Journal of Experimental Psychology: General

Intrinsically Memorable Words Have Unique Associations With Their Meanings

Greta Tuckute, Kyle Mahowald, Phillip Isola, Aude Oliva, Edward Gibson, and Evelina Fedorenko

Online First Publication, March 31, 2025. https://dx.doi.org/10.1037/xge0001742

AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

ISSN: 0096-3445

https://doi.org/10.1037/xge0001742

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

[2] McGovern Institute for Brain Research, Massachusetts Institute of Technology

[3] Department of Linguistics, The University of Texas at Austin

[4] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology

[5] Program in Speech and Hearing Bioscience and Technology, Harvard University

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 uniquely pick out a particular meaning, which means they should have few or no synonyms, and they should be unambiguous. 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, and some words are consistently remembered better than others. 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 number of synonyms outperforms other predictors (such as imageability, frequency, or contextual diversity) of memorability that have been proposed in the past.

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 synonyms ("HAPPY") or words with many meanings ("LIGHT"). 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.

Keywords: memorability, language, semantic memory, word recognition, Bayesian modeling

Supplemental materials: https://doi.org/10.1037/xge0001742.supp

Agnieszka Konopka served as action editor.

Greta Tuckute https://orcid.org/0000-0002-5572-5469

Greta Tuckute and Kyle Mahowald are co-first authors. For comments on this work, the authors 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, the authors thank Barbara Hidalgo-Sotelo, Henrison Hsieh, and members of the Gibson lab. Greta Tuckute was supported by the Amazon Fellowship from the Science Hub (administered by the Massachusetts Institute of Technology (MIT) Schwarzman College of Computing), the International Doctoral Fellowship from the American Association of University Women, the K. Lisa Yang Integrative Computational Neuroscience Center Graduate Fellowship, and the MIT McGovern Institute. Aude Oliva was supported by Multidisciplinary University Research Initiative (MURI) award by the Army Research Office (W911NF-23-1-0277). Evelina Fedorenko was supported by the National Institutes of Health (Awards DC016607 and DC016950) from the National Institute on Deafness and Other Communication Disorders and the National Institute of Neurological Disorders and Stroke (award NS121471) 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.

Greta Tuckute played a lead role in visualization, a supporting role in methodology, and an equal role in formal analysis, investigation, software, writing–original draft, and writing–review and editing. Kyle Mahowald played a supporting role in visualization and an equal role in data curation, formal analysis, investigation, methodology, writing–original draft, and writing–review and editing. Phillip Isola played an equal role in conceptualization, data curation, investigation, methodology, and writing–original draft. Aude Oliva played a supporting role in supervision and an equal role in conceptualization and writing–review and editing. Edward Gibson played an equal role in conceptualization, data curation, investigation, supervision, and writing–review and editing. Evelina Fedorenko played a lead role in supervision and an equal role in conceptualization, data curation, and writing–original draft.

Correspondence concerning this article should be addressed to Greta Tuckute, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 43 Vassar Street, Cambridge, MA 02139, United States, or Kyle Mahowald, Department of Linguistics, The University of Texas at Austin, 305 East 23rd Street, Austin, TX 78712, United States. Email: gretatu@mit.edu or mahowald@utexas.edu

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 article. The remaining words—"precise," "advantage," "manner," and "expression," in case you forgot—were among the most forgettable.

What makes "avalanche" stick in our memory? In particular, what makes a word like "avalanche" stand out in memory so that it is easier to recognize in a recognition-memory test? 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 (Brown & Lewis, 1981; Gorman, 1961; Kinsbourne & George, 1974; Lohnas & Kahana, 2013; Schulman, 1967); concrete and imageable words are both easier to recognize and easier to recall (Gorman, 1961; Khanna & Cortese, 2021; Klaver et al., 2005; Paivio, 1969; Rubin & Friendly, 1986; Walker & Hulme, 1999); and emotionally salient words also enjoy a memory boost in both recognition and recall (Danion et al., 1995; Kensinger & Corkin, 2003; Phelps et al., 1997; Rubin & Friendly, 1986).

Past work has also emphasized the importance of meaning (over the surface form) for the memorability of linguistic strings. In sentence recognition studies, sentence meanings are better retained than their surface-level (lexical and morphosyntactic) properties (Begg & Wickelgren, 1974; Bransford & Franks, 1971; Franks & Bransford, 1972; Katz & Gruenewald, 1974). Furthermore, deeper engagement with the word's meaning—as can be manipulated via tasks at the encoding stage—facilitates subsequent recognition (Jacoby & 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 & Carter-Sobell, 1970). Similarly, in sentence recall studies (using the classic rapid serial visual presentation 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, 2012; Potter et al., 1980; Potter & Lombardi, 1990). 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., Griffiths et al., 2007; Monaco et al., 2007; Steyvers & Malmberg, 2003). 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 & Milson, 1989; Anderson & Schooler, 1991; Chater & Oaksford, 1999; Gershman, 2024; Tenenbaum et al., 2011). Some past research on verbal memory has followed this tradition (Dennis & Humphreys, 2001; McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997; Steyvers & 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 (Annis et al., 2015; Criss & Shiffrin, 2004). "Item-noise models" (e.g., McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997) have emphasized intrinsic, context-independent properties of words, whereas "context-noise models" (e.g., Dennis & 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. For a word to do this effectively, it should ideally 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 in the Ideal Observer Model section.

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 with words associated with fewer topics being recognized more easily. 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 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.

In the present study, we measured memorability in a word recognition task, in which participants see a sequence of words (some of which are repeated) and respond when they recognize a word as previously seen in the experiment. We operationalize memorability as the recognition accuracy—the average proportion of correct responses over all trials that included that word (also see the Overall Design section).

In the Results section, we start by examining the relationship between word memorability and the number of synonyms and

number of meanings. In doing so, we do not intend to ignore the effects of the many other attested influences on lexical memory (e.g., Brown & Lewis, 1981; Paivio, 1971; Steyvers & Malmberg, 2003), and in subsequent analyses, we consider the impact of additional factors above and beyond our two critical predictors. To foreshadow our results, we find that the number of synonyms and the number of meanings matter, alongside many of the other proposed factors that we consider from the literature that show expected effects (e.g., frequency, concreteness, and imageability).

Ideal Observer Model

We propose a Bayesian model of the word recognition task in which a rational agent encounters a word (at the first presentation) and stores a meaning m selected by that word. The agent is then asked, at a later time, whether a word w_c has been previously encountered. At that time, the agent has access to the stored meaning m and the current stimulus w_c and must decide whether the original word that generated m is the same as the current word w_c. (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 quantitative predictions about how a rational agent would act. In particular, we make a simplifying assumption here by reducing the problem to one of only a previous word and a current word, without considering list or ordering effects.) 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 the random variable W takes on the value w_c given the stored meaning m. Applying Bayes' rule, this probability can be written as follows:

$$P(W = w_c|m) = \frac{P(m|W = w_c)P(W = w_c)}{\sum_{w_i} P(m,W = w_i)}. \quad (1)$$

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

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

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

2. $P(m \mid W = w_c)$ is low (i.e., the distribution over a word's meaning has high entropy). This is a case where w_c is ambiguous, that is, it 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 1B(i). 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 1B(ii). 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 reappears when $P(W = w_c)$ is low; conversely, a particular distractor may generate a false positive if it has a high $P(W = w_c)$. The term $P(W = w_c)$ expresses the a priori probability that the random variable W has the specific value w_c, which is equivalent to the probability of seeing a word repeated. In our experimental setup, this value is similar for all critical words in our task, given that any word that 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 $P(W = w_c)$ 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. $P(m \mid W = w_c)$ is high, and $P(m \mid W = w_i)$ for all other word(s) w_i is low. This is a case where w_c has no synonyms and a single meaning (i.e., it is uniquely associated with its meaning). This scenario is illustrated in Figure 1C. Because no competition arises either among the synonyms of the word w_c 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:

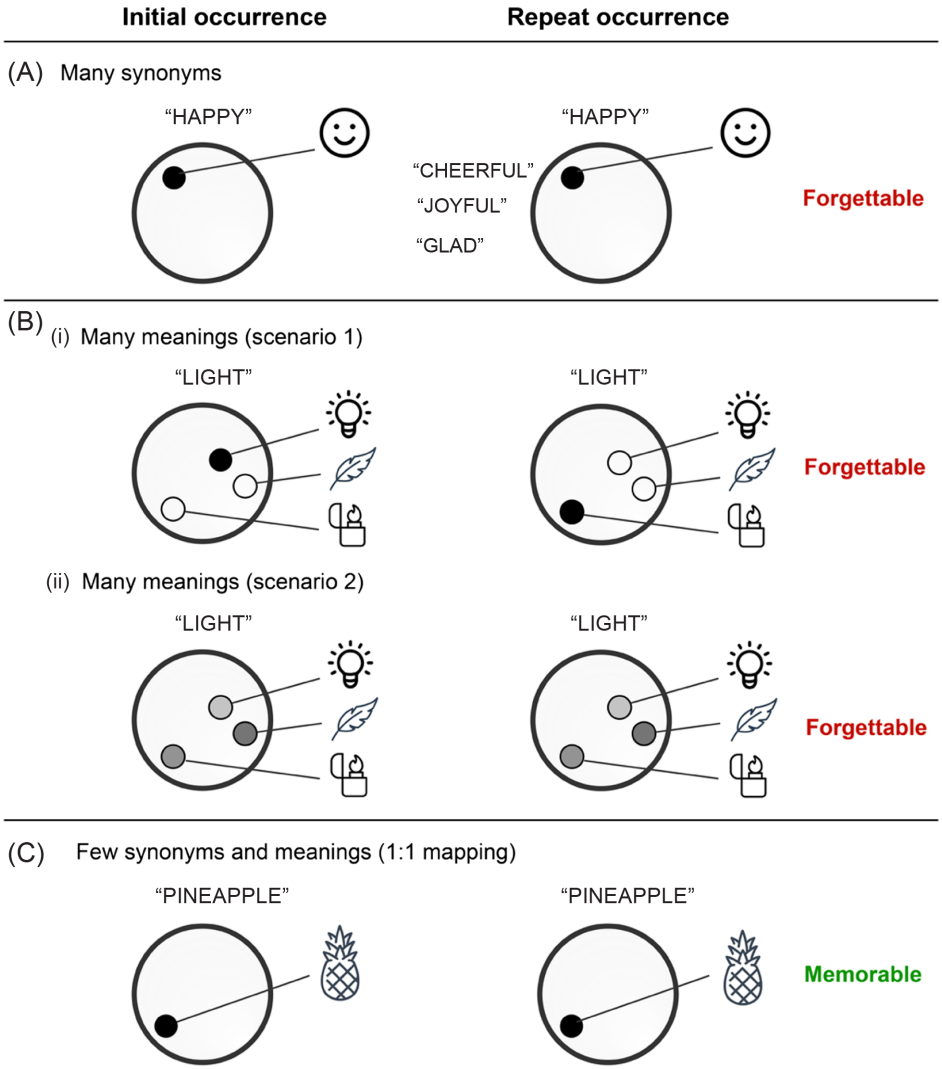
1. Words that have no/few synonyms should be more memorable than words with many synonyms.

2. Unambiguous words should be more memorable than words with two or more meanings.

We do not focus on specific quantitative predictions from the model but rather use the model to motivate these factors. Note that word frequency—shown to affect recognition memory in much past work (e.g., Brown & Lewis, 1981; Schulman, 1967)—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 & Fenk, 2010; Jones et al., 2017; Piantadosi et al., 2012; see also Monaco et al., 2007, for evidence that the frequency effect in recognition memory is related to the structure of the semantic space).

Figure 1

Schematic Illustration of Critical Predictions From the Ideal Observer Model



Note. 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. (A) A word like "HAPPY" with many synonyms (e.g., "CHEERFUL," "JOYFUL," "GLAD") is predicted to be forgettable as any of the synonyms could have generated the relevant meaning. (B) A word like "LIGHT" with many meanings (e.g., a fixture in a house, the opposite of "heavy," a cigarette lighter) is predicted to be forgettable either because a different meaning is activated at the repeat occurrence from the one activated at the initial occurrence (i), or because several meanings are activated (to different degrees), leading to a diffuse memory trace (similar to Steyvers & Malmberg's, 2003 proposal) (ii). (C) 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. See the online article for the color version of this figure.

Method

For the remainder of the article, we use the term "experimental item" to denote the items in the experiment, which are typically words, but sometimes consist of multiword phrases, as clarified under the Materials section.

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 (henceforth "experimental items"), 8,000–9,000 filler experimental items, and over 600 participants (n = 672

in Experiment 1 and $n = 631$ in Experiment 2). Building on past work on image memorability (e.g., Bainbridge et al., 2013; Isola et al., 2014; Isola, Xiao, et al., 2011), the experiments were designed as repeat detection tasks in which participants viewed a long sequence of experimental items, presented one at a time, and were asked to press a key whenever they noticed a repeat (an experimental item that they had already encountered earlier in the sequence; Figure 2). Critical repeats used to measure experimental item memorability occurred at lags of 91–109 experimental items. Approximately one out of every five experimental items was a critical repeat (by design, no critical repeats occurred during the first 90 experimental items). To ensure that participants were paying attention, vigilance repeats (chosen from a set of filler experimental items) occurred at lags of 1–7 experimental items.

For each target experimental item 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 experimental item occurred at most once for any given participant, this measure can be rewritten 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 https://www.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. Six hundred seventy-two 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. Demographic information was not collected. 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, experimental item sequences were broken up into "levels" that each consisted of 120 experimental items (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 nonrepeat trials. When this happened, all data collected on the experimental items at the current level were discarded, but data up to that level were retained, 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. Upon each restart, the sequence was reset so that the participant would never see an experimental item they had seen in a previous session. Participants who received three flags (as noted above) were blocked from further participation in the experiment; this choice was made to exclude "bots" and participants who were not trying to perform the task and answered randomly.

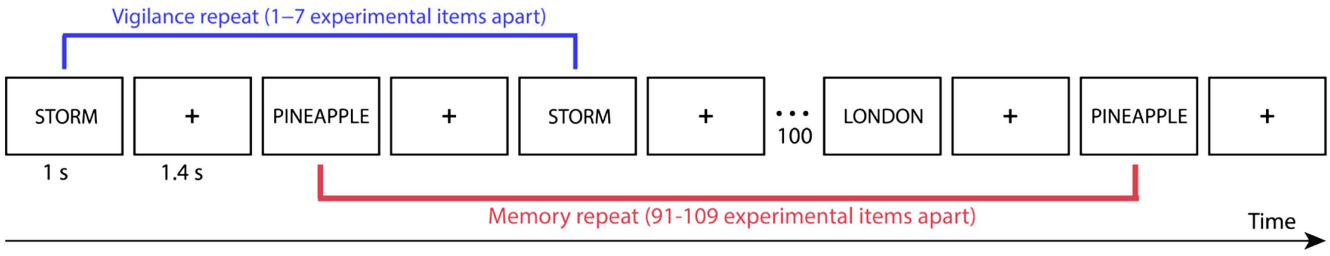
In Experiment 1, participants saw between 4 and 1,353 trials, with a median of 269 and a mean of 445.8. All participants were exposed to at least one target trial. The median participant saw 74 target repeat trials, and 664 participants saw at least one target repeat trial. In Experiment 2, participants 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 (not multiword phrases) was extracted from Subtlex (Brysbaert & New, 2009), such that a word that is twice as frequent in Subtlex was twice as likely to be included in our sample. The experimental items were then semimanually filtered in order to remove offensive experimental

Figure 2
Illustration of the Experimental Paradigm



Note. The procedure was identical across Experiments 1 and 2. Each experimental item was presented for 1 s, with 1.4 s intertrial intervals, giving participants 2.4 s to respond. Participants were asked to press the *r* key when they encountered an experimental item that had already occurred at some point in the experiment. Critical memory repeats (for the 2,222 target experimental items in each experiment) occurred between 91 and 109 experimental items apart, and vigilance repeats—between 1 and 7 experimental items apart. See the online article for the color version of this figure.

items, alternate forms of the same experimental item (color/colour), and experimental items that are clearly morphologically related (happy/happiness). From the remaining set of 11,182 experimental items, 2,222 experimental items were randomly selected for use as targets, and the rest were used as fillers in the experiment.

Experiment 2

A set of 10,320 experimental items 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, and weather. As in Experiment 1, 2,222 experimental items 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.

Procedure

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

Experimental Item Norms

To evaluate our hypothesis, for each of our two critical predictors—the number of synonyms and the number of meanings—we collected (a) human judgments, and, for Experiment 1, (b) 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 experimental item memorability, as well as to compare our hypothesis against some of the earlier proposals in the literature, we collected a set of 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 experimental item were obtained in norming experiments conducted using https://www.amazon.com's Mechanical Turk. For each set of critical materials (2,222 experimental items from Experiment 1 and 2,222 experimental items 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 nine subsets: eight subsets of 250 experimental items each and one subset of 222 experimental items. 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 experimental item: (a) whether the experimental item was a real word of English (the 20 catch trials mentioned above were pseudowords and were included to filter out participants who responded randomly), and (b) the critical judgment. For the number of synonyms experiments, participants were asked to identify how many synonyms each real experimental item has by choosing from among five options (zero synonyms, one synonym, two synonyms, 3–5 synonyms, or more than five synonyms). For the number of meanings experiments, participants were asked to identify how many meanings each real experimental item has by choosing from among four options (one meaning, two meanings, 3–5 meanings, more than five meanings; note that an experimental item may have zero synonyms but not zero meanings, hence only four options here).

Data from participants who provided ratings for more than 80% of pseudowords (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 experimental item (after excluding experimental items with 10 or fewer ratings, as elaborated in the following paragraph) and 250 unique participants for the number of meanings with a median of 27 ratings per experimental item for the number of meanings (after excluding experimental items 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 experimental item (after excluding experimental items with 10 or fewer ratings) and 248 unique participants for the number of meanings with a median of 26 ratings per experimental item for the number of meanings (after excluding experimental items with 10 or fewer ratings).

One hundred twelve of the 2,222 experimental items in Experiment 1 and 52 of the 2,222 experimental items 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 experimental items. Additionally, one experimental item 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 experimental items from Experiment 1. For Experiment 2, five experimental items were not available in the Google n-gram database ("a relux suppressant," "a tangello,"

"Arnold Schwazenegger," "Raphael (Raffaelo) Sanzio," "Scarlett Johanson"), mainly due to inadvertent spelling errors, and were excluded, leading to the exclusion of a total of 57 experimental items from Experiment 2. Thus, Experiment 1 contained 2,109 experimental items for the critical memorability analyses, and Experiment 2 contained 2,165 experimental items for the critical memorability analyses.

To derive a corpus-based estimate of the probability of an experimental item given a meaning, we used the following: (a) GloVe semantic word distinctiveness; and (b) number of synonyms assigned to an experimental item in Wordnet (Sigman & Cecchi, 2002). To estimate GloVe semantic distinctiveness, we obtained a GloVe (Global Vectors for Word Representation; Pennington et al., 2014) vector for each of the target experimental items (pretrained vectors from the Common Crawl corpus, available at https://nlp.stanford.edu/projects/glove/, Version glove.840B.300d.txt), and calculated the mean cosine similarity between this vector and all other experimental item vectors in the set (n = 2,109). This correlation reflects, on average, how similar a given experimental item is to other experimental items 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). Although this measure of semantic distinctiveness was computed based on the experimental items used in the experiment and not all English words, in pilot work we found that this measure was highly correlated with a measure of distinctiveness computed based on a larger set of experimental items. We also note that this measure of semantic distinctiveness does not account for homonymy because the GloVe representation provides a single context-independent representation for each experimental item.

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

Additional Norms

In addition to the norms for the critical predictors, for each experimental item, 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 https://www.amazon.com's Mechanical Turk. In particular, for each set of materials (2,222 experimental items from Experiment 1, and 2,222 experimental items from Experiment 2), five experiments were conducted: one for each attribute of interest. The materials in each set were divided into nine subsets: eight subsets of 250 experimental items each, and one subset of 222 experimental items. 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 experimental item 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; e.g., for imageability, experimental items like "POODLE" and "TOMATO" served as high-imageability catch trials, and experimental items 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 experimental item for concreteness (minimum: 25 ratings), a median of 26 ratings per experimental item for imageability (minimum: 23 ratings), a median of 28 ratings per experimental item for familiarity (minimum: 24 ratings), a median of 28 ratings for valence (minimum: 26 ratings), and a median of 22 ratings per experimental item 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 experimental item for concreteness (minimum: 19 ratings), a median of 26 ratings per experimental item for imageability (minimum: 24 ratings), a median of 27 ratings per experimental item for familiarity (minimum: 22 ratings), a median of 25 ratings for valence (minimum: 22 ratings), and a median of 23 ratings per experimental item for arousal (minimum: 21 ratings). We note that although some of these norms are available in existing databases (e.g., Brysbaert et al., 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 experimental items.[1])

Finally, we obtained a corpus-based frequency measure: for Experiment 1, we used the token log frequency of a word in the Subtlex subtitles corpus (Brysbaert et al., 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 experimental item accuracy as a function of the feature norms of interest. The procedure is cross-validated across participants and experimental items: we fit the linear regression model on half of the participants using half of the experimental items (Experiment 1: 1,055 experimental items 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 experimental items for test set;

[1] For concreteness, the number of overlapping experimental items 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 experimental items 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 Experiments 1 and 2, respectively. Last, for arousal, the number of overlapping experimental items 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 Experiments 1 and 2, respectively.

Experiment 2: 1,082 experimental items for test set, unless otherwise stated).

We demeaned each column of the regressor matrix (i.e., each feature), but we did not normalize the columns to have a 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 independently 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 reported 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 .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 .05).

Besides the feature selection step, the modeling procedure was identical to the remaining model evaluations (as described in the Cross-Validated Model Performance section): we partitioned the data into a train/test set of independent participants and experimental items, 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 reported the median Spearman correlation between predicted accuracy and actual accuracy.

Transparency and Openness

The memorability data for Experiments 1 and 2 are publicly available as csv files in the following repository: https://github.com/gretatuckute/memorable_words/ (Tuckute et al., 2024). Similarly, the repository contains the code to reproduce the figures/tables in the article. The study was not preregistered. The study consists of two independent experiments (Experiments 1 and 2), and the main findings were replicated.

Results

How Memorable Are Word and Multiword Phrases (Experimental Items)?

In Experiment 1, we measured the memorability of 2,222 experimental items sampled from the Subtlex corpus (Brysbaert & New, 2009), which consists of movie transcripts. The experimental items 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 experimental items (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 nonrepeat trials) was 0.09, and the median accuracy (the number of correct repeat detections and correct repeat nondetections divided by the total number of trials) was 0.80 (Figure 3A(iii)).

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

To evaluate the generalizability of the results from Experiment 1, in Experiment 2, we measured the memorability of a new set of 2,222 experimental items, 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 experimental items were consistently better remembered than others (split-half Spearman correlation for accuracy across participants: 0.65, 95% CI [0.63, 0.67] by nonparametric bootstrap).

To allow for an across-experiment comparison for the same materials, 48 experimental items overlapped between the sets of target experimental items in Experiments 1 and 2. The accuracies for these 48 experimental items were strongly correlated between the two experiments ($r = 0.77$, $p << .0001$; Supplemental 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, $p << .0001$; Supplemental Table 1).

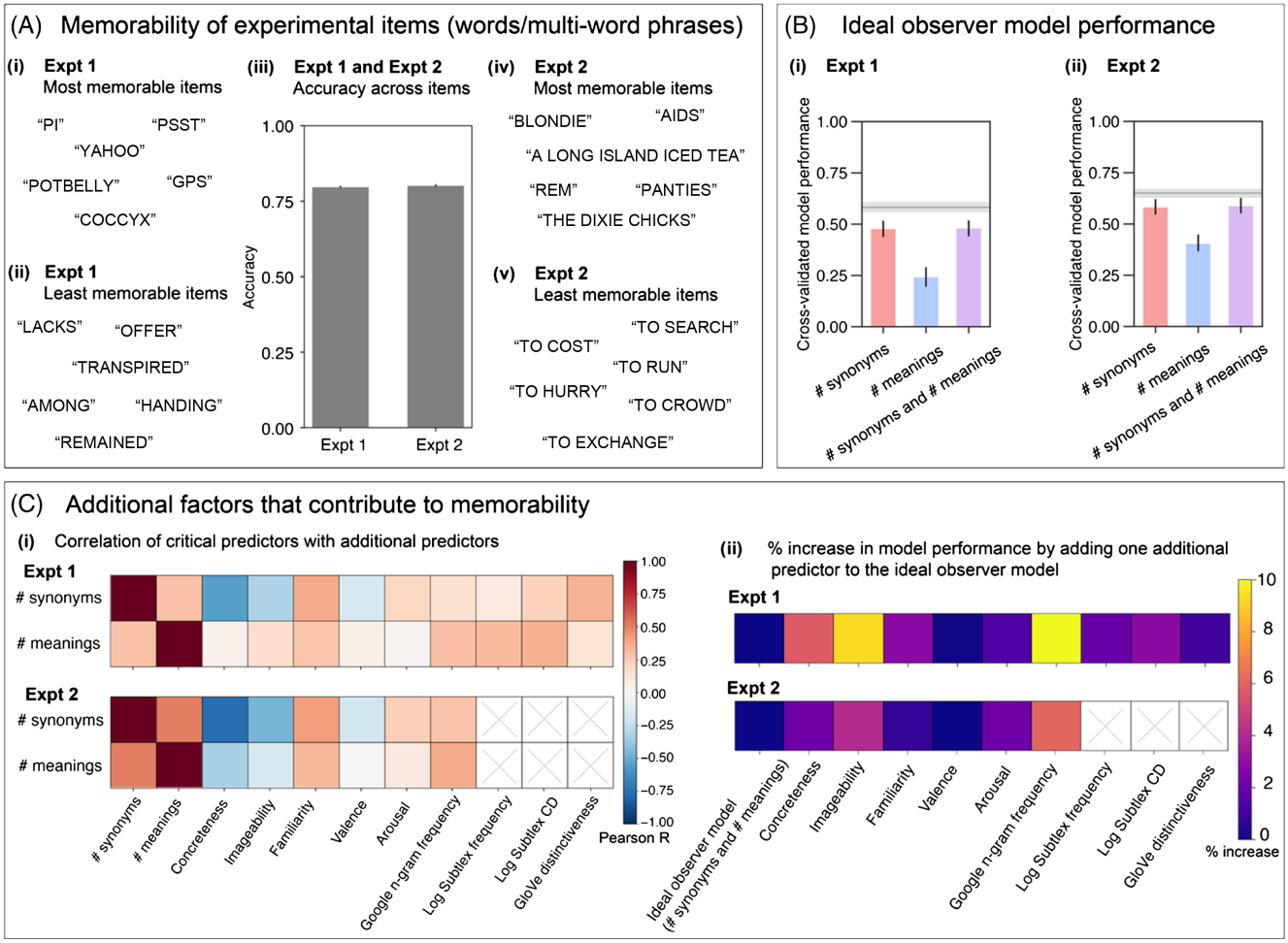
How Well Does the Ideal Observer Model Explain Experimental Item Recognition Performance?

To test the predictions of the proposal put forward here, we examined the relationship between experimental item memorability and the number of synonyms and number of meanings. We report the results based on the human-derived number of synonyms/meanings in the main text because corpus-based measures were not possible to obtain for the materials in Experiment 2. 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 synonym norms is correlated at $r = 0.35$, and the number of meaning norms is correlated at $r = 0.54$. For completeness, we report the results for Experiment 1 based on the corpus-derived measures in Supplemental Figure 2 and Supplemental 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 experimental item accuracy as a function of both the number of synonyms and the number of

Figure 3

Critical Results for Experiments 1 and 2



Note. (A) Memory performance (median recognition accuracy) across all experimental items in Experiment 1 ($n = 2{,}109$ experimental items) and Experiment 2 ($n = 2{,}165$ experimental items) is shown in (iii). Error bars are 95% confidence interval (CI) of the median. The most memorable experimental items for Experiments 1 and 2 are shown in (i) and (iv), respectively; and the least memorable experimental items—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 nonoverlapping participants and experimental items for train/test splits. Error bars are 95% CI across cross-validation splits. The horizontal gray lines denote the interparticipant reliability of the data, as defined by the split-half Spearman correlation for accuracy across participants, which measures the reliability of the data and hence provides a reasonable upper bound on the correlation obtainable from external predictors. (C) The correlations between the critical predictors (human-derived number of synonyms and number of meanings norms, as described in the Experimental Item Norms section), and between the critical predictors and other predictors are shown in (i). The set of noncritical 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 an experimental item in the Google n-gram database in 2013), Subtlex frequency (the overall frequency of an experimental item in the Subtlex movies transcript), Subtlex contextual diversity (CD; a measure of the number of distinct movie transcripts in which an experimental item 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 noncritical predictor is shown in (ii) (see Supplemental Table 7 for the statistical evaluation). Expt = Experiment; CD = contextual diversity. See the online article for the color version of this figure.

meanings. To avoid overfitting, we learned the model coefficients using half of the participants and half of the experimental items and tested the model on the nonoverlapping portion of the data. Iterating this procedure 1,000 times, the median Spearman correlation between the memorability accuracies and model predictions using both the number of synonyms and the number of meanings was 0.48, 95% CI [0.44, 0.52] compared to the interparticipant consistency of 0.58 (the split-half correlation across participants) for Experiment 1, and 0.59, 95% CI [0.55, 0.63] compared to the interparticipant consistency of 0.65 for Experiment 2 (Table 1).

Thus, a simple model of experimental item memorability, based on just two rationally motivated factors—number of synonyms and

Table 1
Ideal Observer Model Performance

Experiment	No. of synonyms, 95% CI	No. of meanings, 95% CI	No. of synonyms and no. of meanings, 95% CI
Expt 1	0.48 [0.44, 0.52]	0.24 [0.19, 0.29]	0.48 [0.44, 0.52]
Expt 2	0.58 [0.55, 0.62]	0.41 [0.37, 0.45]	0.59 [0.55, 0.63]

Note. 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 nonoverlapping participants and experimental items for train/test splits. CI = confidence interval; Expt = experiment.

number of meanings—captures a large portion of the variance in experimental item memorability (cross-validated Spearman correlation of 0.48 and 0.59 in Experiments 1 and 2, compared to the interparticipant consistency of 0.58 and 0.65 in Experiments 1 and 2). 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 experimental item memorability (Supplemental Figure 3 and Supplemental Table 3; cf. Malmberg et al., 2002).

Next, we investigated whether one of 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 (Supplemental Table 4). For Experiment 1, in line with the results based on the cross-validated model performance (Table 1 and 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 and possibly driven by the large number of observations (number of meanings contributed less to the cross-validated analyses reported in Table 1). 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: the number of synonyms is a stronger contributor to memorability compared to the number of meanings.

Do Additional Factors Contribute to Experimental Item Memorability?

To explore the effects on experimental item 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.

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, the number of synonyms showed a strong negative correlation with concreteness ($r = -0.55$ in Experiment 1 and $r = -0.76$ in Experiment 2): experimental items 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): experimental items 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 [$r = 0.80$]; Supplemental 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 a weak negative correlation with valence ($r = -0.17$ in Experiment 1 and $r = -0.19$ in Experiment 2): experimental items that were more frequent and rated as more arousing and less positively charged were rated as having fewer synonyms.

[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 & 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).

The 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 experimental items were rated as having more meanings, as expected given past work (e.g., Fenk-Oczlon & Fenk, 2010; Piantadosi et al., 2012). Similar to the number of synonyms, the number of meanings was also moderately correlated with familiarity ($r = 0.28$ in Experiment 1, and $r = 0.32$ in Experiment 2). (See Supplemental Figure 4 for all predictor correlations.)

Next, we asked whether any of the additional predictors explained 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 the Results; How Well Does the Ideal Observer Model Explain Experimental Item Recognition Performance? section, we investigated both the cross-validated model performance as well as performing the 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 Supplemental 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 (Supplemental Table 7). These tests mirror the patterns from the cross-validated model performance analysis (Figure 3C and Supplemental Tables 5 and 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: $F = 101.58$, $p << .0001$). The Google n-gram frequency feature, reflecting an experimental item's occurrence in a vast number of books, is positively correlated with the number of synonyms (Pearson $r = 0.16$ in Experiment 1 and $r = 0.28$ for Experiment 2) as well as the meanings ($r = 0.30$ in Experiment 1 and $r = 0.38$ for Experiment 2). Yet, the fact that it explains additional variance beyond these two critical predictors suggests that an experimental item's general frequency of use—regardless of its specific meaning—is another significant factor in explaining its memorability.

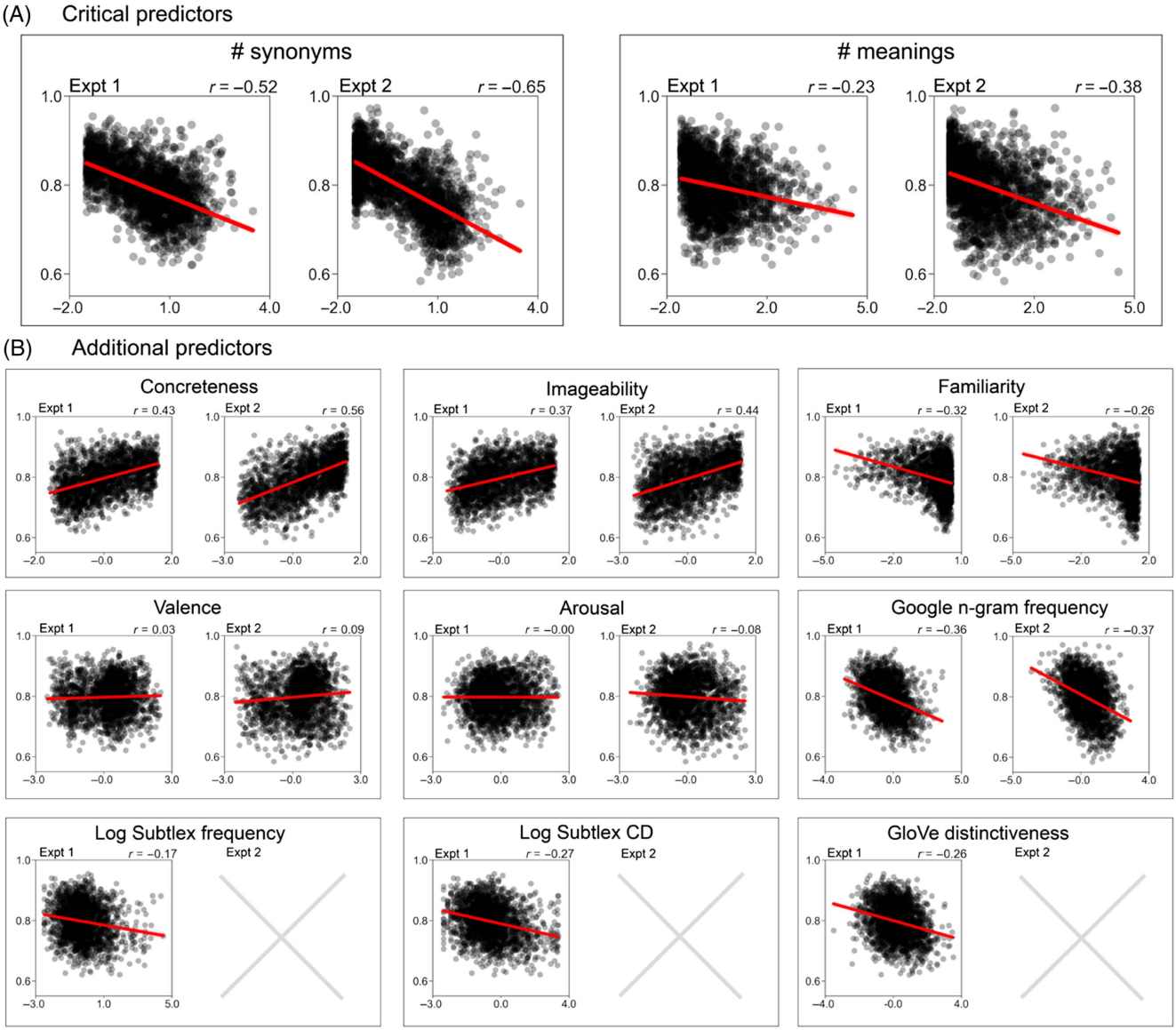
Given that we observed that frequency was a strong additional predictor, we performed a supplementary analysis to investigate whether the ideal observer model could still predict memorability when experimental items were similar in frequency. To do so, we partitioned the experimental items into low-, 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 (Supplemental Figure 5 and Supplemental Tables 8 and 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]; Supplemental Figure 6A and Supplemental Table 10). Furthermore, including the corpus-derived 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]; Supplemental 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 Supplemental 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 by Griffiths et al. (2007; available for a subset of our experimental materials; Supplemental Figure 6B and Supplemental Table 11). However, the topic variability model was not on par with the ideal observer baseline model (Supplemental Figure 6B and Supplemental Table 11; independent two-sided t test between Spearman correlation values across 1,000 cross-validation splits: $p << .0001$, $t = 61.57$, 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 than two related proposals within the Bayesian framework. As noted in the Discussion section, the advantage of the ideal observer model likely stems from the inclusion of the number of synonyms predictor.

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 Supplemental 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 the Forward–Backward Feature Selection section). We partitioned the data into a train/test set of independent participants and experimental items (identical to Results; How Well Does the Ideal Observer Model Explain Experimental Item Recognition Performance? and Results; Do Additional Factors Contribute to Experimental Item Memorability? sections) 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

Figure 4
Correlation of Each Predictor With Recognition Accuracy



Note. 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). Expt = Experiment; CD = contextual diversity. See the online article for the color version of this figure.

folds). Next, we tested the model using these features on the test set and reported 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 CD measures (from Subtlex) and GloVe semantic distinctiveness, that is, a pool of 13 predictors in total. We obtained a maximum explained variance of 0.59, 95% CI [0.55, 0.62] (the subjectwise 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 Supplemental Table 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), that is, a pool of eight predictors in total. We obtained a maximum explained variance of 0.63, 95% CI [0.60 0.67] (the subjectwise 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*, and *Imageability*.

The second most frequently selected feature set was (these features were selected 192/1,000 times): *# Synonyms (human)*, *Google n-gram frequency*, *Imageability*, and *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 nine 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 subjectwise noise ceiling. Mirroring the results in the Results; How Well Does the Ideal Observer Model Explain Experimental Item Recognition Performance? section, the number of synonyms predictor is a stronger predictor of experimental item memorability than the number of meanings predictor (in particular, for Experiment 2).

Discussion

We investigated experimental item memorability (word and multiword phrases) across two large-scale behavioral recognition-memory experiments (n = 672 and 631 participants in Experiments 1 and 2, respectively; n = 2,222 target experimental items in each experiment; n = 3,780 participants used in the norming experiments). The contributions of the current work are fourfold. First, across two large sets of experimental items, we found that memorability is largely an intrinsic property of experimental items: some experimental items are consistently remembered better than others across participants. Second and critically, building on past work (e.g., Griffiths et al., 2007; Steyvers & Malmberg, 2003), we evaluated and provided support for a novel proposal for what makes experimental items memorable—the ideal observer model of experimental item memorability—whereby memorable experimental items have unique associations with their meanings. Third, 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.

Some Experimental Items Are Intrinsically Memorable and Others—Forgettable

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

The intrinsic nature of word memorability is consistent with item-noise models of episodic memory (e.g., McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997), in which items are encoded by their intrinsic properties (in our model, as discussed in the Memorable Experimental Items Have Few Synonyms and Few Meanings section, the features have to do with the word-to-meaning relationship). Conversely, our findings are at odds with context-noise models (Dennis & 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 experimental items 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 interindividual differences in recognition memory for experimental items. For example, are participants with more linguistic experience (as indexed, e.g., by larger vocabularies) better at experimental item recognition memory? What other aspects of individual experiences/cognitive abilities affect experimental item recognition performance? And does experimental item recognition memory correlate across individuals with recognition memory for images? In other words, do experimental item and image memorability both depend on the properties of abstract concept representations, or the particular interface between those representations and verbal/visual-image representations?

Memorable Experimental Items Have Few Synonyms and Few Meanings

In this work, we explored whether experimental items are encoded in memory by their meanings. By formulating an ideal observer model of word memorability, we hypothesized that a memorable experimental item selects a particular meaning in semantic memory. We indeed found that memorable experimental items 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 (Aue et al., 2018; Dennis & Humphreys, 2001; Jones et al., 2017; Shiffrin & Steyvers, 1997; Steyvers & Malmberg, 2003). 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 and 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 & 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 experimental item associated with many meanings)—the number of synonyms predictor (many experimental items associated with one meaning), explains substantially more variance than even the better performing topic variability predictor. 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 the How Well Does the Ideal Observer Model Explain Experimental Item Recognition Performance? section), the number of synonym norms was more reliable than the number of meaning norms. Thus, the number of synonyms predictor was likely a more important predictor of experimental item memorability because our materials varied more along this dimension, and the number-of-synonym estimates were more reliable.

It is worth noting that the 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 words with fewer synonyms (Hino et al., 2002). These findings suggest that a one-to-one mapping between an experimental item form and a meaning not only leads to more robust memory traces, as shown here, but also facilitates retrieval of experimental items from memory.

Finally, our finding that verbal memorability strongly depends on the relationship between the experimental items 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 (Huebner & Gegenfurtner, 2012; Konkle et al., 2010; Lin et al., 2019).

Additional Factors That Affect Experimental Item Memorability

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: the number of synonyms emerged as the strongest predictor in both. The number of synonyms was followed by experimental item frequency, in agreement with prior work demonstrating that lower frequency experimental items are easier to recognize (e.g., Brown & Lewis, 1981; Gorman, 1961; Kinsbourne & 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 & Friendly, 1986; Walker & 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 the number of meanings, familiarity, a corpus-based measure of the number of meanings, and contextual diversity.

Implications of Understanding Experimental Item Memorability

Why might one want to know which experimental items 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 interpersonal 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 levels, including in the education domain. 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 data set presented here, a simple linear model can be used to predict the memorability of 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.

[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 a standard deviation of 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.

Conclusion

In this article, we have offered a simple account, based on rational analysis, as to what makes words and multiword phrases 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 present study investigates the memorability of words in isolation. We acknowledge the importance of context effects on memorability, as well as the 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 more traditional long-term memory paradigm with a separate study and test phase (Goetschalckx et al., 2018), to different retention intervals (Goetschalckx et al., 2018; Isola et al., 2014), and to incidental memory scores where a memory test is administered as a surprise (Goetschalckx et al., 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 be taken into account given the importance of context for word processing.

Extension to Languages Other Than English

The predictions derived from the present 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., C. J. C. Lin & Ahrens, 2010); this variability may affect word memory/recognition across languages.

References

Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cognitive Psychology*, *6*(4), 451–474. https://doi.org/10.1016/0010-0285(74)90021-8

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, *96*(4), 703–719. https://doi.org/10.1037/0033-295X.96.4.703

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, *2*(6), 396–408. https://doi.org/10.1111/j.1467-9280.1991.tb00174.x

Annis, J., Lenes, J. G., Westfall, H. A., Criss, A. H., & Malmberg, K. J. (2015). The list-length effect does not discriminate between models of recognition memory. *Journal of Memory and Language*, *85*, 27–41. https://doi.org/10.1016/j.jml.2015.06.001

Aue, W. R., Fontaine, J. M., & Criss, A. H. (2018). Examining the role of context variability in memory for items and associations. *Memory & Cognition*, *46*(6), 940–954. https://doi.org/10.3758/s13421-018-0813-9

Bainbridge, W. A. (2022). Memorability: Reconceptualizing memory as a visual attribute. In T. F. Brady & W. A. Bainbridge (Eds.), *Visual memory* (pp. 173–187). Routledge. https://doi.org/10.4324/9781003158134-11

Bainbridge, W. A., Isola, P., & Oliva, A. (2013). The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, *142*(4), 1323–1334. https://doi.org/10.1037/a0033872

Begg, I., & Wickelgren, W. A. (1974). Retention functions for syntactic and lexical vs semantic information in sentence recognition memory. *Memory & Cognition*, *2*(2), 353–359. https://doi.org/10.3758/BF03209009

Bird, S., & Loper, E. (2004). NLTK: The natural language toolkit. *Proceedings of the ACL Interactive Poster And Demonstration Sessions* (pp. 214–217). Association for Computational Linguistics. https://aclanthology.org/P04-3031

Borkin, M. A., Vo, A. A., Bylinskii, Z., Isola, P., Sunkavalli, S., Oliva, A., & Pfister, H. (2013). What makes a visualization memorable? *IEEE Transactions on Visualization and Computer Graphics*, *19*(12), 2306–2315. https://doi.org/10.1109/TVCG.2013.234

Bransford, J. D., & Franks, J. J. (1971). The abstraction of linguistic ideas. *Cognitive Psychology*, *2*(4), 331–350. https://doi.org/10.1016/0010-0285(71)90019-3

Brown, J., & Lewis, V. (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), 247–265. https://doi.org/10.2307/1422744

Brysbaert, M., & New, B. (2009). Moving beyond Kucera 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), 977–990. https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods*, *44*(4), 991–997. https://doi.org/10.3758/s13428-012-0190-4

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Bylinskii, Z., Isola, P., Bainbridge, C., Torralba, A., & Oliva, A. (2015). Intrinsic and extrinsic effects on image memorability. *Vision Research*, *116*, 165–178. https://doi.org/10.1016/j.visres.2015.03.005

Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences*, *3*(2), 57–65. https://doi.org/10.1016/S1364-6613(98)01273-X

Criss, A. H., & Shiffrin, R. M. (2004). Context noise and item noise jointly determine recognition memory: A comment on Dennis and Humphreys (2001). *Psychological Review*, *111*(3), 800–807. https://doi.org/10.1037/0033-295X.111.3.800

Danion, J.-M., Kauffmann-Muller, F., Grangé, D., Zimmermann, M.-A., & Greth, P. (1995). Affective valence of words, explicit and implicit memory in clinical depression. *Journal of Affective Disorders*, *34*(3), 227–234. https://doi.org/10.1016/0165-0327(95)00021-E

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review*, *108*(2), 452–478. https://doi.org/10.1037/0033-295X.108.2.452

Fenk-Oczlon, G., & Fenk, A. (2010). Frequency effects on the emergence of polysemy and homophony. *International Journal "Information Technologies and Knowledge"*, *4*(2), 103–109. https://scholar.google.de/citations?view_op=view_citation&hl=en&user=kkjY0pkAAAAJ&sortby=pubdate&citation_for_view=kkjY0pkAAAAJ:Zph67rFs4hoC

Forster, K. I. (1970). Visual perception of rapidly presented word sequences of varying complexity. *Perception & Psychophysics*, *8*(4), 215–221. https://doi.org/10.3758/BF03210208

Franks, J. J., & Bransford, J. D. (1972). The acquisition of abstract ideas. *Journal of Verbal Learning and Verbal Behavior*, *11*(3), 311–315. https://doi.org/10.1016/S0022-5371(72)80092-6

Gershman, S. J. (2024). The rational analysis of memory. In M. Kahana & A. Wagner (Eds.), *Oxford handbook of human memory*. Oxford University Press. https://global.oup.com/academic/product/the-oxford-handbook-of-human-memory-two-volume-pack-9780197746141?cc=us&lang=en&#

Goetschalckx, L., Moors, J., & Wagemans, J. (2019). Incidental image memorability. *Memory*, *27*(9), 1273–1282. https://doi.org/10.1080/09658211.2019.1652328

Goetschalckx, L., Moors, P., & Wagemans, J. (2018). Image memorability across longer time intervals. *Memory*, *26*(5), 581–588. https://doi.org/10.1080/09658211.2017.1383435

Gorman, A. M. (1961). Recognition memory for nouns as a function of abstractness and frequency. *Journal of Experimental Psychology*, *61*(1), 23–29. https://doi.org/10.1037/h0040561

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211–244. https://doi.org/10.1037/0033-295X.114.2.211

Hino, Y., Lupker, S. J., & Pexman, P. M. (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. https://doi.org/10.1037/0278-7393.28.4.686

Huebner, G. M., & Gegenfurtner, K. R. (2012). Conceptual and visual features contribute to visual memory for natural images. *PLOS ONE*, *7*(6), Article e37575. https://doi.org/10.1371/journal.pone.0037575

Isola, P., Parikh, D., Torralba, A., & Oliva, A. (2011). Understanding the intrinsic memorability of images. *Proceedings of the 24th international conference on Neural Information Processing Systems* (pp. 2429–2437).

Isola, P., Xiao, J., Parikh, D., Torralba, A., & Oliva, A. (2014). What makes a photograph memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(7), 1469–1482. https://doi.org/10.1109/TPAMI.2013.200

Isola, P., Xiao, J., Torralba, A., & Oliva, A. (2011). What makes an image memorable? *IEEE conference on Computer Vision and Pattern Recognition (CVPR), 2011* (pp. 145–152).

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General*, *110*(3), 306–340. https://doi.org/10.1037/0096-3445.110.3.306

Jones, M. N., Dye, M., & Johns, B. T. (2017). Context as an organizing principle of the lexicon. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 67, pp. 239–283). Elsevier. https://doi.org/10.1016/bs.plm.2017.03.008

Katz, S., & Gruenewald, P. (1974). The abstraction of linguistic ideas in "meaningless" sentences. *Memory & Cognition*, *2*(4), 737–741. https://doi.org/10.3758/BF03198149

Kensinger, E. A., & Corkin, S. (2003). Memory enhancement for emotional words: Are emotional words more vividly remembered than neutral words? *Memory & Cognition*, *31*(8), 1169–1180. https://doi.org/10.3758/BF03195800

Khanna, M. M., & Cortese, M. J. (2021). How well imageability, concreteness, perceptual strength, and action strength predict recognition memory, lexical decision, and reading aloud performance. *Memory*, *29*(5), 622–636. https://doi.org/10.1080/09658211.2021.1924789

Kinsbourne, M., & George, J. (1974). The mechanism of the word-frequency effect on recognition memory. *Journal of Verbal Learning and Verbal Behavior*, *13*(1), 63–69. https://doi.org/10.1016/S0022-5371(74)80031-9

Klaver, P., Fell, J., Dietl, T., Schür, S., Schaller, C., Elger, C. E., & Fernández, G. (2005). Word imageability affects the hippocampus in recognition memory. *Hippocampus*, *15*(6), 704–712. https://doi.org/10.1002/hipo.20081

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General*, *139*(3), 558–578. https://doi.org/10.1037/a0019165

Light, L. L., & Carter-Sobell, L. (1970). Effects of changed semantic context on recognition memory. *Journal of Verbal Learning and Verbal Behavior*, *9*(1), 1–11. https://doi.org/10.1016/S0022-5371(70)80002-0

Lin, C. J. C., & Ahrens, K. (2010). Ambiguity advantage revisited: Two meanings are better than one when accessing Chinese nouns. *Journal of Psycholinguistic Research*, *39*, 1–19. https://doi.org/10.1007/s10936-009-9120-8

Lin, Q., Yousif, S. R., Scholl, B., & Chun, M. M. (2019). Image memorability is driven by visual and conceptual distinctivenes. *Journal of Vision*, *19*(10), Article 290c. https://doi.org/10.1167/19.10.290c

Lohnas, L. J., & Kahana, M. J. (2013). Parametric effects of word frequency in memory for mixed frequency lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(6), 1943–1946. https://doi.org/10.1037/a0033669

Malmberg, K. J., Steyvers, M., Stephens, J. D., & Shiffrin, R. M. (2002). Feature frequency effects in recognition memory. *Memory & Cognition*, *30*(4), 607–613. https://doi.org/10.3758/BF03194962

Marian, V., Bartolotti, J., Chabal, S., & Shook, A. (2012). CLEARPOND: Cross-linguistic easy-access resource for phonological and orthographic neighborhood densities. *PLOS ONE*, *7*(8), Article e43230. https://doi.org/10.1371/journal.pone.0043230

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review*, *105*(4), 724–760. https://doi.org/10.1037/0033-295X.105.4.734-760

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., the Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–182. https://doi.org/10.1126/science.1199644

Mohammad, S. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 174–184). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1017

Monaco, J. D., Abbott, L. F., & Kahana, M. J. (2007). Lexico-semantic structure and the word-frequency effect in recognition memory. *Learning & Memory*, *14*(3), 204–213. https://doi.org/10.1101/lm.363207

Paivio, A. (1969). Mental imagery in associative learning and memory. *Psychological Review*, *76*(3), 241–263. https://doi.org/10.1037/h0027272

Paivio, A. (1971). *Imagery and verbal processes*. Holt, Rinehart and Winston.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

Pereira, F., Gershman, S., Ritter, S., & Botvinick, M. (2016). A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive Neuropsychology*, *33*(3–4), 175–190. https://doi.org/10.1080/02643294.2016.1176907

Phelps, E. A., LaBar, K. S., & Spencer, D. D. (1997). Memory for emotional words following unilateral temporal lobectomy. *Brain and Cognition*, *35*(1), 85–109. https://doi.org/10.1006/brcg.1997.0929

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280–291. https://doi.org/10.1016/j.cognition.2011.10.004

Potter, M. C. (2012). Conceptual short term memory in perception and thought. *Frontiers in Psychology*, *3*, Article 113. https://doi.org/10.3389/fpsyg.2012.00113

Potter, M. C., Kroll, J. F., & Harris, C. (1980). Comprehension and memory in rapid sequential reading. In R. S. Nickerson (Ed.), *Attention and performance VIII* (pp. 395–418). Erlbaum.

Potter, M. C., & Lombardi, L. (1990). Regeneration in the short-term recall of sentences. *Journal of Memory and Language*, *29*(6), 633–654. https://doi.org/10.1016/0749-596X(90)90042-X

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, *34*(6), 909–957. https://doi.org/10.1111/j.1551-6709.2009.01092.x

Rubin, D. C., & Friendly, M. (1986). Predicting which words get recalled: Measures of free recall, availability, goodness, emotionality, and pronunciability for 925 nouns. *Memory & Cognition*, *14*(1), 79–94. https://doi.org/10.3758/BF03209231

Schulman, A. I. (1967). Word length and rarity in recognition memory. *Psychonomic Science*, *9*(4), 211–212. https://doi.org/10.3758/BF03330834

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM-retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*(2), 145–166. https://doi.org/10.3758/BF03209391

Sigman, M., & Cecchi, G. A. (2002). Global organization of the Wordnet lexicon. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(3), 1742–1747. https://doi.org/10.1073/pnas.022341799

Steyvers, M., & Malmberg, K. J. (2003). The effect of normative context variability on recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(5), 760–766. https://doi.org/10.1037/0278-7393.29.5.760

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285. https://doi.org/10.1126/science.1192788

Tuckute, G., Mahowald, K., Isola, P., Oliva, A., Gibson, E., & Fedorenko, E. (2024). *Memorable words* (zenodo.8349726; Version v1) [Data set]. https://zenodo.org/doi/10.5281/zenodo.8349726

Walker, I., & Hulme, C. (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), 1256–1271. https://doi.org/10.1037/0278-7393.25.5.1256

Received January 13, 2023
Revision received December 26, 2024
Accepted January 18, 2025 ■