	look the part over	examine the part	2.7 (3.1)/ 2.5 (1.9)	3.2 (2.7)/ 4.8 (3)	4.1 (3.2)/ 3.8 (2.8)
They will	boil down the soup	shoot the hawk	shoot down the hawk	shoot the hawk down	kill the hawk	7.8 (2.5)/ 8.1 (0.9)	8.2 (1)/ 8.2 (1.9)	7.7 (1.3)/ 7.5 (0.8)
He will	see through the rumors	think the thoughts	think through the thoughts	think the thoughts through	contemplate the thoughts	7.3 (1.8)/ 7.5 (1.2)	5.8 (1.5)/ 7 (1.3)	7.2 (1.8)/ 7.6 (2.2)
We will	pull off the sticker	mark the pages	mark off the pages	mark the pages off	check the pages	8 (0.9)/ 7.2 (1.8)	7.2 (1.6)/ 6.2 (2.9)	8.1 (1.9)/ 5.3 (2.7)
She will	throw out the trash	buy the store	buy out the store	buy the store out	empty the store	5.6 (2.8)/ 4.4 (3.1)	5.7 (2.7)/ 5 (2.7)	3 (2.1)/ 4.3 (3.2)
We will	dress up the doll	charge the phone	charge up the phone	charge the phone up	power the phone	8.7 (0.5)/ 8.3 (1.3)	8.3 (1.3)/ 8.8 (0.4)	7.7 (1.6)/ 6.3 (2.6)
He will	strike down the balance	hold the job	hold down the job	hold the job down	keep the job	6.6 (2.2)/ 8 (1.7)	7.6 (1.7)/ 7.1 (2.1)	8.4 (1)/ 8.5 (1.1)
They will	catch up the students	mark the lumber	mark up the lumber	mark the lumber up	measure the lumber	6.5 (2.3)/ 6.4 (2.5)	7.4 (2.5)/ 5.6 (2.9)	6 (2.2)/ 3.5 (2.4)
She will	break off the stick	pair the geese	pair off the geese	pair the geese off	partner the geese	7.6 (2.4)/ 8.4 (1)	6.4 (2.6)/ 8 (1.2)	8.5 (0.7)/ 8.5 (0.7)
We will	call forth the juror	bring the gifts	bring forth the gifts	bring the gifts forth	reveal the gifts	7.2 (1.8)/ 8 (1.3)	7.8 (1.4)/ 7.9 (1.3)	5.5 (1.1)/ 4.5 (1.6)
She will	drive away the candidate	give the prizes	give away the prizes	give the prizes away	donate the prizes	7.9 (1.3) /8.3 (1.1)	8.4 (.8)/ 8.3 (1.1)	6.2 (2.6)/ 7.8 (1.9)
We will	bust out the moves	bail the prisoner	bail out the prisoner	bail the prisoner out	free the prisoner	7.4 (2)/ 7.6 (2.5)	8.5 (1.3)/ 7.7 (1.3)	7.5 (1.4)/ 7.9 (1.2)
He will	lock up the bicycle	cut the meat	cut up the meat	cut the meat up	slice the meat	8.8 (0.4)/ 8.4 (1.1)	8.7 (0.5)/ 8.5 (1)	8.9 (0.3)/ 8.6 (0.7)
He will	eat up the spaghetti	cover the research	cover up the research	cover the research up	conceal the research	6.9 (2)/ 3.7 (3.2)	7.1 (1.5)/ 4 (3.2)	5.2 (2.1)/ 6.2 (3)
They will	wring out the rag	find the word	find out the word	find the word out	discover the word	6.7 (1.9)/ 7.6 (1.1)	7.5 (1)/ 7.4 (2)	7.9 (1.4)/ 8.3 (1.3)
She will	patch up the tire	set the boundaries	set up the boundaries	set the boundaries up	outline the boundaries	8 (2.2)/ 8.4 (0.5)	8.5 (0.8)/ 8.4 (1)	7.2 (1.6)/ 7.6 (1.4)
She will	finish up the game	close the café	close up the café	close the café up	shut the café	8.22 (1.7)/ 7.5 (2.3)	8.6 (0.5)/ 8.4 (1.3)	8.4 (0.8)/ 8.2 (1.1)

Note. V represents 'verb'; Aux represents 'auxillary'.

Received September 8, 2016
Revision received December 21, 2016
Accepted December 23, 2016 ■