Journal of Memory and Language 87 (2016) 38–58

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Part and whole linguistic experience affect recognition memory for multiword sequences

Cassandra L. Jacobs [a,*], Gary S. Dell [a], Aaron S. Benjamin [a], Colin Bannard [b]

[a] University of Illinois at Urbana-Champaign, United States
[b] University of Liverpool, United Kingdom

ARTICLE INFO

Article history:
Received 17 January 2015
revision received 27 October 2015
Available online 19 November 2015

Keywords:
Phrase frequency
Compositionality
Word frequency
Recognition memory

ABSTRACT

Low frequency words (like *wizard*) are better remembered in recognition memory than high frequency words like *tree*. Previously studied low frequency words are endorsed more often than high-frequency words, and unstudied low frequency lures attract fewer false alarms than high frequency lures. In order to evaluate whether repeated experience of phrases has the same effect as that of words, we tested whether infrequent combinations of words (like *psychic nephew*) are better recognized than frequent word combinations (like *alcoholic beverages*). In contrast to single words, people were more biased to endorse high-frequency phrases, but phrase frequency did not affect discrimination between studied and unstudied phrases. When high and low frequency nouns were embedded in adjective-noun phrases of equal frequency (e.g. *handsome wizard* and *premature tree*), people were better able to recognize phrases containing low frequency than high frequency nouns. Taken together, the high frequency phrase bias and the low frequency embedded-noun advantage suggest that the recognition of word sequences calls on prior experience with both the specific phrase and its component words.

Introduction

Researchers have carried out thousands of experiments in which word frequency is manipulated with the goal of understanding how words are processed, produced, and remembered. For the most part, this research demonstrates that low frequency words are less easily acquired, comprehended, and produced than more common words (see Ellis (2002) for a complete review). More recently, the question of whether multi-word sequences (or phrases) might exhibit frequency effects has been assessed. As with common words, high-frequency phrases are associated with benefits in reading time (Bannard, 2006; Smith & Levy, 2013), phrase decision reaction time (Arnon & Snider, 2010), greater fluency and speed of production (Arnon & Priva, 2013; Bannard & Matthews, 2008; Janssen & Barber, 2012) and recall memory (Tremblay & Baayen, 2010).

Phrase frequency effects are of interest because they tell us about the cognitive mechanisms implicated in the production and comprehension of word sequences. The findings cited above indicate that the combination matters. A phrase is not just a list of words. More importantly, these results are analogous to the discovery in morphology that people are sensitive to the frequency of whole words, and the inference that word processing involves some knowledge of the whole as well as of the component morphemes (Bien, Levelt, & Baayen, 2005). However there remain many questions about the exact nature of the mechanisms involved. There are two main issues that arise: compositionality and abstraction. In this paper, we

* Corresponding author at: 603 E. Daniel St., Department of Psychology, University of Illinois Urbana-Champaign, Champaign, IL 61820, United States.
E-mail address: cljacob2@illinois.edu (C.L. Jacobs).

http://dx.doi.org/10.1016/j.jml.2015.11.001
0749-596X/

present five recognition memory experiments that address these issues.

The compositionality issue concerns the representation of a phrase, by which we loosely mean the mental/neural codes implicated in producing and understanding it, and whether these codes are a predictable superset of the representational spaces involved in the production and comprehension of its parts. So, a person's knowledge of the phrase *red house* may be compositional, derived solely from their knowledge of its component words, *red* and *house*. If the phrase is not compositional, but instead holistic, a language user's representation of it might be largely separate from their representation of the component words. Phrases vary in the extent to which their meaning is predictable from their parts, with the meaning of *red house* being much more predictable than the meaning of *red herring*. A phrase with an unpredictable meaning therefore may seem to require a largely disjoint representation. It is also plausible that such representations might also be employed for more predictable phrases as well. Indeed, the discovery of phrase-frequency effects has occasionally been taken to indicate that the representation of phrases is holistic. However, while such results indicate that speakers do encode knowledge of the sequences, they do not address the question of whether combination-specific knowledge is utilized instead of or in addition to word knowledge when processing phrases.

The issue of abstraction concerns how we encode multiple instances of the same phrase. A phrase could be represented either as a collection of episodic memories, each containing a token of that phrase, or as a single abstract encoding of the type with an associated strength. In the episodic approach, the particular episodes in which a phrase is experienced are kept distinct, and effects of the phrase frequency would be attributed to the number of such episodes. In particular, any processing benefits that accrue to common phrases would be attributed to the greater availability of relevant memories to guide the processing (e.g. Goldinger, 2004; Hintzman, 1988). Alternatively, in the abstractionist approach, each phrase type is a single representation such as a node in a lexical-semantic network (e.g. MacKay, 1982). If *red house* had been experienced a number of times, a node would represent the phrase type, with its strength (e.g. resting level of activation) proportional to its frequency. Of course, the abstractionist approach does not deny the existence of episodic knowledge about phrases. It simply assumes that the abstraction exists in addition to episodic memories, and it is this abstraction that plays the major role in how the phrase is processed, rather than the episodes.

Some accounts of word and phrase frequency effects are neither clearly episodic nor explicitly abstractionist in the sense that they have a single node for each word or phrase. Multi-level connectionist models (e.g. Seidenberg & McClelland, 1989) occupy an interesting middle ground in this respect. Each experience changes the weights in a network (as with an episode) and yet these alterations are not stored separately, but rather are superimposed. The resulting superposition is somewhat like an abstraction, but it is not easily recognized as such and is certainly not a single node. A related class of models, the naive discrimination learning models (e.g. Baayen, Hendrix, & Ramscar, 2013; Baayen, Milin, Đurđević, Hendrix, & Marelli, 2011), also lacks discrete episodes and explicit representations of abstract items. For example, one such model by Baayen et al. (2011) consists of an input layer of letters and letter pairs and an output layer of semantic features. The model learns input–output mapping for words or phrases by applying the Rescorla–Wagner (1972) equations to probabilistic information obtained from corpora. Even though it lacks explicit words or phrases, its behavior (e.g. mapping accuracy) reflects both word and phrase frequency.

As we noted, benefits for high frequency phrases have been clearly demonstrated in comprehension and production tasks, and in memory recall. In our studies, we turn to a different memory task in order to address phrase frequency from a new perspective: the yes–no recognition task. Importantly, the high frequency advantage apparent in linguistic tasks and recall is not evident in recognition memory; in fact, low frequency words are recognized better. We more easily pick out *panther* when it was studied and reject it when it was not studied than a higher frequency word like *cat*. That is, low frequency words attract more hits and fewer false alarms than high frequency words. This pair of results is one manifestation of a broader category of what are called mirror effects, effects in which a particular class of items or condition of study for a set of items leads to them being more easily discriminated (Glanzer & Adams, 1985). The mirror effect allows us to derive predictions about frequency effects for phrases in recognition, and thus examine the cognitive mechanisms implicated in their processing.

In the next section we review studies of word frequency in language processing and acquisition. Next, we discuss the degree to which the high frequency word advantage is reflected in larger sequences of linguistic units, such as multi-word sequences. Finally, we review the mirror effect in yes–no recognition memory and consider its implications for multi-word sequences.

The high frequency word advantage

High frequency words are easier to process than low frequency words. The language processing system is adaptive and thus learns to process more probable events with greater facility (Dell & Jacobs, 2015; Forster & Chambers, 1973; Jusczyk, 1997; Lively, Pisoni, & Goldinger, 1994; Saffran, Newport, & Aslin, 1996). For example, identification of high frequency words is more robust under both noisy (Howes, 1952) and clear (e.g. Forster & Chambers, 1973) conditions.

When reading words in text, reading times scale inversely with the logarithmic frequency of the word that is being read, with the most common words in the language being barely read at all or even skipped entirely (e.g. Demberg & Keller, 2008; Howes & Solomon, 1951; Rayner, 1998; Smith & Levy, 2013). When a text contains low frequency words, comprehension suffers (Diana & Reder, 2006; Freebody & Anderson, 1983; Marks, Doctorow, & Wittrock, 1974). In production, uncommon words are retrieved more slowly during picture naming (e.g. Oldfield & Wingfield, 1965) and produced less

accurately (Dell, 1990; Kittredge, Dell, Verkuilen, & Schwartz, 2008). In short, the deck is stacked against low frequency words in linguistic tasks.

High frequency words are also easier to acquire. Children respond from a very early age to highly probable content words like *milk*, producing them reliably early in development (Tomasello, 1998). Familiar words contribute to the refinement of phonological categories (Martin, Peperkamp, & Dupoux, 2012; Swingley, 2009) and the acquisition of syntax (Fisher, Gertner, Scott, & Yuan, 2010).

The child also notes the frequency of recurring phoneme combinations to pick words out of the speech stream (Saffran et al., 1996). Algorithms that attempt to simulate this process, however, sometimes fail to find words or morphemes, and instead under-segment, treating multiword sequences, collocations, and frequent phrases as big words (Feldman, Griffiths, Goldwater, & Morgan, 2013; Goldwater, Griffiths, & Johnson, 2009). The word segmentation literature sees this result as a failure, but their results raise the interesting possibility that high frequency phrases may be discovered in the same way that common words are, a claim that brings us to the question of phrase frequency effects. If there are common attributes to the representations of these under-segmented phrases and entire words, then "erroneously" treating common phrases as single words may sometimes be useful behavior in language acquisition (Tomasello, 2006) and potentially in its ongoing use by mature language users.

The high frequency phrase advantage

People are sensitive to the frequencies of word sequences, as they are to individual words. One of the first studies to demonstrate phrase frequency effects was Bannard and Matthews (2008). In that study, phrases such as *a drink of milk* and *a drink of tea*, which were matched for semantic class, word frequency, and two-word (bigram) frequency were presented to young children. These phrases differed only in their phrase frequency as measured in a corpus of child-directed speech. *Of milk* is about as common as *of tea*, and *milk* and *tea* are also comparably common in a corpus of British English. However, *a drink of milk* is more common than *a drink of tea*. Recordings of these phrases were played to young children, who were asked to repeat them; they made fewer errors when repeating the more frequent phrases, and were quicker in doing so. These results suggest that children's experience of particular phrases, as well as of words, have a measurable impact on the representations that underlie their developing linguistic abilities.

Adult production seems sensitive to phrase frequency as well. In particular, prosodic measures such as duration reflect the frequencies of multiword combinations as well as the frequencies of the component words. In one study (Arnon & Priva, 2013), frequent 3-word sequences (trigrams) were produced with shorter duration than infrequent trigrams, even when considering word frequencies within those phrases as well. That is, the more frequent a *drink of milk* would have a shorter duration than a *drink of tea*, just as Bannard and Matthews (2008) found in children. Shaoul, Harald Baayen, and Westbury (2014) used a phrase Cloze completion task to assess implicit knowledge of phrase frequencies. The endings that speakers provided to the incomplete phrases mirrored the phrase frequency distributions that have been observed in corpora. Speech onset times are also sensitive to phrase frequencies. Janssen and Barber (2012) constructed phrases such as *blue car* and *red car* in Spanish, which differed in their phrase frequencies. They asked participants to name pictures that could be described by these phrases. High frequency phrases, but not necessarily phrases containing high frequency words, were initiated more quickly.

Multiword sequences can impact comprehension as well. Smith and Levy (2013) examined whether word and phrase frequencies jointly influence reading times for text, confirming previous findings in reading (Bannard, 2006). They found that there were contributions of word, bigram, and trigram frequencies, such that the more frequent each of these components were, the more quickly those words and word sequences were read. Furthermore, reading times were logarithmic with respect to phrase frequency, an effect that had been robustly demonstrated with words (Howes & Solomon, 1951; Rayner, 1998). This result occurred despite the fact that their statistical model contained no syntactic information, just information about the lexical sequences. Other work on sentence processing suggests that models using information about phrases only can explain reading times just as well as models with syntax (Frank, Bod, & Christiansen, 2012).

Taken together, these results demonstrate that language users represent phrases. At this point, though, there is uncertainty as to the degree to which such representations are holistic. There are many more possible phrases than words. For a vocabulary of N words, there are n^2 bigrams, n^3 trigrams, etc. Thus, a language user often confronts a phrase for the first time, and its meaning will have to be constructed compositionally from its parts (i.e. its words) and from context (e.g. Medin & Shoben, 1988; Smith & Osherson, 1984). Furthermore, there is greater difficulty in estimating the frequencies of phrases, especially those in the lowest frequency ranges (Evert, 2005; Piantadosi, 2014). Given this, efficient encoding of language then might involve representing phrases in a way that makes use of the knowledge of their component words and separately representing only the information that is not contained within the word-level representations (such as frequency of occurrence of the combination).

The contribution of individual words to the fluency of phrase processing is difficult to assess using the previously employed methods. For both words and phrases, higher frequency linguistic events are easier to process and produce than lower frequency events, and the correlation between the frequency of phrases and their component words can make them hard to tease apart. As we noted, a major exception to the general linguistic advantage for high frequency events is apparent in tests of recognition memory, a topic to which we turn now.

A paradox of word frequency

Low frequency words have long been documented to do better on recognition memory tests than high frequency

words. Specifically, low frequency words are better identified when they were studied (more hits) and better rejected when they were not studied (fewer false alarms). Crucially, because of the increase in hits *and* the decrease in false alarms to low frequency words, the mirror effect represents a situation that cannot be strictly explained by any one class of words attracting more yes responses than high frequency words, since any such advantage would not play out in opposite advantages for studied and unstudied items.

The word frequency mirror effect is in itself part of a broader set of mirror effect phenomena. In general, words with strange meanings, odd letter combinations, or which occur in only a few contexts in real life, all exhibit the mirror effect (Glanzer & Adams, 1985; Malmberg, Steyvers, Stephens, & Shiffrin, 2002; Seamon & Murray, 1976; Steyvers & Malmberg, 2003; Zechmeister, 1972). The mirror effect has also been demonstrated for faces varying in typicality (Vokey & Read, 1992) and picture-word pairs that have unusual labels (Bloom, 1971). Malmberg et al. (2002); see also Shiffrin & Steyvers, 1996) attribute the effect to "feature frequency," a conceptualization that suggests that the mirror effect generalizes to any arbitrary distinctive features that are attended to in processing a stimulus, with rare features providing a benefit to memory.

Some accounts of the word-frequency effect in recognition appeal to the impoverished episodic representation of low frequency words. Because people experience low frequency words fewer times than high frequency words, they have more memories of high frequency words. These multiple memories lead generally to the high-frequency advantage in most language processing tasks. But this benefit comes at a cost to memory. We have seen many *cats* but few *panthers* and, consequently, are better able to recover the particular contexts in which we experienced *panther*. It is this recovery of the context of the studied list that is crucial for a recognition decision (Reder et al., 2000). Other accounts have emphasized that the amount of change that our memorial representations undergo is greater for a low frequency word when it is encoded (Benjamin, 2003; Reder et al., 2000). As a result, unstudied low frequency words seem especially novel in comparison to unstudied high frequency words (Benjamin, Bjork, & Schwartz, 1998; Brown, Lewis, & Monk, 1977). Thus, low frequency words benefit from a one-two punch in a recognition memory test. The first effect is that it is easier to recover the studied episode for a low frequency word, leading to more hits. The second effect is that unstudied low frequency words will look especially unfamiliar, leading to fewer false alarms.

If phrases are represented holistically, low frequency phrases should garner more hits and fewer false alarms in a recognition memory test, much like low frequency words. In Experiment 1 we test whether phrase frequency induces a mirror effect in recognition memory in the same way that word frequency does.

Experiment 1a

In this experiment, participants studied 26 adjective-noun phrases that varied in phrase frequency. The studied phrases were sampled from a set of 52. After a 30-min retention interval, participants saw the complete set of 52 phrases and judged whether the phrase had been previously studied or not.

Method

Participants

Participants were 40 undergraduate students from the University of Illinois who acquired no language before the age of 5 other than English. Participants received course credit for their participation in this experiment.

Materials

52 adjective-noun phrases served as stimuli. These were extracted from the Google 1T n-gram corpus (Brants & Franz, 2006) using word lists for each category extracted from the part-of-speech-tagged British National Corpus (BNC). In these phrases, both the nouns and the adjectives had a restricted frequency range of 19–23.5 (taking the $\log_2$ transform of their frequency in the corpus) (The British National Corpus, 2007). The phrases exhibited a relatively broad phrase frequency range ($5.4 < \log_2$ (phrase frequency) < 19.7; see Fig. 1). The most common phrases (*rheumatoid arthritis*, *alcoholic beverages*) were approximately as common as the least common adjective (*decadent*) and noun (*grasslands*) in our dataset. The stimulus set may be found in Table A1 of Appendix.

The items were selected so that phrase frequency in the materials did not correlate with word frequency (adjectives with phrases, $r = -0.09$, $p = 0.54$; nouns with phrases, $r = 0.17$, $p = 0.23$), which is not normally the case because a common phrase naturally makes its words more common. We also verified the lack of correlation between the two word frequencies ($r = 0.09$, $p = 0.50$). Phrase frequency, noun frequency, and adjective frequency were not

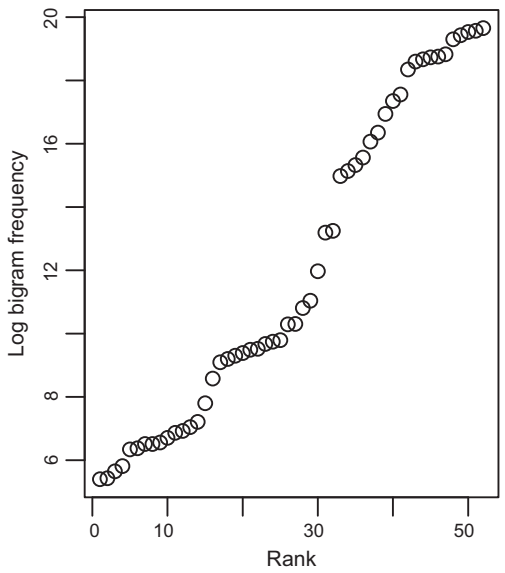

Fig. 1. Log-frequency rank plot illustrating the uniform distribution of the phrase frequency range for the stimulus set in Experiment 1.

correlated with adjective or noun lengths in this stimulus set, and phrase frequency was not correlated with total phrase length. Phrase frequency was neither correlated with orthographic neighborhood density ($r = -0.05$, $p = .73$) nor orthographic probability ($r = 0.04$, $p = .76$), which were calculated using the CLEARPOND database (Marian, Bartolotti, Chabal, & Shook, 2012).

Procedure

Each participant received a different set of 26 phrases to study. So that phrase frequency was varied to an even extent in each study set, a randomly seeded sampler selected items using a median split based on the items' phrase frequency, following a method used in prior work (Benjamin, 2003; Tullis & Benjamin, 2012). For each participant, we took random subsets of the top and bottom halves of the phrase frequency range, obtaining 13 phrases from the top, and 13 phrases from the bottom.

Participants were told, "You will be presented with pairs of words that combine to make meaningful phrases that you should memorize. You should try to remember as many of the pairs of words as you can." They were not given further specification about what type of memory test they would complete. The studied phrases were presented in random order. Each phrase was presented at the center of a computer screen for 1 s, with a 1 s inter-stimulus interval. After the study phase, participants put together a puzzle of St. Basil's Cathedral for 30 min.

At test, all 52 phrases were presented, again in random order. Each phrase was presented at the center of the screen while participants made a recognition judgment. To make their judgment, they pressed the "p" key if the item was "old" and the "q" key if the item was "new". Participants could take as much time as they wanted to make a response.

Results

The recognition judgments were analyzed using logit mixed effects models. Responses were modeled as a function of whether or not the item being responded to was in the studied list (studied status), the log phrase frequency (phrase frequency), and the interaction of these variables (Wright, Horry, & Skagerberg, 2009). The random effects included random slopes for the effect of studied status on response bias and intercepts by participants. We also included random intercepts for items. All analyses were completed in the R package `lmer` version 1.1–6 using the optimizer `bobyqa` to prevent non-convergence problems (Bates, Maechler, Bolker, & Walker, 2014; Powell, 2009). All coefficients represent changes in log odds of a yes versus a no response as a function of the predictor.

Participants demonstrated the ability to correctly identify studied items ($\beta = 1.63$, $z = 16.92$, $p < .001$). If the predicted greater accuracy for low frequency phrases had occurred, then the interaction coefficient between phrase frequency and studied status would be negative and significantly different from zero. In fact, this interaction was not found and actually was slightly positive ($\beta = 0.08$, $z = 1.28$, $p = .20$). What we saw instead was only a bias for participants to say that they had studied high frequency phrases, regardless of whether the phrase had been studied or not ($\beta = 0.39$, $z = 4.55$, $p < .001$). We illustrate the bias effect that phrase frequency has on hits and false alarms in Fig. 2. The full model is reported in Table 1. Random effects are reported in Appendix in Tables A3 and A4.

Discussion of Experiment 1a

The lack of a phrase frequency mirror effect suggests that the effect of phrase frequency on participants' representation of their language is different from the effect of word frequency, at least with respect to recognition memory. This effect is surprising, because many stimuli benefit from some kind of "unusualness" in recognition memory tasks (Glanzer & Adams, 1985; Malmberg et al., 2002; Seamon & Murray, 1976; Steyvers & Malmberg, 2003; Vokey & Read, 1992). We note, though, that the bias to respond "yes" as phrase frequency increases does suggest that frequency influences performance. Therefore speakers do somehow encode information about the frequency of the word combinations.

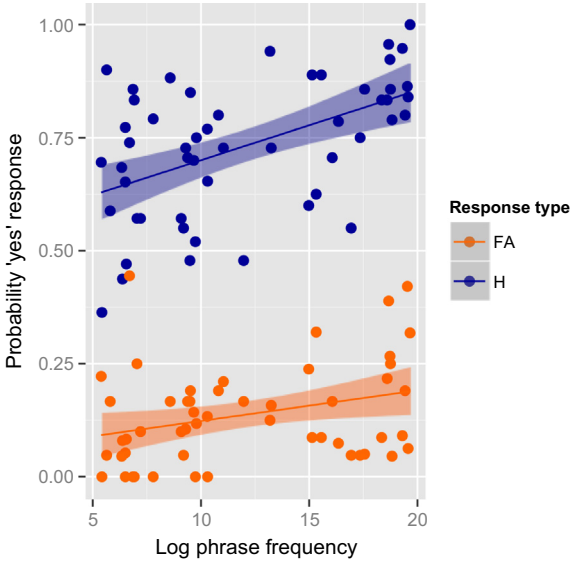

Fig. 2. Hit rates and false alarm rates to phrases for Experiment 1a as a function of phrase frequency, collapsed across participant variance. The shaded areas correspond to one standard error around the regression line. Participants make more hits and false alarms to high frequency phrases.

Table 1
Summary of Experiment 1a fixed effects.

Predictor	Parameter estimates		Wald's test	
	Log-odds	S.E.	*Z*	*p*
(Intercept)	−0.46	0.12	−3.95	<.001
Old or new status	1.63	0.10	16.92	<.001
Phrase frequency (bias)	0.39	0.09	4.55	<.001
Phrase frequency by old–new status	0.08	0.06	1.28	.20

Note: Significance obtained at $p < .05$.

Experiment 1b

The unexpected results of Experiment 1a motivated an attempted replication. In Experiment 1b, we looked again for a phrase frequency mirror effect. We also sought to rule out the possibility that the bias toward saying "yes" to high frequency phrases was due to the simultaneous presentation of the two words of the phrases at study and test. Presenting the words in sequence could encourage the separate processing of the individual words and possibly nullify the phrase frequency bias effect of Experiment 1a. Because of these concerns, we repeated Experiment 1a in all respects, except that the two words of the phrases were presented in sequence during study and at test.

Method

Participants

Participants were 40 undergraduate students from the University of Illinois who acquired no language before the age of 5 other than English. Participants received $8 for their participation in this experiment.

Materials and procedure

The only difference between Experiment 1a and Experiment 1b was the manner in which the phrases were presented at study and at test. In Experiment 1b, we presented phrases word by word, instead of simultaneously. At the beginning of every study trial, we presented the adjective at the center of the screen for 450 ms, followed by a 50 ms blank screen before presenting the noun at the center of the screen for 450 ms. Words within phrases never appeared together. There was a 1 s intertrial interval before the presentation of the next phrase. To the extent that participants chose to encode the pairs of words as phrases, it would likely be due to the longer intertrial interval between phrases than the interstimulus interval between words in a phrase. After study, participants again put together a puzzle for 30 min.

Presentation of the phrases at test followed a similar design. Participants were asked to respond as to whether the phrases presented were ones they had studied or not. The rate of presentation of the words within the phrases was the same as at study, with the adjective and noun on the screen at separate times. In addition, judgments were solicited only after both words had been presented and removed from the screen. Only the cues as to what response to make ("p" for "old" and "q" for "new") were on the screen during the response. Participants were told to judge whether they had studied the entire phrase. They were allowed to take as much time as they needed to make a response.

Results

Analysis followed as in Experiment 1a. We again found no low frequency advantage, as evidenced by the lack of interaction between phrase frequency and studied status ($\beta = 0.05$, $z = 0.65$, $p = .26$). Also as before, we found that participants were biased toward saying that they had studied high frequency phrases, though this effect was somewhat weaker in this experiment than in Experiment 1a ($\beta = 0.27$, $z = 2.31$, $p < .05$). We illustrate the phrase bias in Fig. 3. The full model is reported in Table 2. Random effects and random effects correlations are presented in Appendix in Table A5.

Discussion of Experiment 1b

In Experiment 1b, we replicated the findings of Experiment 1a. There is evidence in both experiments that participants use phrase frequency to make their judgments about whether a phrase was studied or not (evidenced in a bias to say "yes" to more common phrases), but phrase frequency does not appear to impact people's ability to discriminate studied from unstudied phrases. The fact that the results of 1a replicated even when the words are presented individually suggested that the phrases are processed as units.

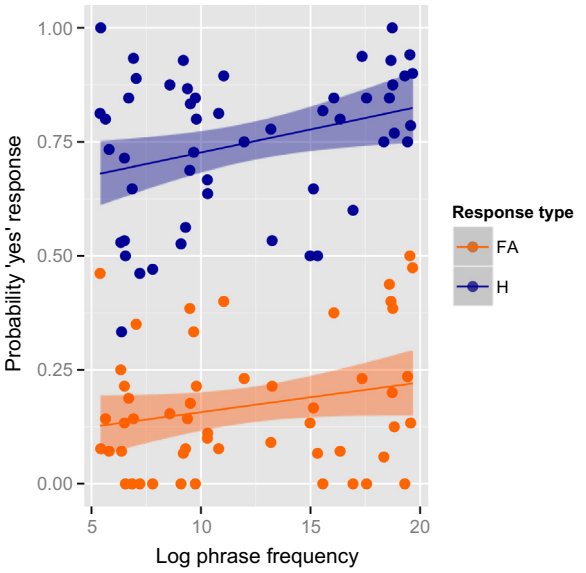

Fig. 3. Hit rates and false alarm rates to word-by-word presented phrases for Experiment 1b as a function of phrase frequency, collapsed across participant variance. The shaded areas correspond to one standard error around the regression line. As in Experiment 1a, participants show more hits and false alarms to high frequency phrases.

Table 2
Summary of Experiment 1b fixed effects.

Predictor	Parameter estimates		Wald's test	
	Log-odds	S.E.	Z	p_z
(Intercept)	0.09	0.18	0.62	.73
Old or new status	1.77	0.13	14.14	<.001
Phrase frequency (bias)	0.27	0.12	2.31	<.05
Phrase frequency by old–new status	0.05	0.08	0.65	.26

Note: Significance obtained at $p < .05$.

Interim discussion

Given the results of Experiments 1a and 1b, we propose the following model of phrase frequency effects, which is outlined in Fig. 4. We base our model on episodic accounts of the word frequency effect in recognition memory, most specifically Reder et al. (2000), as other models of frequency effects (e.g. Baayen et al., 2013) have not been developed for recognition memory. Specifically, we propose that each experience with a multiword sequence leaves a trace (with each episode represented by a star in the figure). For example, consider the phrase *psychic nephew*. Each experience with this particular phrase results in another episodic token. This includes the experience of studying *psychic nephew* in an experimental list (represented by the red star). The Reder et al. model accounts for the higher hit rates for low frequency items and higher false alarm rates for high frequency items using two mechanisms. Because an individual episodic memory has fewer competitors for low frequency items, the study episode for that item is more likely to be chosen at test. Higher false alarm rates for unstudied high frequency words arise because the baseline activation of the word is higher. In sum, when high frequency words are studied, the individual episode is more difficult to retrieve among competitor episodes, but when the high frequency word is new, the baseline activation or familiarity of that item is already high, leading to a bias to say yes to that item even when it was unstudied. Critically for phrase memory, these episodic tokens can be retrieved from memory not just from a cue that matches the entire phrase, but from a cue that matches part of it, such as the noun. So, the tokens can be thought of sets of as features that represent the experience of the phrase, with a featural cue having the capacity to retrieve an entire episode. Crucially, words act as features. Thus, this model takes a compositional, episodic approach. Relevant phrasal episodes are retrieved because the episodes contain their words, and the influence of frequency is attributed to the multiplicity of episodes.

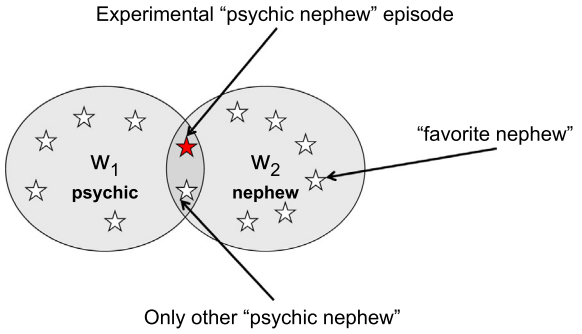

Fig. 4. Schematic for the representation of the episodic memories involved in the multiword phrase *psychic nephew*. Each word within the phrase is represented as a circle containing episodic memories (stars). So, the word *nephew* might have many other memories, such as *favorite nephew*. When a phrase is processed at study, the red star is placed at the intersection of the two words. The process of recognition at test requires retrieving the red star that was experienced during study. At test, all memories associated with the words compete for retrieval, so words with many more memories make that phrase more difficult to locate. Because phrases are generally impoverished (the intersection between two words is often very unpopulated), phrases have very few competitors from the same phrase, so word frequency becomes more important. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Why is there no benefit for low frequency phrases? In a recognition test, each word of a test phrase serves as a retrieval cue. So, when *psychic nephew* is presented at test, it has the potential to retrieve all episodes with *psychic* and all episodes with *nephew*, as indicated by circles in Fig. 4. We assume, consistent with prior work (e.g. Smith & Osherson, 1984), that nouns contribute more to the meaning of an adjective-noun phrase, so episodes that overlap in just the noun will be more retrievable than those that share only the adjective. As in the model of Reder et al. (2000), recognition judgments are determined in part by whether or not the critical episode (the red star representing that the phrase that was studied in the experiment) has been retrieved. Finding that episode is more difficult when many other episodes are active. Because words are far more frequent than phrases on average, the main determinant of the number of interfering episodes will be the frequencies of the *words* within a phrase, particularly the noun, and not the phrase itself. Because there are relatively few memories of the whole phrase, they contribute few interfering episodic tokens. In fact, for many phrases, the number of possible episodic tokens is possibly zero (e.g. *psychic nephew*), so only word-level information would be available for use during search for the critical episode.

Why is there a bias to say “yes” for high frequency phrases? Reder et al. (2000)’s word-frequency model assumes that there exist abstract representations of word types in addition to episodic memories. When a word is more frequent, this representation is stronger and contributes to a feeling of familiarity, and thus to a bias to say yes. We can borrow the same account for phrase frequency. This requires that there be an abstract holistic representation of the phrase (sensitive to frequency) in addition to the hypothesized phrasal episodes. The representation of strength does not necessarily have to be a property of a phrasal “node.” For example, it could be an association strength between, say, the adjective and the noun, such as might be acquired from a model that learns through prediction error about subsequent words, given previous ones. Alternately, we can dispense with abstractions and hypothesize that somehow, the participant is able to discern phrasal familiarity from the set of retrieved episodes that match on both words (e.g. the number of episodes that contain both *psychic* and *nephews*). Our data do not distinguish between this episodic and abstractionist account of the bias to say “yes”.

The model we outline generates a specific prediction from its assumption that phrasal episodes have a compositional nature: The frequency of the words within a phrase should affect the amount of interference that is generated at test, such that phrases containing low frequency words should be better recognized than phrases containing high frequency words, leading to a word frequency mirror effect. We specifically expect to see a contribution of noun frequency to phrase memory because of the greater contribution of the noun to phrase meaning. A phrase with a frequent noun should tend to attract fewer hits and more

false alarms than a comparable phrase with a less common noun. In the next section we run a combined analysis of Experiments 1a and 1b to look for preliminary evidence of a noun frequency mirror effect.

Cross-experiment analysis

Norming study

Phrases, like words, have conceptual properties associated with them that may enhance or obscure memory for those phrases. In particular, phrases differ from monomorphemic words by having meanings that can be composed, or which are idiomatic (e.g. *red house* versus *red herring*). However, like words, phrases may be familiar concepts or not. It is necessary to ask, therefore, whether the effects of phrase frequency that we saw in Experiment 1a and 1b might be in part due to the relationship between these factors and phrase frequency. To account for these factors, we conducted an additional norming study with 50 participants from the University of Illinois course credit subject pool. Each participant rated each of the 52 phrases for concreteness (e.g. "This phrase denotes a real-world entity"; Paivio, Yuille, & Madigan, 1968), imageability (e.g. "I can easily picture what this phrase describes."), and compositionality (e.g. "Are alcoholic beverages beverages that are alcoholic?"; Szabo, 2013) on a three-point scale ("Not at all", "Somewhat", and "Definitely"). We then averaged over all 50 participants for each of the 52 items to obtain concreteness, imageability, and compositionality scores to use as control variables.

Analysis

We first constructed a null model using the results of the norms to predict memory performance, and then introduced our key predictors: phrase frequency, noun frequency, and adjective frequency. Concreteness and imageability were highly correlated ($r = .93$), while compositionality was less strongly correlated with concreteness ($r = .62$) and imageability ($r = .69$). Due to these correlations, we only added imageability and compositionality, and their potential interactions with studied status, to the null model. Because this analysis uses the data from both Experiments 1a and 1b, we added *Experiment* as a fixed effect. Experiment did not significantly interact with any terms in the model, so we did not retain the higher order interactions, and only include Experiment as an additive term in the null model. We then conducted a stepwise additive model building procedure to test for differential effects of phrase frequency, noun frequency, and adjective frequency on hits and false alarms. Random effects terms with near-perfect correlations were removed (Baayen, Davidson, & Bates, 2008); please see Table A6 in Appendix for the full random effects structure.

First, we introduced phrase frequency and its possible interaction with studied status. This significantly improved model fit over the null model (χ^2 (7) = 62.48, $p < .001$).We then found that the addition of a noun frequency main effect term as well as the interaction of noun frequency with studied status again improved model fit (χ^2 (11) = 19.89, $p < .05$). Finally, we asked whether adjective frequency contributed anything to model fit. The adjective terms did not significantly improve the likelihood of the model (χ^2 (2) = 2.56, $p = .28$), so adjective frequency and its interaction with studied status were not included in the final model. The final model is presented in Table 3.

Altogether, the results suggest that phrase and noun frequency contribute to recognition memory judgments. First, participants are more likely to say "yes" to a higher frequency phrase than a lower frequency phrase, regardless of whether the item was actually studied or not ($\beta = 0.36$, Wald $Z = 3.77$, $p < .001$). This result shows that the phrase-frequency bias effect identified in each of the two experiments is robust when phrasal differences in compositionality and imageability are taken into account.

Critically, phrases containing uncommon nouns show a benefit to recognition memory, as evidenced by a noun frequency mirror effect ($\beta = -0.41$, Wald $Z = -3.98$, $p < .001$). This was exactly what we predicted from our model. This suggests that memory for phrases depends at least in part on the distinctiveness of the component parts, specifically the nouns, which are central to the meaning of the phrase and have been implicated in prior research as an "anchor" in memory (Lockhart, 1969; Mata, Percy, & Sherman, 2013; Morris & Reid, 1972; Richardson, 1978; Yuille, Paivio, & Lambert, 1969). We note one additional finding from the final model: As has been reported previously in the literature (Paivio, 1971), increasing imageability led to greater phrase discriminability ($\beta = 0.18$, Wald $Z = 2.62$, $p < .01$).

The presence of a noun frequency mirror effect provides preliminary support for our account. It generally suggests that knowledge of words contributes to the processing of phrases, and thus that phrasal representations are not entirely holistic.

Experiment 2

The phrases used in Experiment 1 were taken from the Google *n*-gram corpus as described. While this tells us that they occurred on the Internet with some frequency, many of the infrequent phrases (e.g. "chrome throttle" or "psychic nephew") would not be encountered frequently in daily life, and consequently we cannot be sure that they

Table 3
Summary of Experiments 1a and 1b combined analysis.

Predictor	Parameter estimates		Wald's test	
	Log-odds	S.E.	*Z*	p_z
Intercept	−0.26	0.11	−2.31	<.05
Studied status	1.71	0.08	22.11	<.001
Phrase frequency	0.36	0.09	3.77	<.001
Experiment	0.66	0.18	3.75	<.001
Noun frequency	0.05	0.10	0.52	.30
Compositionality	−0.08	0.12	−0.66	.25
Imageability	−0.06	0.12	−0.51	.30
Phrase frequency * Studied status	0.08	0.05	1.52	.06
Noun frequency * Studied status	−0.23	0.06	−3.98	<.001
Imageability * Studied status	0.18	0.07	2.62	<.01
Compositionality * Studied status	0.01	0.07	0.14	.55

Note: Significance obtained at $p < .05$.

are meaningful to participants. This may put them at an encoding disadvantage, as has been seen in recognition memory for pseudowords (e.g. Diana & Reder, 2006). We therefore tested whether our key effects hold for another set of phrases where even the "low frequency" phrases are likely to be familiar and meaningful to participants.

To that end, we developed an additional stimulus set from the spoken portion of the Corpus of Contemporary American English (COCA; Davies, 2008), which consists primarily of publically broadcast material from the news, on talk shows, etc. These phrases therefore represent more easily recognizable phrases. We gathered a total set of 112 phrases (56 in a high frequency phrase list and 56 in a low phrase frequency list) meeting several criteria, which we discuss below.

All the phrases we gathered from COCA were compositional (nonidiomatic) adjective-noun phrases varying in their frequency of occurring in the subset of the database containing spoken English. We calculated the spoken frequencies of these phrases from the years 2009 to 2012, which represents more a recent and ecologically valid sample of the language the typical freshman undergraduate would experience while watching the news from the beginning of middle school through the most recent collection of data in COCA. Noun and adjective length did not significantly correlate with phrase frequency (Pearson's $r = -.11$, $p = .28$ and $r = -0.14$, $p = .16$).

Nouns and adjectives were deliberately selected to be higher in frequency than in Experiment 1 in order to increase the chances that the participants actually knew all of the words within the phrase, with the least common adjective and noun occurring 200 times more often than the least common phrase. Frequencies for the adjectives and nouns were restricted to the same range, from 1031 to 4021 and from 1026 to 4037, respectively out of the entire corpus from 2009 to 2012. As such, all nouns and adjectives were within a single power of 2 in COCA frequency. The lowest frequency phrases were "poor credit", "southern food", "fantastic panel", and "nice hair". The highest frequency phrases were "foreign language", "presidential candidate", "middle class", and "grand jury". Log_2 frequencies of the counts ranged from 2.32 to 9.57. These phrases are listed in Tables A2–A4 of Appendix.

Procedure

The procedure was the same as in Experiment 1a.

Results

The analysis proceeded as in Experiment 1. We replicated the key results of that experiment.[1] There was a main effect of studied status, suggesting that participants were highly accurate ($B = 2.60$, $z = 9.15$, $p < .001$. There was no interaction between studied status and phrase frequency ($B = -0.14$, $z = -1.07$, $p = .28$), indicating that there was no frequency-related mirror effect. Crucially, there was a main effect of log phrase frequency on whether participants were likely to call a phrase old or new ($B = 0.19$, $z = 2.09$, $p < .05$). When a phrase was high frequency (e.g. "foreign language") participants were more likely to say it was studied than a low frequency phrase (e.g. "angry crowd") regardless of whether the phrase had been studied or not. These results are summarized below in Table 4 and plotted below in Fig. 5. The random effects and random effects correlations are reported in Table A7 of Appendix.

Discussion of Experiment 2

The results of this experiment demonstrate that the bias to endorse high-frequency phrases as having been studied is not an artifact of the stimuli from Experiment 1, some of which may have been nonsensical to some subjects. We see the same pattern of results in this experiment as we

Table 4
Summary of Experiment 2 fixed effects.

Predictor	Parameter estimates		Wald's test	
	Log-odds	S.E.	Z	p_z
(Intercept)	−2.06	0.25	−8.35	<.001
Old or new status	2.60	0.28	9.15	<.001
Phrase frequency (bias)	0.19	0.09	2.09	<.05
Phrase frequency by old–new status	−0.14	0.13	−1.10	.28
Noun frequency (bias)	−0.02	0.09	−0.23	.64
Noun frequency by old–new status	0.05	0.11	0.47	.82

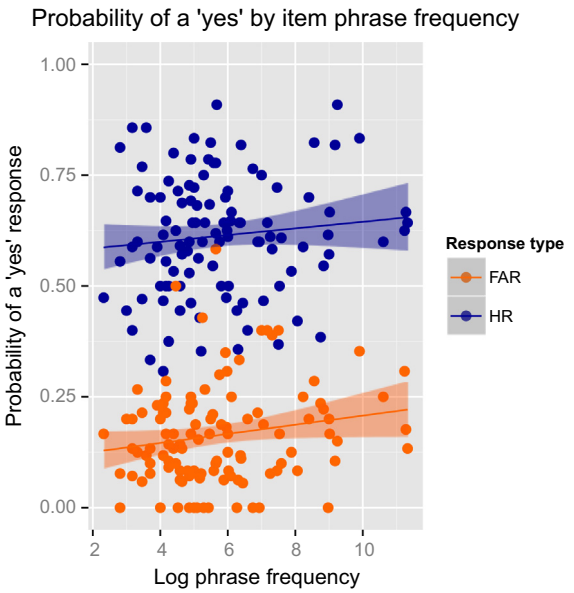

Fig. 5. Hit rates and false alarm rates to word-by-word presented phrases for Experiment 2 as a function of phrase frequency in COCA, collapsed across participant variance. The shaded areas correspond to one standard error around the regression line. As in Experiment 1a and 1b, participants show more hits and false alarms to high frequency phrases.

[1] The contribution of the noun to phrase memory with the materials from the COCA corpus was unsurprisingly quite small, as noun (and adjective) frequencies were quite restricted. There was neither a main effect of noun frequency ($B = -0.02$, $z = -0.23$, $p = .82$) on recognition responses, nor an interaction between noun frequency and studied status ($B = 0.05$, $z = 0.47$, $p = .63$).

do in Experiment 1: high frequency phrases are more likely to garner "yes" responses regardless of whether the phrase was studied or not. Furthermore, this experiment, like Experiment 1, failed to show the frequency-based mirror effect that is typically observed in recognition memory experiments for single words.

Experiment 3a

The results from Experiments 1a and 1b suggest that noun frequency controls our ability to discriminate studied from unstudied phrases. On the other hand, phrase frequency seems to have some effect on the impression of familiarity for the phrase without affecting accuracy, as seen in both Experiment 1 and 2. The importance of the noun for phrase memory is not without precedent: some work suggests that letter frequencies can lead to the mirror effect, with words with uncommon letters garnering more hits and fewer false alarms (Malmberg et al., 2002). In our case, the uncommonness of the noun contributes to the discriminability of a phrase in recognition memory. The next two experiments sought to confirm this finding. In Experiment 3a we determined the strength of the relationship between word frequency and recognition with single words (nouns), and then in Experiment 3b embedded those same words in phrases with the goal of providing a definitive test of our prediction. Because we did not explicitly manipulate phrase frequency in these experiments, the results of these two experiments can only speak to the role of the noun in phrase memory as a test of our account.

Method

Participants

Participants were 30 undergraduate students from the University of Illinois who acquired no language before the age of 5 other than English. Participants received $8 for their participation in this experiment.

Materials

Eighty-eight nouns from a set of ninety-six nouns used by Balota, Burgess, Cortese, and Adams (2002) served as the stimuli, which had been controlled for concreteness/imageability and word length. The full set of nouns was not used because there are additional constraints based on phrase construction that will be clarified when we introduce Experiment 3b. Words in the dataset spanned a continuous frequency range of 16.9–28.4 in $\log_2$ and included, for example, *tree*, *wizard*, and *anvil*. All nouns were concrete with the exception of *nation*. The materials for Experiment 3a are found in the "noun" column of Tables A3 and A4 of Appendix.

Procedure

The 88 nouns were repartitioned based on their frequencies in the Google corpus into "high" and "low" frequency categories based on a median split. This split resulted in some items from the Balota et al. (2002) materials, which had been assigned to "low" and "high" frequency categories, switching frequency categories. A random sample of each half of the high and half of the low frequency nouns comprised the study materials, for a total of 44 study items.

As in Experiment 1a, each noun was presented for 1 s, followed by a 1 s inter-stimulus interval. Due to the greater number of items at study and at test than in Experiment 1, there was no retention interval prior to starting the test. Participants then completed a yes–no recognition test where the nouns were presented and remained on the screen until participants responded. Participants could take as much time as needed to make a response.

Results

We again modeled participant responses to each item as a function of whether the noun was studied or not, noun frequency, and the interaction of those two terms. The most important result was that there was a strong noun frequency mirror effect, such that low frequency nouns received significantly more hits and fewer false alarms ($\beta = -0.48$, $z = -6.93$, $p < .001$). Unlike the two previous experiments, participants did not exhibit a bias to respond positively (or negatively) as a function of the frequency of the test item ($\beta = -0.06$, $z = -0.63$, $p = .27$). These results are summarized in Table 5. Random effects and correlations are presented in Appendix in Table A8. A visual inspection reveals a strong relationship between hit and false alarm rates and noun frequency, which we include in Fig. 6.

Discussion of Experiment 3a

The results of this experiment demonstrate that the word frequency mirror effect for our items is robust. Given this, the nouns were then incorporated into adjective-noun phrases to evaluate the degree to which this relationship holds when those phrases do not vary in adjectival or phrase frequency. Specifically, we looked for an effect of noun frequency even when the study of those nouns is incorporated in phrases.

Experiment 3b

Method

Participants

Participants were 30 undergraduate students from the University of Illinois who acquired no language before

Table 5
Summary of Experiment 3a fixed effects.

Predictor	Parameter estimates		Wald's test	
	Log-odds	S.E.	Z	p_z
(Intercept)	−0.42	0.15	−2.77	<.001
Old or new status	1.86	0.14	13.59	<.001
Noun frequency (bias)	−0.06	0.10	−0.63	.27
Noun frequency by old–new status	−0.48	0.07	−6.93	<.001

Note: Significance obtained at $p < .05$.

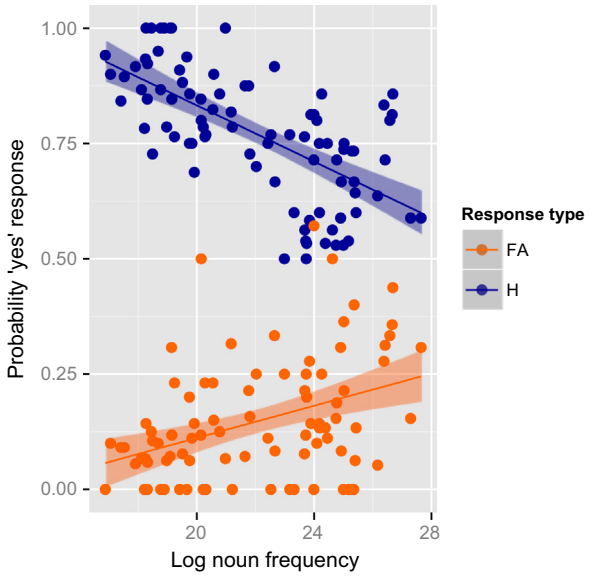

Fig. 6. Hit rates and false alarm rates to nouns for Experiment 3a as a function of noun frequency, collapsed across participant variance. The shaded areas correspond to one standard error around the regression line. Participants make more hits and fewer false alarms to low frequency words.

the age of 5 other than English. Participants received $8 for their participation in this experiment.

Materials

The items used in this experiment were adjective-nouns phrases containing the nouns from Experiment 3a. We created these phrases using a corpus of part-of-speech tagged adjective-noun phrases within the Google 1T *n*-gram corpus (Brants & Franz, 2006). The adjectives and nouns were identified using part of speech labels available from the BNC. The process of excluding nouns that did not occur in our subset of the Google corpus limited the set of nouns used to Experiment 3a to 88. We used 88 adjective-noun phrases in Experiment 3b that contained the nouns tested in Experiment 3a.

We chose the adjectives in these phrases from a very narrow frequency distribution (within a unit of $\log_2$ frequency). Moreover, when these adjectives were combined with the nouns, the resulting phrases also had a very narrow frequency distribution (within a factor of 2). There were no significant correlations between any of adjective, noun, or phrase frequencies, and the means and ranges of all frequencies are almost identical. This was equally true when we divided the nouns into *high* and *low* frequency halves. We present these summary statistics in Table 6. Such resulting phrases included *handsome wizard* (containing a low frequency noun) and *premature tree* (containing a high frequency noun). These are available in Tables A3 and A4 of Appendix.

Table 6
Ranges of ($\log_2$) frequencies by noun frequency category in Experiment 2.

Noun frequency	Mean adjective frequency	Adjective frequency range	Mean phrase frequency	Phrase frequency range
Low	22.28	21.5–23.5	7.38	6.75–8
High	22.33		7.4	

Procedure

Participants studied and were tested on adjective-noun phrases containing the nouns from Experiment 3a. Materials were sampled for each participant in the same way as in Experiment 3a. This experiment followed the same study and test procedures as in Experiment 1a, so participants studied and then were tested on phrases with both the adjective and noun presented simultaneously. As in the prior experiment, there was no retention interval.

Results

The analysis of this experiment was the same as in Experiment 3a. Because we only manipulated noun frequency, while holding the other factors constant, the only frequency factor that was considered was noun frequency. Crucially, and as predicted by our account, there was a noun frequency mirror effect, such that phrases containing low frequency nouns got more hits and fewer false alarms than phrases containing high frequency nouns ($\beta = -0.25$, $z = -2.52$, $p < .05$), although this effect was considerably more modest than in Experiment 3a, in which the nouns were presented and tested alone. Also, as in the previous experiment, they showed no frequency-related response bias; that is, they were not significantly more likely to say that they had seen phrases containing high frequency nouns (e.g. *premature tree*) than low frequency nouns (e.g. *handsome wizard*; $\beta = 0.07$, $z = 0.69$, $p = .25$). We summarize these results in Table 7 and the data are pictured in Fig. 7. Random effects and correlations are presented in Table A9 in Appendix.

As a final test, we assessed whether the nouns' memorability in Experiment 3a was a predictor of performance on phrases containing those nouns in Experiment 3b. Because of the nature of our model, we predict that phrases containing more memorable nouns should be better recognized. We assessed this using a simple linear regression analysis relating the discriminability (d'; Verde, Macmillan, & Rotello, 2006) of the phrases to the discriminability of the nouns. We found a reliable relationship between noun memorability and phrase memorability, with phrases containing more memorable nouns being

Table 7
Summary of Experiment 3b fixed effects.

Predictor	Parameter estimates		Wald's test	
	Log-odds	S.E.	Z	p_z
(Intercept)	−0.83	0.17	−4.98	<.001
Old or new status	2.10	0.17	12.17	<.001
Noun frequency (bias)	0.07	0.10	0.69	.25
Noun frequency by old–new status	−0.25	0.10	−2.52	<.05

Note: Significance obtained at $p < .05$.

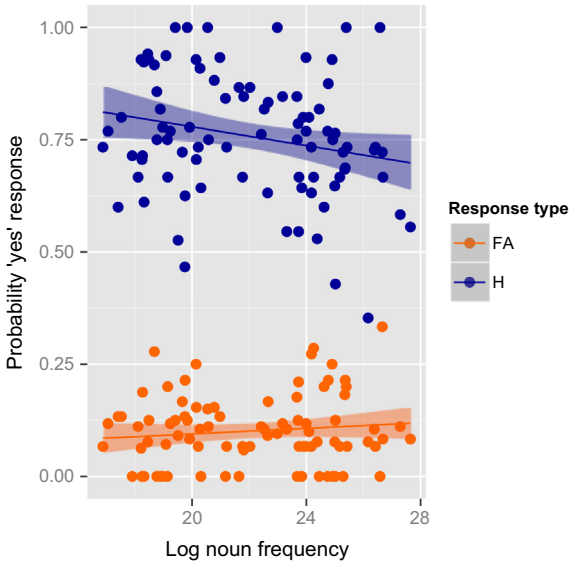

Fig. 7. Hit rates and false alarm rates to phrases for 3b as a function of noun frequency, collapsed across participant variance. The shaded areas correspond to one standard error around the regression line. Participants make more hits and fewer false alarms to phrases containing low frequency words.

Table 8
Summary of analysis relating phrase discriminability to noun discriminability.

Predictor	Parameter estimates		*t*	*p*
	Pearson's *r*	S.E.		
(Intercept)	−0.82	0.16	−5.06	<.001
Noun discriminability	−0.24	0.10	−2.48	<.05

Note: Significance obtained at $p < .05$.

better recognized (Pearson's $r = -0.24$, SE = 0.10, $p < .05$), summarized in Table 8 and Fig. 8.

Discussion of Experiment 3b

This experiment suggests that, as with letters within words (Malmberg et al., 2002; Zechmeister, 1972), words within phrases can provide a cue to memory about whether that phrase was studied or not. Furthermore, the results confirm the effects seen in the joint analysis of Experiments 1a and 1b, where we found a small noun frequency mirror effect. This result was a key prediction of the theoretical position outlined earlier.

General discussion

Experiments 1a, 1b, and 2 tested whether phrase frequency can generate a mirror effect in the same way as word frequency does. The presence of a mirror effect in this case would suggest that phrases are stored at least somewhat separately from their component parts.

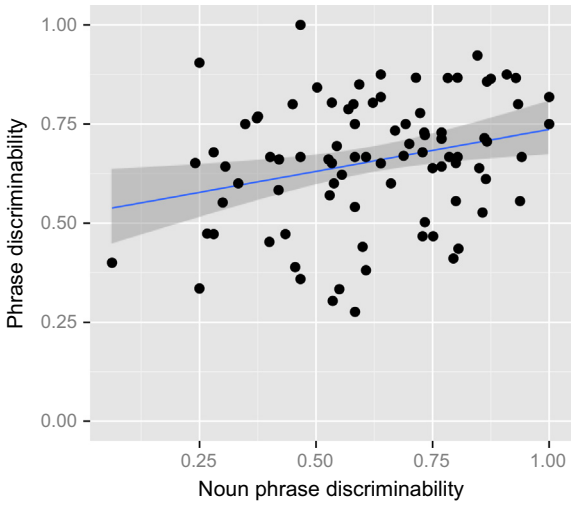

Fig. 8. The discriminability of a phrase as a function of the discriminability of the noun within the phrase. Phrases containing nouns that are more memorable (higher discriminability) are discriminated better as well, as predicted by the model.

In contrast to the standard word-frequency effect in recognition, we found that high frequency phrases were as accurately remembered as low frequency phrases. We also found a bias toward saying that high frequency phrases were studied. The lack of a phrase frequency mirror effect is problematic for the interpretation of phrase frequency effects as indicating completely holistic phrasal representations, and for the findings that any perceptually salient feature can generate the mirror effect (Bloom, 1971; Glanzer & Adams, 1985; Malmberg et al., 2002; Seamon & Murray, 1976; Steyvers & Malmberg, 2003; Vokey & Read, 1992). At the same time, the fact that we get a strong phrase frequency bias effect, where participants were more likely to say that they had studied high frequency phrases regardless of whether they actually had, does suggest that phrase frequency is reflected in language users' representations.

In addition to the absence of a phrase frequency mirror effect, we found a noun frequency mirror effect in the analysis of Experiments 1a and 1b. Phrases containing low frequency nouns garnered more hits and fewer false alarms. This result suggests that recognizing word combinations calls on knowledge of the component words, since individual words can provide a cue for recognition of phrases. In fact, we predicted this finding from an episodic model of phrase recognition that we developed to explain why we found no phrase-frequency mirror effect. The key feature of this model is that there are many more episodic tokens containing the components of a test phrase, such as the noun, than contain the whole phrase. Hence, the amount of interference that is experienced when attempting to recover the relevant episode for a studied phrase is greater for phrases containing common nouns. The fact that there are many more activated episodes containing a word of the phrase than there are episodes that contain the whole phrase explains why phrase frequency is an impotent variable with regard to

memory accuracy. The interference that makes it difficult to find the studied episode comes overwhelmingly from other episodes containing the noun, rather than those corresponding to the whole phrase.

Experiment 3 followed up on the results of Experiment 1 to fully establish the existence of the predicted noun frequency mirror effect. We explicitly manipulated noun frequency within phrases while holding phrase and adjective frequency constant. The results of Experiment 3 showed that the frequency of the words within the phrase can generate a mirror effect, confirming the noun frequency mirror effect of Experiment 1. The presence of a noun frequency mirror effect is in line with many other results on feature frequency or feature salience (Malmberg et al., 2002; Vokey & Read, 1992), but the absence of a phrase frequency mirror effect is not.

What could give rise to a phrase frequency bias effect, but not a phrase frequency mirror effect? Perhaps knowledge of phrases is different from knowledge of words, as we proposed in our model. In the model, episodes for multiword sequences have a relational structure, composed of the words within it. Such sequences may or may not possess syntactic relations between the words (Frank et al., 2012; Arnon & Priva, 2013). A word provides information about the meaning of the phrase, while a letter largely does not (Monaghan, Christiansen, & Fitneva, 2011). This is an important difference between words within phrases compared to letters within words (e.g. Malmberg et al., 2002; Zechmeister, 1972), In general, words relate to phrases *non-arbitrarily*, while the characteristics of a word, such as its orthography, are incidental (Evert, 2005). In our model, this is represented by the potential to retrieve phrasal episodes through words. *Lantern* can help retrieve *red lantern*, but "d" is not a particularly good cue for retrieving *red*.

It is also possible that phrase frequency may not be a reliable cue to memory within the frequency ranges we examined. Recall that the most frequent phrase in Experiment 1 was just as common as the least frequent adjective and noun. The fact that participants could not use phrase frequency to guide their accuracy could simply be due to the fact that the phrases we were using were not actually comparable to higher frequency words. That is, the absence of a phrase frequency mirror effect might reflect the fact that there are not enough phrase episodes to create interference in the first place. Several studies have shown that recognition judgments are not enhanced for very low frequency words relative to low frequency words (Mandler, Goodman, & Wilkes-Gibbs, 1982; Mulligan, 2001; Rao & Proctor, 1984; Reder et al., 2000). This line of reasoning suggests that, at higher phrase frequency ranges than the ones we tested here, once word combinations have become "stickier", we should begin to see the mirror effect.

In fact, this expectation is completely consistent with our model. Because our "high" frequency phrases do not have a large number of individually stored episodes, they cannot create interference in the same way that words can. If one could garner or create much more frequent phrases, there is the potential to create interference during recognition judgments and a phrase-frequency mirror effect. In this sense, the lack of a phrase frequency effect with our range of phrase frequencies is fundamentally the result of the fact that the phrase frequencies are a lot lower than the frequencies of their words (and that the words are a critical part of the phrasal representations).

It is important to note that the phrase frequency range that we used and its relation to the word frequency range is generally true for real life texts. While all word and combinations of word frequencies follow a loosely Zipfian distribution – log frequency is linear with log rank for at least a large part of the frequency range (Evert, 2005; Ha, Sicilia-Garcia, Ming, & Smith, 2002; Zipf, 1949; though see Piantadosi, 2014 for a more critical review of this method) – there are many more phrases than words, making the task of encoding and remembering them much more computationally intensive (Baayen et al., 2013). Furthermore, the use of phrases may be more contextually constrained than that of words, which is important given the contribution of contextual diversity to recognition memory judgments (Steyvers & Malmberg, 2003). Based on the research presented here, we propose that encoding and remembering phrases is accomplished using episodic representations of whole phrases that can be retrieved compositionally, that is, via their individual component words.

To review, we found evidence of memory traces for whole phrases but also of a compositional component to their representation. The memory for phrases is revealed by the bias effect: participants were more likely to endorse high frequency than low frequency phrases, regardless of whether they were studied (Experiments 1 and 2). At the same time, phrases containing low frequency words were better recognized than phrases containing high frequency words, suggesting that the individual components of phrases play a role in the recognition of the entire phrase (Experiment 2).

While our results accord well with the memory literature (Malmberg et al., 2002; Rao & Proctor, 1984), they are hard to fit with proposals that phrases are represented entirely holistically (Arnon & Priva, 2013; Janssen & Barber, 2012). Our proposal is similar to existing models that combine information about words and multiword sequences to predict reading times, language production, and acquisition (Baayen et al., 2013; Bannard & Matthews, 2008; Smith & Levy, 2013). The degree to which we use both information about phrases and words seems to depend somewhat on the task. We may rely on phrase-level information only when it is beneficial to do so, as needed during the processing of known, but non-literal word combinations like *red herring*. At the same time, much of language involves novel combinations of words, meaning that word-level information is useful to both memory and language processing. There may be points where language statistics make it particularly advantageous to ignore word-level information and engage in phrase processing.

Acknowledgments

The first author is supported by an NSF Graduate Research Fellowship. This project was funded in part by NIH DC-000191 to the second author.

Appendix

Tables A1–A9

Table A1
Phrases from the Google 1T n-gram corpus (Brants & Franz, 2006) used in Experiments 1 and 2.

Phrase		Log_2 phrase frequency	Log_2 adjective frequency	Log_2 noun frequency
Adjective	Noun			
Low frequency phrases				
simultaneous	transduction	5.39	21.48	19.70
downstream	subcontractors	5.42	21.57	19.98
naughty	tot	5.64	21.88	20.14
abandoned	arena	5.80	22.36	22.71
accompanying	visions	6.33	22.31	20.91
packaged	hunts	6.37	21.72	19.43
chrome	throttle	6.50	21.53	20.22
optimum	staining	6.50	21.69	20.45
flaming	bounds	6.55	19.67	21.65
predominant	organ	6.70	20.15	22.39
psychic	nephew	6.85	21.10	20.43
transgenic	allele	6.91	20.17	19.88
inhaled	compounds	7.04	19.60	22.64
programmable	fuse	7.20	20.82	20.55
sleek	fleece	7.79	20.91	20.68
piercing	headache	8.57	21.04	21.47
metropolitan	zones	9.09	21.69	22.61
decadent	era	9.19	19.22	23.28
commanding	brigade	9.29	20.23	19.95
distinct	affinity	9.38	23.20	21.07
routine	expressions	9.48	23.32	22.56
untreated	asthma	9.51	20.18	22.02
painful	consciousness	9.66	22.27	22.50
tangled	headset	9.74	19.34	21.38
intense	cultivation	9.79	22.76	20.93
perennial	grasslands	10.29	20.41	19.034
High frequency phrases				
thick	bundles	10.30	23.50	20.49
vibrant	acidity	10.80	21.61	19.28
polynomial	curves	11.04	21.11	22.09
cherished	traditions	11.97	19.88	22.31
passionate	embrace	13.18	21.58	21.71
accumulated	surplus	13.24	21.61	22.18
conditional	expectation	14.97	21.83	21.80
Relentless	pursuit	15.13	19.84	21.84
unsecured	tenant	15.32	21.64	21.81
roman	numerals	15.56	20.28	19.25
interior	decoration	16.06	23.48	21.41
contaminated	soils	16.35	21.81	21.83
undue	hardship	16.94	20.31	20.60
outer	shell	17.35	22.83	23.43
dining	hall	17.55	23.44	23.09
mashed	potatoes	18.34	19.37	21.71
respiratory	tract	18.59	22.01	21.93
cystic	fibrosis	18.67	19.37	19.85
cerebral	palsy	18.73	20.98	19.39
monoclonal	antibody	18.75	19.99	22.03
bald	eagle	18.82	22.00	21.54
nitric	oxide	19.30	19.75	21.74
myocardial	infarction	19.42	20.37	19.93
coronary	artery	19.53	21.29	21.35
alcoholic	beverages	19.56	21.34	21.55
rheumatoid	arthritis	19.65	19.93	21.79

Table A2
Phrases derived from COCA (Davies, 2008) used in Experiment 2.

Phrase		Log_2 phrase frequency	Log_2 adjective frequency	Log_2 noun frequency
Adjective	Noun			
Low frequency phrases				
poor	credit	2.32	12.82	12.79
southern	food	2.81	12.36	13.81
fantastic	panel	2.81	11.18	12.19
nice	hair	3	13.37	12.69
incredible	pain	3.17	12.28	12.75
safe	space	3.17	12.81	13.04
available	flight	3.17	12.78	12.42
controversial	statement	3.32	11.71	13.19
violent	weather	3.32	12.04	12.41
similar	incident	3.46	12.45	12
particular	church	3.46	13.12	13.35
local	airport	3.58	13.57	12.09
open	relationship	3.7	12.77	13.59
heavy	heart	3.7	12.09	13.86
likely	suspect	3.7	12.64	11.35
impossible	dream	3.91	11.87	12.18
wonderful	trip	4	13.51	12.58
british	actor	4	12.66	12.56
serious	nature	4.09	13.83	12.44
major	bank	4.09	13.94	12.61
crazy	talk	4.09	12.45	13.43
sad	truth	4.17	12.05	13.53
successful	mission	4.17	12.66	12.84
simple	rule	4.17	12.91	11.94
global	recession	4.17	12.14	11.8
angry	crowd	4.25	12.62	11.97
late	term	4.25	12.53	12.82
guilty	pleasure	4.25	13	12.2
normal	behavior	4.39	12.74	12.43
fresh	blood	4.39	12.05	13.2
strong	opinion	4.46	13.73	13.05
healthy	weight	4.46	12.03	12.25
super	model	4.52	11.57	11.9
funny	feeling	4.58	12.76	12.47
necessary	step	4.58	12.43	12.31
positive	test	4.58	12.63	12.57
lucky	break	4.64	11.82	14.74
current	governor	4.64	12.59	13.46
actual	cost	4.81	11.87	12.55
easy	solution	4.86	12.97	12.16
civil	union	4.86	13.19	13.19
horrible	mistake	4.86	11.68	12.58
fair	deal	4.91	12.44	13.61
international	agreement	4.91	13.77	12.94
clear	winner	4.91	13.42	11.68
famous	speech	4.91	12.52	13.38
effective	treatment	4.95	12.23	12.71
white	neighborhood	5	13.57	12.298
sexual	act	5	12.85	12.73
legal	strategy	5.04	13.46	12.46
senior	officer	5.09	12.89	12.95
military	background	5.13	14.14	12.05
quick	action	5.17	12.63	13.62
full	picture	5.21	13.48	13.45
short	film	5.25	12.79	13.47
High frequency phrases				
liberal	agenda	5.29	11.94	12.31
dangerous	drug	5.32	12.88	13.72
afghan	border	5.43	10.87	12.6
commercial	success	5.46	14.21	12.75
physical	violence	5.49	12.18	13.43
emotional	response	5.55	12.07	12.81
innocent	victim	5.58	11.92	12.42
terrible	accident	5.64	12.8	12.21
prime	example	5.64	12.87	12.67

Table A2 (*continued*)

Phrase		Log_2 phrase frequency	Log_2 adjective frequency	Log_2 noun frequency
Adjective	Noun			
iraqi	freedom	5.67	13.51	12.7
extraordinary	amount	5.75	12.1	13.18
specific	threat	5.81	12.55	12.9
amazing	experience	5.88	12.83	13.39
beautiful	song	5.93	13.34	13.08
private	plane	5.95	13.32	12.9
certain	type	5.98	13.88	12.9
personal	choice	6	13.56	13.13
social	network	6	13.57	12.5
entire	industry	6.02	13.16	13.54
fine	art	6.09	13.23	12.56
powerful	message	6.11	12.58	13.63
independent	investigation	6.27	12.44	13.56
smart	move	6.3	11.96	11.74
significant	progress	6.34	12.59	12.24
main	course	6.38	12.6	12.64
correct	answer	6.39	12.63	13.48
supreme	leader	6.44	13.18	13.57
enormous	pressure	6.58	12.18	13.12
red	tape	6.74	12.13	12.8
financial	reform	6.88	12.91	13.22
tough	love	6.93	13.58	13.44
perfect	storm	7	12.58	12.44
religious	right	7.06	12.46	13.09
close	attention	7.17	12.72	13.57
dead	heat	7.26	13.37	11.55
hot	seat	7.31	12.76	12.16
single	parent	7.46	13.27	11.98
critical	condition	7.5	12.51	11.81
low	income	7.53	12.03	12.42
recent	study	7.58	13.05	12.68
early	age	7.88	13.22	13.56
wrong	direction	8.06	12.78	12.4
central	park	8.23	12.74	12.16
popular	vote	8.4	12.73	13.44
congressional	budget	8.56	12.36	13.7
regular	basis	8.75	12.07	12.47
common	ground	8.84	12.72	13.3
free	market	8.97	13.58	13.5
natural	gas	9	12.18	12.68
nuclear	weapon	9.02	13.47	11.68
illegal	immigration	9.18	12.54	12.06
gay	marriage	9.25	12.41	12.96
economic	growth	9.9	13.92	12.52
Foreign	Language	10.6	13.82	12.86
Presidential	Candidate	11.24	13.49	13.45
middle	class	11.27	13.13	13.16
grand	jury	11.32	12.18	13.72

Table A3
Phrases and nouns derived from Balota et al. (2002) and used in Experiments 3a and 3b.

Phrase		Log_2 phrase frequency	Log_2 adjective frequency	Log_2 noun frequency
Adjective	Noun			
High frequency nouns				
adjacent	nation	7.11	23.04	24.74
ambitious	library	7.81	21.55	25.33
artificial	home	7.70	22.48	28.37
awesome	valley	7.09	23.29	22.64
beloved	chicken	7.59	21.78	23.16
beneficial	sun	7.07	22.59	24.38
biological	garden	7.23	23.43	24.25
bold	rose	7.65	22.80	23.67
burning	palace	7.99	23.25	21.54

(*continued on next page*)

Table A3 (*continued*)

Phrase		Log_2 phrase frequency	Log_2 adjective frequency	Log_2 noun frequency
Adjective	Noun			
cooling	floor	7.56	22.63	24.99
cycling	town	7.75	21.85	25.42
destructive	baby	7.22	21.43	25.17
downstream	field	7.42	21.57	26.41
emerging	road	7.95	23.03	25.36
endless	cloud	8.08	21.96	22.42
engaged	father	7.21	23.49	24.92
failing	hotel	6.90	22.46	26.58
gentle	snake	6.85	22.27	21.56
governing	village	7.84	22.74	23.99
grounded	world	7.29	21.66	27.64
hanging	dress	7.28	22.77	23.68
hazardous	car	7.42	22.70	26.68
inspiring	college	6.91	21.46	25.39
instructional	kitchen	6.97	22.16	24.17
insured	truck	7.27	22.28	23.84
invisible	mouth	7.58	22.01	24.18
jumping	cow	7.43	21.72	22.02
literary	radio	8.00	22.57	25.28
magnificent	plane	7.93	21.93	23.72
metallic	wheel	7.12	21.43	23.71
patented	bottle	7.83	21.44	23.31
premature	tree	6.96	21.42	25.01
provincial	street	6.85	22.33	25.01
refurbished	engine	7.88	21.41	25.34
rejected	picture	7.09	22.92	26.16
rolled	bread	8.08	22.32	22.98
specialized	pool	8.02	22.90	24.77
stainless	key	6.98	22.67	26.37
sterling	cup	7.42	22.11	23.99
sticky	book	7.02	21.45	27.28
stolen	jacket	7.24	22.35	22.66
striking	beach	7.77	22.37	24.62
surfing	market	7.26	21.95	26.66
surprised	cat	7.48	23.21	24.09
teenage	king	7.25	23.18	23.74
tough	bear	7.12	23.38	23.88
toxic	stream	7.09	22.50	24.90
versatile	ball	7.15	21.83	24.45
Low frequency nouns				
adjustable	anvil	6.84	22.45	18.11
bald	vulture	6.87	22.00	17.52
bitter	pecan	7.97	21.76	18.21
blind	owl	7.68	23.27	20.54
brilliant	sleuth	6.84	22.83	17.05
circular	parasol	7.24	22.01	16.87
complementary	valet	7.42	21.64	19.13
copper	altar	8.05	22.70	20.77
crowded	isle	7.14	21.43	19.24
cute	otter	7.02	23.40	18.49
deadly	dungeon	7.82	21.83	19.65
decorative	gourd	7.86	21.75	17.90
delicate	sequin	7.61	21.71	18.26
dried	eel	7.52	22.14	18.76
elegant	lily	7.22	22.67	20.14
expanding	cavern	7.53	22.87	19.75
extraordinary	gem	7.04	22.64	21.17
fake	cobra	6.80	22.74	19.41
fancy	loft	7.30	22.21	20.27
golden	plum	7.82	22.92	19.75
grey	bonnet	7.90	22.52	18.97
handsome	wizard	7.81	21.44	21.64
indigenous	spa	7.19	22.12	22.52
lively	lass	6.89	21.54	18.45
miniature	tripod	7.97	21.44	20.31
nasty	beggar	6.96	22.28	18.31
occasional	jaguar	6.76	22.26	18.88

Table A3 (*continued*)

Phrase		Log_2 phrase frequency	Log_2 adjective frequency	Log_2 noun frequency
Adjective	Noun			
offshore	wharf	6.82	22.33	18.77
ordinary	flea	7.43	23.43	20.21
polished	flask	7.76	21.66	19.14
portable	keg	6.85	23.47	18.24
relaxing	harp	7.40	21.91	19.91
robust	vine	7.07	22.52	20.15
sacred	urn	7.96	22.06	19.83
shallow	crevice	7.18	21.86	17.41
silly	dwarf	7.38	22.27	20.35
slim	vase	7.93	21.89	20.57
spinning	galaxy	7.35	21.40	21.21
stuffed	boar	7.24	21.40	18.68
stylish	yacht	7.45	22.88	20.97
tan	tunic	7.28	21.59	18.52
thin	tablet	7.31	23.42	21.77
tropical	olive	7.80	22.66	21.80
vintage	banjo	7.99	22.98	19.51
wooden	silo	7.01	22.76	18.32
yearly	monsoon	7.41	21.83	19.09

Table A4
Random effects, Experiment 1a.

Random effects		Variance	sd
Participant	Intercept	0.31	0.55
	Studied status	0.26	0.51
Item	Intercept	0.19	0.43

Table A5
Random effects and random effects correlations, Experiment 1b.

Random effects		Variance	sd
Participant	Intercept	0.71	0.84
	Phrase frequency	0.09	0.31
	Studied status	0.33	0.57
	Phrase frequency * Studied status	0.01	0.01
Item	Intercept	0.31	0.56

	Intercept	Phrase frequency	Studied status
Random effects correlations			
Phrase frequency	−0.07		
Studied status	0.36	0.25	
Phrase frequency * Studied status	0.58	−0.84	−0.14

Table A6
Random effects and random effects correlations, Experiment 1a and Experiment 1b combined analysis.

Random effects		Variance	sd
Participant	Intercept	0.44	0.66
	Phrase frequency	0.05	0.23
	Studied status	0.25	0.50
	Noun frequency	0.02	0.13
	Noun frequency * Studied status	0.005	0.07
Item	Intercept	0.21	0.46

	Intercept	Phrase frequency	Studied status	Noun frequency
Random effects correlations				
Phrase frequency	0.06			
Studied status	−0.07	0.42		
Noun frequency	−0.58	−0.85	−0.34	
Noun frequency * Studied status	−0.41	−0.85	−0.67	0.92

Table A7
Random effects and random effects correlations, Experiment 2.

Random effects		Variance	sd
Participant	Intercept	1.49	1.22
	Phrase frequency	0.02	0.13
	Studied status	1.92	1.39
	Phrase frequency * Studied status	0.13	0.36
Item	Intercept	0.19	0.43

	Intercept	Studied status	Phrase frequency
Random effects correlations			
Studied status	−0.55		
Phrase frequency	−0.32	0.87	
Phrase frequency * Studied status	−0.14	−0.40	−0.80

Table A8
Random effects and random effects correlations, Experiment 3a.

Random effects		Variance	sd
Participant	Intercept	0.50	0.71
	Noun frequency	0.09	0.29
	Studied status	0.41	0.64
	Noun frequency * Studied status	0.004	0.07
Item	Intercept	0.13	0.35

	Intercept	Studied status	Noun frequency
Random effects correlations			
Studied status	−0.38		
Noun frequency	0.58	0.27	
Noun frequency * Studied status	−0.75	0.36	−0.80

Table A9
Random effects and random effects correlations, Experiment 3b.

Random effects		Variance	sd
Participant	Intercept	0.55	0.74
	Noun frequency	0.07	0.27
	Studied status	0.63	0.79
	Noun frequency * Studied status	0.07	0.27
Item	Intercept	0.09	0.31

	Intercept	Noun frequency	Studied status
Random effects correlations			
Noun frequency	−0.32		
Studied status	−0.62	0.80	
Noun frequency * Studied status	0.24	−0.57	−0.68

References

Arnon, I., & Priva, U. C. (2013). More than words: The effect of multi-word frequency and constituency on phonetic duration. *Language and Speech, 56*, 349–371.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multiword phrases. *Journal of Memory and Language, 62*, 67–82.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390–412. http://dx.doi.org/10.1016/j.jml.2007.12.005.

Baayen, R. H., Hendrix, P., & Ramscar, M. (2013). Sidestepping the combinatorial explosion: An explanation of n-gram frequency effects based on naive discriminative learning. *Language and Speech, 56*, 329–347.

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review, 118*, 438–481.

Balota, D. A., Burgess, G. C., Cortese, M. J., & Adams, D. R. (2002). The word-frequency mirror effect in young, old, and early-stage Alzheimer's disease: Evidence for two processes in episodic recognition performance. *Journal of Memory and Language, 46*, 199–226.

Bannard, C. (2006). *Acquiring phrasal lexicons from corpora* (Unpublished doctoral thesis). University of Edinburgh. Edinburgh, Scotland, UK.

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning. *Psychological Science, 19*, 241–248.

Bates, D., Maechler, M., Bolker, B. & Walker, S., (2014). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-6 <http://CRAN.R-project.org/package=lme4>.

Benjamin, A. S. (2003). Predicting and postdicting the effects of word frequency on memory. *Memory & Cognition, 31*, 297–305.
Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: when retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General, 127*(1), 55.
Bien, H., Levelt, W. M. J., & Baayen, R. H. (2005). Frequency effects in compound production. *Proceedings of the National Academy of Sciences, 102*, 17876–17881.
Bloom, S. W. (1971). Recognition memory for pictures and their word labels. Unpublished doctoral dissertation, University of Rochester, Rochester, NY.
Brants, T. & Franz, A. (2006). Web 1T 5-gram version 1 LDC2006T13. Web Download. Philadelphia: Linguistic Data Consortium.
Brown, J., Lewis, V. J., & Monk, A. F. (1977). Memorability, word frequency and negative recognition. *The Quarterly Journal of Experimental Psychology, 29*(3), 461–473.
Davies, M. (2008). The corpus of contemporary American English: 450 million words, 1990-present <http://corpus.byu.edu/coca/>.
Dell, G. S. (1990). Effects of frequency and vocabulary type on phonological speech errors. *Language and Cognitive Processes, 5*, 313–349.
Dell, G. S., & Jacobs, C. L. (2015). Successful speaking: Cognitive mechanisms of adaptation in language production. In G. Hickok & S. L. Small (Eds.), *The neurobiology of language* (pp. 209–220). Elsevier.
Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition, 109*, 193–210.
Diana, R. A., & Reder, L. M. (2006). The low-frequency encoding disadvantage: Word frequency affects processing demands. *Journal of Experimental Psychology: Learning, Memory and Cognition, 32*, 805–815.
Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition, 24*, 143–188.
Evert, S. (2005). The statistics of word cooccurrences (Doctoral dissertation, Dissertation, Stuttgart University).
Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological Review, 120*, 751–778.
Fisher, C., Gertner, Y., Scott, R. M., & Yuan, S. (2010). Syntactic bootstrapping. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*, 143–149.
Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior, 12*, 627–635.
Frank, S. L., Bod, R., & Christiansen, M. H. (2012). How hierarchical is language use? *Proceedings of the Royal Society B: Biological Sciences, 279*, 4522–4531.
Freebody, P., & Anderson, R. C. (1983). Effects of vocabulary difficulty, text cohesion, and schema availability on reading comprehension. *Reading Research Quarterly*, 277–294.
Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition, 13*, 8–20.
Goldinger, S. D. (2004). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*, 251–279.
Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition, 112*, 21–54.
Ha, L. Q., Sicilia-Garcia, E. I., Ming, J., & Smith, F. J. (2002). Extension of Zipf's law to words and phrases. *Proceedings of the 19th international conference on Computational linguistics* (Vol. 1, pp. 1–6). Association for Computational Linguistics. August.
Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95*, 528–551.
Howes, D. (1952). On the relation between the intelligibility and frequency of occurrence of English words. *The Journal of the Acoustical Society of America, 29*, 296–305.
Howes, D. H., & Solomon, R. L. (1951). Visual duration threshold as a function of word-probability. *Journal of Experimental Psychology, 41*, 401–410.
Janssen, N., & Barber, H. A. (2012). Phrase frequency effects in language production. *PLoS ONE, 7*, e33202.
Jusczyk, P. W. (1997). *The discovery of spoken language*. Cambridge, MA: MIT Press.
Kittredge, A. K., Dell, G. S., Verkuilen, J., & Schwartz, M. F. (2008). Where is the effect of frequency in word production? Insights from aphasic picture-naming errors. *Cognitive Neuropsychology, 25*, 463–492.
Lively, S. E., Pisoni, D. B., & Goldinger, S. D. (1994). Spoken word recognition. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 265–318). San Diego, CA: Academic Press.
Lockhart, R. S. (1969). Retrieval asymmetry in the recall of adjectives and nouns. *Journal of Experimental Psychology, 79*, 12–17.
MacKay, D. G. (1982). The problems of flexibility, fluency, and speed–accuracy trade-off in skilled behavior. *Psychological Review, 89*, 483–506.
Malmberg, K. J., Steyvers, M., Stephens, J. D., & Shiffrin, R. M. (2002). Feature frequency effects in recognition memory. *Memory & Cognition, 30*, 607–613.
Mandler, G., Goodman, G. O., & Wilkes-Gibbs, D. L. (1982). The word-frequency paradox in recognition. *Memory & Cognition, 10*, 33–42.
Marian, V., Bartolotti, J., Chabal, S., & Shook, A. (2012). CLEARPOND: Cross-linguistic easy-access resource for phonological and orthographic neighborhood densities. *PloS One, 7*(8), e43230.
Marks, C. B., Doctorow, M. J., & Wittrock, M. C. (1974). Word frequency and reading comprehension. *The Journal of Educational Research*, 259–262.
Martin, A., Peperkamp, S., & Dupoux, E. (2012). Learning phonemes with a proto-lexicon. *Cognitive Science, 37*, 103–124.
Mata, A., Percy, E. J., & Sherman, S. J. (2013). Adjective–noun order as representational structure: Native-language grammar influences perception of similarity and recognition memory. *Psychonomic Bulletin & Review, 2*, 193–197.
Medin, D. L., & Shoben, E. J. (1988). Context and structure in conceptual combination. *Cognitive Psychology, 20*, 158–190.
Monaghan, P., Christiansen, M. H., & Fitneva, S. A. (2011). The arbitrariness of the sign: Learning advantages from the structure of the vocabulary. *Journal of Experimental Psychology: General, 140*, 325–347.
Morris, P. E., & Reid, R. L. (1972). Imagery and the recall of adjectives and nouns from meaningful prose. *Psychonomic Science, 27*, 117–118.
Mulligan, N. W. (2001). Word frequency and memory: Effects on absolute versus relative order memory and on item memory versus order memory. *Memory & Cognition, 29*, 977–985.
Oldfield, R. C., & Wingfield, A. (1965). Response latencies in naming objects. *The Quarterly Journal of Experimental Psychology, 17*, 273–281.
Paivio, A. (1971). *1magery and verbal processes*. NJ Erlbaum: Hillsdale.
Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology, 76*, 1–25.
Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review, 21*, 1112–1130.
Powell, M. J. (2009). The BOBYQA algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge*.
Rao, K. V., & Proctor, R. W. (1984). Study-phase processing and the word frequency effect in recognition memory. *Journal of Experimental Psychology: Learning, Memory and Cognition, 10*, 386–394.
Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*, 372–422.
Reder, L. M., Nhouyvanisvong, A., Schunn, C. D., Ayers, M. S., Angstadt, P., & Hiraki, K. (2000). A mechanistic account of the mirror effect for word frequency: A computational model of remember–know judgments in a continuous recognition paradigm. *Journal of Experimental Psychology: Learning, Memory and Cognition, 26*, 294–320.
Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton Century Crofts.
Richardson, J. T. E. (1978). Word-order and imagery in the recall of adjective-noun phrases. *International Journal of Psychology, 13*, 179–184.
Saffran, J., Newport, E., & Aslin, R. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language, 35*, 606–621.
Seamon, J. G., & Murray, P. (1976). Depth of processing in recall and recognition memory: Differential effects of stimulus meaningfulness and serial position. *Journal of Experimental Psychology: Human Learning and Memory, 2*, 680–687.
Seidenberg, M. S., & McClelland, J. L. (1989). A distributed developmental model of word recognition and naming. *Psychological Review, 96*, 523–568.
Shaoul, C., Harald Baayen, R., & Westbury, C. F. (2014). N-gram probability effects in a cloze task. In S. Wulff, D. Titone (Eds.), *The mental lexicon* (Vol. 9(3), pp. 437–472).
Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM – Retrieving effectively from memory. *Psychonomic Bulletin & Review, 4*, 145–166.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition, 128*, 302–319.
Smith, E. E., & Osherson, D. N. (1984). Conceptual combination with prototype concepts. *Cognitive Science, 8*, 337–361.
Steyvers, M., & Malmberg, K. J. (2003). The effect of normative context variability on recognition memory. *Journal of Experimental Psychology: Learning, Memory and Cognition, 29*(5), 760–766.
Swingley, D. (2009). Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*, 3617–3632.
Szabo, Z. (2013). Compositionality. In *The stanford encyclopedia of philosophy* <http://plato.stanford.edu/archives/fall2013/entries/compositionality/>.
The British National Corpus, 2007. version 3 (BNC XML Edition). Distributed by Oxford University computing services on behalf of the BNC Consortium <http://www.natcorp.ox.ac.uk/>.
Tomasello, M. (Ed.). (1998). *The new psychology of language: Cognitive and functional approaches to language structure*. Mahwah, NJ: Erlbaum.
Tomasello, M. (2006). Acquiring linguistic constructions. In D. Kuhn & R. Siegler (Eds.), *Handbook of child psychology*. Wiley.
Tremblay, A., & Baayen, R. H. (2010). Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. *Perspectives on formulaic language: Acquisition and communication*, 151–173.
Tullis, J. G., & Benjamin, A. S. (2012). The effectiveness of updating metacognitive knowledge in the elderly: Evidence from metamnemonic judgments of word frequency. *Psychology and Aging, 27*, 683–690.
Verde, M. F., Macmillan, N. A., & Rotello, C. M. (2006). Measures of sensitivity based on a single hit rate and false-alarm rate: The accuracy, precision, and robustness of d', Az, and A'. *Perception & Psychophysics, 68*, 643–654.
Vokey, J. R., & Read, J. D. (1992). Familiarity, memorability, and the effect of typicality on the recognition of faces. *Memory & Cognition, 20*, 291–302.
Wright, D. B., Horry, R., & Skagerberg, E. M. (2009). Functions for traditional and multilevel approaches to signal detection theory. *Behavior Research Methods, 41*, 257–267.
Yuille, J. C., Paivio, A., & Lambert, W. E. (1969). Noun and adjective imagery and order in paired-associate learning by French and English subjects. *Canadian Journal of Psychology/Revue canadienne de psychologie, 23*, 459–466.
Zechmeister, E. B. (1972). Orthographic distinctiveness as a variable in word recognition. *American Journal of Psychology, 85*, 425–430.
Zipf (1949). Human behavior and the principle of least effort, 1–24.