Memorable words are monogamous: The role of synonymy and homonymy in word recognition memory

Kyle Mahowald[1,+], Phillip Isola[2,*,+], Evelina Fedorenko[3,4], Edward Gibson[5], and Aude Oliva[6]

[1]Massachusetts Institute of Technology, Brain and Cognitive Sciences, Cambridge, MA, 02139, USA
[2]UC Berkeley, EECS, Berkeley, CA, 94709, USA
[3]Harvard Medical School, Department of Psychiatry, Boston, MA, 02115, USA
[4]Massachusetts General Hospital, Department of Psychiatry, Charlestown, MA, 02129, USA
[5]Massachusetts Institute of Technology, Brain and Cognitive Sciences, Cambridge, MA, 02139, USA
[6]Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, 02139, USA
[*]isola@eecs.berkeley.edu
[+]these authors contributed equally to this work

ABSTRACT

What makes a word memorable? Prior research has identified numerous factors: word frequency, concreteness, imageability, and valence have all been shown to affect recognition performance. One important dimension that has not received much attention is the nature of the relationship between words and meanings. Under the hypothesis that words are encoded primarily by their meanings, and not by their surface forms, this relationship should be central to determining word memorability. In particular, rational analysis suggests that people will more easily remember words that convey a large amount of information about their intended meaning and that have few alternatives – that is, memorable words will be those with few possible meanings and synonyms. To test this hypothesis, we ran two large-scale recognition memory experiments (each with 2,222 words, 600+ participants). Memory performance was overall high, on par with memory for pictures in a similar paradigm. Critically, however, not all words were remembered equally well. Consistent with our proposal, the best recognized words had few meanings and few synonyms. Indeed, the most memorable words had a one-to-one relationship with their meanings. Estimates of memorability derived from this rational account explain a large amount of the variance in word memorability.

Introduction

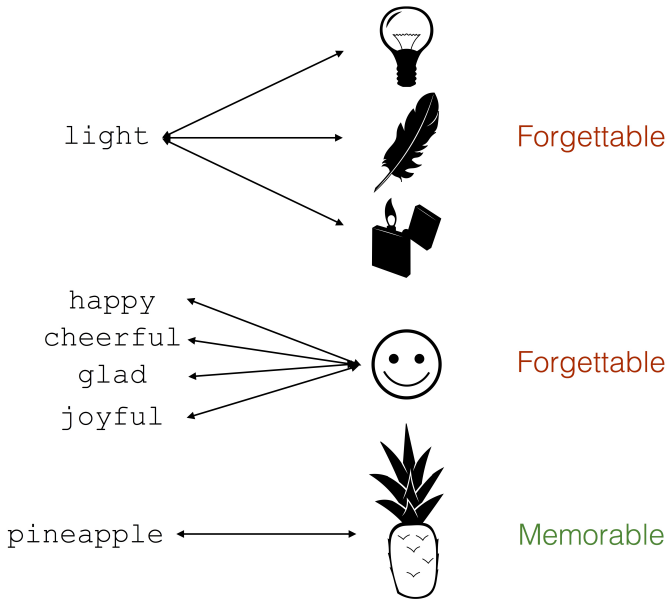
An avalanche of precise, lucid vocabulary has an advantage as a manner of expression. Perhaps this comes as no surprise. We all know that effective word choice is critical to clear communication. Less obvious is the impact word choice has on memory. Consider the first sentence of this paragraph. "Avalanche,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

What makes "avalanche" stick in our head? The literature provides several possible explanations: less familiar, lower-frequency words are easier to recognize but more difficult to recall[1,2,3,4]; imageable and concrete words are both easy to recognize and easy to recall[5,6,7]; and negative words[8] and emotionally salient words[9,10] also enjoy a memory boost. Semantic context[11,12] and semantic prompts[13] have also been shown to be important. On the other hand, syntactic and lexical properties of words are far more poorly retained than meaning[14–16].

Building on existing work on semantic effects of verbal memorability, here we focus on the functional task of recognition memory. Our analysis follows from the idea that during recognition, the memory system is asked to identify whether or not a currently observed word form matches a representation in memory. We assume that in solving this problem, the brain approximates the Bayes optimal inference.

This is the same general approach as has been taken by several past authors, including Shiffrin & Steyvers (1997)[17], Dennis & Humphreys (2001)[18], and Steyvers & Malmberg (2003)[19]. Using the REM model, Shiffrin & Steyvers suggested that speakers performing a recognition task are, in effect, computing a probability that a given stimulus is "new" or "old" by accessing vectors of stored features. Exactly what those stored features are has been a matter of substantial debate[20,21].

Figure 1. Schematic illustration of a word with many meanings (*light*), a word with many synonyms that all mean approximately the same thing (*happy*), and a word that is roughly one-to-one in its form-meaning mapping (*pineapple*).



"Item-noise models" argue that the features are intrinsic properties of individual words, whereas "context-noise models" claim the features are properties of the context in which the words are presented. However, neither class of model makes strong claims as to which intrinsic or contextual properties are encoded.

Whereas past models have tended to be noncommittal about the exact form of the encoded features, here we explore a simple, concrete possibility: that words are encoded by their meaning. Given this encoding hypothesis, and assuming a Bayesian ideal observer, we derive the implications as to which words will be memorable. This model motivates two novel predictors of memorability: memorable words should be those with few synonyms and few meanings. Across two large-scale experiments, we demonstrate that these measures are indeed highly predictive of recognition rates.

Ideal Observer Model

In order to formalize these ideas, we propose a simple Bayesian model of the task in which a rational agent sees a word and stores not the word itself but a meaning m selected by that word. The agent is then asked, at a later time, whether a new word w is a word that has already been seen. At recall time, the participant has access to the stored meaning m and the new stimulus w and must decide whether the original word that generated m is the same as the new word w. Note that we are not proposing such a process as the actual way that humans play this game, which is presumably more complex, but are merely using it as a tool for deriving simple predictions about how such an agent would act.

Formulating the problem this way allows us to design an optimal decision rule, which tells us how an optimal agent would perform the "have you seen this word before?" task, given that the agent stores meanings and not exact wordforms. Formally, the agent must assess the probability that the newly observed word w is the same as the originally observed word. We can express this as the probability that a random variable W (which represents the original word) takes on the value w given stored meaning m. Effectively, this model assumes that the sequence consists of two words: a "current word" and a "previous word" that gave rise to the stored meaning m. Applying Bayes' rule, this probability can be written as:

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

This formula has an intuitive interpretation. The agent is assessing "out of all the possible ways I could have ended up with this memory, what are the chances w is the culprit?"

Given this decision rule, which words will be most memorable? We operationalize memorability as the expected value of $P(W = w \mid m)$, taking expectation over all cases in which W does in fact equal w. This is equivalent to the hit rate for w if the

agent plays a probability matching strategy:

$$\mathscr{M}(w) = \mathbb{E}_{P(m|W=w)}[P(W=w|m)] = P(W=w)\sum_{m}\frac{P(m|W=w)^2}{P(m)} \quad (2)$$

According to this model, there are three reasons a word might not be remembered, despite having been seen: (1) $P(m \mid W=w)$ is generally small (i.e. the distribution has high entropy). In simple terms, this is the case in which w has many meanings. The agent will be uncertain as to which meaning would have been encoded. (2) $P(m)$ is generally large for all the meanings of w. This corresponds to the case where w has many synonyms. That is, for each meaning of w, there are other common words that have the same meaning. These alternatives compete with w as the cause of the memory. (3) $P(W=w)$ is small. This term can be thought of as the *a priori* probability that the random variable W has the specific value w. Given our model assumptions of a current word and a previous word, this is equivalent to the probability of seeing a repeat and is the same for all words, provided that any word which appears once in the task has the same probability of being repeated. (This assumption is in fact true of the task.)

The first two scenarios are illustrated in Figure 1. On the first row, we have case #1. Because "light" can refer to several meanings (e.g., lightweight, bright, a lighter), the probability of "light" being encoded as any given one of these must be relatively low. On the second row we have case #2. There are many synonyms for "happy" (e.g., "cheerful", "glad", "joyful"). If the agent indeed only encodes the meaning, and not the wordform, then the agent cannot know for sure which of the synonyms it actually saw.

Row three of Figure 1 shows a third case, in which a word has a one-to-one relationship with its meaning. In this case, there is no competition among either multiple meanings or multiple synonyms – the word is in a *monogamous* relationship with its meaning. The agent can be certain that if it remembers pineapple, then it must have seen the word "pineapple". From the above analysis, we derive two predictions for what makes a word memorable:

1. It will be easier to remember words that have few meanings relative to words with many meanings.
2. It will be easier to remember words that have few synonyms as opposed to words with many synonyms.

These two predictors are reminiscent of the "fan effect"[22], in which recognition times increase in proportion to the number of distinct attributes attached to a particular concept. Here we have a "fan" of associations between wordforms and meanings. Steyvers & Malmberg (2003)[19] also modeled word recognition memory as related to a form of fan effect. They proposed that words that occur in more diverse contexts leave more diffuse memory traces. Their model follows from the assumption that words are encoded in memory by features of their context, whereas ours follows from the assumption that words are encoded by their meanings (in reality, of course, both may be true, as suggested by Criss & Shiffrin 2001[20]). Below, we compare our model to that of Steyvers & Malmberg, and find that ours explains more variance in memory performance.

Unlike many models of word recognition memory, word frequency does not directly figure into our model. As discussed above, much previous research shows that rare words are better remembered in recognition tasks[1,2,3]. Here, this frequency effect is something of an epiphenomenon and falls out of the fact that words with few synonyms and few meanings tend to be rare.[23] makes a similar argument, showing that the frequency effect in recognition memory is related to the structure of the semantic space.

It is also worth noting that the prior term $P(W=w)$ is sensitive to the task characteristics. Under a different experimental set-up in which the probability of a word repeating was in some way related to its real world probability (i.e., words like *the* and *a* are repeated over and over), $P(W=w)$ would be higher for words with high real-world frequency. Here, however, we assume that it is uniform over all words since no word appears more than twice over the course of the experiment and the probability of a repeat is not sensitive to the probability of the word.

Experiments

We ran two large-scale behavioral experiments to test the factors that make a word memorable. Whereas most previous work on recognition memory was conducted with relatively small numbers of participants and items, we ran two experiments with 2,222 words each and tested over 1,000 participants in total. The large scale of this study makes these results more likely to generalize to other lexical items and other participants. The experiments were constructed as repeat detection tasks in which participants viewed a long series of words and were asked to press a key whenever they noticed a repeat. To ensure that participants were paying attention, vigilance repeats occurred at lags of 1-7 words. Critical repeats, used to measure word memorability, occurred at lags of 91-109 words. Approximately one out of every five words was a critical repeat. A schematic of our experiment is shown in Figure 2. For each word in the experiment, we empirically defined three measures of memorability: hit rate (proportion of trials on which a repeat was correctly detected), false alarm rate (proportion of trials on which a repeat was incorrectly claimed), and accuracy [(correct detections + correct passes) / (total number of times word was presented)].

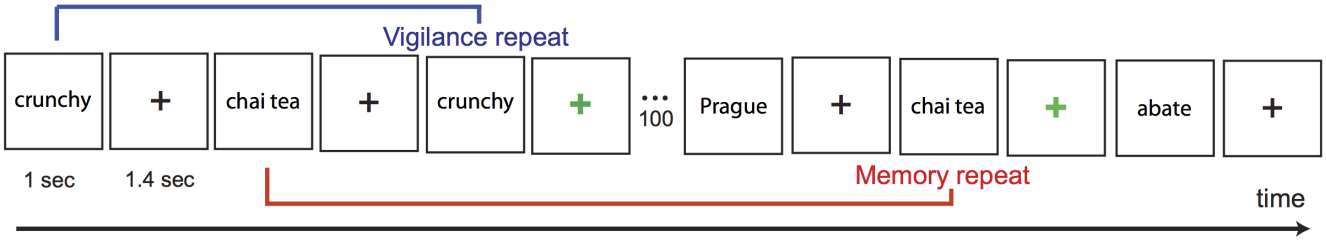


Figure 2. Schematic of the experimental set-up showing how often words repeat and how long words are shown.

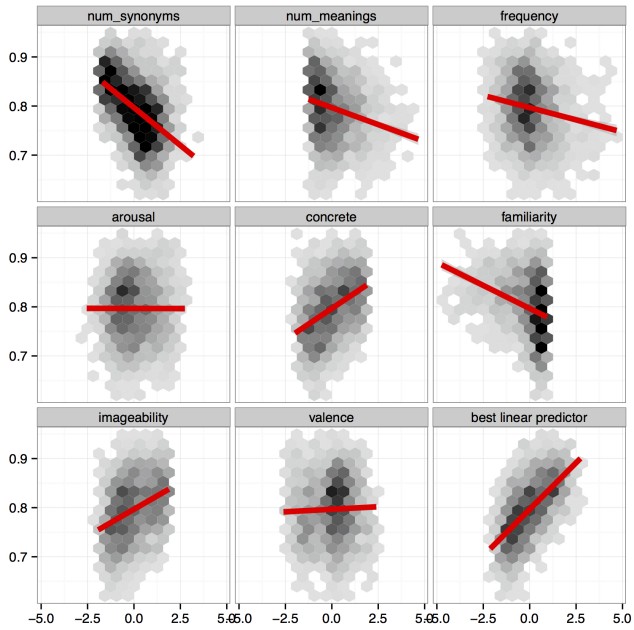


Figure 3. Correlations between z-scored attributes shown (x-axis) and task accuracy (y-axis). The darkness of the hexagons represents the amount of data at those coordinates such that darker areas of the plot have more data. The red lines are lines of best fit.

Experiment 1: Subtlex words

We measured the memorability of 2,222 words sampled from the Subtlex corpus[24], which consists of movie transcripts. The words sampled from this corpus are intended to represent a sampling of words that one might encounter in everyday speech–including a mix of low-frequency and high-frequency words. Figure 4 shows the most and least memorable words measured in our experimental task.

Overall memory performance was high. The mean hit rate over words was .68, the mean false alarm rate was .10, and the mean accuracy was .80. Although accuracy was high, some words were consistently better remembered than others (split-half Spearman correlation for accuracy across participants: .58 [95% CI of .56, .60 by non-parametric bootstrap]). This consistency indicates that there is a reliable signal of word-intrinsic memorability, which varies substantially between different words. This property can be said to be an intrinsic factor since it is stable over randomized observer and context in which the word is presented. Note that our results are consistent with item noise models of word memory, in which items are encoded by their intrinsic features (in our model, the features are word meanings). However, our experiment does not argue against an additional role for context noise. In all our experiments, context was randomized and averaged out of subsequent analysis.

To test our model predictions, we collected estimates of the number of meanings and number of synonyms for each word using an online survey as well as through Wordnet. All four measures showed robust Spearman rank correlations with recognition accuracy in the predicted directions, as summarized in Table 1. Table 1 also shows correlations between recognition memory and 5 other norms obtained through human survey (valence, imageability, familiarity, concreteness, and arousal) and 3

automatically obtained norms based on statistics of usage, measured over online corpora (GloVe uniqueness: a measure of semantic uniqueness derived from GloVe semantics; Subtlex contextual diversity: a measure of the number of distinct movie transcripts in which a word appears in the Subtlex database; Subtlex token frequency: the overall frequency of a word in the Subtlex movies transcript database). Note that Subtlex contextual diversity is the predictor proposed by Steyvers & Malmberg 2003[19]. As can be seen in Table 1, this norm alone only explains a portion of the variance; our new norms based on number of meanings and number of synonyms provide quantitatively better explanations.

Table 1. Spearman correlations for each predictor in Experiment 1

Predictor	Accuracy	Hit rate	False alarm rate
# synonyms (human rating)	-0.54	-0.45	0.26
# meanings (human rating)	-0.27	-0.16	0.27
# synonyms (Wordnet)	-0.37	-0.31	0.21
# meanings (Wordnet)	-0.39	-0.32	0.22
GloVe uniqueness	-0.39	-0.29	0.26
Subtlex contextual diversity	-0.26	-0.16	0.22
Subtlex token frequency	-0.14	-0.06	0.18
Valence	0.05	0.06	-0.02
Imageability	0.37	0.37	-0.05
Familiarity	-0.34	-0.22	0.28
Concreteness	0.44	0.39	-0.14
Arousal	-0.01	0.04	0.10

In order to predict memorability as a function of both number of synonyms and number of meanings, we fit a linear regression predicting per-word task accuracy as a function of these two factors. To avoid overfitting, we learned the model coefficients using half the participants and half the words and tested the model on the non-overlapping set of data such that the model was tested on only words and participants not included in the training set. Iterating this procedure 1000 times, the mean Spearman correlation between the memorability score and model prediction was .48 [95% CI .44, .52] out of a theoretical maximum of .58 (the split-half correlation across participants).

This simple model of memorability, based on just two rationally motivated factors – number of synonyms and number of meanings – captures a large portion of the variance in word memorability. We get similar results using estimates that do not rely on explicit human judgments (instead obtaining number of synonyms and meanings using Wordnet and computing contextual diversity using GloVe semantics and Subtlex).

To get the best possible linear predictor of memorability, we also included several other norms that are known to affect recognition memory: Subtlex frequency, valence (positive or negative), imageability (how easily the word produces a mental image), familiarity, concreteness, and arousal. Imageability and concreteness (which themselves are highly correlated) both correlate well with hit rate (.35 and .38, respectively), but neither is highly correlated with false alarm rate. Familiarity (which is highly correlated with frequency) performs qualitatively similarly to frequency and is likely measuring something similar. When we combine valence, imageability, familiarity, concreteness, and arousal with the 3 measures used in the rational analysis model to create a linear model, with the parameters learned on held-out training data as above, the Spearman correlation between memorability score and the model prediction is .57 – which is almost the theoretical maximum of .58. These factors

Most memorable words (highest accuracy): pineapple, Madonna, tampons, Eisenhower, yahoo, vagina, coccyx, potbelly, GPS, whorehouse, PSST, pi

Least memorable words (lowest accuracy): lacks, offer, among, transpired, handing, remained, fortunes, fought, remind, constantly, reluctance, concepts

Figure 4. Most and least memorable words in Experiment 1, based on accuracy.

thus appear to be explaining most of the explainable variance in the model.

In sum, the measures predicted by the model as likely to influence memory explain much of the variance in performance. When combined with other factors known to influence memory performance, almost all of the possible variance in our measurements is explained.

Experiment 2: Words across semantic categories

In order to replicate the results from Experiment 1 in a new set of words, as well as to test the effect of semantic categories, we ran the same experiment described in Experiment 1 using a new set of words that were hand-selected in order to have a wide range of possible semantic categories, such as chemical elements, geography, living people, and drinks. Figure 4 lists the most and least memorable words in this experiment. Notice that the same general trends apply as in Experiment 1: specific, precise words are on top, vague and interchangeable words are at the bottom. Despite the fact that Experiment 2 used a very different lexicon from Experiment 1, the results are strikingly similar. People's ability to remember words was extremely consistent with their performance in Experiment 1: mean hit rate .68; mean false alarm rate .10; accuracy .80. Again some words were consistently better remembered than others (split-half Spearman correlation for accuracy across participants: .65 [95% CI of .63, .67 by non-parametric bootstrap]). The explanation for why certain words were remembered better than others also follows the same pattern as in Experiment 1: monogamous words were again memorable, and the other norms we tested correlated with memorability in the same direction as in Experiment 1. The full set of correlations is shown in Table 2.

Following the same procedure as in Experiment 1, we fit a linear regression to predict recognition accuracy as a function of various norms. Using just number of synonyms and number of meanings we reach a Spearman correlation between the empirical memorability score and model prediction of .59 [95% CI .56, .62] out of a theoretical maximum of .63 (the split-half correlation across participants and items). Adding in all the additional word norms described in Experiment 1 raises the correlation to .63, which is almost the theoretical maximum of .65. The fact that these norms explain almost all the variance in both experiments, despite that the lexicons in the two experiments differed greatly, suggests that our findings are robust to context differences in the distribution of words encountered during the two experiments.

We also observed significant effects of syntactic category and semantic category. Nouns were remembered more easily (mean accuracy .83) than adjectives and verbs (mean accuracy .75 for both), potentially due to the increased concreteness of nouns. The most memorable semantic categories were famous landmarks (e.g., *Statue of Liberty*, *Machu Picchu*), games (e.g., *Simon Says*, *Legend of Zelda*), and common names (e.g., *Emmaneul*, *Jackie*). The least memorable categories were weather (e.g., *frost*, *heat wave*), building components (e.g., *dome*, *kitchen*), and general human nouns (e.g., *scribe*, *speaker*).

Experiment 2 replicates the results of Experiment 1. The model, trained on a held-out data set and cross-validated, explains most of the possible variance that can be explained in the data. Moreover, the measures of synonymy and homonymy, derived from rational analysis, are significant predictors of task performance.

Table 2. Spearman correlations for each predictor in Experiment 2

Predictor	Accuracy	Hit rate	False alarm rate
# synonyms (human rating)	-0.64	-0.56	0.32
# meanings (human rating)	-0.45	-0.35	0.31
Google Books log frequency	-0.21	-0.14	0.21
Valence	0.12	0.11	-0.02
Imageability	0.44	0.44	-0.11
Familiarity	-0.30	-0.20	0.28
Concreteness	0.57	0.53	-0.22
Arousal	-0.05	0.00	0.11

Discussion

In many past experiments, memory for words has been shown to be weaker than memory for pictures, a phenomenon known as the "picture superiority effect"[25,26]. Interestingly, in our experiments word memory was on par with picture memory performance in similar experiments[27]. Our experimental setup is nearly identical to that of Isola et al., 2011 ([27]), except that they used pictures as stimuli while we used words. Participants in their experiment correctly detected repeats 68% of the time and false alarmed 11% of the time. The corresponding numbers in both our current experiments are 68% and 10%. Clearly,

there does not appear to be an advantage for picture memory compared to word memory here. This perhaps suggests that encoding is happening, at least in part, at a conceptual level and is perhaps independent of modality.

The idea that people store meanings and not necessarily specific item forms is consistent with the long-standing idea that what is encoded in episodic memory is not just the word itself but the set of attributes associated with the word[28]. Specifically, our model is related to the generation-recognition theory of memory[29,30], which posits that the reason that semantic cues help recall of a target is because they produce a series of related responses, one of which is the target word. Similar effects have been found in visual memory. Konkle et al. found that the semantics of the image have a greater affect on memory performance than low-level perceptual features[31].

In this paper, we have provided an account of which words are memorable, and we have introduced tools that allow us to automatically predict how memorable a newly encountered word will be. We have additionally offered a simple theory, based on rational analysis, as to why these words are memorable. Our model posits that words are encoded in memory by their meaning, and this gives raise to competition during recognition between different words with the same meaning. As a result, words that are monogamous with their meaning are most memorable.

Materials and Methods

Participants

Participants were recruited using Amazon's Mechanical Turk crowd-sourcing platform. Only workers with a U.S. IP and an approval rating of $> 95\%$ were allowed to participate. 676 participants took part in Experiment 1 and 634 participants took part in Experiment 2. All experiments were conducted with approval from and in accordance with the Committee on the Use of Humans as Experimental Subjects at the Massachusetts Institute of Technology. Participants gave informed consent before starting each experiment.

Materials

Experiment 1: We took a frequency-weighted sample of 13,980 words from Subtlex[24], such that a word which is twice as frequent in Subtlex was twice as likely to be included in our sample. We then semi-manually went through the words to remove offensive words, alternate forms of the same word (color/colour), and words that were clearly morphologically related (happy/happiness). We then randomly sampled to have 11,182 words remaining for the experiment.

Experiment 2: We presented nouns with the word THE and verbs with TO in order to make it clear what part of speech was being used and capture any potential syntactic category effects.

Procedure

Each word was shown for 1 second followed by a 1.4 second fixation. Participants were asked to press the *r* key when a word occurred that they had already seen. Vigilance repeats (used to make sure that participants were paying attention) occurred at lags of 1-7 words. Critical repeats occurred 91-109 trials after the first presentation. Critical repeats only occurred on a random subset of 2,222 words used in the experiment. See Figure 2 for a summary of the experimental set-up.

Word-level statistics

Norms were obtained for each word in the data set in the following categories by asking for estimates from a separate set of Mechanical Turk workers: imageability, concreteness, arousal, familiarity, valence. Norms were calculated by averaging together the participants' ratings. After excluding participants who appeared to be guessing randomly instead of doing the task, we were left with an average of 26 ratings per word.

For the participant-supplied data on number of meanings and number of synonyms, we first asked participants to identify if the word was a word or not. There were a number of pseudo-words mixed in. Participants who did not correctly identify 80% of real words as words were removed from the task. 112 words in the task were identified as real words by 10 or fewer

Most memorable words (highest accuracy): Blondie, Long Island Iced Tea, panties, AIDS, R.E.M., The Dixie Chicks, Jennifer Anniston, David Hasselhoff, toilet bowl brush, mahi-mahi, Mike Tyson, Eminem
Least memorable words (lowest accuracy): cost, search, hurry, crowd, run, exchange, concern, shake, remain, disagree, leave

Figure 5. Most and least memorable words in Experiment 2, based on accuracy.

participants. These words were excluded from the analyses since many participants in the memory task likely would also not know these words. All critical analyses are qualitatively the same with or without excluding these words.

Thus, we obtained human ratings for norms in the following categories: *imageability, concreteness, valence, familiarity, arousal, number of meanings, number of synonyms*.

We also used corpus data to obtain several other estimates related to the probability of a word given a meaning (related to number of synonyms for a word) and the probability of a meaning given a word (related to number of meanings for a word).

We obtained a corpus-based frequency measure using Subtlex [1]:

- token frequency of word in the Subtlex subtitles database

For a corpus-based estimate for probability of a word given a meaning, we used the following:

- GloVe semantic word uniqueness. Using software to obtain co-occurrence vectors for words (GloVe: Global Vectors for Word Representation[32]), we obtained a vector of co-occurrence for each word in the database (pre-trained vectors from Wikipedia corpus, available at http://nlp.stanford.edu/projects/glove/). We then calculated the mean Spearman correlation between this vector and all other word vectors in our database. This correlation reflects, on average, how similar a given word w is to other words in the database in terms of its co-occurrence characteristics.
- Wordnet number of synonyms (number of synonyms assigned to a word in Wordnet)

For a corpus-based estimate for probability of a meaning given a word, we used the following:

- Subtlex contextual diversity (the unique number of movies in which a word appears–words that have more meanings typically appear in more diverse settings)
- Wordnet number of meanings (number of meanings listed for a word in Wordnet)

The human ratings for number of synonyms, the GloVe semantic uniqueness measure, and the Wordnet number of synonyms all reflect quantities intended to approximate probability of a word given its meaning. The human ratings for number of meanings, the Subtlex contextual diversity measure, and the Wordnet number of meanings all reflect approximations for the probability of a meaning given a word.

For Experiment 2, 52 words, which were known to 10 or fewer raters, were excluded. Because the items in this experiment were sometimes phrases, we were not able to obtain Subtlex or other corpus-based data on the items in this experiment. We did, however, use the Google Ngram corpus to calculate log frequencies for each item.

References

1. Garlock, V. M., Walley, A. C. & Metsala, J. L. Age-of-acquisition, word frequency, and neighborhood density effects on spoken word recognition by children and adults. *Journal of Memory and language* **45**, 468–492 (2001).
2. Jescheniak, J. D. & Levelt, W. J. Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **20**, 824 (1994).
3. Kinsbourne, M. & George, J. The mechanism of the word-frequency effect on recognition memory. *Journal of Verbal Learning and Verbal Behavior* **13**, 63–69 (1974).
4. Lohnas, L. J. & Kahana, M. J. Parametric effects of word frequency in memory for mixed frequency lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **39**, 1943–1946 (2013).
5. Klaver, P. *et al.* Word imageability affects the hippocampus in recognition memory. *Hippocampus* **15**, 704–712 (2005).
6. Paivio, A. Mental imagery in associative learning and memory. *Psychological review* **76**, 241 (1969).
7. Walker, I. & Hulme, C. 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**, 1256 (1999).
8. Maratos, E. J., Allan, K. & Rugg, M. D. Recognition memory for emotionally negative and neutral words: An erp study. *Neuropsychologia* **38**, 1452–1465 (2000).
9. Kensinger, E. A. & Corkin, S. Memory enhancement for emotional words: Are emotional words more vividly remembered than neutral words? *Memory & cognition* **31**, 1169–1180 (2003).

[1] For all measures from Subtlex, we use case-insensitive measures.

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

11. Jacoby, L. L. & Dallas, M. On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General* **110**, 306 (1981).

12. Light, L. L. & Carter-Sobell, L. Effects of changed semantic context on recognition memory. *Journal of verbal learning and verbal behavior* **9**, 1–11 (1970).

13. Bahrick, H. P. Measurement of memory by prompted recall. *Journal of Experimental Psychology* **79**, 213 (1969).

14. Bransford, J. D. & Franks, J. J. The abstraction of linguistic ideas. *Cognitive psychology* **2**, 331–350 (1971).

15. Franks, J. J. & Bransford, J. D. The acquisition of abstract ideas. *Journal of Verbal Learning and Verbal Behavior* **11**, 311–315 (1972).

16. Begg, I. & Wickelgren, W. A. Retention functions for syntactic and lexical vs semantic information in sentence recognition memory. *Memory & Cognition* **2**, 353–359 (1974).

17. Shiffrin, R. M. & Steyvers, M. A model for recognition memory: Rem—retrieving effectively from memory. *Psychonomic Bulletin & Review* **4**, 145–166 (1997).

18. Dennis, S. & Humphreys, M. S. A context noise model of episodic word recognition. *Psychological review* **108**, 452 (2001).

19. Steyvers, M. & Malmberg, K. J. The effect of normative context variability on recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **29**, 760 (2003).

20. Criss, A. H. & Shiffrin, R. M. Context noise and item noise jointly determine recognition memory: a comment on dennis and humphreys (2001). (2004).

21. Annis, J., Lenes, J. G., Westfall, H. A., Criss, A. H. & Malmberg, K. J. The list-length effect does not discriminate between models of recognition memory. *Journal of Memory and Language* **85**, 27–41 (2015).

22. Anderson, J. R. & Reder, L. M. The fan effect: New results and new theories. *Journal of Experimental Psychology: General* **128**, 186 (1999).

23. Monaco, J. D., Abbott, L. & Kahana, M. J. Lexico-semantic structure and the word-frequency effect in recognition memory. *Learning & Memory* **14**, 204–213 (2007).

24. Brysbaert, M. & New, B. Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods* **41**, 977–990 (2009).

25. Paivio, A. Imagery and verbal processes. (1971).

26. Standing, L. Learning 10000 pictures. *The Quarterly journal of experimental psychology* **25**, 207–222 (1973).

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

28. Tulving, E. & Thomson, D. M. Retrieval processes in recognition memory: Effects of associative context. *Journal of Experimental Psychology* **87**, 116 (1971).

29. Bahrick, H. P. Two-phase model for prompted recall. *Psychological Review* **77**, 215 (1970).

30. Martin, E. Generation-recognition theory and the encoding specificity principle. (1975).

31. Konkle, T., Brady, T. F., Alvarez, G. A. & Oliva, A. Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General* **139**, 558 (2010).

32. Pennington, J., Socher, R. & Manning, C. D. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)* **12** (2014).

Acknowledgements

We thank members of Tedlab, Josh Tenenbaum and members of Cocosci, members of the Computational Vision and Perception Lab, Sam Gershman. K.M. was supported by the Department of Defense through an NDSEG graduate fellowship. We thank Barbara Hidalgo-Sotelo, Henrison Hsieh, and members of TedLab for her help with constructing the materials for Expt 2. E.F. was supported by NIH award HD-057522. This work was partly supported by research awards from Google and Xerox to A.O.

Author contributions statement

All authors conceived the experiments, P.I., K.M., E.F., and E.G. generated the materials and conducted the experiments, K.M. and P.I. analyzed the results. All authors wrote the manuscript.

Competing financial interests

The authors declare no competing financial interests.