Intrinsically memorable words have unique associations with their meanings

Greta Tuckute*[1,2], Kyle Mahowald*[3], Phillip Isola[4], Aude Oliva[4], Edward Gibson[1], and Evelina Fedorenko[1,2,5]

[1]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA.
[2]McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA, USA.
[3]Department of Linguistics, The University of Texas at Austin, Austin, TX, USA.
[4]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA.
[5]Program in Speech and Hearing Bioscience and Technology, Harvard University, Boston, MA, USA.

* Co-first authors

Author contributions

	GT	KM	PI	AO	EG	EF
Conceptualization			x	x	x	x
Design and materials creation		x	x		x	x
Experimental script creation			x			
Data collection		x	x		x	
Data analysis	x	x	x			
Formal statistical analysis	x	x				
Figures	x	x	x			
Writing	x	x	x			x
Editing + comments				x	x	
Overall supervision				x	x	x

Acknowledgments: For comments on this work, we thank members of the Fedorenko, Gibson, and Oliva labs, Josh Tenenbaum and members of Cocosci, and Sam Gershman. For help with constructing the materials for Experiment 2, we thank Barbara Hidalgo-Sotelo, Henrison Hsieh, and members of the Gibson lab. G.T. was supported by the Amazon Fellowship from the Science Hub (administered by the MIT Schwarzman College of Computing) and the International Doctoral Fellowship from American Association of University Women (AAUW). E.F. was supported by NIH awards DC016607 and DC016950 from NIDCD and award NS121471 from NINDS, and by research funds from the McGovern Institute for Brain Research, the Brain and Cognitive Sciences department, and the Simons Center for the Social Brain at MIT.

ABSTRACT

What makes a word memorable? An important claim from past work is that words are encoded by their meanings, and not their forms. If true, then, following rational analysis, memorable words should convey a lot of information about their meaning (i.e., uniquely pick out a particular meaning). Words with multiple meanings have been shown to be forgettable, but the other side of the word-meaning relationship (whether a meaning can be expressed by multiple words) has not been previously explored. Across two large-scale recognition-memory experiments (2,222 target words and >600 participants each, plus 3,780 participants for the norming experiments), we found that memory performance is overall high, on par with memory for images in similar paradigms. Critically, the most memorable words indeed have a one-to-one relationship with their meanings—with number of synonyms being a stronger contributor than number of meanings—and this property explains >80% of explainable variance in memorability.

PUBLIC SIGNIFICANCE STATEMENT

We all know the feeling of hearing or reading words that really stick in our memory. Here, we show that certain words are consistently remembered better than others, despite our differences in our exposure to language and our patterns of language use. Specifically, words that pick out a unique meaning in semantic memory (like "*PINEAPPLE*") are more memorable than words with many meanings ("*LIGHT*") or words with many synonyms ("*HAPPY*"). We develop a Bayesian model that explains these findings and makes predictions for new words across languages. Understanding which words lead to longer-lasting memory traces can enable more effective information sharing.

INTRODUCTION

An avalanche of precise, lucid vocabulary has an advantage as a manner of expression. Perhaps this comes as no surprise. Effective word choice is indisputably critical to clear communication. Less obvious is the impact that word choice has on subsequent memory. Consider the first sentence of this paragraph. “Avalanche”, “lucid”, and “vocabulary” are among the most memorable words as measured by the experiments described later in this paper. The remaining words – “precise”, “advantage”, “manner”, and “expression”, in case you forgot – were among the most forgettable.

What makes “avalanche” stick in our memory? The literature provides several possible explanations. Studies where words are presented in isolation have found that less familiar, lower-frequency words are easier to recognize though more difficult to recall (Gorman 1961; Schulman 1967; Kinsbourne and George 1974; Brown and Lewis 1981; Lohnas and Kahana 2013); concrete and imageable words are both easier to recognize and easier to recall (Gorman 1961; Paivio 1969; Rubin and Friendly 1986; Walker and Hulme 1999; Klaver et al. 2005; Khanna and Cortese 2021); and emotionally salient words also enjoy a memory boost in both recognition and recall (Rubin and Friendly 1986; Danion et al. 1995; Phelps, LaBar, and Spencer 1997; Kensinger and Corkin 2003).

Past work has also emphasized the importance of meaning (over the surface form) for memorability of linguistic strings. In sentence recognition studies, sentence meanings are better retained than their surface-level (lexical and morpho-syntactic) properties (Bransford and Franks 1971; Franks and Bransford 1972; Katz and Gruenewald 1974; Begg and Wickelgren 1974). Furthermore, deeper engagement with the word’s meaning—as can be manipulated via tasks at the encoding stage—facilitates subsequent recognition (Jacoby and Dallas 1981). Moreover, for ambiguous words (e.g., “jam”), narrowing in on a particular meaning via context (e.g., “strawberry jam”) leads to better subsequent memory, but only when the word is used in the same meaning (e.g., “raspberry jam”) compared to a different meaning (e.g., “traffic jam”) (Light and Carter-Sobell 1970). Similarly, in sentence recall studies (using the classic rapid serial visual presentation, RSVP, paradigm; Forster 1970), the sentence meaning is typically well remembered whereas the surface properties of the sentence, and its composite words, are often forgotten (Potter, Kroll, and Harris 1980; Potter and Lombardi 1990; Potter 2012). Expanding on the importance of meaning for the encoding and retention of linguistic information, we here explore the effects on memorability of the *relationship* between words and their meanings—or, how uniquely a given word is associated with a particular meaning (e.g., Steyvers and Malmberg 2003; Griffiths, Steyvers, and Tenenbaum 2007; Monaco, Abbott, and Kahana 2007). We build on these past studies and examine word memorability through the lens of Bayesian optimal inference by performing a large-scale evaluation of a novel two-factor hypothesis about what makes words stick in memory.

This general approach is rooted in rational models of cognition whereby human behavior approximates optimal solutions to problems in the environment (Anderson and Milson 1989; Anderson and Schooler 1991; Chater and Oaksford 1999; Tenenbaum et al. 2011; Gershman 2021). Some past research on verbal memory had followed this tradition (Shiffrin and Steyvers 1997; McClelland and Chappell 1998; Dennis and Humphreys 2001; Steyvers and Malmberg

2003). For example, in the 'retrieving effectively from memory' (REM) model, Shiffrin and Steyvers (1997) suggested that performing a recognition task effectively requires computing a probability that a given stimulus (e.g., word) is "new" or "old" by accessing vectors of stored features (for previously encountered words) and comparing them to the current word's features. However, the nature of these features has been a matter of debate (Criss and Shiffrin 2004; Annis et al. 2015). "Item-noise models" (e.g., Shiffrin and Steyvers 1997; McClelland and Chappell 1998) have emphasized intrinsic, context-independent properties of words, whereas "context-noise models" (e.g., Dennis and Humphreys, 2001) have focused on the context in which the words occur, and neither class of model makes strong claims about which particular features (be they intrinsic or contextual) are encoded and later used during recognition.

Here we explore a simple idea: given that—as discussed above—words appear to be encoded by their meanings, we hypothesize that a memorable word unambiguously selects a particular meaning in the conceptual space. To be able to do so, a word should have no synonyms (i.e., other words that can express the same meaning) and it should only have a single meaning (i.e., be unambiguous), as elaborated below in Ideal Observer Model.

The idea that words with more meanings would be less memorable is reminiscent of the "fan effect" (Anderson 1974; Monaco et al. 2007), in which recognition times for an item increase in proportion to the number of its distinct attributes. Ambiguous words can be construed as having a "fan" of associations, one for each of their meanings. Steyvers and Malmberg (2003) also modeled word recognition memory as related to a form of the fan effect. They proposed that words that occur in more diverse contexts in our past experience with them (as estimated in their study by the number of different documents a given word occurs in) leave more diffuse memory traces and provided some empirical support for this hypothesis. Relatedly, Griffiths et al. (2007) demonstrated that the number of distinct *topics* a word is associated with—a measure that corresponds well with the number of meanings or senses of a word—is an even better predictor of human recognition performance. Below, we directly compare our two-factor model to that of Steyvers and Malmberg (2003) and Griffiths et al. (2007) and show that our model explains more variance in memory performance, presumably because the "fan effect" only captures one side of the word-to-meaning relationship: the one (word) to many (meanings) component, but not the many (words) to one (meaning) component. In fact, as our data show, the latter (number of synonyms) component—a factor that, to our knowledge, has not been proposed before—explains more variance in word memorability. The proposed approach here is conceptually similar to cue competition approaches in discriminative learning (e.g., Ramscar et al. 2010), but we leave it to future work to integrate that class of models with ours.

Ideal Observer Model

We propose a Bayesian model of the word recognition task in which a rational agent encounters a word W and stores a meaning m selected by that word. The agent is then asked, at a later time, whether a word w has been previously encountered. At that time, the agent has access to the stored meaning m and the new stimulus w and must decide whether the original word W that generated m is the same as the new word w. (Note that we are not proposing that such a process fully explains how humans perform this task but are merely using this model as a tool for

deriving testable predictions about how a rational agent would act.) Formally, the agent must assess the probability that the new word *w* is the same as the originally encountered word, which can be expressed as the probability that *W* takes on the value *w* given the stored meaning *m*. Applying Bayes' rule, this probability can be written as:

$$P(W = w|m) = \frac{P(m|W = w)P(W = w)}{\sum_{w'} P(m, W = w')} \quad (1)$$

This formula has an intuitive interpretation. The agent is assessing "out of all the possible ways I could have ended up with this memory, what are the chances that *w* generated it?" Given this decision rule, memorability can be operationalized as the expected value of $P(W = w \mid m)$.

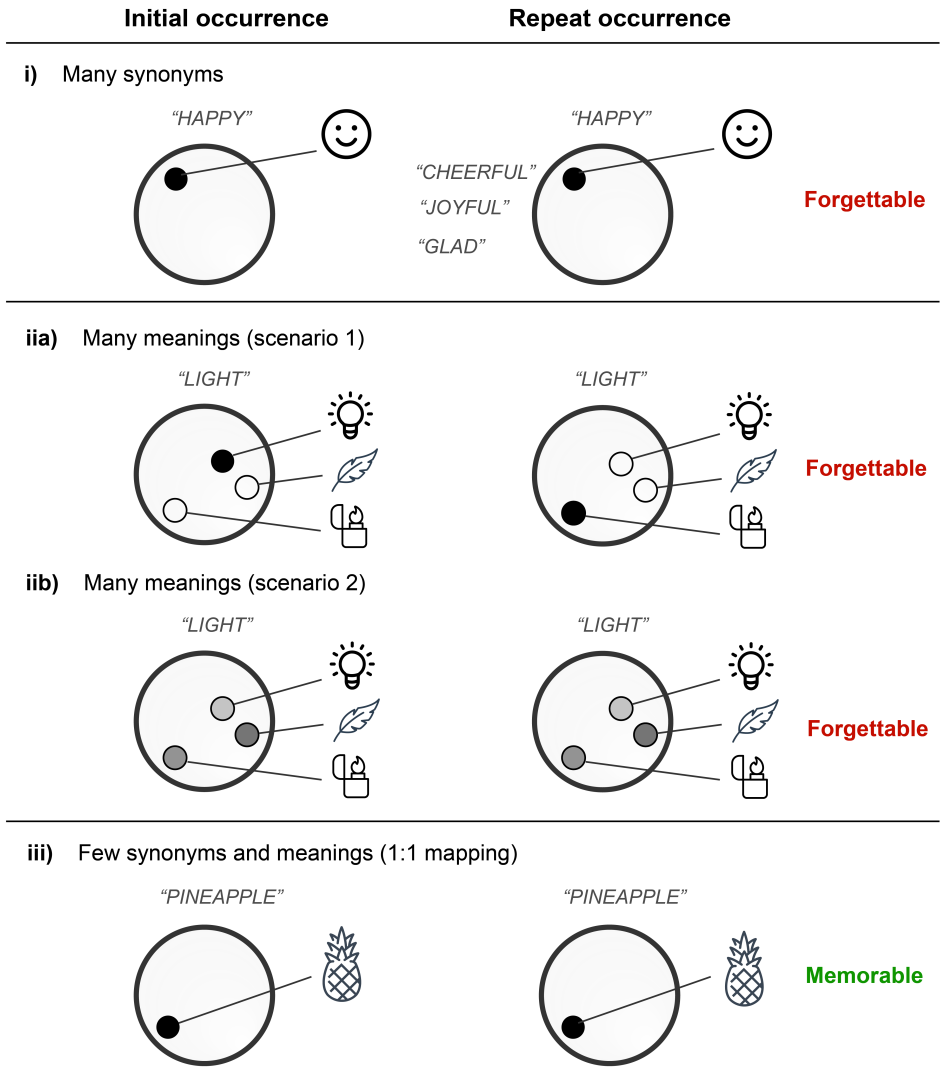


Figure 1. Schematic of the ideal observer model. The first column illustrates the representation of the initial occurrence of a word, and the second column illustrates hypothesized

representations of the repeat occurrence of the same word. In each case, the large circle represents semantic memory, and smaller circles within it represent the portion(s) of the semantic space that is/are activated by particular words, with darker circles corresponding to stronger activation. **(i)** A word like "HAPPY" with many synonyms (e.g., "CHEERFUL", "JOYFUL", "GLAD", etc.) is predicted to be forgettable as any of the synonyms could have generated the relevant meaning. **(ii)** A word like "LIGHT" with many meanings (e.g., a fixture in a house, the opposite of "heavy", a cigarette lighter, etc.) is predicted to be forgettable either because a different meaning is activated at the repeat occurrence from the one activated at the initial occurrence (iia), or because several meanings are activated (to different degrees), leading to a diffuse memory trace (similar to Steyvers & Malmberg's 2003 proposal) (iib). **(iii**) An unambiguous word with no synonyms (i.e., a word that has a unique association with its meaning), like "PINEAPPLE", is predicted to be memorable.

According to this model, a word that has been encountered might not be remembered for two reasons:

(1) $P(m \mid W = w)$ is high, but $P(m \mid W = w')$ for some other word(s) w' is also high. This is a case where w has one or more synonyms, such that the meaning m can be expressed by several words, which will compete with w as the cause of the memory. This scenario is illustrated in Figure 1 (i). Because "HAPPY" has many synonyms (e.g., "CHEERFUL", "JOYFUL", "GLAD"), any one of these words could have generated the relevant meaning.

(2) $P(m \mid W = w)$ is low (i.e., the distribution over a word's meaning has high entropy). This is a case where w is ambiguous, i.e., has more than one meaning. An ambiguous word, like "LIGHT" (which has several meanings: e.g., a fixture in a house, the opposite of "heavy", a cigarette lighter), can be forgettable for two reasons. One possibility is that a single meaning is accessed and robustly activated during the initial encounter, but a *different* meaning is accessed during the repeat occurrence. This scenario is illustrated in Figure 1 (iia). The word "LIGHT" would not be recognized because no memory trace exists of the meaning of "cigarette lighter", accessed during the repeat occurrence (because the meaning "a fixture in a house" was accessed during the first encounter). Another possibility is that during the initial encounter multiple meanings are accessed and each is activated to some degree, but not as strongly as a meaning of an unambiguous word. This scenario is illustrated in Figure 1 (iib). The word "LIGHT" would not be recognized because none of the meanings were sufficiently robustly activated during the initial encounter, leading to a weak subsequent memory of all of the meanings (a diffuse memory trace) and/or competition among the different meanings.

It is worth noting that a word may also not be remembered if it re-appears when $P(W = w)$ is low (conversely, a particular distractor may generate a false positive if it has a high $P(W = w)$). The term $P(W = w)$ expresses the *a priori* probability that the random variable W has the specific value w, which is equivalent to the probability of seeing a word repeated. In our experimental set-up, this value is similar for all critical words in our task given that any word which appears once in the task has the same probability of repeating, and no word appears more than twice over

the course of the experiment. Thus, we do not expect the prior probability $P(W = w)$ to play a big role, although it is possible that participants have prior expectations about the kinds of words that are likely to appear in experiments.

On the other hand, a word that has been encountered is likely to be remembered when:

(3) $P(m \mid W = w)$ is high, and $P(m \mid W = w')$ for all other word(s) w' is low. This is a case where w has no synonyms and a single meaning (i.e., it is uniquely associated with its meaning). This scenario is illustrated in Figure 1 (iii). Because no competition arises either among the synonyms of the word w for the cause of the memory of meaning m, or among its multiple meanings, the agent can be certain that if they have a memory representation of the meaning "PINEAPPLE", then the word "PINEAPPLE" must have been encountered given that only "PINEAPPLE" could have given rise to this memory.

From the above analysis, we derive two predictions for properties that should make a word memorable:

1. Words that have no/few synonyms should be more memorable than words with many synonyms.
2. Unambiguous words should be more memorable than words with two or more meanings.

Note that word frequency—shown to affect recognition memory in much past work (e.g., (39–41)Schulman 1967; Brown and Lewis 1981)—does not directly figure into our proposal. Instead, better recognition memory for rare words falls out of the fact that words with few synonyms and few meanings tend to be low-frequency words (Fenk-Oczlon and Fenk 2010; Piantadosi, Tily, and Gibson 2012; Jones, Dye, and Johns 2017; see also Monaco et al. 2007, for evidence that the frequency effect in recognition memory is related to the structure of the semantic space).

METHODS

Overall design

We evaluated the ideal observer model of word memorability in two large-scale behavioral recognition memory experiments, each with 2,222 target words, 8,000-9,000 filler words, and over 600 participants (n=672 in Experiment 1 and n=631 in Experiment 2). Building on past work on image memorability (e.g., Isola, Xiao, et al. 2011; Bainbridge, Isola, and Oliva 2013; Isola et al. 2014), the experiments were designed as repeat detection tasks in which participants viewed a long sequence of words, presented one at a time, and were asked to press a key whenever they noticed a repeat (a word that they have already encountered earlier in the sequence; Figure 2). Critical repeats, used to measure word memorability, occurred at lags of 91-109 words. Approximately one out of every five words was a critical repeat (by design, no critical repeats occurred during the first 90 words). To ensure that participants were paying attention, vigilance repeats (chosen from a set of filler words) occurred at lags of 1-7 words.

For each target word in the experiment, we empirically defined three measures of memorability: hit rate (proportion of trials on which a repeat was correctly detected; given that a repeat of any

given target word occurred at most once for any given participant, this measure can be re-written as "proportion of participants who correctly detected a repeat"), false alarm rate (proportion of trials on which a repeat was incorrectly claimed—or proportion of participants who incorrectly claimed a repeat), and accuracy [(hits (correct detections) + correct rejections on initial presentation) / (hits + correct rejections on initial presentation + missed detections + false alarms on initial presentation)].

Participants

Participants were recruited using Amazon.com's Mechanical Turk crowd-sourcing platform. Only workers with a U.S. IP address and an approval rating of > 95% were allowed to participate. 672 participants took part in Experiment 1, 631 participants took part in Experiment 2, and 3,780 participants took part in the norming studies, as elaborated below. The experiments were conducted with approval from and in accordance with the Committee on the Use of Humans as Experimental Subjects at the Massachusetts Institute of Technology. Participants gave informed consent before starting each experiment and were compensated for their time.

The participant exclusion procedure for the critical memory experiments was identical to the one reported by Isola, Xiao, et al. (2011). In brief, word sequences were broken up into 'levels' that each consisted of 120 words (and lasted 4.8 min). At the end of each level, the participant saw their correct response average score for that level and was allowed to take a short break. Participants could complete at most 30 levels and were able to exit the experiment at any time, including in the middle of a level. Worker performance was continuously monitored within each level, resetting at the end of the level. The experiment ended if a participant fell below a 50% success rate on the last 10 vigilance repeat trials or above a 50% error rate on the last 30 non-repeat trials. When this happened, all data collected on the words in the current level were discarded and the participant was flagged. Participants were allowed to restart the experiment as many times as they wished until they completed the maximum of 30 levels. However, participants who received three flags (as noted above) were blocked from further participation in the experiment. Upon each restart, the sequence was reset so that the participant would never see a word they had seen in a previous session.

In Experiment 1, participants saw between 4 and 1,353 trials, with a median of 264.5 and a mean of 445.8. The median participant saw 74 target repeat trials, and 645 participants saw at least one target repeat trial (and thus contributed to the memorability analyses). In Experiment 2, participants saw between 2 and 1,255 trials, with a median of 297 and a mean of 465.5. The median participant saw 86 target repeat trials, and 619 participants saw at least one target repeat trial.

Materials

Experiment 1: A frequency-weighted sample of 13,980 words was extracted from Subtlex (Brysbaert and New 2009), such that a word which is twice as frequent in Subtlex was twice as likely to be included in our sample. The words were then semi-manually filtered in order to remove offensive words, alternate forms of the same word (color/colour), and words that are clearly morphologically related (happy/happiness). From the remaining set of 11,182 words,

2,222 words were randomly selected for use as targets and the rest were used as fillers in the experiment.

Experiment 2: A set of 10,320 words was manually constructed to span diverse semantic domains and major parts of speech. The nouns were selected from the following 31 semantic categories: building components, calendar items, chemical elements, clothing, common names, containers, diseases and conditions, drinks, earth formations, family relations, famous landmarks, famous people, food, furniture, games, geography, household items, human body, human dwellings, living things, medicine, musical instruments, mythical creatures, people, precious stones, reading material, sports, tools, vehicles, weapons, weather. As in Experiment 1, 2,222 words were randomly selected for use as targets, and the rest were used as fillers. Nouns were presented with a determiner ("the" or "a"/"an"; the most commonly used determiner was chosen based on a set of native speakers' intuitions) and verbs with "to" in order to make it clear what part of speech was being used and to capture any potential syntactic category effects. We note that we use the term "word" to denote an experimental item, irrespective of whether the item consisted of a multi-word phrase (e.g., "high heels" or single token words prefixed with a determiner ("to help").

Procedure

Each word was presented for 1 second (in uppercase letters in Experiment 1, and in lowercase letters, except for proper nouns, in Experiment 2) followed by a 1.4 second fixation; thus participants had 2.4 seconds to respond. Participants were asked to press the **r** key when a word occurred that they had already seen. Critical repeats occurred on a subset of the 2,222 target words used in each experiment. In Experiment 1, each target word was seen twice by at least 56 and at most 109 participants (median=79; mean=78.9); in Experiment 2, each target word was seen twice by at least 49 and, at most, 118 participants (median=77; mean=77.7). Vigilance repeats (included to make sure that participants were paying attention) occurred on a separate subset of the words and were chosen randomly for each participant.

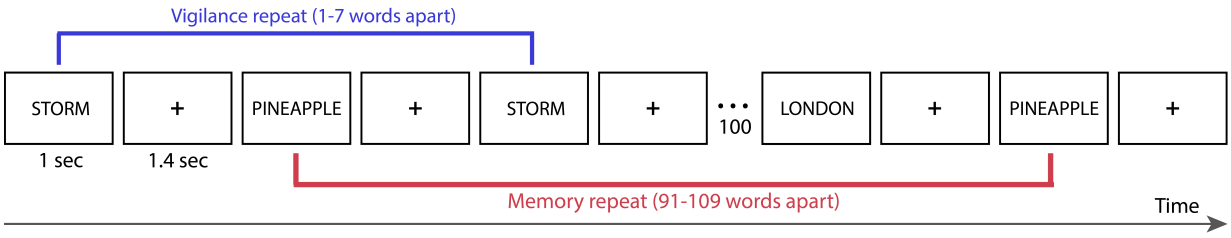


Figure 2. Illustration of the experimental paradigm. The procedure was identical across Experiments 1 and 2. Each word was presented for 1s, with 1.4s inter-trial intervals, giving participants 2.4s to respond. Participants were asked to press the **r** key when they encounter a word that had already occurred at some point in the experiment. Critical memory repeats (for the 2,222 target words in each experiment) occurred between 91 and 109 words apart, and vigilance repeats—between 1 and 7 words apart.

Word norms

To evaluate our hypothesis, for each of our two critical predictors—the number of synonyms and the number of meanings—we collected i) human judgments, and, for Experiment 1, ii) corpus-based estimates. For Experiment 2, items often consisted of multiple words/phrases (e.g., "bipolar disorder", "Christmas Eve" or "high heels"), which made accurate corpus estimates difficult to obtain.

To explore the relationship between the critical predictors and some of the factors that have been argued or shown in the past to affect word memorability, as well as to compare our hypothesis against some of the earlier proposals in the literature, we collected a set human and corpus norms for several additional predictors.

Norms for the two critical predictors:

Human estimates for the number of synonyms and the number of meanings for each word were obtained in norming experiments, conducted using Amazon.com's Mechanical Turk. For each set of critical materials (2,222 words from Experiment 1, and 2,222 words from Experiment 2), two experiments were conducted: one eliciting the number of synonyms judgments, and one—the number of meanings judgments. The materials in each set were divided into 9 subsets: 8 subsets of 250 words each, and 1 subset of 222 words. Each subset was presented to a different group of 30 participants, and also included a set of 20 catch trials, as detailed below. Thus, across the two sets of materials, 1,080 participants were tested (540 for the number-of-synonyms experiment, and 540 for the number-of-meanings experiment).

In each experiment, participants were asked to answer two questions about each word: i) whether the word was a real word of English (the 20 catch trials mentioned above were pseudowords and were included to filter out participants who respond randomly), and ii) the critical judgment. For the number of synonyms experiments, participants were asked to identify how many synonyms each real word has by choosing from among 5 options (0 synonyms, 1 synonym, 2 synonyms, 3-5 synonyms, more than 5 synonyms). For the number of meanings experiments, participants were asked to identify how many meanings each real word has by choosing from among 4 options (1 meaning, 2 meanings, 3-5 meanings, more than 5 meanings; note that a word may have 0 synonyms but not 0 meanings, hence only 4 options here).

Data from participants who provided ratings for more than 80% of non-words (suggesting they were not paying attention) were removed when computing the critical judgment norms. For Experiment 1, these exclusions left a total of 226 unique participants for number of synonyms with a median of 23 ratings per word (after excluding words with 10 or fewer ratings, as elaborated in the following paragraph) and 250 unique participants for number of meanings with a median of 27 ratings per word for the number of meanings (after excluding words with 10 or fewer ratings). For Experiment 2, these exclusions left a total of 223 unique participants for number of synonyms with a median of 23 ratings per word (after excluding words with 10 or fewer ratings) and 248 unique participants for number of meanings with a median of 26 ratings per word for the number of meanings (after excluding words with 10 or fewer ratings).

112 of the 2,222 words in Experiment 1, and 52 of the 2,222 words in Experiment 2 were identified as real words by 10 or fewer participants and were excluded from the critical

memorability analyses given that many participants in the memory experiments would also be unlikely to know these words. Additionally, one word from Experiment 1 (“BOUDOIRS”) was not present in the initial GloVe database used for this study (version glove.6B.50d.txt) and was therefore excluded from the analyses, leading to the exclusion of a total of 113 words from Experiment 1. For Experiment 2, five words were not available in the Google n-gram database (“a relux suppressant”, “a tangello”, “Arnold Schwazenegger", "Raphael (Raffaelo) Sanzio", "Scarlett Johanson"), mainly due to spelling errors that permeated the memory experiment, and were excluded, leading to the exclusion of a total of 57 words from Experiment 2. Thus, Experiment 1 contained 2,109 words for the critical memorability analyses, and Experiment 2 contained 2,165 words for the critical memorability analyses.

To derive a corpus-based estimate of the probability of a word given a meaning, we used the following: a) GloVe semantic word distinctiveness; and b) number of synonyms assigned to a word in Wordnet (Sigman and Cecchi 2002). To estimate GloVe semantic distinctiveness, we obtained a GloVe (Global Vectors for Word Representation; Pennington, Socher, and Manning (2014)) vector for each of the target words (pre-trained vectors from the Common Crawl corpus, available at http://nlp.stanford.edu/projects/glove/, version glove.840B.300d.txt), and calculated the mean Cosine similarity between this vector and all other word vectors in the set (n=2,109). This correlation reflects, on average, how similar a given word is to other words in terms of its co-occurrence characteristics (which have been shown to serve as a useful proxy for word meanings; e.g., Pereira et al. 2016). We note that this measure of semantic distinctiveness does not account for homonymy because the GloVe representation provides a single context-independent representation for each word.

And to derive a corpus-based estimate of the probability of a meaning given a word, we used a combination of the following two measures: a) Subtlex contextual diversity (the unique number of movie transcripts in which a word appears; words that have more meanings typically appear in more diverse settings); b) number of meanings listed for a word in Wordnet (Sigman and Cecchi 2002).

Additional norms:

In addition to the norms for the critical predictors, for each word, we obtained norms for five attributes that have been shown to affect word memorability in past work: concreteness, imageability, familiarity, valence, and arousal. These norming experiments were set up in a similar way to the ones for the critical predictors and also conducted using Amazon.com’s Mechanical Turk. In particular, for each set of materials (2,222 words from Experiment 1, and 2,222 words from Experiment 2), five experiments were conducted: one for each attribute of interest. The materials in each set were divided into 9 subsets: 8 subsets of 250 words each, and 1 subset of 222 words. Each subset was presented to a different group of 30 participants, and also included a set of 20 catch trials, as detailed below. Thus, across the two sets of materials, 2,700 participants contributed data (540 for each of the five attributes).

In each experiment, participants were asked to rate each word on a scale of 1-5 (the 20 catch trials mentioned above were chosen to serve as extremes: 10 on each side of the scale; for

example, for imageability, words like “POODLE” and “TOMATO” served as high-imageability catch trials and words like “ELUSIVE” and “RELATE” served as low-imageability catch trials).

Data from participants who did not rate at least 8/10 of the high and 8/10 of the low catch trials, and who assigned less than a 1-point difference in the expected direction for the catch trials were removed when computing the critical judgment norms. For Experiment 1, these exclusions left a total of 1,218 unique participants with a median of 26 ratings per word for concreteness (minimum: 25 ratings), a median of 26 ratings per word for imageability (minimum: 23 ratings), a median of 28 ratings per word for familiarity (minimum: 24 ratings), a median of 28 ratings for valence (minimum: 26 ratings), and a median of 22 ratings per word for arousal (minimum: 18 ratings). For Experiment 2, these exclusions left a total of 1,201 unique participants with a median of 24 ratings per word for concreteness (minimum: 19 ratings), a median of 26 ratings per word for imageability (minimum: 24 ratings), a median of 27 ratings per word for familiarity (minimum: 22 ratings), a median of 25 ratings for valence (minimum: 22 ratings), and a median of 23 ratings per word for arousal (minimum: 21 ratings). We note that although some of these norms are available in existing databases (e.g., Brysbaert, Warriner, and Kuperman 2014; Mohammad 2018), we chose to collect our own norms in order to obtain a unified set of norms for all items in Experiments 1 and 2. (As expected, our collected norms are strongly correlated with the norms in published databases for the sets of overlapping words[1].)

Finally, we obtained a corpus-based frequency measure: for Experiment 1, we used token log frequency of a word in the Subtlex subtitles corpus (Brysbaert, New, and Keuleers 2012); and for Experiment 2, because items often consisted of multiple words/phrases, we used the Google n-gram corpus (year 2013) to calculate log frequencies for each item (Michel et al. 2011). For completeness, we also obtained the Google n-gram frequencies for Experiment 1.

Modeling

Cross-validated model performance

For an unbiased evaluation of how predictive certain features (e.g., number of synonyms, number of meanings) are of memorability recognition accuracy, we fit a linear regression predicting per-word accuracy as a function of the feature norms of interest. The procedure is cross-validated across participants and words: we fit the linear regression model on half of the participants using half of the words (Experiment 1: 1,055 words for train set; Experiment 2: 1,083 for train set, unless otherwise stated—e.g., in analyses on subsets of the data, these numbers are lower) and test the model on the other half of the participants (Experiment 1: 1,054 words for test set; Experiment 2: 1,082 words for test set, unless otherwise stated).

[1] For concreteness, the number of overlapping words between our sets and the Brysbaert et al. (2014) database was 1,394 for Experiment 1 and 428 for Experiment 2, and the norms were correlated at r = 0.94 and 0.93 for Experiments 1 and 2, respectively. For valence, the number of overlapping words between our sets and the Mohammad (2018) database was 1,226 for Experiment 1 and 383 words for Experiment 2, and the norms were correlated at Pearson r = 0.86 and 0.89 for Experiment 1 and 2, respectively. Lastly, for arousal, the number of overlapping words between our sets and the Mohammad (2018) database was the same, and the norms were correlated at r = 0.71 and r = 0.70 for Experiment 1 and 2, respectively.

We demeaned each column of the regressor matrix (i.e., each feature), but we did not normalize the columns to have unit norm. Similarly, we demeaned the target vector (i.e., memorability recognition accuracy). The demeaning was performed on the train set and the same transformation was applied on the test set. This ensured independence (no data leakage) between the train and test sets. We performed this procedure 1,000 times and report the median Spearman correlation between the predicted recognition accuracy and the actual recognition accuracy.

Forward-backward feature selection

For an assumption-neutral evaluation of which features emerge in the best possible linear model of memorability, we implemented a forward-backward feature selection method. The feature selection method allows for features to be included/excluded in a linear model based on the p-values associated with a given feature. The forward step consists of inclusion of the feature associated with the lowest (i.e., most significant) p-value (if less than the inclusion threshold of 0.01). The backward step consists of exclusion of the feature associated with the greatest (i.e., least significant) p-value (if greater than the exclusion threshold of 0.05).

Besides the feature selection step, the modeling procedure was identical to the remaining model evaluations (as described in "Cross-validated model performance"): we partitioned the data into a train/test set of independent participants and words, demeaned the regressors and targets, and ran the feature selection procedure on the training data, obtaining a set of features for optimal fit to the training data. Next, we tested the model using these features on the test set and report the median Spearman correlation between predicted accuracy and the actual accuracy.

Transparency and openness

The memorability data for Experiment 1 and Experiment 2 is publicly available as csv files in the following repository: https://github.com/gretatuckute/memorable_words/. Similarly, the repository contains the code to reproduce the figures/tables in the manuscript.

RESULTS

1. How memorable are words?

In Experiment 1, we measured the memorability of 2,222 words sampled from the Subtlex corpus (Brysbaert and New, 2009), which consists of movie transcripts. The words sampled from this corpus are intended to represent a typical range of words that one might encounter in everyday life, including a mix of low-frequency and high-frequency words. Overall memory performance was high and similar to what has been previously reported for images (Isola, Xiao, et al. 2011). The median hit rate over words (computed as the number of correct repeat detections divided by the total number of critical repeat trials) was 0.69, the median false alarm rate (the number of incorrect repeat detections divided by the total number of non-repeat trials) was 0.09, and the median accuracy (the number of correct repeat detections and correct repeat non-detections divided by the total number of trials) was 0.80 (Figure 3A (iii)).

Although accuracy was high, some words were consistently better remembered than others (split-half Spearman correlation for accuracy across participants: 0.58 [95% CI of 0.56, 0.60 by non-parametric bootstrap]). This consistency indicates that there exists a reliable signal of word-intrinsic memorability, which varies substantially between words (Figure 3A shows the most and least memorable words).

To evaluate the generalizability of the results from Experiment 1, in Experiment 2, we measured the memorability of a new set of 2,222 words, which were hand-selected so as to span a wide range of semantic categories. Despite the fact that Experiment 2 used a very different lexicon from Experiment 1 and a new set of participants, the results were strikingly similar. The median hit rate was 0.69, the median false alarm rate was 0.09, and the median accuracy was 0.80. As in Experiment 1, some words were consistently better remembered than others (split-half Spearman correlation for accuracy across participants: 0.65 [95% CI of 0.63, 0.67 by non-parametric bootstrap]).

To allow for an across-experiment comparison for the same materials, 48 words overlapped between the sets of target words in Experiments 1 and 2. The accuracies for these 48 words were strongly correlated between the two experiments ($r=0.77$, $p<<.0001$; SI Figure 1). Similarly, the correlations for the human-derived norms (number of synonyms, number of meanings, concreteness, imageability, familiarity, valence, arousal) between both experiments were very high (in the range of $r=0.85$-0.98, $p<<.0001$; SI Table 1).

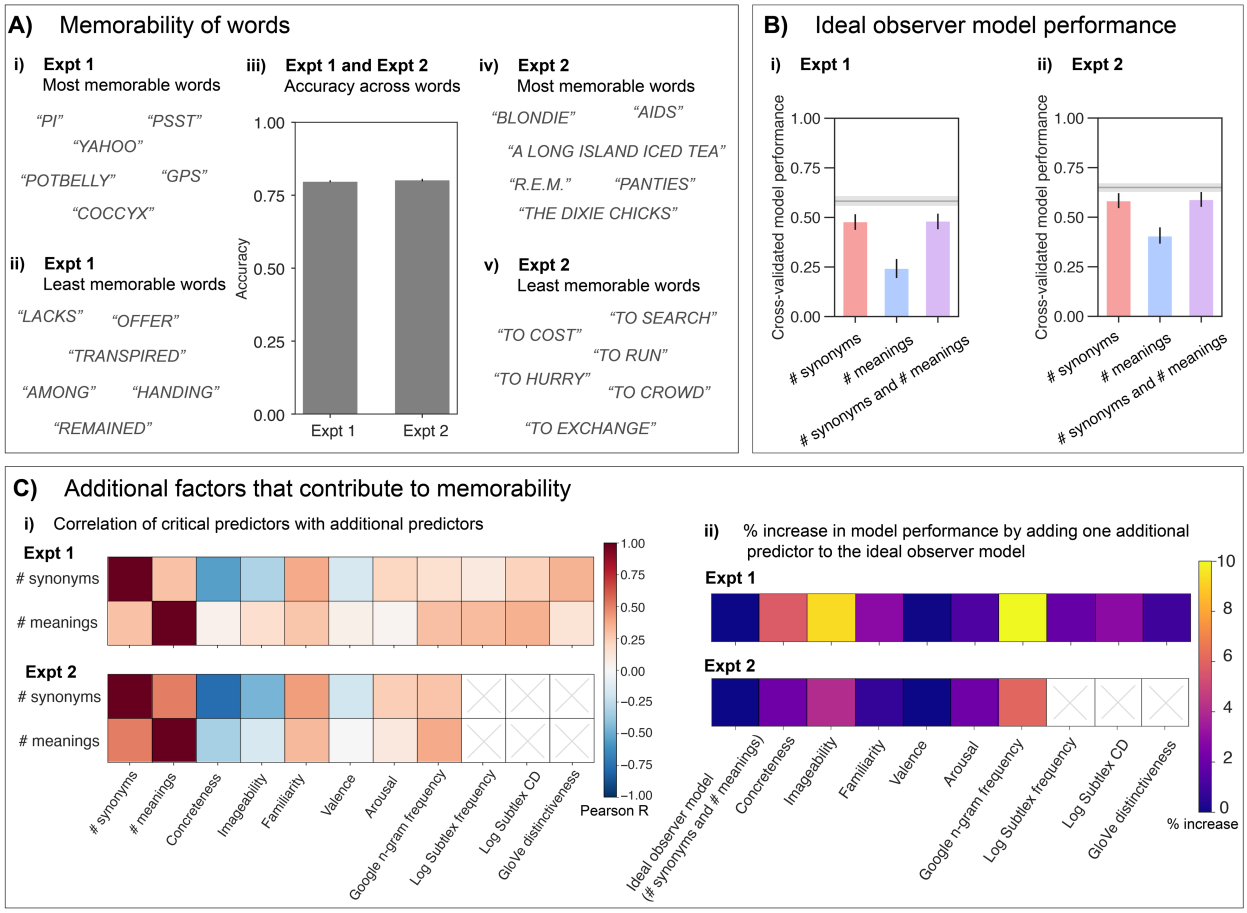


Figure 3. Critical results for Experiments 1 and 2. (A) Memory performance (median recognition accuracy) across all words in Experiment 1 (n=2,109 words) and Experiment 2 (n=2,165 words) is shown in (iii). Error bars are 95% CI of the median. The most memorable words for Experiments 1 and 2 are shown in (i) and (iv), respectively; and the least memorable words—in (ii) and (v). **(B)** Model performance using the number of synonyms, number of meanings, or both (the ideal observer model) as predictors of memory performance (see also Table 1). Median Spearman rank correlation between predicted and observed recognition accuracy for Experiments 1 and 2 is shown in (i) and (ii), respectively. Correlations were computed across 1,000 cross-validation folds using data from a set of non-overlapping participants and words for train/test splits. Error bars are 95% CI across cross-validation splits. The horizontal grey lines denote the theoretical maximum correlation, as defined by the split-half Spearman correlation for accuracy across participants. **(C)** The correlations between the critical predictors (human-derived number of synonyms and number of meanings norms, as described in Section "Word norms"), and between the critical predictors and other predictors are shown in (i). The set of non-critical predictors includes five human judgment norms (concreteness, imageability, familiarity, valence, and arousal) and some corpus-based norms. For Experiment 1, we included four corpus-based norms: Google n-gram frequency (the overall frequency of a word in the Google n-gram database in 2013), Subtlex frequency (the overall frequency of a word in the Subtlex movies transcript), Subtlex contextual diversity (CD) (a measure of the number of distinct movie transcripts in which a word appears in the Subtlex corpus), and GloVe distinctiveness (a measure of semantic distinctiveness derived from corpus-based GloVe co-occurrence statistics); for Experiment 2, we included Google n-gram frequency. Percent increase

in model performance associated with each non-critical predictor is shown in (ii) (see SI Table 7 for the statistical evaluation).

2. How well does the ideal observer model explain word recognition performance?

To test the predictions of the proposal put forward here, we examined the relationship between word memorability and the number of synonyms and number of meanings. Because corpus-based measures were not possible to obtain for the materials in Experiment 2, we report the results based on the human norms in the main text. It is important to note that the human-derived norms were highly reliable: The split-half Spearman correlation for the rating values across participants was 0.86 [95% CI 0.81, 0.89] for synonyms and 0.74 [95% CI 0.69, 0.77] for meanings for Experiment 1, and 0.92 [95% CI 0.89, 0.93] for synonyms and 0.78 [95% CI 0.75, 0.81] for meanings for Experiment 2. (In Experiment 1, where we were able to obtain corpus-based measures in addition to the human ratings, we found that the corpus-based measures and human ratings are correlated: the number of synonyms norms are correlated at r=0.35, and the number of meaning norms are correlated at r=0.54. For completeness, we report the results for Experiment 1 based on the corpus-derived measures in SI Figure 2 and SI Table 2.).

Each of the two critical measures (number of synonyms, number of meanings) was predictive of memory recognition accuracy in both Experiments 1 and 2, as summarized in Figure 3B and Table 1. We fit a linear regression predicting per-word accuracy as a function of both number of synonyms and number of meanings. To avoid overfitting, we learned the model coefficients using half of the participants and half of the words and tested the model on the non-overlapping portion of the data. Iterating this procedure 1,000 times, the median Spearman correlation between the memorability accuracies and model predictions using both number of synonyms and number of meanings was 0.48 [95% CI 0.44, 0.52] out of a theoretical maximum of 0.58 (the split-half correlation across participants) for Experiment 1, and 0.59 [95% CI 0.55, 0.63] out of a theoretical maximum of 0.65 for Experiment 2 (Table 1).

Experiment	# synonyms	# meanings	# synonyms and # meanings
Expt 1	0.48 [95% CI 0.44, 0.52]	0.24 [95% CI 0.19, 0.29]	0.48 [95% CI 0.44, 0.52]
Expt 2	0.58 [95% CI 0.55, 0.62]	0.41 [95% CI 0.37, 0.45]	0.59 [95% CI 0.55, 0.63]

Table 1. Ideal observer model performance. Rows: experiments (1, 2); columns: predictors (number of synonyms, number of meanings, both; based on human norms). Values are the median Spearman correlation between memory performance (recognition accuracy) and model predictions across 1,000 cross-validation splits using data from a set of non-overlapping participants and words for train/test splits.

Thus, a simple model of word memorability, based on just two rationally motivated factors—number of synonyms and number of meanings—captures a large portion of the variance in word memorability (0.48/0.58 = ~83% and 0.59/0.65 = ~91% of explainable variance in Experiments 1 and 2, respectively). In a control analysis, we showed that in sharp contrast with our critical predictors, form-based predictors based on orthographic and phonological features (Marian et al.

2012) explain almost no variance in word memorability (SI Figure 3, SI Table 3) (cf. Malmberg et al. (2002)).

Next, we investigated whether one the critical predictors (number of synonyms or number of meanings) was a stronger predictor of memorability. We evaluated this in two ways. First, as shown in Figure 3B, we evaluated how much variance a model with only one of the critical predictors explains on the held-out data across 1,000 splits (Table 1). We observed that the model with *only* number of synonyms as a predictor performs on par with the model with both predictors (0.48 with just number of synonyms vs. 0.48 with both predictors for Experiment 1, and 0.58 with just number of synonyms vs. 0.59 with both predictors for Experiment 2). Thus, the explained variance does not seem to benefit substantially from adding number of meanings as a predictor.

And second, we performed a statistical comparison of how much benefit there is of adding each of the critical predictors to the full model using a likelihood ratio test comparing the full model to a model without the critical predictor (SI Table 4). For Experiment 1, in line with the results based on the cross-validated model performance (Table 1, Figure 3B), we observed that adding number of synonyms as an additional predictor to the number of meanings model provides a large improvement in model fit ($F=651.24$, $p<<.0001$). Adding number of meanings as an additional predictor to the number of synonyms model yields some model fit improvement ($F=20.37$, $p<<.0001$), albeit much smaller. The trends were similar for Experiment 2: adding number of synonyms as an additional predictor to the number of meanings model provides a large improvement in model fit ($F=1,073.26$, $p<<.0001$), compared to a significant but much smaller improvement when adding number of meanings as an additional predictor to the number of synonyms model ($F=11.88$, $p<.001$). Finally, because the number of synonyms and the degree of lexical ambiguity may vary between syntactic categories, following a reviewer's suggestion, we included part of speech[2] as an additional predictor in the baseline model: doing so provided only a modest improvement in model fit (Experiment 1: $F=19.87$, Experiment 2: $F=39.33$, both $p<<.0001$).

Thus, both methods for comparing the contributions of the two critical predictors yield a similar answer: number of synonyms is a stronger contributor to memorability compared to number of meanings.

3. Do additional factors contribute to word memorability?

To explore the effects on word memorability of the factors that have been argued in the past to be important, as well as to compare our hypothesis against some of the earlier proposals in the literature, we performed several analyses.

[2] For Experiment 1, the part of speech (POS) composition of the materials was: adjectives: 121 items, nouns: 1,836 items, verbs: 56, adverbs: 56, and other: 17 (obtained using the NLTK Python library (Bird and Loper 2004)). For Experiment 2, the POS composition of the materials was: adjectives: 304 items, nouns: 1,265 items, and verbs: 596 (POS estimates were manually assigned during the material construction procedure).

First, we examined the relationship (via Pearson correlations) between each of our critical predictors and each of the additional predictors (Figure 3C). The latter set included five predictors obtained from human ratings (concreteness, imageability, familiarity, valence, and arousal) and corpus-based norms (for Experiment 1, frequency and contextual diversity (CD) measures (from Subtlex and Google n-gram) and GloVe distinctiveness; for Experiment 2, the Google n-gram frequency). As can be seen in Figure 3C (i) and focusing on relationships that were consistent between the two experiments, number of synonyms showed a strong negative correlation with concreteness (r=-0.55 in Experiment 1, and r=-0.76 in Experiment 2): words that were rated as more concrete were rated as having fewer synonyms. In addition, number of synonyms showed a moderate negative correlation with imageability (r=-0.31 in Experiment 1, and r=-0.47 in Experiment 2): words that were rated as more imageable were rated as having fewer synonyms. (Concreteness and imageability were strongly correlated for the materials in both Experiment 1 (r=0.85) and Experiment 2 (r=0.80) (SI Figure 4).) Further, number of synonyms was also moderately correlated with familiarity (r=0.38 in Experiment 1, and r=0.42 in Experiment 2). Finally, number of synonyms showed weak positive correlations with frequency (r=0.10 for Subtlex frequency and r=0.16 for Google n-gram frequency in Experiment 1, and r=0.28 in Experiment 2) and arousal (r=0.21 in Experiment 1 and r=0.24 in Experiment 2) and, and a weak negative correlation with valence (r=-0.17 in Experiment 1, and r=-0.19 in Experiment 2): words that were more frequent and rated as more arousing and less positively charged were rated as having fewer synonyms.

Number of meanings was moderately correlated with frequency (r=0.32 for Subtlex frequency and r=0.30 for Google n-gram frequency in Experiment 1, and r=0.38 in Experiment 2): more frequent words were rated as having more meanings, as expected given past work (e.g., Fenk-Oczlon and Fenk 2010; Piantadosi et al. 2012). Similar to number of synonyms, number of meanings was also moderately correlated with familiarity (r=0.28 in Experiment 1, and r=0.32 in Experiment 2). (See SI Figure 4 for all predictor correlations.).

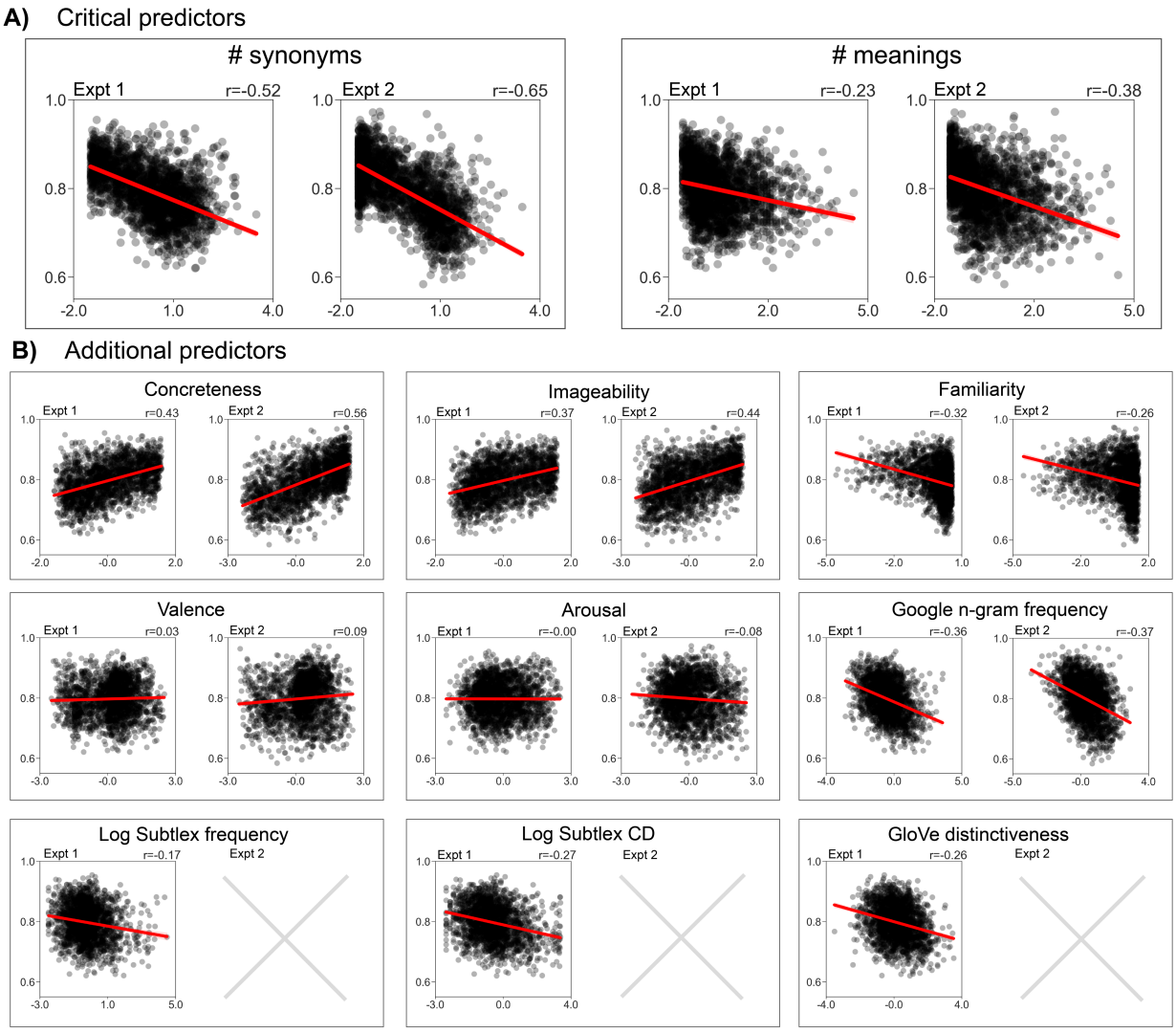


Figure 4. Correlation of each predictor with recognition accuracy. On each plot, the x-axis shows the z-scored predictor and the y-axis shows recognition accuracy. The red line is the line of best fit. The Pearson correlation is reported in the top right of each plot. **(A)** Critical predictors (human-derived number of synonyms and number of meanings). **(B)** Additional predictors. (See Figure 3C for more information on the relationship between the critical and the additional predictors.)

Next, we asked whether any of the additional predictors explain any variance above and beyond the two critical predictors. For Experiment 1, there were nine additional predictors (concreteness, imageability, familiarity, valence, arousal, Google n-gram frequency, Subtlex frequency, Subtlex contextual diversity, and GloVe distinctiveness, as evidenced in Figure 3C). For Experiment 2, there were six additional predictors (concreteness, imageability, familiarity, valence, arousal, and Google n-gram frequency). The correlation of each predictor with memory recognition accuracy is shown in Figure 4. To formally evaluate the contribution of each predictor, we compared the ideal observer baseline model that only includes the two critical predictors (human-derived number of synonyms and number of meanings) to a set of models each including one additional

predictor. Similar to Section 2, we investigated both the cross-validated model performance as well as performing a likelihood ratio test by comparing the ideal observer model against a model with the additional predictor. The percent increase in cross-validated model performance by adding each of the additional predictors is shown in Figure 3C (ii) and SI Tables 5 and 6.

For Experiment 1, the largest increase in explained variance from additional predictors stems from Google n-gram frequency (10.24% increase with a model performance of 0.53 [95% CI 0.49, 0.57]), followed by imageability (9.30% increase with a model performance of 0.53 [95% CI 0.49, 0.56]). Similarly, for Experiment 2, the largest increase in explained variance from additional predictors stems from Google n-gram frequency (6.07% increase with model performance of 0.62 [95% CI 0.59, 0.66]), followed by imageability (4.07% increase with model performance of 0.61 [95% CI 0.58, 0.65]). To statistically evaluate how much benefit additional predictors have, we used likelihood ratio tests (SI Table 7). These tests mirror the patterns from the cross-validated model performance analysis (Figure 3C, SI Tables 5, 6). For both experiments, adding Google n-gram frequency to the ideal observer baseline model results in the largest improvement in model fit (Experiment 1: $F=224.81$, $p<<.0001$; Experiment 2: $F=142.06$, $p<<.0001$) followed by imageability (Experiment 1: $F=210.18$, $p<<.0001$; Experiment 2: $F=101.58$, $p<<.0001$).

Given that we observed that frequency was a strong additional predictor, we performed a supplementary analysis to investigate whether the ideal observer model could still predict memorability when words were similar in frequency. To do so, we partitioned the words into low-, medium-, and high-frequency subsets and found that the ideal observer model still explained most of the explainable variance within these partitions. Hence, lexical frequency cannot account for the high performance of the ideal observer model (SI Figure 5, SI Tables 8, 9).

Finally, we compared our model's performance to two previous proposals, both within the framework of Bayesian optimal inference: those of Steyvers and Malmberg (2003) and Griffiths et al. (2007). To do so, we used the data from Experiment 1. We found that compared to our baseline model (model performance: 0.48 [95% CI 0.44, 0.5]), a model that only includes the corpus-derived contextual diversity predictor, as proposed by Steyvers and Malmberg (2003), explains substantially less variance (0.23 [95% CI 0.18, 0.28]; SI Figure 6A, SI Table 10). Furthermore, including the corpus-derived contextual diversity (CD) measure as an additional predictor only shows a small increase in performance above the baseline model (0.50 [95% CI 0.46, 0.53]; SI Table 6). So despite the fact that the contextual diversity measure is moderately correlated with both of our critical predictors (number of meanings: $r=0.35$; number of synonyms: $r=0.23$; Figure 3C (i) and SI Figure 3), this measure alone only explains a fraction of the variance. Griffiths et al. (2007) demonstrated that the number of topics a given word is associated with (what they termed 'topic variability') is a better predictor of human recognition performance than contextual diversity. We replicated this finding here using the topic variability scores released in Griffiths et al. (2007) (available for a subset of our experimental materials) (SI Figure 6B, SI Table 11). However, the topic variability model was not on par with the ideal observer baseline model (SI Figure 6B, SI Table 11) (independent two-sided t-test between Spearman correlation values across 1,000 cross-validation splits: $p<<.001$, $t=61.57$, Cohen's $d=2.75$). In summary, the two-factor model based on the number of synonyms and the number of

meanings provides a quantitatively better account of recognition memory that two related proposals within the Bayesian framework. As noted in the Discussion, the advantage of the ideal observer model likely stems from the inclusion of the number of synonyms predictor.

4. What is the best possible model of memorability?

Finally, we tested whether our critical predictors emerge in the best possible linear model of memorability in an assumption-neutral manner. To do so, we performed a forward-backward feature selection analysis using all the features visualized in SI Figure 4 (same set of features as in Figure 4 with the addition of corpus-based synonym- and meaning norms for Experiment 1). This approach allows features to be included/excluded in the memorability model based on the p-values associated with a given feature with no top-down assumptions (see Methods: Modeling; Forward-backward feature selection). We partitioned the data into a train/test set of independent participants and words (identical to Sections 2 and 3) and ran the feature selection procedure on the training data, obtaining a set of features for optimal fit to the training data (n=1,000 cross-validation folds). Next, we tested the model using these features on the test set and report the median Spearman correlation between memorability accuracy and predicted memorability.

For Experiment 1, the feature pool consisted of human- and corpus-based meaning and synonym norms as well as the five norms obtained from human ratings (concreteness, imageability, familiarity, valence, and arousal), corpus-based frequency (from Subtlex and Google) and contextual diversity (CD) measures (from Subtlex), and GloVe semantic distinctiveness, i.e., a pool of 13 predictors in total. We obtained a maximum explained variance of 0.59 [95% CI 0.55, 0.62] (the subject-wise noise ceiling was 0.58 for Experiment 1). The most frequently occurring models across 1,000 cross-validation splits included the following six predictors (these features were selected 276/1,000 times):

 # Synonyms (human), Google n-gram frequency, Imageability, Familiarity, # Meanings (Wordnet), Subtlex frequency.

The second most frequently selected feature set was (these features were selected 80/1,000 times):

 # Synonyms (human), Google n-gram frequency, Imageability, Familiarity, # Meanings (Wordnet), Subtlex frequency, Subtlex CD.

Furthermore, across the 1,000 models, # Synonyms (human or corpus-based) was selected every time (in fact, it was selected as the first predictor every time); # Meanings (human or corpus-based) was selected 217 times. (See Table SI 12 for predictor inclusion numbers for all predictors).

For Experiment 2, the feature pool consisted of human synonym- and meaning norms as well as the five norms obtained from human ratings, and frequency (from Google n-gram), i.e., a pool of 8 predictors in total. We obtained a maximum explained variance of 0.63 [95% CI 0.60 0.67] (the subject-wise noise ceiling was 0.65 for Experiment 2). The most frequently occurring models included the following three predictors (these features were selected 515/1,000 times):

 # Synonyms (human), Google n-gram frequency, Imageability.

The second most frequently selected feature set was (these features were selected 192/1,000 times):

 # Synonyms (human), Google n-gram frequency, Imageability, Arousal.

As in Experiment 1, across the 1,000 models, # Synonyms (human) was selected every time (and it was selected as the first predictor every time); # Meanings (human) was selected 9 times.

Thus, by using an assumption-neutral, cross-validated approach to estimate which features contribute most to memorability, we find that, across the two experiments, our two critical predictors—number of synonyms and number of meanings—along with frequency, imageability, and familiarity predict the data up to the subject-wise noise ceiling. Mirroring the results in Section 2, the number of synonyms predictor is a stronger predictor to word memorability than the number of meanings predictor (in particular, for Experiment 2).

DISCUSSION

We investigated word memorability across two large-scale behavioral recognition-memory experiments (n=672 and 631 participants in Experiments 1 and 2, respectively; n=2,222 target words in each experiment; n=3,780 participants used in the norming experiments). The contributions of the current work are four-fold. First, we found that recognition memory for words is on par with recognition memory for images, contrary to accounts of superior memory for images. Second, across two large sets of words, we found that memorability is largely an intrinsic property of words: some words are consistently remembered better than others across participants. Third and critically, building on past work (e.g., Steyvers & Malmberg, 2003 and Griffiths et al., 2007), we evaluated and provided support for a novel proposal for what makes words memorable—the ideal observer model of word memorability—whereby memorable words have unique associations with their meanings. Finally, we systematically evaluated several additional factors that have been argued to affect word memorability in past work and found some support for frequency, imageability, familiarity, and arousal. We elaborate on these findings below. In addition, by making the memorability data and all the behavioral norming data and Corpus measures publicly available, we hope to help move the field of word memory research forward, allowing for streamlined testing of novel proposals.

Words are as memorable as pictures

In many past experiments, memory for words has been shown to be weaker than memory for pictures in both recognition (Shepard 1967; Standing 1973; Nelson, Metzler, and Reed 1974) and recall (Paivio, Rogers, and Smythe 1968; Standing 1973; Potter, Valian, and Faulconer 1977)—a phenomenon known as the “picture superiority effect”. The superior memory for pictures has been explained by the “dual coding account” (Paivio, Allan 1971), which posits that images activate both pictorial and verbal codes as opposed to only verbal codes for words, making

images easier to remember due to two memory traces. Other accounts of the picture superiority effect have emphasized increased perceptual, i.e. "stimulus surface", features of images (Nelson et al. 1974) and increased conceptual engagement (Smith and Magee 1980; Stenberg 2006).

Interestingly, in our experiments, word memory was on par with picture memory performance in similar experiments (Isola, Xiao, et al. 2011). Our experimental set-up is nearly identical to that of Isola et al. (2011), except for the use of words rather than pictures as stimuli. Participants in their experiment correctly detected repeats 68% of the time and false alarmed 11% of the time on average. The corresponding numbers in our experiments are almost identical: 68% and 10%. Thus, there does not appear to be a memory advantage for pictures compared to words, at least in this kind of recognition memory set-up. Several prior studies have also failed to observe a picture superiority effect. For example, Schacter et al. (1999) reported a recognition task where participants had to distinguish among sets of semantically related items, and found similar performance for pictures and words. Others have not found the picture superiority effect in tasks that require retrieving semantic information (Amrhein, McDaniel, and Waddill 2002; Seifert 1997) and in speeded recognition tasks (Boldini et al. 2007). Why some past studies found better memory for pictures than words is unclear, but the evidence in tandem suggests that the picture superiority effect is not ubiquitous across memory recognition experiments.

Some words are intrinsically memorable and others—forgettable

In both experiments, we found that memorability of words is consistent across participants. This result suggests that that word memorability depends, in large part, on stimulus properties, and is a critical prerequisite for our ability to ask our critical research question: i.e., what makes words memorable? This result also mirrors the findings from the image memorability literature (Isola, Xiao, et al. 2011; Isola, Parikh, et al. 2011; Bainbridge et al. 2013; Borkin et al. 2013; Isola et al. 2014; Bylinskii et al. 2015; Bainbridge 2022).

The intrinsic nature of word memorability is consistent with item-noise models of episodic memory (e.g. Shiffrin and Steyvers 1997; McClelland and Chappell 1998), in which items are encoded by their intrinsic properties (in our model, as discussed in the next section, the features have to do with the word-to-meaning relationship). Conversely, our findings are at odds with context-noise models (Dennis and Humphreys 2001), which assume that only context variability (i.e., the diversity of contexts in which a test item appeared), *not* item-level information, affects recognition memory.

The consistency across participants in which words they found memorable is interesting given that individuals vary substantially in terms of the amount and kind of linguistic input they get across their lifetimes. Future work may investigate inter-individual differences in recognition memory for words. For example, are participants with more linguistic experience (as indexed, for example, by larger vocabularies) better at word recognition memory? What other aspects of individual experiences / cognitive abilities affect word recognition performance? And does word recognition memory correlate across individuals with recognition memory for images? In other words, does word and image memorability both depend on the properties of abstract concept representations, or on the particular interface between those representations and verbal / visual-image representations?

The ideal observer model explains most of the explainable variance in word recognition performance

In this work, we explored whether words are encoded in memory by their meanings. By formulating an ideal observer model of word memorability, we hypothesized that a memorable word selects a particular meaning in semantic memory. We indeed found that memorable words are the ones with few synonyms and few meanings, with the number of synonyms being a more important predictor of memorability.

Prior work on word recognition memory has focused on factors that relate to one of our two critical predictors: the number of word meanings. In particular, many researchers have investigated how the number of different contexts in which a word occurs—which should generally be higher for words with many meanings—affects memory performance (Shiffrin and Steyvers 1997; Dennis and Humphreys 2001; Steyvers and Malmberg 2003; Jones, Dye, and Johns 2017; Aue, Fontaine, and Criss 2018). Steyvers and Malmberg (2003) operationalized *contextual diversity* as the number of different texts (in a large language corpus) in which a given word appears. They found that words that occur in fewer contexts (distinct text documents) were better recognized than words that occur in many contexts. Of course, words that occur in different texts may still be used in similar semantic contexts. As a result, Griffiths et al. (2007) improved on Steyvers & Mamberg's idea by suggesting that word memorability should depend on the number of different topics a given word is associated with—what they call *topic variability*. This measure may therefore be more likely to capture the different meanings or distinct senses of a word. In our data, we find that both contextual diversity (Steyvers and Malmberg 2003) and topic variability (Griffiths et al. 2007) affect word recognition performance, with topic variability showing better performance[3]. Critically, however, our ideal observer model, which includes—in addition to the number of meanings predictor (one word associated with many meanings)—the number of synonyms predictor (many words associated with one meaning) explains substantially more variance than even the better-performing topic variability predictor ($p<<.001$). In follow-up analyses, we investigated why the number of synonyms predictor was superior to the number of meanings predictor. First, the variance of the number of meaning norms was lower than for the number of synonym norms in both experiments[4]. In other words, because the range of number of synonyms across experimental items was greater than the range of number of meanings, the number of meanings predictor was necessarily less powerful in distinguishing among items. Second, we quantified the reliability of the synonym and meaning norms, and although both were highly reliable (see Results; How well

[3] Interestingly, the better performance of the topic variability predictor does not appear to be due to it being a better estimate of the number of meanings: contextual diversity and topic variability correlate similarly strongly with our human-based norms for the number of meanings ($r=0.35$ and $r=0.32$, respectively). Instead, the topic variability measure appears to show a stronger correlation with our number of synonyms estimate ($r=0.52$; cf. 0.25 for the correlation between contextual diversity and the number of synonyms measure). Note that these correlation values were obtained using a subset of the words ($n=1,366$) for which topic variability norms were available.

[4] For Experiment 1, the average number of synonyms was 1.71 with a standard deviation of 1.01 whereas the average number of meanings was 1.42 with standard deviation 0.33. For Experiment 2, the average number of synonyms was 1.46 with a standard deviation of 1.19 whereas the average number of meanings was 1.35 with a standard deviation of 0.32.

does the ideal observer model explain word recognition performance?), the number of synonym norms were more reliable than the number of meaning norms. Thus, the number of synonyms predictor was likely a more important predictor of word memorability because our materials varied more along this dimension, and the number-of-synonym estimates were more reliable.

It is worth noting that number of synonyms has also been shown to affect lexical access (e.g., in lexical decision and naming tasks): response latencies are slower for words with many synonyms than for word with fewer synonyms (Hino, Lupker, and Pexman 2002). These findings suggest that a one-to-one mapping between a word form and a meaning not only leads to more robust memory traces, as shown here, but also facilitates retrieval of words from memory.

The ideal observer model provides a straightforward explanation for why phrases and sentences are generally remembered better than single words: tapping into particular parts of the conceptual space becomes easier when words are combined into phrases because doing so results in semantically narrower, more specific meanings. This hypothesis should be evaluated directly in future work.

Finally, our finding that verbal memorability strongly depends on the relationship between the words and their meanings mirrors findings from the literature on image memorability, where conceptual features, in addition to perceptual ones, have been shown to drive recognition of images (Konkle et al. 2010; Huebner and Gegenfurtner 2012; Lin et al. 2019).

The role for additional factors in word recognition performance

We investigated which features emerge in the best possible linear model of memorability in an assumption-neutral manner. We performed a forward-backward feature selection that included our critical predictors (number of synonyms and number of meanings) and other predictors that have been shown to affect word memorability in prior work. The patterns were similar for both experiments: number of synonyms emerged as the strongest predictor in both. Number of synonyms was followed by word frequency, in agreement with prior work demonstrating that lower-frequency words are easier to recognize (e.g., Brown and Lewis 1981; Gorman 1961; Kinsbourne and George 1974; Schulman 1967). The next predictor that was shared between experiments was imageability, also in line with prior work (e.g., Gorman 1961; Klaver et al. 2005; Paivio 1969; Rubin and Friendly 1986; Walker and Hulme 1999). For Experiment 2, the most frequently selected feature set for the best possible linear model of memorability consisted of these three features (number of synonyms, frequency, and imageability). For Experiment 1, the best feature set additionally included human estimates of number of meanings, familiarity, a corpus-based measure of the number of meanings, and contextual diversity.

Implications of understanding word memorability

Why might one want to know which words are memorable? A lot of what we learn about the world we learn through language from other humans rather than through our direct experience. Language is also used to build inter-personal relationships, maintain international political order, and bring about social change. Formulating messages in a way that powerfully and precisely activates the relevant conceptual structures in other people's minds is therefore of critical

importance at both personal and societal level. Of course, the memories we retain are at the conceptual level, but word choices can help optimize the initial semantic encoding. As a result, understanding which words lead to longer-lasting traces in memory can enable more precise and effective information sharing.

Limitations and future directions

The underlying mechanisms of how different features affect word memorability are yet unclear. In future work, these features can be manipulated in a targeted manner to investigate and understand their precise effects on verbal memorability in more naturalistic settings. Moreover, of course, most linguistic messages do not consist of single, isolated words. A critical future direction will be to extend the current account to sentences. Understanding how well the memorability of individual words explains the memorability of longer linguistic strings can illuminate fundamental aspects of language processing and complex meaning construction in the mind and brain.

We have also introduced data and tools that allow for formulating predictions about the memorability of newly encountered words. Based on the word memorability dataset presented here a simple linear model can be used to predict the memorability for any English word. Additionally, because the critical predictors (number of synonyms and number of meanings) can be automatically estimated from text corpora using the ever-improving tools from Natural Language Processing, the simple Bayesian model presented here can be used to make predictions about any word in any language where large corpora are available.

Conclusion

In this paper, we have offered a simple account, based on rational analysis, as to what makes words memorable. Our model posits that because words are encoded in memory by their meanings, words that uniquely pick out a particular meaning in semantic memory (i.e., unambiguous words with no synonyms) should be the memorable ones. We evaluated this idea across two large-scale experiments. The scale of our study makes the results more likely to generalize to other words and other participants. Thus, building on some classic findings, this work lays a theoretical and empirical foundation for future work on verbal memorability.

CONSTRAINTS ON GENERALITY

Ecological validity and effects of context

The current study investigates memorability of words in isolation. We acknowledge the importance of context effects on memorability, as well as learning and processing of words. The experimental paradigm is not designed to investigate either ordering effects (words are presented as randomly ordered lists) or effects of natural linguistic context. However, we find a high intrinsic memorability of words: some words are consistently remembered better than others across participants, even when presented in isolation, out-of-context. Moreover, this recognition paradigm has been shown to be highly robust and—at least for images—memorability estimates

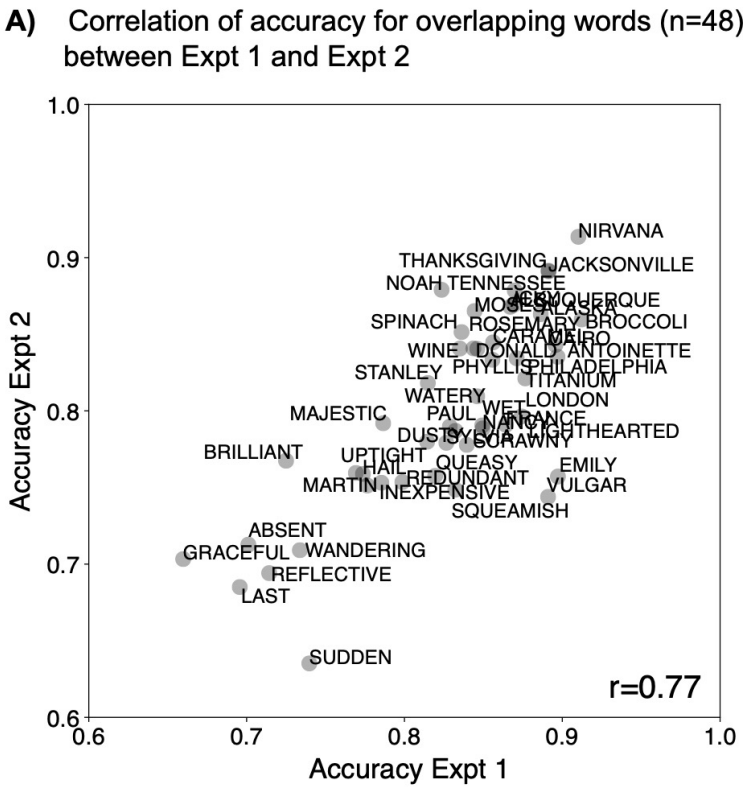
from the recognition task match incidental memory scores where a memory test is administered as a surprise (Goetschalckx, Moors, and Wagemans 2019). Thus, we believe that studying memory using words in isolation is still informative about human cognition, although the limitations of this particular approach should of course to be taken into account given the importance of context for word processing.

Extension to languages other than English

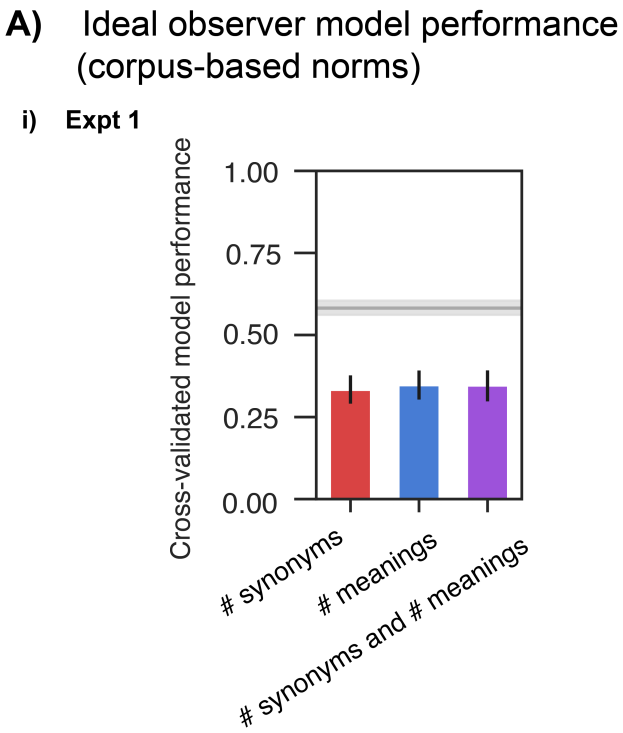
The predictions derived from the current study can be applied and tested in other languages given the availability of norms for the number of synonyms and meanings. We note that languages vary in the extent to which synonymy and ambiguity exist, and in how much contextual information is required to determine a word's meaning (e.g., Lin & Ahrens, 2010); this variability may affect word memory/recognition across languages.

SUPPLEMENTAL INFORMATION (SI)

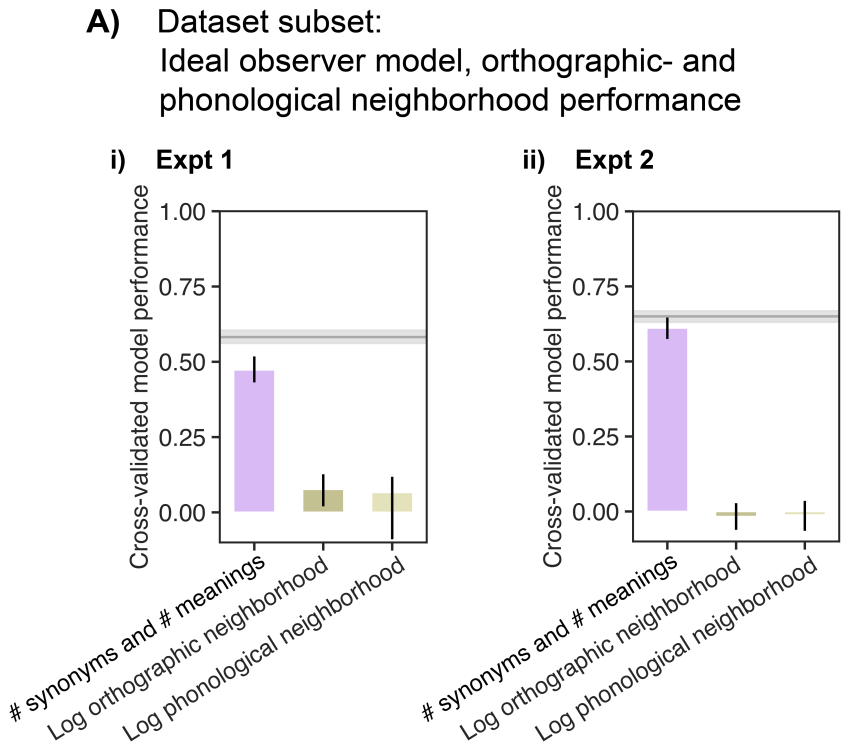
SI Figures



SI Figure 1. Correlation (Pearson R) of memory recognition accuracy for overlapping words (n=48) between Experiment 1 and Experiment 2.

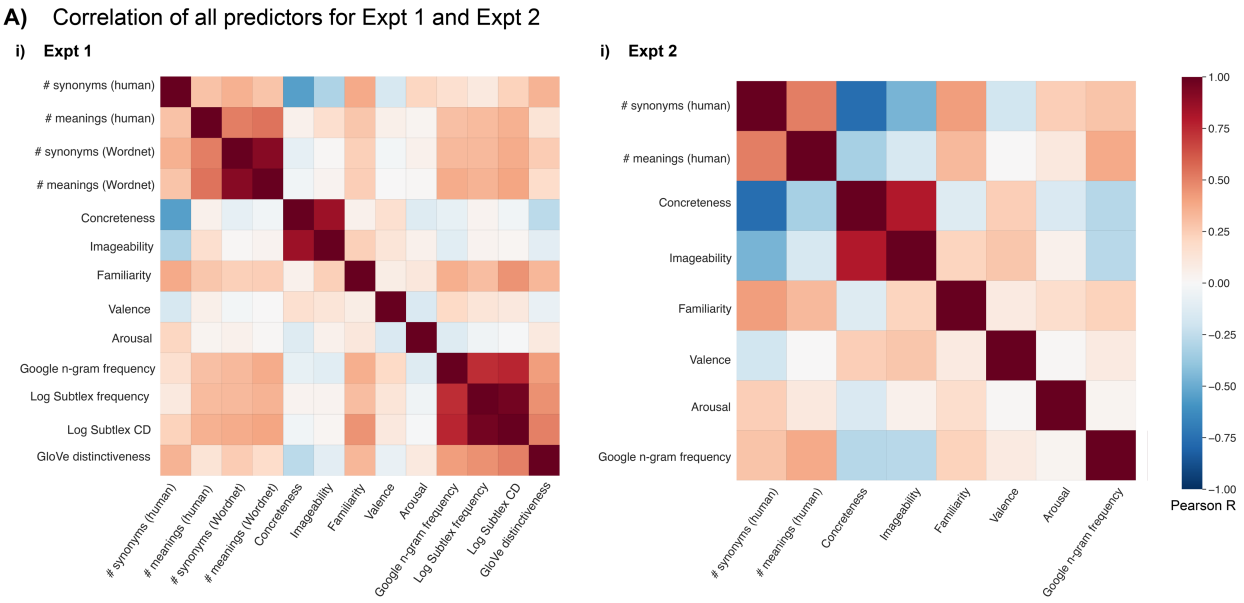


SI Figure 2. Ideal observer model performance using corpus-based norms. For Experiment 1, we were able to estimate norms for number of synonyms and meanings that do not rely on explicit human judgments (instead obtaining number of meanings and synonyms using Wordnet (Miller, 1990)). Bars show the corpus-based number of synonyms, number of meanings, or both (the ideal observer model) as predictors of memory performance (see also SI Table 2). Median Spearman rank correlation between predicted and observed recognition accuracy for Experiment 1 is shown in (i). Correlations were computed across 1,000 cross-validation folds using data from a set of non-overlapping participants and words for train/test splits. Error bars are 95% CI across cross-validation splits. The horizontal grey lines denote the theoretical maximum correlation, as defined by the split-half Spearman correlation for accuracy across participants.

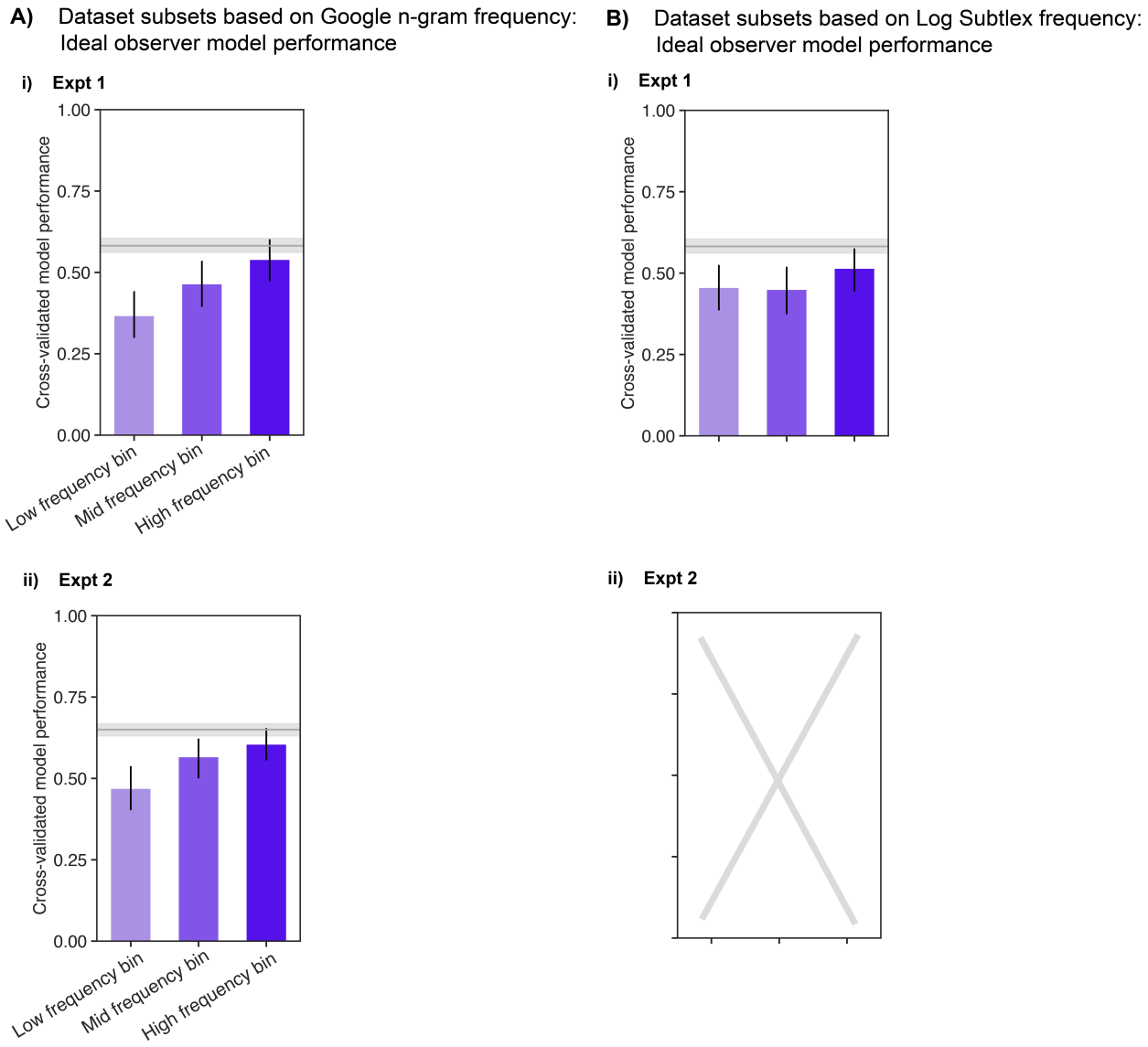


SI Figure 3. Comparison of the ideal observer model and orthographic and phonological neighborhood predictors (derived from the CLEARPOND database; Marian et al. 2012). A word's neighbors were defined as those that differ from it in either one letter (for orthographic neighborhoods) or one phoneme (for phonological neighborhoods). Neighbors consisted of words that differed only by the addition, deletion, or substitution of a single letter/phoneme. Bars show model performance as quantified by the median Spearman rank correlation between predicted and observed recognition accuracy for Experiments 1 and 2 (in panels (i) and (ii), respectively) (see also SI Table 3). Correlations were computed across 1,000 cross-validation folds using data from a set of non-overlapping participants and words for train/test splits. Error bars are 95% CI across cross-validation splits. The horizontal grey lines denote the theoretical maximum correlation, as defined by the split-half Spearman correlation for accuracy across participants.

We were able to estimate the orthographic and phonological norms for a subset of critical words (Experiment 1: n=1,873 words; 937 words in the train set and 936 words in the test set; Experiment 2: n=1,720 words; 860 words in the train set and 860 words in the test set). For the multi-word items in Experiment 2, such as "high heels", we obtained the orthographic/phonological feature for each sub-word, i.e. "high" and "heels" separately and averaged the values of interest to obtain the final predictor (which might explain the chance-level performance of these predictors for Experiment 2).

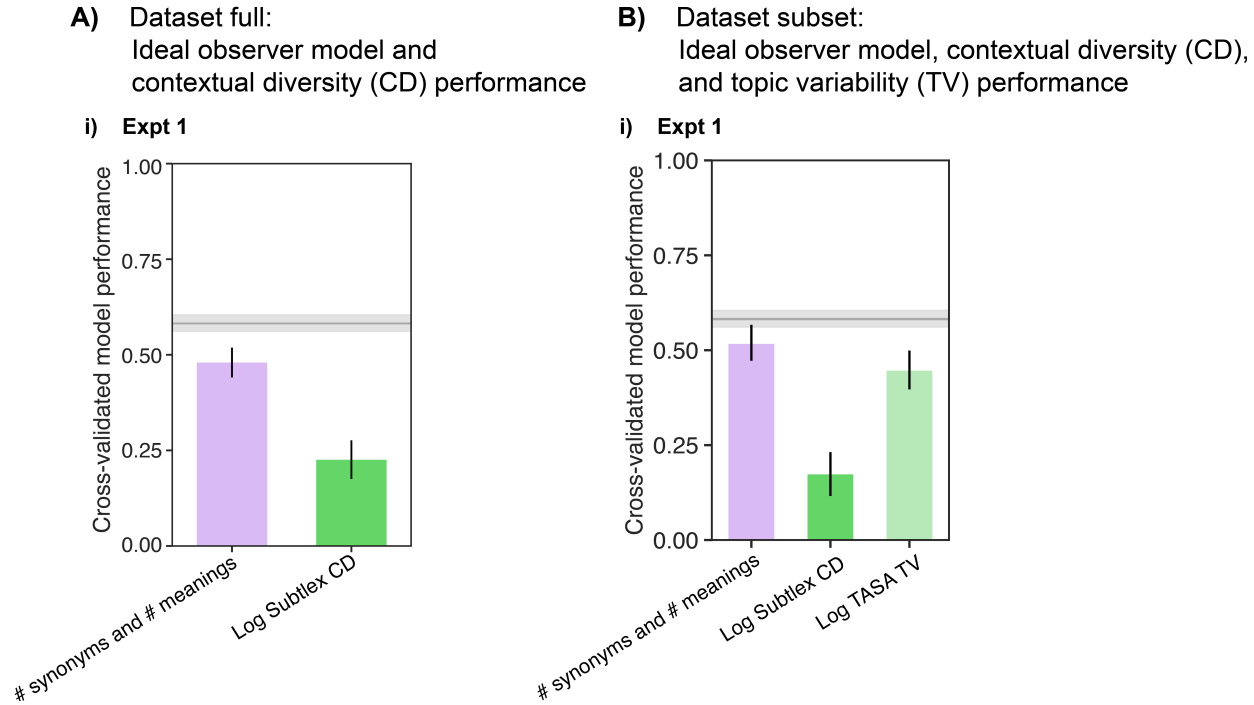


SI Figure 4. Pearson correlation among all predictors for Experiment 1 (n=13 predictors) and Experiment 2 (n=8 predictors).



SI Figure 5. Performance of the ideal observer model on data subsets that were obtained by partitioning the words according to their word frequency (either Google n-gram frequency **(A)** or Subtlex frequency **(B)**).
For Experiment 1, the data (n=2,109 words) was partitioned into three frequency bins with 703 words each (352 words in the train set and 351 words in the test set). For Experiment 2, the data (n=2,165 words) was partitioned into three frequency bins with 722 (low frequency bin), 721 (mid frequency bin), and 722 (high frequency bin) words each (361 words in the train set and 361/360 words in the test set).
Bars show model performance as quantified by the median Spearman rank correlation between predicted and observed recognition accuracy. Correlations were computed across 1,000 cross-validation folds using data from a set of non-overlapping participants and words for train/test splits. Error bars are 95% CI across cross-validation splits. The horizontal grey lines denote the theoretical maximum correlation, as defined by the split-half Spearman correlation for accuracy across participants. **(A)** Performance of the ideal observer model (number of synonyms, number of meanings) using Google n-gram frequency to partition the data into bins (see also SI Table 8).

For Experiment 1, the mean Google n-gram frequency for the low bin was -14.87 (SD: 0.93), for the mid bin -12.9 (SD: 0.45), and for the high bin -10.82 (SD: 1.11). For Experiment 2, the mean Google n-gram frequency for the low bin was -17.33 (SD: 1.26), for the mid bin -14.66 (SD: 0.69), and for the high bin -11.83 (SD: 1.25). **(A)** Performance of the ideal observer model using Log Subtlex frequency to partition the data into bins (see also SI Table 9). For Experiment 1, the mean Log Subtlex frequency for the low bin was 1.38 (SD: 0.31), for the mid bin 2.02 (SD: 0.13), and for the high bin 2.84 (SD: 0.56).



SI Figure 6. Comparison of the ideal observer model and contextual diversity and topic variability predictors using data from Experiment 1. We were able to estimate norms for Subtlex contextual diversity (CD) as the unique number of movie transcripts in which a word appears) (Brysbaert and New 2009) for all critical words (n=2,109 words); topic variability (TV) norms were extracted directly from Griffiths et al. (2007) using the model fitted on 1,700 topics, with help from Mark Steyvers, and were available for a subset of words (n=1,366 words). Bars show model performance as quantified by the median Spearman rank correlation between predicted and observed recognition accuracy. Correlations were computed across 1,000 cross-validation folds using data from a set of non-overlapping participants and words for train/test splits. Error bars are 95% CI across cross-validation splits. The horizontal grey lines denote the theoretical maximum correlation, as defined by the split-half Spearman correlation for accuracy across participants. **(A)** Comparison of the ideal observer model (number of synonyms, number of meanings) to CD as a predictor of memory performance (see also SI Table 10). This analysis was performed on the full set of materials (n=2,109 words; 1,055 words in the train set and 1,054 words in the test set). **(B)** Comparison of the ideal observer model (number of synonyms, number of meanings) to CD and TV as predictors of memory performance (see also SI Table

11). This analysis performed on a subset of the materials for which we were able to obtain TV norms (n=1,366 words; 683 words in the train set and 683 words in the test set).

SI Tables

Norm	Pearson R	p-value
# meanings (human)	0.85	1.13e-14
# synonyms (human)	0.97	7.23e-31
Concreteness	0.96	5.05e-27
Imageability	0.94	4.32e-23
Familiarity	0.86	3.27e-15
Valence	0.98	9.76e-33
Arousal	0.87	1.21e-15

SI Table 1. Correlation of human-derived norms for overlapping words (n=48) between Experiment 1 and Experiment 2.

Experiment	# synonyms (Wordnet)	# meanings (Wordnet)	# synonyms and # meanings (Wordnet)
Expt 1	0.33 [95% CI 0.29, 0.38]	0.35 [95% CI 0.30, 0.39]	0.35 [95% CI 0.30, 0.39]

SI Table 2. Ideal observer model performance (corpus-based norms). For Experiment 1, we quantified the number of synonyms and meanings using a corpus (WordNet (Miller 1995)). Row: experiment (1), columns: predictors (number of synonyms, number of meanings, both; based on corpus norms). Values are the median Spearman correlation between memorability (recognition accuracy) and model predictions across 1,000 cross-validation splits using data from a set of non-overlapping participants and words for train/test splits.

Expt	# synonyms and # meanings	Log orthographic neighborhood	Log phonological neighborhood
Expt 1	0.47 [95% CI 0.43, 0.52]	0.08 [95% CI 0.02, 0.13]	0.07 [95% CI -0.09, 0.12]
Expt 2	0.61 [95% CI 0.57, 0.65]	-0.02 [95% CI -0.06, 0.03]	-0.01 [95% CI -0.06, 0.04]

SI Table 3. Ideal observer model (human-based norms), orthographic and phonological neighborhood performance (derived from the CLEARPOND database; Marian et al. 2012). For Experiment 1, we were able to estimate the orthographic and phonological norms for n=1,873 words of the target words. For Experiment 2, we were able to estimate the orthographic and phonological norms for n=1,720 words of the target words.
Rows: experiment (1 or 2), columns: predictors (number of synonyms and number of meanings, log orthographic neighborhood, and log phonological neighborhood). Values are the median Spearman correlation between memorability (recognition accuracy) and model predictions across 1,000 cross-validation splits using data from a set of non-overlapping participants and words for train/test splits using the available subset of words.

Expt	Model	Model with additional predictor	F-statistic	Pr(>F)	SSR difference
Expt 1	Accuracy ~ Num_Synonyms	Accuracy ~ Num_Synonyms + Num_Meanings	20.37	6.74e-06	5.13e-02
Expt 2	Accuracy ~ Num_Synonyms	Accuracy ~ Num_Synonyms + Num_Meanings	11.88	5.78e-04	3.04e-02
Expt 1	Accuracy ~ Num_Meanings	Accuracy ~ Num_Synonyms + Num_Meanings	651.24	2.16e-125	1.64e+00
Expt 2	Accuracy ~ Num_Meanings	Accuracy ~ Num_Synonyms + Num_Meanings	1073.26	1.74e-191	2.75e+00

SI Table 4. Statistical comparison of individual predictors (number of synonyms and number of meanings) of the ideal observer model. ANOVA comparisons of a model ("Model" column) vs. a model with an additional model term, i.e., additional predictor ("Model with additional predictor" column). F statistic and Pr(>F) denote the values for significance of adding model terms. Difference is sum of squares of residuals (SSR) denotes the difference in SSR from the model without additional model terms.

Expt	Baseline	Concreteness	Imageability	Familiarity	Valence	Arousal	Google n-gram frequency	Log Subtlex frequency	Log Subtlex CD	GloVe distinctiveness
Expt 1	0.00%	5.72%	9.30%	2.86%	0.08%	1.36%	10.24%	1.89%	2.91%	0.96%
Expt 2	0.00%	2.13%	4.07%	0.70%	-0.17%	2.14%	6.07%			

SI Table 5. Percent change in cross-validated model performance by adding each individual predictor (columns) to the ideal observer baseline model. Rows are respectively Experiment 1 and Experiment 2.

Expt	Baseline	Concreteness	Imageability	Familiarity	Valence	Arousal	Google n-gram frequency	Log Subtlex frequency	Log Subtlex CD	GloVe distinctiveness
Expt 1	0.48 [95% CI 0.44, 0.52]	0.51 [95% CI 0.47, 0.55]	0.53 [95% CI 0.49, 0.56]	0.50 [95% CI 0.45, 0.53]	0.48 [95% CI 0.44, 0.52]	0.49 [95% CI 0.45, 0.52]	0.53 [95% CI 0.49, 0.57]	0.49 [95% CI 0.45, 0.53]	0.50 [95% CI 0.46, 0.53]	0.49 [95% CI 0.44, 0.52]
Expt 2	0.59 [95% CI 0.55, 0.63]	0.60 [95% CI 0.57, 0.64]	0.61 [95% CI 0.58, 0.65]	0.59 [95% CI 0.55, 0.63]	0.59 [95% CI 0.55, 0.62]	0.60 [95% CI 0.56, 0.64]	0.62 [95% CI 0.59, 0.66]			

SI Table 6. Cross-validated model performance by adding each individual predictor (columns) to the ideal observer baseline model (same model as reported in Table 1 in the main text). Rows are respectively Experiment 1 and Experiment 2.

Expt	Model	Model with additional predictor	F-statistic	Pr(>F)	SSR difference
Expt 1	Accuracy ~ Num_Synonyms + Num_Meanings	Accuracy ~ Num_Synonyms + Num_Meanings + Concreteness	128.21	6.80e-29	3.05e-01
Expt 2	Accuracy ~ Num_Synonyms + Num_Meanings	Accuracy ~ Num_Synonyms + Num_Meanings + Concreteness	46.07	1.47e-11	1.16e-01
Expt 1	Accuracy ~ Num_Synonyms + Num_Meanings	Accuracy ~ Num_Synonyms + Num_Meanings + Imageability	210.18	1.81e-45	4.82e-01
Expt 2	Accuracy ~ Num_Synonyms + Num_Meanings	Accuracy ~ Num_Synonyms + Num_Meanings + Imageability	101.58	2.23e-23	2.49e-01
Expt 1	Accuracy ~ Num_Synonyms + Num_Meanings	Accuracy ~ Num_Synonyms + Num_Meanings + Familiarity	38.91	5.35e-10	9.63e-02
Expt 2	Accuracy ~ Num_Synonyms + Num_Meanings	Accuracy ~ Num_Synonyms + Num_Meanings + Familiarity	2.24	1.35e-01	5.73e-03
Expt 1	Accuracy ~ Num_Synonyms + Num_Meanings	Accuracy ~ Num_Synonyms + Num_Meanings + Valence	5.96	1.47e-02	1.50e-02
Expt 2	Accuracy ~ Num_Synonyms + Num_Meanings	Accuracy ~ Num_Synonyms + Num_Meanings + Valence	2.40	1.21e-01	6.14e-03
Expt 1	Accuracy ~ Num_Synonyms + Num_Meanings	Accuracy ~ Num_Synonyms + Num_Meanings + Arousal	35.91	2.43e-09	8.90e-02
Expt 2	Accuracy ~ Num_Synonyms + Num_Meanings	Accuracy ~ Num_Synonyms + Num_Meanings + Arousal	26.01	3.69e-07	6.59e-02
Expt 1	Accuracy ~ Num_Synonyms + Num_Meanings	Accuracy ~ Num_Synonyms + Num_Meanings + Google n-gram frequency	224.81	2.32e-48	5.12e-01
Expt 2	Accuracy ~ Num_Synonyms + Num_Meanings	Accuracy ~ Num_Synonyms + Num_Meanings + Google n-gram frequency	142.06	9.12e-32	3.42e-01
Expt 1	Accuracy ~ Num_Synonyms + Num_Meanings	Accuracy ~ Num_Synonyms + Num_Meanings + Log Subtlex frequency	27.63	1.62e-07	6.88e-02
Expt 1	Accuracy ~ Num_Synonyms + Num_Meanings	Accuracy ~ Num_Synonyms + Num_Meanings + Log Subtlex CD	51.99	7.75e-13	1.28e-01
Expt 1	Accuracy ~ Num_Synonyms + Num_Meanings	Accuracy ~ Num_Synonyms + Num_Meanings + GloVe distinctiveness	17.64	2.78e-05	4.41e-02

SI Table 7. Statistical comparison of individual predictors (number of synonyms and number of meanings) of the ideal observer model. ANOVA comparisons of a model ("Model" column) vs. a model with an additional model term, i.e., additional predictor ("Model with additional predictor" column). F statistic and Pr(>F) denote the values for significance of adding model

terms. Difference is sum of squares of residuals (SSR) denotes the difference in SSR from the model without additional model terms.

Expt	Low frequency bin	Mid frequency bin	High frequency bin
Expt 1	0.37 [95% CI 0.30, 0.44]	0.47 [95% CI 0.39, 0.54]	0.54 [95% CI 0.47, 0.60]
Expt 2	0.47 [95% CI 0.40, 0.54]	0.57 [95% CI 0.50, 0.62]	0.61 [95% CI 0.56, 0.65]

SI Table 8. Ideal observer model (human-based norms) performance on data subsets that were obtained by partitioning the words according to their Google n-gram word frequency.
For Experiment 1, the data (n=2,109 words) was partitioned into three frequency bins with 703 words each. For Experiment 2, the data (n=2,165 words) was partitioned into three frequency bins with 722 (low frequency bin), 721 (mid frequency bin), and 722 (high frequency bin) words each.
Rows: experiment (1 or 2), columns: predictors (number of synonyms and number of meanings). Values are the median Spearman correlation between memorability (recognition accuracy) and model predictions across 1,000 cross-validation splits using data from a set of non-overlapping participants and words for train/test splits using the words in the frequency bin of interest (low, mid, or high).

Expt	Low frequency bin	Mid frequency bin	High frequency bin
Expt 1	0.46 [95% CI 0.39, 0.53]	0.45 [95% CI 0.37, 0.52]	0.51 [95% CI 0.44, 0.58]

SI Table 9. Ideal observer model (human-based norms) performance on data subsets that were obtained by partitioning the words according to their Log Subtlex word frequency.
For Experiment 1, the data (n=2,109 words) was partitioned into three frequency bins with 703 words each.
Rows: experiment (1), columns: predictors (number of synonyms and number of meanings). Values are the median Spearman correlation between memorability (recognition accuracy) and model predictions across 1,000 cross-validation splits using data from a set of non-overlapping participants and words for train/test splits using the words in the frequency bin of interest (low, mid, or high).

Expt	Log Subtlex CD
Expt 1	0.23 [95% CI 0.18, 0.28]

SI Table 10. Performance of the Log Subtlex Contextual Diversity (CD) predictor on the words from Experiment 1 (n=2,109 words).
Rows: experiment (1), columns: predictor (Log Subtlex CD). Values are the median Spearman correlation between memorability (recognition accuracy) and model predictions across 1,000 cross-validation splits using data from a set of non-overlapping participants and words for train/test splits.

Expt	# synonyms and # meanings	Log Subtlex CD	Log TASA TV
Expt 1	0.52 [95% CI 0.47, 0.57]	0.18 [95% CI 0.12, 0.23]	0.45 [95% CI 0.40, 0.50]

SI Table 11. Ideal observer model (human-based norms), Log Subtlex Contextual Diversity (CD), and Log TASA Topic Variability (TV) model performance. For Experiment 1, we were able to estimate the Log TASA TV norms for n=1,366 words of the target words.
Rows: experiment (1), columns: predictor (number of synonyms and number of meanings, Log Subtlex CD, Log TASA TV). Values are the median Spearman correlation between memorability (recognition accuracy) and model predictions across 1,000 cross-validation splits using data from a set of non-overlapping participants and words for train/test splits.

Predictor	Expt1 Predictor inclusion	Expt1 Predictor inclusion as first predictor	Expt2 Predictor inclusion	Expt2 Predictor inclusion as first predictor
# synonyms (human)	1000	1000	1000	1000
# meanings (human)	138	0	9	0
# synonyms (Wordnet)	79	0		
# meanings (Wordnet)	952	0		
Concreteness	11	0	23	0
Imageability	994	0	973	0
Familiarity	977	0	63	0
Valence	3	0	80	0
Arousal	98	0	304	0
Google n-gram frequency	985	0	1000	0
Log Subtlex frequency	869	0		
Log Subtlex CD	247	0		
GloVe distinctiveness	9	0		

SI Table 12. Quantification of how many times each predictor was included in the best possible model of memorability using a forward-backward feature selection approach (across 1,000 cross-validation splits).

References

Amrhein, Paul C., Mark McDaniel, and Paula Waddill. 2002. "Revisiting the Picture-Superiority Effect in Symbolic Comparisons: Do Pictures Provide Privileged Access?" *Journal of Experimental Psychology: Learning, Memory, and Cognition* 28(5):843–57. doi: 10.1037/0278-7393.28.5.843.

Anderson, J. R., and R. Milson. 1989. "Human Memory: An Adaptive Perspective." *Psychological Review* 96(4):703.

Anderson, J. R., and L. J. Schooler. 1991. "Reflections of the Environment in Memory." *Psychological Science* 2(6):396.

Anderson, John Robert. 1974. "Retrieval of Propositional Information from Long-Term Memory." *Cognitive Psychology* 6(4):451–74. doi: 10.1016/0010-0285(74)90021-8.

Annis, Jeffrey, Joshua Guy Lenes, Holly A. Westfall, Amy H. Criss, and Kenneth J. Malmberg. 2015. "The List-Length Effect Does Not Discriminate between Models of Recognition Memory." *Journal of Memory and Language* 85:27–41. doi: 10.1016/j.jml.2015.06.001.

Aue, William R., Jessica M. Fontaine, and Amy H. Criss. 2018. "Examining the Role of Context Variability in Memory for Items and Associations." *Memory & Cognition* 46(6):940–54. doi: 10.3758/s13421-018-0813-9.

Bainbridge, Wilma A. 2022. "Memorability: Reconceptualizing Memory as a Visual Attribute." in *Visual Memory*. Routledge.

Bainbridge, Wilma A., Phillip Isola, and Aude Oliva. 2013. "The Intrinsic Memorability of Face Photographs." *Journal of Experimental Psychology: General* 142(4):1323.

Begg, Ian, and Wayne A. Wickelgren. 1974. "Retention Functions for Syntactic and Lexical vs Semantic Information in Sentence Recognition Memory." *Memory & Cognition* 2(2):353–59.

Bird, Steven, and Edward Loper. 2004. "NLTK: The Natural Language Toolkit." Pp. 214–17 in *Proceedings of the ACL Interactive Poster and Demonstration Sessions*. Barcelona, Spain: Association for Computational Linguistics.

Boldini, Angela, Riccardo Russo, Sahiba Punia, and S. E. Avons. 2007. "Reversing the Picture Superiority Effect: A Speed—Accuracy Trade-off Study of Recognition Memory." *Memory & Cognition* 35(1):113–23. doi: 10.3758/BF03195948.

Borkin, Michelle A., Azalea A. Vo, Zoya Bylinskii, Phillip Isola, Shashank Sunkavalli, Aude Oliva, and Hanspeter Pfister. 2013. "What Makes a Visualization Memorable?" *IEEE Transactions on Visualization and Computer Graphics* 19(12):2306–15. doi: 10.1109/TVCG.2013.234.

Bransford, John D., and Jeffery J. Franks. 1971. “The Abstraction of Linguistic Ideas.” *Cognitive Psychology* 2(4):331–50.

Brown, John, and Vivian Lewis. 1981. “The Relationship between Item Retrievability and Item Discriminability and Their Interactions with Item Frequency and Personal Significance.” *The American Journal of Psychology* 94(2):247. doi: 10.2307/1422744.

Brysbaert, Marc, and Boris New. 2009. “Moving beyond Kučera and Francis: A Critical Evaluation of Current Word Frequency Norms and the Introduction of a New and Improved Word Frequency Measure for American English.” *Behavior Research Methods* 41(4):977–90. doi: 10.3758/BRM.41.4.977.

Brysbaert, Marc, Boris New, and Emmanuel Keuleers. 2012. “Adding Part-of-Speech Information to the SUBTLEX-US Word Frequencies.” *Behavior Research Methods* 44(4):991–97. doi: 10.3758/s13428-012-0190-4.

Brysbaert, Marc, Amy Beth Warriner, and Victor Kuperman. 2014. “Concreteness Ratings for 40 Thousand Generally Known English Word Lemmas.” *Behavior Research Methods* 46(3):904–11. doi: 10.3758/s13428-013-0403-5.

Bylinskii, Zoya, Phillip Isola, Constance Bainbridge, Antonio Torralba, and Aude Oliva. 2015. “Intrinsic and Extrinsic Effects on Image Memorability.” *Vision Research* 116:165–78. doi: 10.1016/j.visres.2015.03.005.

Chater, N., and M. Oaksford. 1999. “Ten Years of the Rational Analysis of Cognition.” *Trends in Cognitive Sciences* 3(2):57–65.

Criss, Amy H., and Richard M. Shiffrin. 2004. “Context Noise and Item Noise Jointly Determine Recognition Memory: A Comment on Dennis and Humphreys (2001).” *Psychological Review* 111(3):800–807. doi: 10.1037/0033-295X.111.3.800.

Danion, Jean-Marie, Françoise Kauffmann-Muller, Danielle Grangé, Marie-Agathe Zimmermann, and Philippe Greth. 1995. “Affective Valence of Words, Explicit and Implicit Memory in Clinical Depression.” *Journal of Affective Disorders* 34(3):227–34. doi: 10.1016/0165-0327(95)00021-E.

Dennis, Simon, and Michael S. Humphreys. 2001. “A Context Noise Model of Episodic Word Recognition.” *Psychological Review* 108(2):452–78. doi: 10.1037/0033-295X.108.2.452.

Fenk-Oczlon, Gertraud, and August Fenk. 2010. “FREQUENCY EFFECTS ON THE EMERGENCE OF POLYSEMY AND HOMOPHONY.” *International Journal Information Technologies and Knowledge* 4(2):8.

Forster, Kenneth I. 1970. “Visual Perception of Rapidly Presented Word Sequences of Varying Complexity.” *Perception & Psychophysics* 8(4):215–21. doi: 10.3758/BF03210208.

Franks, Jeffery J., and John D. Bransford. 1972. “The Acquisition of Abstract Ideas.” *Journal of Verbal Learning and Verbal Behavior* 11(3):311–15.

Gershman, Samuel J. 2021. “The Rational Analysis of Memory.” in *Oxford handbook of human memory.* Oxford University Press Oxford, UK.

Goetschalckx, Lore, Jade Moors, and Johan Wagemans. 2019. “Incidental Image Memorability.” *Memory (Hove, England)* 27(9):1273–82. doi: 10.1080/09658211.2019.1652328.

Gorman, Aloysia M. 1961. “Recognition Memory for Nouns as a Function of Abstractness and Frequency.” *Journal of Experimental Psychology* 61(1):23–29. doi: 10.1037/h0040561.

Griffiths, Thomas L., Mark Steyvers, and Joshua B. Tenenbaum. 2007. “Topics in Semantic Representation.” *Psychological Review* 114(2):211–44. doi: 10.1037/0033-295X.114.2.211.

Hino, Yasushi, Stephen J. Lupker, and Penny M. Pexman. 2002. “Ambiguity and Synonymy Effects in Lexical Decision, Naming, and Semantic Categorization Tasks: Interactions between Orthography, Phonology, and Semantics.” *Journal of Experimental Psychology. Learning, Memory, and Cognition* 28(4):686–713.

Huebner, Gesche M., and Karl R. Gegenfurtner. 2012. “Conceptual and Visual Features Contribute to Visual Memory for Natural Images.” *PLOS ONE* 7(6):e37575. doi: 10.1371/journal.pone.0037575.

Isola, Phillip, Devi Parikh, Antonio Torralba, and Aude Oliva. 2011. “Understanding the Intrinsic Memorability of Images.” Pp. 2429–37 in *Proceedings of the 24th International Conference on Neural Information Processing Systems*, *NIPS’11*. Red Hook, NY, USA: Curran Associates Inc.

Isola, Phillip, Jianxiong Xiao, Devi Parikh, Antonio Torralba, and Aude Oliva. 2014. “What Makes a Photograph Memorable?” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36(7):1469–82. doi: 10.1109/TPAMI.2013.200.

Isola, Phillip, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2011. “What Makes an Image Memorable?” Pp. 145–52 in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE.

Jacoby, Larry L., and Mark Dallas. 1981. “On the Relationship Between Autobiographical Memory and Perceptual Learning.” *Journal of Experimental Psychology* 110(3):306–40.

Jones, Michael N., Melody Dye, and Brendan T. Johns. 2017. “Context as an Organizing Principle of the Lexicon.” Pp. 239–83 in *Psychology of Learning and Motivation*. Vol. 67. Elsevier.

Katz, Stuart, and Paul Gruenewald. 1974. “The Abstraction of Linguistic Ideas in ‘Meaningless’ Sentences.” *Memory & Cognition* 2(4):737–41. doi: 10.3758/BF03198149.

Kensinger, Elizabeth A., and Suzanne Corkin. 2003. “Memory Enhancement for Emotional Words: Are Emotional Words More Vividly Remembered than Neutral Words?” *Memory & Cognition* 31(8):1169–80.

Khanna, Maya M., and Michael J. Cortese. 2021. “How Well Imageability, Concreteness, Perceptual Strength, and Action Strength Predict Recognition Memory, Lexical Decision, and Reading Aloud Performance.” *Memory (Hove, England)* 29(5):622–36. doi: 10.1080/09658211.2021.1924789.

Kinsbourne, Marcel, and James George. 1974. “The Mechanism of the Word-Frequency Effect on Recognition Memory.” *Journal of Verbal Learning and Verbal Behavior* 13(1):63–69.

Klaver, Peter, Jürgen Fell, Thomas Dietl, Simone Schür, Carlo Schaller, Christian E. Elger, and Guillén Fernández. 2005. “Word Imageability Affects the Hippocampus in Recognition Memory.” *Hippocampus* 15(6):704–12.

Konkle, Talia, Timothy F. Brady, George A. Alvarez, and Aude Oliva. 2010. “Conceptual Distinctiveness Supports Detailed Visual Long-Term Memory for Real-World Objects.” *Journal of Experimental Psychology. General* 139(3):558–78. doi: 10.1037/a0019165.

Light, Leah L., and Linda Carter-Sobell. 1970. “Effects of Changed Semantic Context on Recognition Memory.” *Journal of Verbal Learning and Verbal Behavior* 9(1):1–11.

Lin, Qi, Sami R. Yousif, Brian Scholl, and Marvin M. Chun. 2019. “Image Memorability Is Driven by Visual and Conceptual Distinctivenes.” *Journal of Vision* 19(10):290c. doi: 10.1167/19.10.290c.

Lohnas, Lynn J., and Michael J. Kahana. 2013. “Parametric Effects of Word Frequency in Memory for Mixed Frequency Lists.” *Journal of Experimental Psychology: Learning, Memory, and Cognition* 39(6):1943–46.

Malmberg, Kenneth J., Mark Steyvers, Joseph D. Stephens, and Richard M. Shiffrin. 2002. “Feature Frequency Effects in Recognition Memory.” *Memory & Cognition* 30(4):607–13. doi: 10.3758/BF03194962.

Marian, Viorica, James Bartolotti, Sarah Chabal, and Anthony Shook. 2012. “CLEARPOND: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities.” *PLOS ONE* 7(8):e43230. doi: 10.1371/journal.pone.0043230.

McClelland, James L., and Mark Chappell. 1998. “Familiarity Breeds Differentiation: A Subjective-Likelihood Approach to the Effects of Experience in Recognition Memory.” *Psychological Review* 105(4):724.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, and others. 2011. “Quantitative Analysis of Culture Using Millions of Digitized Books.” *Science* 331(6014):176–82.

Miller, George A. 1995. “WordNet: A Lexical Database for English.” *Communications of the ACM* 38(11):39–41. doi: 10.1145/219717.219748.

Mohammad, Saif. 2018. “Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words.” Pp. 174–84 in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics.

Monaco, Joseph D., LF Abbott, and Michael J. Kahana. 2007. “Lexico-Semantic Structure and the Word-Frequency Effect in Recognition Memory.” *Learning & Memory* 14(3):204–13.

Nelson, Thomas O., Jacqueline Metzler, and David A. Reed. 1974. “Role of Details in the Long-Term Recognition of Pictures and Verbal Descriptions.” *Journal of Experimental Psychology* 102(1):184–86. doi: 10.1037/h0035700.

Paivio, Allan. 1969. “Mental Imagery in Associative Learning and Memory.” *Psychological Review* 76(3):241.

Paivio, Allan. 1971. *Imagery and Verbal Processes*. the University of Michigan: Holt, Rinehart and Winston.

Paivio, Allan, T. B. Rogers, and Padric C. Smythe. 1968. “Why Are Pictures Easier to Recall than Words?” *Psychonomic Science* 11(4):137–38. doi: 10.3758/BF03331011.

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. “Glove: Global Vectors for Word Representation.” Pp. 1532–43 in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics.

Pereira, Francisco, Samuel Gershman, Samuel Ritter, and Matthew Botvinick. 2016. “A Comparative Evaluation of Off-the-Shelf Distributed Semantic Representations for Modelling Behavioural Data.” *Cognitive Neuropsychology* 33(3–4):175–90. doi: 10.1080/02643294.2016.1176907.

Phelps, Elizabeth A., Kevin S. LaBar, and Dennis D. Spencer. 1997. “Memory for Emotional Words Following Unilateral Temporal Lobectomy.” *Brain and Cognition* 35(1):85–109. doi: 10.1006/brcg.1997.0929.

Piantadosi, S. T., H. Tily, and E. Gibson. 2012. “The Communicative Function of Ambiguity in Language.” *Cognition* 122(3):280–91. doi: 10.1016/j.cognition.2011.10.004.

Potter, Mary C. 2012. “Conceptual Short Term Memory in Perception and Thought.” *Frontiers in Psychology* 3. doi: 10.3389/fpsyg.2012.00113.

Potter, Mary C., Judith F. Kroll, and Carol Harris. 1980. “Comprehension and Memory in Rapid Sequential Reading.” Pp. 395–418 in *Attention and Performance VIII*. Hillsdale, NJ: Erlbaum.

Potter, Mary C., and Linda Lombardi. 1990. “Regeneration in the Short-Term Recall of Sentences.” *Journal of Memory and Language* 29(6):633–54. doi: 10.1016/0749-596X(90)90042-X.

Potter, Mary C., Virginia V. Valian, and Barbara A. Faulconer. 1977. “Representation of a Sentence and Its Pragmatic Implications: Verbal, Imagistic, or Abstract?” *Journal of Verbal Learning and Verbal Behavior* 16(1):1–12. doi: 10.1016/S0022-5371(77)80002-9.

Rubin, David C., and Michael Friendly. 1986. “Predicting Which Words Get Recalled: Measures of Free Recall, Availability, Goodness, Emotionality, and Pronunciability for 925 Nouns.” *Memory & Cognition* 14(1):79–94.

Schulman, Arthur I. 1967. “Word Length and Rarity in Recognition Memory.” *Psychonomic Science* 9(4):211–12. doi: 10.3758/BF03330834.

Seifert, Lauren S. 1997. “Activating Representations in Permanent Memory: Different Benefits for Pictures and Words.” *Journal of Experimental Psychology: Learning, Memory, and Cognition* 23(5):1106–21. doi: 10.1037/0278-7393.23.5.1106.

Shepard, Roger N. 1967. “Recognition Memory for Words, Sentences, and Pictures.” *Journal of Verbal Learning and Verbal Behavior* 6(1):156–63. doi: 10.1016/S0022-5371(67)80067-7.

Shiffrin, Richard M., and Mark Steyvers. 1997. “A Model for Recognition Memory: REM—Retrieving Effectively from Memory.” *Psychonomic Bulletin & Review* 4(2):145–66. doi: 10.3758/BF03209391.

Sigman, M., and G. A. Cecchi. 2002. “Global Organization of the Wordnet Lexicon.” *Proceedings of the National Academy of Sciences* 99(3):1742–47. doi: 10.1073/pnas.022341799.

Smith, Marilyn C., and Lochlan E. Magee. 1980. “Tracing the Time Course of Picture–Word Processing.” *Journal of Experimental Psychology: General* 109(4):373.

Standing, Lionel. 1973. “Learning 10000 Pictures.” *Quarterly Journal of Experimental Psychology* 25(2):207–22. doi: 10.1080/14640747308400340.

Stenberg, Georg. 2006. “Conceptual and Perceptual Factors in the Picture Superiority Effect.” *European Journal of Cognitive Psychology* 18(6):813–47. doi: 10.1080/09541440500412361.

Steyvers, Mark, and Kenneth J. Malmberg. 2003. “The Effect of Normative Context Variability on Recognition Memory.” *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29(5):760–66. doi: 10.1037/0278-7393.29.5.760.

Tenenbaum, J. B., C. Kemp, T. L. Griffiths, and N. D. Goodman. 2011. “How to Grow a Mind: Statistics, Structure, and Abstraction.” *Science* 331(6022):1279–85.

Tweedy, James R., Robert H. Lapinski, and Roger W. Schvaneveldt. 1977. “Semantic-Context Effects on Word Recognition: Influence of Varying the Proportion of Items Presented in an Appropriate Context.” *Memory & Cognition* 5(1):84–89. doi: 10.3758/BF03209197.

Walker, Ian, and Charles Hulme. 1999. “Concrete Words Are Easier to Recall than Abstract Words: Evidence for a Semantic Contribution to Short-Term Serial Recall.” *Journal of Experimental Psychology: Learning, Memory, and Cognition* 25(5):1256.