Research

Lexico-semantic structure and the word-frequency effect in recognition memory

Joseph D. Monaco,[1,3] L. F. Abbott,[1] and Michael J. Kahana[2]

[1]*Center for Neurobiology and Behavior, Department of Physiology and Cellular Biophysics, Columbia University College of Physicians and Surgeons, Kolb Research Annex, New York, New York, 10032-2695, USA;* [2]*Department of Psychology,* University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA

The word-frequency effect (WFE) in recognition memory refers to the finding that more rare words are better recognized than more common words. We demonstrate that a familiarity-discrimination model operating on data from a semantic word-association space yields a robust WFE in data on both hit rates and false-alarm rates. Our modeling results suggest that word frequency is encoded in the semantic structure of language, and that this encoding contributes to the WFE observed in item-recognition experiments.

Old–new item recognition is the task of deciding whether or not test items were presented on a previous study list. Performance is quantified as the probability of old responses to (old) study items (hit rate [HR]) and to (new) nonstudy items (false-alarm rate [FAR]). One of the most prominent phenomena observed in this task is the word-frequency effect (WFE): rare or low-frequency (LF) words are better recognized than common or high-frequency (HF) words (Schulman 1967; Shepard 1967). The recognition WFE is a mirror effect (Glanzer and Adams 1985, 1990): it consists of an HR effect and an opposite, but approximately equal FAR effect. The cause of the WFE and other mirror effects has been the subject of extensive study, but no consensus view has been established (e.g., Murdock 1998; Stretch and Wixted 1998; Reder et al. 2000).

Both single- and dual-process models have been proposed to explain the WFE. The former perform familiarity discrimination (FD) based on similarity measures such as global feature matching. These models typically require some additional transformation, such as log-likelihood computation, to achieve the required symmetry between old- and new-item familiarity distributions (Murdock 1998). To explain the WFE, certain differences between LF and HF words must be assumed. These may include the modulation of attentionally marked features (Glanzer et al. 1993), diagnostic content (Shiffrin and Steyvers 1997), or representative feature variability (McClelland and Chappell 1998). In the end, such models produce a unidimensional scalar value for the strength or familiarity of a given stimulus that allows further analysis with signal-detection theory. HR and FAR calculations can be made by integrating thresholded familiarity distributions, and threshold-independent performance may be quantified with receiver operating characteristics (ROCs) (see Wickens 2002). Dual-process models, however, rely on differential contributions of recollective and familiarity-based processes to explain the performance differences. Recollection, a recall-like process, is characterized as less error prone than a global-matching familiarity process (Guttentag and Carroll 1997; Reder et al. 2000).

In humans, both familiarity and recollection appear to depend on the medial temporal lobe (MTL) (see Levy et al. 2004; de Zubicaray et al. 2005). Electrophysiological studies in monkey have shown perirhinal cortex (PRC), specifically, to have a substantial proportion of familiarity-sensitive neurons (Miller et al. 1991; Li et al. 1993; Xiang and Brown 1998; Brown and Bashir 2002). Theoretically, it is known that a familiarity signal can be read out from a simple autoassociative neural network by computing its internal energy (Amit 1989). Indeed, the evaluation of network energy may approximate the familiarity signal evident in perirhinal neurons (Bogacz et al. 2001a; Brown and Bashir 2002) and has been used to determine theoretical limits on recognition capacity (Bogacz et al. 2001b; Bogacz and Brown 2002). Thus, we set out to create a recognition model that uses network energy as a readout of stimulus familiarity. For this purpose, we used input vectors from a word-association space (WAS) (Steyvers et al. 2004). The WAS is an empirical model of semantic similarity based on normative data from free-association experiments (Nelson et al. 2004). Simulating old–new recognition experiments with this model, we found that word frequency produces discriminable signal distributions such that LF words tend to be more familiar than HF words. Further, coupling this output with a particular decision-making strategy exhibited a WFE mirror effect. These results have novel implications for the roles of distinct retrieval processes in recognition memory.

Model

We present a simple item-recognition model, where the familiarity of a probe stimulus is read out as the internal energy of a network trained on a set of activity vectors corresponding to WAS word representations. This is coupled with an experimental protocol emulating a typical word-recognition experiment (see Methods: Experiment simulation). Importantly, all study and test words are trained initially and then followed by retraining of the study list. Retraining corresponds here to the subject having recently experienced a word in the context of an experimental study list.

Familiarity as network energy

In the Bogacz et al. (2001a) FD model, item vectors are associatively encoded into a Hopfield network (Hopfield 1982). The familiarity signal is simply the internal energy of the network when activated with a probe stimulus (Bogacz et al. 2001a). Hopfield networks are fully connected recurrent networks of binary units. The network weights are trained on an input set ξ_N^P of P N–dimensional activity vectors, such that $\xi_i^\mu \in \{-1,+1\}$ for all $i \in \{1...N\}$ and $\mu \in \{1...P\}$. That is, each unit is either active (+1) or inactive (–1) for a given input vector. If we denote the weight matrix as $W = [w_{ij}]_{i,j=1}^N$, its elements are computed according to an associative Hebbian learning rule:

[3]Corresponding author.
E-mail joe@neurotheory.columbia.edu; fax (212) 543-5410.
Article is online at http://www.learnmem.org/cgi/doi/10.1101/lm.363207.

14:204–213 ©2007 by Cold Spring Harbor Laboratory Press ISSN 1072-0502/07; www.learnmem.org

$$w_{ij} = \frac{1}{N}\sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu = \frac{1}{N}\,\xi_i \cdot \xi_j, \quad \text{for } i \neq j, \tag{1}$$

where $w_{ii} = 0$ for $i \in \{1...N\}$. Once trained, we are only interested in the internal energy of the network when presented with a given stimulus, so no network dynamics are involved here. This internal energy calculation is distinct from recollective processes that use some form of network relaxation to fully recall the features of stored items (Amit 1989). For a probe stimulus vector $X = [x_i]_{i=1}^N$, the internal energy[1] is computed as

$$\mathbf{E}(X) = -\frac{1}{2}\sum_{i=1}^{N} x_i \sum_{j=1}^{N} x_j w_{ij} = -\frac{1}{2} X \cdot W \cdot X^T. \tag{2}$$

Note that more familiar stimuli will have lower energies than less familiar stimuli. A probe X will thus be associated with the familiarity quantity $\mathbf{E}(X)$ for a network trained on a given input set. In this form, the only free variables of the FD process are the size of the network, N, and the set of input vectors, ξ_N^P. FD such as this is more efficient and has a much higher capacity than associative recall. Allowing an error rate up to 0.01, the recall capacity of the network is $0.145N$ (Amit 1989), whereas its recognition capacity is $0.023N$[2] (Bogacz et al. 2001b).

Semantically structured input

Recognition models operating on correlated input spaces (Bogacz and Brown 2003; Norman et al. 2005) have been studied that benchmark behavioral data (Norman and O'Reilly 2003). However, recent empirical models of semantic space, such as the WAS model of Steyvers et al. (2004) provide a basis for constructing an input set with a similarity structure derived from behavioral word-association data. The basis for the WAS is a free-association data set containing the probabilities with which subjects named a given word as the first associate of a cue word (Nelson et al. 2004). These data can be taken as a measure of direct associative strength among over 5000 words. Indirect or second-order associative strengths can also be calculated from the data set. To create the WAS, singular-value decomposition (SVD) was applied to these direct and indirect associations so as to place words in a reduced 400-dimensional space. This was constrained so that the cosine between any two-word vectors[2] reflects their mutual associative strength. The dimensional reduction revealed latent, higher-order semantic relations within the data set. Importantly, 400 dimensions were found to be the lowest dimensionality that remains highly predictive of experimental data such as free-recall intrusion rates, extralist cued recall, and semantic similarity ratings in recognition (Steyvers et al. 2004). The resultant WAS shares gross structural characteristics, small-world but not scale-free, with other semantic networks (Steyvers and Tenenbaum 2005). Thus, it is now possible to operate on input vectors whose similarity relations approximate the lexico-semantic space of English speakers.

Here, we used a set of 1748 WAS vectors for which we have the associated Kučera-Francis word frequency (WF) (see Kučera and Francis 1967). In the cosine similarity matrix for the 100 most common and the 100 most rare words in the set (Fig. 1A), it is evident that common words tend to be similar to other common words and rare words tend to be similar to other rare words. Similarities between rare and common words tend to be lower than similarities within frequency groups. An intuitive reason for such differential encoding of frequency is that rare words tend to have a single definition, while common words may have many definitions and usages. This reasoning predicts that rare synonyms will be clustered in semantic space, whereas more common synonyms will be placed at a semantic "centroid" of multiple distinct meanings. That is, LF words will be encoded into clusters and HF words will tend to occupy the space between such clusters. We will refer to this as the "tight clustering" hypothesis for LF words.

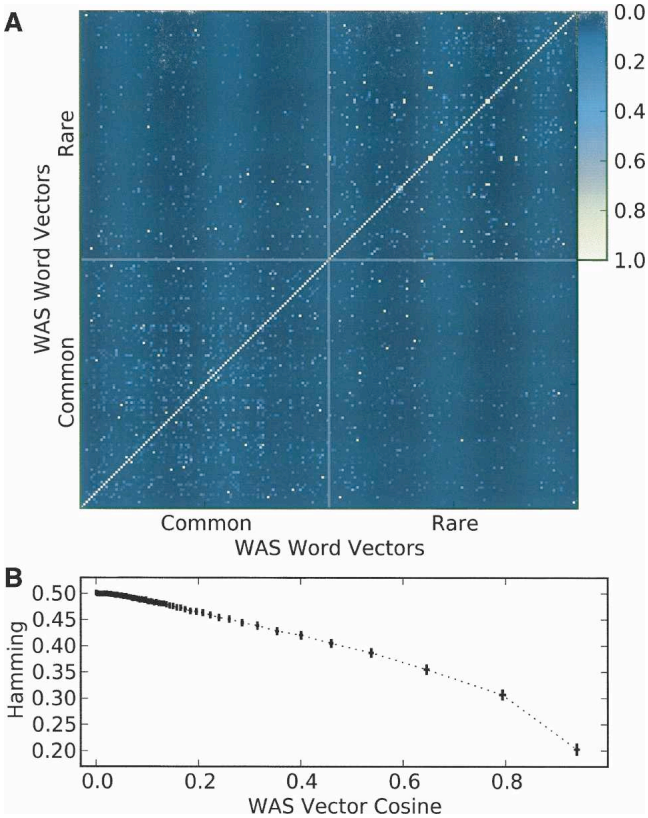

Figure 1. (*A*) Similarity matrix of the WAS vectors for the 100 highest and lowest frequency words in the set. The color of each pixel denotes the cosine of the angle between the *i*th and *j*th vectors. The *bottom-left* and *top-right* quadrants represent the cosine similarity among pairs whose members are both HF or LF words, respectively. The white diagonal signifies identity. The symmetric off-diagonal quadrants represent cosines between HF and LF words. (*B*) Normalized Hamming distances between binarized WAS vectors decrease monotonically with the cosine of the corresponding WAS vectors. Every point represents the mean and error ($\alpha = 10^{-5}$) for each bin in a cosine-sorted partition (600 bins) of all vector pairs.

Interpreting the recognition model

The two functional components of our FD model are the Hebbian learning of a Hopfield network and the dimensional reduction of a word-association matrix. In considering this combination, we have to interpret the necessary combination of the assumptions inherent in both. We must properly frame the limitations of the results and emphasize that they comprise, at most, a high-level explanation. A simple network computation on a carefully chosen input set will not explain the intricacies of human recognition memory for semantic stimuli, yet may provide insight into some aspect thereof. We argue for the specificity and functional plausibility of the model components as well as their composition.

First, we assume that using the energy (Equation [2]) of a

[1] Here, we let $\mathbf{E}(\cdot)$ be the function mapping a stimulus or set of stimuli to an energy value or energy distribution, respectively. Statistical expectations are noted by $\langle\cdot\rangle$ brackets.

[2] A vector cosine is the inner product between V and U normalized to [0, 1], $\cos(\theta_{V,U}) = (V \cdot U)/(\|V\|\,\|U\|)$.

trained Hopfield network (Equation [1]) as a familiarity signal captures some salient characteristic of familiarity processing in the perirhinal cortex. Bogacz et al. (2001a,b) provide support for this assumption by arguing from a standpoint of functionality and efficiency as well as from modeling results. Also, neurons responding differentially to familiar stimuli have been found consistently within monkey PRC (Miller et al. 1991; Li et al. 1993; Sobotka and Ringo 1993). This difference is characterized by a reduction in stimulus-induced activity for familiar stimuli and rapid familiarity discrimination (on the order of 100 msec), but neural responses for recency and novelty have also been found (Fahy et al. 1993; Xiang and Brown 1998; Brown and Bashir 2002). Despite this functional diversity, only familiarity-sensitive neurons are considered here. Furthermore, evidence from ablation and impairment studies indicate that the PRC acts independently from other inferotemporal (IT) mnemonic systems such as that of the hippocampal formation (Gaffan 1994; Aggleton and Brown 1999; Murray and Bussey 1999). This independence suggests that PRC is the site of the neural substrate for familiarity judgments (for reviews, see Yonelinas 2002; Rugg and Yonelinas 2003).

Autoassociative neural networks such as the binary Hopfield network produce stimulus-dependent attractors (Hopfield 1982). Reading out the internal network energy as a stimulus-familiarity signal is much more efficient than involving recollective processes (Bogacz et al. 2001a; Bogacz and Brown 2003), which typically involve relaxing the network to reconstruct an attractor state (Murdock 1982; Humphreys et al. 1989). For random vectors, this FD process has a very high storage capacity, as does human memory (Standing 1973), and enables a rapid network response. It is also more robust than other network architectures; e.g., encoding via feedforward competitive synaptic processes can exhibit forgetting after a relatively small number of subsequent stimuli (Sohal and Hasselmo 2000). Bogacz et al. (2001b) use both a Hopfield network and a multilayer spike-response model to argue that perirhinal FD neurons may form an autoassociative network in order to exhibit such efficiency and robustness. From this, we posit that it is reasonable that the energy computation of a Hopfield network is, at least, a useful abstraction of the FD processing performed by PRC neurons.

Second, we assume that the WAS is at least approximately isomorphic to the space of neural representations of the semantic features of words for speakers of English. This amounts to the assumption that behavioral associativity refects the neural encoding of semantic similarity. That the WAS serves well as a predictive model for known human memory effects recommends it as a useful inference of semantic space. Further, the WAS has structural characteristics consistent with being isomorphic to real semantic representations. Thus, this assumption is assuredly a simplification, but it is likely to yield salient semantic information.

Third, we assume that the semantic WAS vectors serve as appropriate inputs to the FD model of PRC. This assumption allows us to posit the combination of the two components as a unified model of recognition memory. Supporting this, several clinical studies indicate a role for PRC in associative memory for semantic content and lexical processing (for review, see Murray and Bussey 1999). Further, neurons in the perirhinal and other IT areas in monkeys demonstrate the ability to represent abstract object categories (Erickson et al. 2000; Miller 2000; Miller et al. 2003). Such abstraction is a hallmark of semantic information processing and indicates that the PRC has access to semantic features among its inputs.

Finally, although we investigate the recognition WFE with this model, it can only describe effects due to familiarity processing of semantically structured input data. There are certainly nonsemantic contributions to the WFE that are not within this scope; e.g., context variability, and orthographic and phonological features (Malmberg et al. 2002; Steyvers and Malmberg 2003). Qualified as such, we will refer to this bipartite recognition model as WAS–FE.

Energy and semantic attractors

If a meaningful stimulus is ultimately represented as a binary pattern of activation across N perirhinal neurons, then we can think of this stimulus as an N-dimensional vector of features that are either present in the stimulus (+1) or not (-1). In Hopfield learning (Equation [1]), these component features are pairwise associated according to their correlation: the strength of the synapse between two neuronal units is directly and linearly related to the number of patterns for which the units carry the same activity. Synaptic weights are simply inner products of across-pattern activity vectors. Internal network energy (Equation [2]), then, is an outer product measure of how well the pairwise bit structure of a given activity vector aligns with the pairwise correlations stored in the weights of the network. This is opposed to recognition models based on summed similarity or global matching (Shiffrin and Steyvers 1997; Zaki and Nosofsky 2001; Kahana and Sekuler 2002; Kahana et al. 2005). However, a summed-similarity recognition model using inputs derived from perceptual preprocessing of natural stimuli has been able to match experimental similarity and recognition data (Lacroix et al. 2006).

Small-world structure is characterized by short minimum-path lengths, but also by hub-like connectivity (Watts and Strogatz 1998). Considering the WAS as a small-world network (Steyvers and Tenenbaum 2005), there must be subsets of vectors that significantly share pairwise activity. Each of these groups, or clusters, will bias those synaptic weights corresponding to their respective set of shared features. Clusters of feature-sharing vectors will form attractors commensurate with their size and mutual similarity. Thus, a probe vector may yield a low energy by matching features characteristic of different attractors in the network: there is a combinatoric aspect to the diversity of such "spurious" attractors (Amit 1989). The strongest attractors, though, will correspond to groups of words with substantial semantic similarities.

Results

Initially, we binarized[3] the WAS vectors in our wordset. This allows proper operation of the Hopfield learning rule (Equation [1]) and energy computation (Equation [2]). The normalized Hamming distances between these binary vectors decreased monotonically with cosine similarities of the corresponding continuous WAS vectors (Fig. 1B). This indicates that the binarization significantly preserves similarity relations between vectors.

Random inputs

In the initial item-recognition experiment we used an input set of unbiased random vectors. The energy distributions for old and new items were binomial with means of -399 and -199 (Fig. 2A, left column), matching theoretical means[4] of $\mu_\Lambda = -400$ and $\mu_\Phi = -200$. Study lists consisted of 100 vectors ($L = 100$) and results using two reference pool sizes are shown in Figure 2A: $P = 400$ (top row) and $P = 1600$ (bottom row). We refer to these as the low-load and high-load training conditions, respectively. They demonstrate the effects of adding noise to the network in

[3] The elements of WAS vectors are symmetrically distributed around zero, so that taking the sign of each vector element produces a set of binary vectors with, on average, unbiased activity levels.

[4] These means are offset —$N/2$ due to the initial training of both study and test vectors (see Methods: Experiment Simulation).

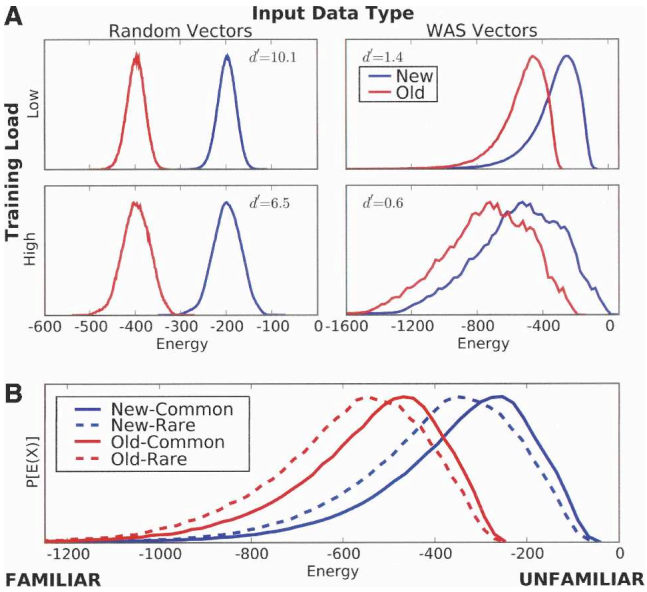

Figure 2. Effects of semantic input vectors and training load on resultant energy distributions. (*A*) Increasing training load for both random (*left*) and semantic (*right*) vectors increases overlap between energy distributions (100 study items, 400/1600 new items for low/high (*top*/*bottom*) training load). The energies for semantic inputs, however, have load-dependent means, non-Gaussian distributions, and worse discriminability than in the random case. (*B*) WF-sorted partitioning of word vectors results in discriminable familiarity distributions (150 study items, 600 new items). The LF distributions are more familiar than HF words for both old and new items. The "rare" and "common" bins here are the least and most frequent thirds of the lists, respectively.

the form of additional stored vectors. In all cases, the old and new distributions have equal variance. The low- and high-load conditions had standard deviations s_E of 20 and 31, respectively. This increase in spread decreases d' from 10.1 to 6.5 (Equation [4]) in the high-load condition; however, both values indicate perfect discrimination. Although we could have degraded the model's performance by reducing the magnitude of the weight update for study-list items (Equation [5]), these results serve as an input control for the simulations below.

Semantic inputs

The energy distributions resulting from the semantic input set (Fig. 2A, right column) differ substantially from the random input condition. The distributions are non-normal, negatively skewed (i.e., biased toward increasing familiarity), and their statistics have changed significantly. The means are lower than those in the random input case. In the low-load condition, mean old and new energies are -524 and -326, respectively, and the high-load case shows -777 and -578, respectively. So, not only are the distributions exhibiting enhanced familiarity, the means are load dependent. The more semantic vectors we store in the network, the more negative the energy distributions become. Furthermore, they exhibit much lower d' distances than in the random case. The discriminability as measured by d' decreases from 1.4 at $P = 400$ (Fig. 2A, top right) to 0.61 at $P = 1600$ (Fig. 2A, bottom right). This 57% reduction in separation compares with a corresponding 36% decrease for random inputs. Finally, the high-load condition produces energy distributions with noise-like irregularities that are not evident in the other cases. These were not investigated, but they may be the result of capacity effects or structural heterogeneity of the input space.

Statistical changes such as these could be expected for any sufficiently nonrandom input set. However, there are systematic differences in the energy distributions among WF classes. We found that vectors representing LF words tend to have lower energies, and thus enhanced familiarity, than those of HF words (Fig. 2B). This was observed for frequency classes in both old and new energy distributions. Figure 2B shows the distributions for the thirds of the study list and reference pool with the highest and lowest frequencies. This effect of increasing familiarity with decreasing WF was observed robustly across the full range of possible list and reference pool sizes. For the data shown here, based on $L = 150$ and $P = 600$, all four distributions exhibited standard deviation $s_E = 192$, and both WF-dependent effects were discriminable at $d' = 0.23$. This effect is present in the new distribution and, as such, does not depend on item study (Equation [5]).

ROCs computed from the semantic familiarity distributions in Figure 2A are presented in Figure 3. The WF dependence of the ROCs is shown for both the low-load (Fig. 3A) and the high-load (Fig. 3B) conditions. That is, for each training condition, the "common" and "rare" ROCs compare Λ_1 (old–HF) and Λ_6 (old–LF), respectively, to the reference pool. These are the distribution comparisons used to assess item-recognition performance (see below and Discussion). In both conditions, LF words yield better old–new discrimination than HF words. There are two load effects. First, the low-load ROCs (Fig. 3A) indicate better overall performance, evident as higher HRs and lower FARs, than the high-load ROCs (Fig. 3B). Second, the WF-dependence of effect is greater in the high-load than in the low-load condition. That is, the ROCs in Figure 3B are more separated than those in Figure 3A. Both load effects are a result of the increase in energy variance and decrease in d' distances evident in Figure 2A. In the low-load condition, the d' distances are 1.2 and 1.5 for common and rare words, respectively. For high-load, the relatively low d' of 0.34 for common words more than doubles to 0.78 for rare words. This Hopfield FD model has a theoretical recognition capacity of 3.7×10^3 random vectors (Bogacz et al. 2001b). Here, storing 1.7×10^3 semantic vectors is severely detrimental to FD performance, indicating that the correlations inherent in the semantic inputs reduce the effective capacity of the network (Appendix C of Bogacz and Brown 2003).

Word frequency effect

We observed the WFE mirror effect across the possible range of the list lengths for the study list and reference pool. We based recognition performance on a WF-based decision criterion (see Discussion). Mean HR–FAR trends for the low-load condition are

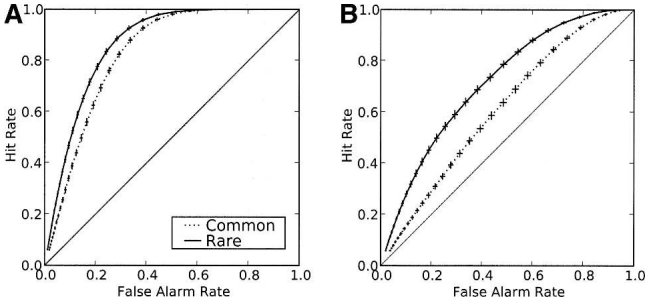

Figure 3. Operating characteristics for item-recognition performance under low (*A*; $P = 400$) and high (*B*; $P = 1600$) training load. The study list is composed of 100 items in both conditions. Performance across word frequency is assessed using a six-partition of the study list indexed and sorted by Kučera-Francis frequencies. The common and rare ROC curves represent the performance of the highest and lowest frequency bins, respectively. The $P = 400$ case (*A*) demonstrates better baseline performance but a smaller frequency effect than the $P = 1600$ case (*B*).

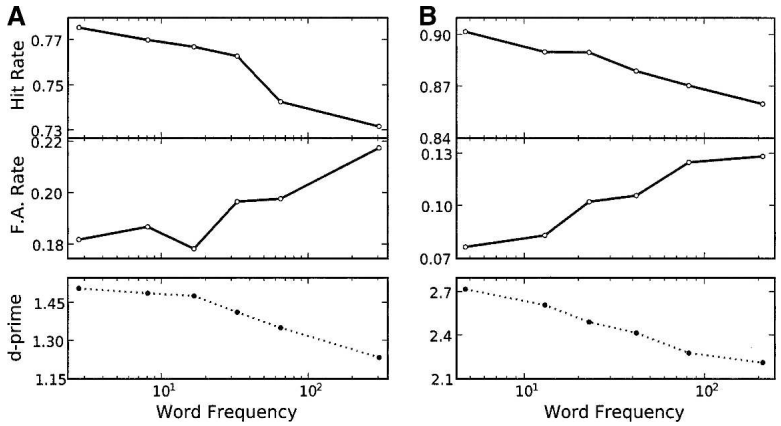

Figure 4. Word-frequency mirror effect from 2000 trials of item-recognition experiment simulations (*A*; $L = 100$, $P = 400$) and as seen in human memory experiments (*B*; data from Schwartz et al. 2005). The HR and FAR effects compose the mirror effect (*top*) and are due to changes in discriminability (*bottom*). Model HR, FAR, and d' data have 95% confidence intervals of mean $\pm$ 4.2 $\times 10^{-3}$, 1.5×10^{-3}, and 1.2×10^{-2}, respectively. The discriminability of the experimental data was estimated from signal detection theory.

shown in Figure 4A with d' distances. In this condition, HR decreases from 0.77 for the lowest-frequency words to 0.73 for the highest-frequency words. Similarly, the FAR increases from 0.18 to 0.22. Human recognition data collected by Schwartz et al. (2005) are shown for comparison in Figure 4B. The experimental d' distances (Fig. 4B, bottom) were calculated using an unbiased estimator from detection theory (Wickens 2002). The experimental HR decreases from 0.90 to 0.86, while the FAR increases from 0.077 to 0.13. These data approximately match the trends observed in the model data.

Note, however, the differences in absolute magnitude of the HRs, FARs, and d' distances between the model and experimental data in Figure 4. The absolute d' distances could be manually tuned with the addition of a coefficient in the study rule (Equation [5]), but we chose not to do this. Scaling up the model d' data would increase HRs and decrease FARs to better match the experimental data. For our purposes, it is sufficient that we observe a qualitatively correct WFE.

Semantic clustering

The above supports the tight clustering hypothesis for LF words, so we performed two simple neighborhood analyses of the continuous WAS data set. Consider an even partition of the entire WAS such that each bin contains a distinct WF class. Figure 5A shows, for each of the WF bins, the mean neighborhood population counts for word-centered hyperspheres of varying radii in cosine space. We count all neighbors, regardless of word frequency. Counts are shown for radii up to $d_{cos} = 0.062$, as that is sufficient to illustrate the WF dependence of the number of close neighbors as d_{cos} approaches zero. The more rare words (blue solid line) have more neighbors on average than common words (red lines) for most of this range, and especially around $d_{cos} = 0.01$–0.02. This indicates that LF words tend to have more close neighbors than HF words.

Next, Figure 5B addresses the WF composition of those neighbors. For each vector ξ, we compute the quantity $\nu(\xi)$ as an average of the WF of its neighbors (Equation [6]). The distributions of these values for each frequency class are shown as 15-bin histograms in Figure 5B. The LF and HF distributions have means of 1.13 and 1.74, respectively, and a distance of $d' = 0.93$ standard deviations. From Figure 5A, LF words tend to have closer neighbors than HF words. Given the null hypothesis that semantic similarity and WF are not correlated, the $\cos(\theta)$ coefficients in Equation (6) would dictate that these distributions be orientated oppositely from those in Figure 5B. That is, even though LF words have more high-similarity neighbors—so that neighbor WF values are more strongly weighted—$\nu(\xi)$ is distributed to lower WF than those of common words. Thus, the dominant factor in these $\nu(\xi)$ averages must be the WF component, indicating that LF words are tightly clustered with other LF words, whereas HF words are more diffusely distributed.

Discussion

Here, we bring together a simple model of familiarity-based recognition (Bogacz et al. 2001a,b) and a recent model of semantic similarity (Steyvers et al. 2004) and demonstrate a word-frequency effect. The only free parameters for the resulting model (WAS–FE) are the lengths of the study and test lists, across which the WFE is robust, only differing in magnitude. Notably, and as discussed below, the observed WFE is a mirror effect when decisions are determined using a stimulus-dependent criterion.

Input structure effects

The input-type effect evident in Figure 2A is attributable to nonrandom structure of the semantic input space. From the low means for the new distributions, we can infer that the vectors in the semantic input space tend to be near network attractors. The presence of spurious attractors, as well as learned attractors, further contributes to lower energies across the space. In fact, the large systematic decrease in probe energies indicates that much of the input space is likely spanned by basins of attraction. The observed shape of the energy distributions, then, is a function of the number, density, and spatial distribution of vector clusters in WAS.

Vectors populating a space with large-scale structure carry redundant information in their correlations. For the Hopfield FD mechanism, it can be shown that input redundancy reduces the effective capacity of the network (Bogacz and Brown 2003, Appendix C). This is evident in the large drop in discriminability between the random and semantic input spaces and between the low- and high-load training conditions (see d' in Fig. 2A). So, processing raw semantic information is inefficient, but results in much more realistic (i.e., measurably worse than perfect) recognition performance. Stimulus decorrelation is thought to occur downstream of IT cortex in the dentate gyrus (O'Reilly and McClelland 1994; Kesner et al. 2000), so presumably the FD neurons in PRC have access to the original stimulus features.

Word-frequency effects

On the hypothesis that the WFE is supported by semantic coding differences and thus might be evident within WAS–FE, we sorted and partitioned the task lists into WF bins using a normative frequency measure (Kučera and Francis 1967). These WF-differentiated bins resulted in separable energy distributions for both old and new lists (Fig. 2B), an effect observed robustly across list length. Notably, this frequency effect is in the observationally correct direction of increased familiarity for rarer words. That is, as in Figure 2B, the energies for each "Rare" (LF) bin are distributed more negatively than those of the corresponding "Common" (HF) bin.

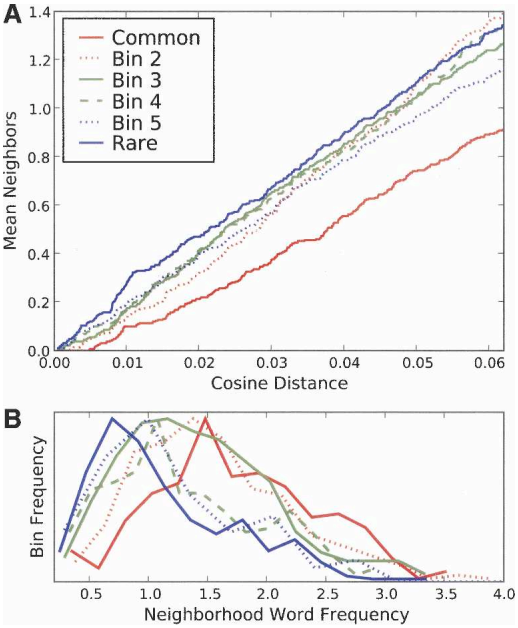

Figure 5. Word frequency-dependent clustering of word vectors in the WAS. Using a six-partition of the word set, the mean population size for a neighborhood of a given radius in cosine space shows that rare words have more close neighbors than common words (*A*; the abscissa is $d_{cos} = 1 - \cos(\theta)$. The WF composition of those neighbors is indicated by the relative distributions of WF–cos(θ) convolutions (Equation 6) for different WF bins (*B*). The close neighbors of rare words tend to be other rare words.

There must be some structural or statistical characteristic of the WAS underlying this effect: aside from the network mechanism, there simply is nothing else in WAS–FE to cause it. Specifically, the random input condition (Fig. 2A) serves as control for the structure in the semantic inputs. Cumulative population counts across cosine space (Fig. 5A) show that LF words have more close neighbors than HF words. Further, distributions of a neighborhood frequency average (Equation [6]), demonstrate the required WF dependence and separation to show that words tend to be colocated with neighboring words of the self-same WF class (Fig. 5B). These neighborhood effects support the tight-clustering hypothesis that LF words tend to cluster with other LF words, while HF words are coded more diffusely. Considering attractor formation on small-world inputs, this is sufficient cause for the type of relative familiarity differentiation observed in Figure 2B.

Among single-process recognition models, there has not been a consistent approach for the structural representation of semantic stimuli. For instance, the retrieving effectively from memory (Shiffrin and Steyvers 1997) model assigns higher diagnostic content to LF words by spreading out the distribution of feature values for LF words. This is based on the assumption that HF words share features to a higher degree than LF words. However, the subjective-likelihood model (McClelland and Chappell 1998) approaches WF differentiation by injecting more noise into the feature vectors of HF words to represent the higher degree of contextual variability for more frequent words. Lastly, the attention-likelihood theory (Glanzer et al. 1993; Malmberg and Nelson 2003) does not rely on structural differences to demonstrate the WF mirror effect. Instead, it uses the hypothesis that fewer features of HF words are attended to by the subject. This effectively reduces the semantic information in HF words, analogous to adding noise or placing vectors in smaller clusters. Algorithmically, these models combine features matching with the computation (or estimation) of log-likelihood ratios.

One intuitive line of thought is that a semantic attractor represents an atom of semantic content, a "sense." LF words tend to be associated with a very small number of different senses. HF words, however, may be associated with many distinct word senses. Semantic encoding would then place HF words in the space between their several senses. This leaves most rare words proximal to strong semantic attractors, with corresponding high-familiarity judgments, and more common words are placed away from these energy minima. To assess the validity of this idea, we repeated the neighborhood analyses for a WAS four-partition based on the number of senses a word has in the WordNet database (Fellbaum 1998). The cumulative population count Figure 6A shows a significant decline in close neighbors only for the most-senses words. The $\nu'(\xi)$ (i.e., neighborhood senses average) distributions (Fig. 6B) show only the most-senses words have a slight tendency to be encoded with other high-senses words. Finally, plotting WF against WordNet senses (Fig. 6C) reveals that only the most-senses words have any correlation with word frequency. This indicates that the number of word senses does contribute explanatory power to our structural observations, but this is limited to the HF/most-senses domain.

Decision process and performance

The WFE is fundamentally a behavioral effect of recognition performance, so a decision-making process is needed. The human

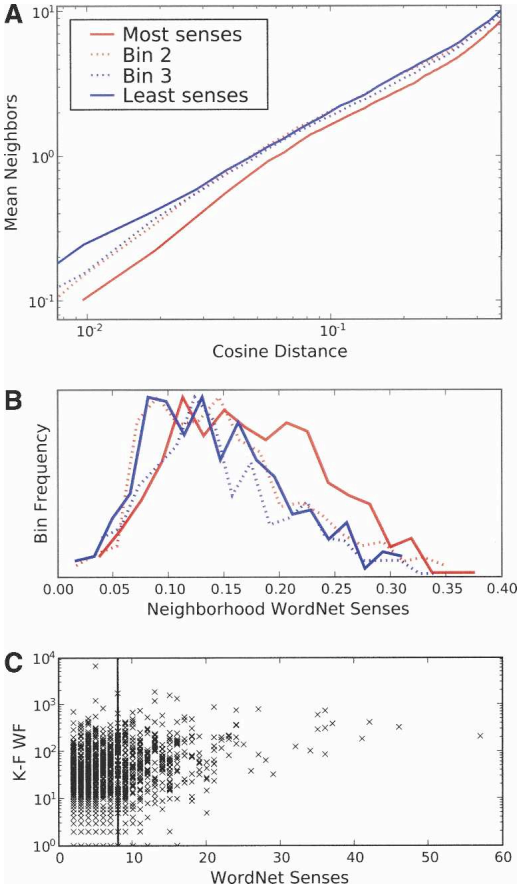

Figure 6. Clustering analysis redone in terms of the number of WordNet senses for a given word (*A,B*), instead of the K-F normative word frequency. The wordset used is the intersection of WAS words with K-F values and those with WordNet results. (*C*) WordNet senses are plotted against WF for comparison, with the blue vertical line indicating the lowest number of WordNet senses for Bin 1, which contains the words with the highest number of senses.

WFE is a mirror effect (Glanzer and Adams 1990; Glanzer et al. 1993), meaning that, for LF probes, subjects are better at both accepting targets as old and rejecting lures as new. Many different decision processes could be devised, but we will explore what would be necessary with the restriction of a simple process within the scope of and commensurate with the results presented so far.

We classify WAS–FE as a single-process signal-detection model of recognition: LF and HF stimuli are not processed differently and a single scalar energy value is the only output. This familiarity signal is noisy and a decision must be made whether a given probe was studied (old) or not (new). We consider a simple threshold process (see Methods: Signal detection) with a decision criterion below which a probe is judged old, otherwise new. For simplicity, we will consider as decision criterion, the midpoint of the empirical means of the energy distributions. Similarly, but for random vectors, Bogacz et al. (2001a,b) used the midpoint of the theoretical means. The question becomes that of which two distributions, in particular, are being compared in the decision process.

The signal detection comparison here is between a study list (Λ) and a reference pool (Φ), either of which may be broken down in WF classes. Thus, there are 2×2 possible comparisons: the WF bins or aggregate study list against the WF bins or aggregate reference pool. In the notation used above and in Methods: Experiment Simulation, these are $\Lambda_i - \Phi_i$, $\Lambda_i - \Phi$, $\Lambda - \Phi_i$, and $\Lambda - \Phi$, respectively, where i traverses WF bins. The $\Lambda - \Phi$ comparison is not WF dependent, and thus, meaningless in terms of the WFE. With a means-based criterion, the $\Lambda_i - \Phi_i$ comparison will not produce either component of the mirror effect, because the WF dependence of both distributions is the same. A fixed, WF-independent criterion could be used, but the resultant FAR effect would not "mirror" the HR trend. This is typical of the fundamental difficulty with single-process signal-detection models, as the familiarity effect needs to be reversed for new items to achieve a mirror effect (Glanzer et al. 1993). For instance, the attention-likelihood theory of the WFE mirror effect uses a log-likelihood ratio to bring about this required symmetry (Murdock 1998).

We are left to consider decisions involving a mixed comparison: WF bins of one distribution against the aggregate of the other. The $\Lambda - \Phi_i$ comparison falters on two counts. First, if only one distribution is going to receive the benefit of WF information, it does not make sense for it to be the distribution for items that have not been recently experienced. Second, it results in a performance decrease with rarity, because the increasingly negative distributions have more overlap with the Λ distribution. The $\Lambda_i - \Phi$ comparison addresses both counts: cognitively and intuitively, it makes sense that the subject has information regarding the WF classes of recently studied stimuli, and performance increases with rarity because the old distributions are farther from the Φ distribution. There are different possible forms for a stimulus-dependent criterion shift, but most allow that the criterion must increase in the signal direction "with the memorability of old items" (Hirshman 1995). Thus, we can tentatively delimit certain requirements for both the discimination comparision ($\Lambda_i - \Phi$) and the decision criterion (Equation [3]). The resultant WFE has mirrored HR and FAR effects that match recognition data (Fig. 4).

This decision process, however, is not entirely satisfactory. The study lists here are mixed, containing words randomly sampled from the data set, so the decision criterion needs to be adjusted on a per-stimulus basis. As discussed, this is necessary to achieve proper FAR trends. However, this type of criterion shift in recognition has been controversial (Miller and Wolford 1999; Roediger and McDermott 1999; Wickens and Hirshman 2000; Wixted and Stretch 2000). However, some recent cases are able to demonstrate that subjects modulate their decision criteria online according to stimulus class to optimize performance (Heit et al. 2003; Benjamin and Bawa 2004). This strategic use of multiple criteria may be driven by self-knowledge of the category-dependent memorability differences of probe stimuli (Strack et al. 2005). The multiple-criterion decision process required by WAS–FE is in line with these observations. Note also that we intentionally constrained our decision-making process to be simple, plausible, and within the scope of WAS–FE.

Finally, criterion-independent performance is illustrated by the ROCs. For both small (Fig. 3A) and large (Fig. 3B) list sizes, the trial-averaged ROC for the bin of LF words has higher HRs and lower FARs than that of HF words. These characteristics correspond to the $\Lambda_6 - \Phi$ and $\Lambda_1 - \Phi$ comparisons, respectively. They largely resemble those of other recognition models, except that they are not symmetric around the negative diagonal. These ROC examples also demonstrate two performance effects of the number of trained stimuli. The low-load condition (Fig. 3A) shows high absolute performance, but a relatively small WFE; the high-load condition (Fig. 3B), however, shows worse overall performance, but a larger difference between LF and HF words. These are capacity effects of the attractor-based FD mechanism. Larger stored lists entail higher synaptic load and reduced recognition accuracy. Further, we can infer that such capacity effects hurt the performance of HF words more than LF words. This WF dependence may be a result of the sparse encoding of HF words in the WAS: weaker attractors are more sensitive to the perturbations of overlearning than the strong LF-word attractors. So, for a given reference pool size, increases in study-list length push the network closer to capacity, decreasing HRs and increasing FARs, regardless of WF class. This means that WAS–FE exhibits a list-length mirror effect, which previous single-process models have also demonstrated (Shiffrin et al. 1990; Shiffrin and Steyvers 1997; McClelland and Chappell 1998). Conversely, for constant study-list length, this predicts better overall recognition performance and a smaller WFE for subjects with relatively less background experience (e.g., children versus adults).

Role of contextual information

Dual-process recognition theories use asymmetric recollective processing as the basis of the HR effect for LF words; the FAR effect is due to error-prone familiarity processing of similarly encoded HF words (Guttentag and Carroll 1994, 1997; Reder et al. 2000; Arndt and Reder 2002). This account is supported by evidence from pharmacological dissociation of recollection (Hirshman et al. 2002; Mintzner 2003), but not to the ultimate exclusion of single-process accounts (Malmberg et al. 2004). Indeed, it seems that both familiarity and recollection are involved, but the exact nature of their interaction is not yet definitively characterized (for review, see Yonelinas 2002).

As a decision-making recognition model, WAS–FE is not purely a single-process familiarity model. In the decision comparison $\Lambda_i - \Phi$, words from the study context are treated categorically as members of their respective WF class. However, nonstudy probes are not likewise differentiated. The contextual distinction consists of the subject having formed stimulus categories, such as frequency, only for recently studied stimuli. These categories then inform the decision process. Both IT cortex and PRC are implicated in highly plastic category formation (Erickson et al. 2000; Miller 2000; Miller et al. 2003), thus the formation of such WF categories of recent semantic stimuli is plausible. Also, further episodic information could allow discrimination between, for example, several study lists in a session. This could be modeled within the framework of WAS–FE by in-

tegrating a representation of a time-varying context signal (e.g., Howard and Kahana 2002).

Dual-process models typically use differential recruitment of recollective processing. Physiologically, this would be evident as WF-modulation of activity in regions such as the hippocampal formation and MTL. However, WAS–FE predicts that such areas, including PRC, differentiate old–new responses, but do not exhibit frequency dependence. It also predicts that the area responsible for semantic representation and processing shows WF-modulated activity. Using event-related fMRI at retrieval, de Zubicaray et al. (2005) sought to test predictions such as these and found two main effects. First, recollection-specific MTL regions with significant old–new responses did not show WF modulation. Second, the LF word HR advantage was associated with left lateral temporal cortex (LTC) activation. Evidence suggests that LTC but not MTL structures are necessary for lexico-semantic information processing (Levy et al. 2004; de Zubicaray et al. 2005). Thus, LTC is well positioned as a possible semantic input region for familarity processing in PRC. De Zubicaray et al. (2005) suggest that these results are consistent with context-noise models of the recognition WFE, but they are also consistent with our account here. Recently, EEG techniques have been able to dissociate verbal from nonverbal retrieval (Hwang et al. 2005), indicating the possibility of investigations using higher temporal-resolution methods. More such studies are needed to complement the large body of behavioral data.

Conclusion

In the present work, we take advantage of an empirically determined model of semantic space to demonstrate a benchmark effect of human memory. Using the WAS as an input space for a Hopfield model of perirhinal familiarity processing, we found a word-frequency effect on familiarity distributions that can be explained as a function of the small-world structure of the semantic space. This structure, characterized by tight local clustering of rare words, implies that word frequency is nonintuitively encoded into the semantic structure of language. We argue that the model components plausibly capture the salient features, respectively, of semantic representation and neurobiological familiarity processing. Thus, we suggest that lexico-semantic structure forms a causal basis for the recognition WFE. Further, we show that a frequency-dependent criterion shift produces a WFE mirror effect without requiring log-likelihood computations to bring about old–new symmetry. This entails a role for dual-process involvement in recognition contrary to previous models but consistent with some recent imaging data. Finally, we hope to have demonstrated the utility of relatively simple, but specific and salient models of complex biological systems and likewise the importance of establishing an appropriate interpretative context.

Methods

Signal detection

For a criterion λ, a vector X is determined to be old if $\mathbf{E}(X) < \lambda$, otherwise it is judged new. The distribution of energies from trained vectors is distinct from that of untrained probes. Consider a random and unbiased set ξ_P^N of stored vectors. The distribution of synaptic weights in W can be approximated by a Gaussian distribution with $\mu_W = 0$ and $\sigma_W^2 = P/N^2$. The energy distributions for both untrained probes and stored vectors have $\sigma_E^2 = P/2$. The expected energy value of an untrained probe X is $\langle \mathbf{E}(X) \rangle = 0$, while that of a stored vector ξ^μ is $\langle \mathbf{E}(\xi^\mu) \rangle = -N/2$. Here, the logical decision criterion would be $\lambda = -N/4$, the midpoint between the old and new energy distributions. This is the criterion used by Bogacz et al. (2001b, 2002) in their signal–noise analysis of capacity. For the semantically structured inputs considered here, λ is chosen as the midpoint between the empirical means of the distributions. To determine HRs and FARs, we used a WF-based multiple-criterion decision strategy (see Discussion) with means-based thresholds,

$$\lambda_i = [\langle \mathbf{E}(\Phi) \rangle + \langle \mathbf{E}(\Lambda_i) \rangle]/2, \tag{3}$$

where $\mathbf{E}(\Phi)$ is the energy distribution of the reference pool. The HR is the fraction of stored vectors with $\mathbf{E} < \lambda$, while the FAR is the probability for an untrained vector to have $\mathbf{E} < \lambda$. The discriminability between the old $\langle \mathbf{E}(\xi^o) \rangle$ and new $\langle \mathbf{E}(\xi^n) \rangle$ energy distributions is computed as the distance between means in standard deviations,

$$d'\langle \mathbf{E}(\xi^n), \mathbf{E}(\xi^o) \rangle, = \frac{\langle \mathbf{E}(\xi^n) \rangle - \langle \mathbf{E}(\xi^o) \rangle}{\sqrt{\sigma_n^2 + \sigma_o^2/2}}. \tag{4}$$

For the experimental data in Figure 4B, d' was calculated based on an unbiased estimator assuming underlying Gaussian distributions: $\hat{d}' = z(\langle HR \rangle) - z(\langle FAR \rangle)$. ROCs are constructed by plotting HR against FAR for a range of possible decision criteria. Better performance is indicated by an ROC curve farther from the chance function (where HR = FAR and $d' = 0$) in the direction of higher HRs and lower FARs.

Experiment simulation

In an old–new item-recognition experiment, the subject studies a list of known items from a training set. At test, the subject is shown a list of probe items, some of which had appeared in the training set (old items) and others that had not (new items). The task is to judge whether each item is old or new.

Experimental subjects here are defined by Θ, the random subset of word vectors on which the Hopfield network is initially trained (Equation [1]). Each vector in Θ is associated with a WF value corresponding to the word that it represents. Using these frequencies to index the vectors, Θ is sorted and evenly partitioned into six bins with Θ_1 containing the highest-frequency subset and Θ_6 containing the lowest-frequency subset of words. The study list for the task is defined by Λ, which is a random subset of Θ, such that an approximately equal number of study items are chosen from each WF bin. That is, Λ comprises random subsets $\Lambda_i \subset \Theta_i$, for $i \in \{1...6\}$, such that the length of the study list is $L = \sum_{i=1}^{6} |\Lambda_i|$ where $|\Lambda_i|$ is the number of vectors in Λ_i. The study list is presented to the model by retraining the network on all of the study vectors. If ξ_Λ is a matrix containing the study vectors row-wise, then W is updated as

$$W \rightarrow W + \frac{1}{N} \xi_\Lambda^T \xi_\Lambda \tag{5}$$

and then zeroing out diagonal terms. This procedure is analogous to strengthening the pre-existing neural representation of the items in a study list attended to by a subject. Specifically, this operation doubles every weight component resulting from the initial training of Λ as part of Θ. All items are studied equally. The training-set vectors not chosen for the study list composed the reference pool, Φ. Therefore, the size of the pool is the size of the training set Θ minus the study-list length $|\Lambda|$. The items in the study list and the reference pool serve as the old and new probes during test, respectively. We calculate the internal energy (Equation [2]) of each vector in Θ using the updated weight matrix. We then compare the resulting sample energy distributions by calculating d' distances (Equation [4]), ROC curves, HRs, and FARs. This process, starting with a new random Θ chosen from our binarized WAS, is repeated for 2000 trials. That is, one trial here is analogous to a new subject performing a single recognition task. Across-trial means and confidence intervals were computed for Figure 3 and Figure 4. Representative energy histograms were created by accumulating energy vectors across trials and computing frequencies for 100 equally spaced bins across the range of energies.

Neighborhood measures

First, we computed the WF dependence of the mean number of neighbors in WAS space. Consider the metric $d_{\cos} = 1 - \cos(\theta)$ so

that close neighbors have a cosine approaching 1 and a d_{cos} approaching 0. For every WAS vector, our algorithm traversed the range of possible d_{cos} from 0 to 1 while counting the number of vectors within that distance from the given vector. The mean populations were computed for every d_{cos} radius for each of six WF classes (Fig. 5A). Second, to address the question of the WF composition of neighbors, we used a cos(θ)-weighted frequency measure. We calculated the quantity,

$$v(\xi^{\mu}) = \sum_{\varphi=1}^{1,748} \left(\frac{\xi^{\mu} \cdot \xi^{\varphi}}{\|\xi^{\mu}\| \, \|\xi^{\varphi}\|} \right) f_{KF}(\xi^{\varphi}) - f_{KF}(\xi^{\mu}) \quad (6)$$

for all WAS vectors ξ^{μ}, where $f_{KF}(\cdot)$ is the function mapping a vector to its associated Kučera-Francis WF value. This measure quantifies the expectation of the WF for neighbors of a given vector. The distributions of these convolutions for the word vectors of each of the six frequency classes are shown as 15-bin histograms in Figure 5B.

Acknowledgment

This work was supported by grants MH062196, MH55687, and MH61975 to M.J.K.; and MH58754 and an NIH Director's Pioneer Award, part of the NIH Roadmap for Medical Research, through grant 5-DP1-OD114-02 to L.F.A.

References

Aggleton, J. and Brown, M.W. 1999. Episodic memory, amnesia and the hippocampal-anterior thalamic axis. *Behav. Brain Sci.* **22:** 425.

Amit, D.J. 1989. *Modeling brain function.* Cambridge University Press. Cambridge, U.K.

Arndt, J. and Reder, L.M. 2002. Word frequency and receiver operating characteristic curves in recognition memory: Evidence for a dual-process interpretation. *J. Exp. Psychol. Learn. Mem. Cogn.* **28:** 830–842.

Benjamin, A.S. and Bawa, S. 2004. Distractor plausibility and criterion placement in recognition. *J. Mem. Lang.* **51:** 159–172.

Bogacz, R. and Brown, M.W. 2002. Capacity of perirhinal cortex network for recognising frequently repeating stimuli. *Neurocomputing* **44-46:** 337–342.

Bogacz, R. and Brown, M.W. 2003. Comparison of computational models of familiarity discrimination in the perirhinal cortex. *Hippocampus* **13:** 494–524.

Bogacz, R., Brown, M.W., and Giraud-Carrier, C. 2001a. A familiarity discrimination algorithm inspired by computations of the perirhinal cortex. *Lect. Notes Comput. Sci.* **2036:** 428–441.

Bogacz, R., Brown, M.W., and Giraud-Carrier, C. 2001b. Model of familiarity discrimination in perirhinal cortex. *J. Comput. Neurosci.* **10:** 5–23.

Brown, M.W. and Bashir, Z.I. 2002. Evidence concerning how neurons of the perirhinal cortex may effect familiarity discrimination. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **357:** 1083–1095.

de Zubicaray, G.I., McMahon, K.L., Eastburn, M.M., Finnigan, S., and Humphreys, M.S. 2005. fMRI evidence of word frequency and strength effects in recognition memory. *Brain Res. Cogn. Brain Res.* **24:** 587–598.

Erickson, C.A., Jagadeesh, B., and Desimone, R. 2000. Clustering of perirhinal neurons with similar properties following visual experience in adult monkeys. *Nat. Neurosci.* **3:** 1143–1148.

Fahy, F., Riches, I., and Brown, M.W. 1993. Neuronal activity related to visual recognition memory: Long-term memory and the encoding of recency and familiarity information in the primate anterior and medial inferior temporal and rhinal cortex. *Exp. Brain Res.* **96:** 457–472.

Fellbaum, C. (Ed.). 1998. *WordNet: An electronic lexical database.* Bradford Books, Cambridge, MA.

Gaffan, D. 1994. Dissociated effects of perirhinal cortex ablation, fornix transection and amygdalectomy–evidence for multiple memory-systems in the primate temporal-lobe. *Exp. Brain Res.* **99:** 411–422.

Glanzer, M. and Adams, J.K. 1985. The mirror effect in recognition memory. *Mem. Cognit.* **13:** 8–20.

Glanzer, M. and Adams, J.K. 1990. The mirror effect in recognition memory: Data and theory. *J. Exp. Psychol. Learn. Mem. Cogn.* **16:** 5–16.

Glanzer, M., Adams, J.K., Iverson, G.J., and Kisok, K. 1993. The regularities of recognition memory. *Psychol. Rev.* **100:** 546–567.

Guttentag, R.E. and Carroll, D. 1994. Identifying the basis for the word-frequency effect in recognition memory. *Memory* **2:** 255–273.

Guttentag, R.E. and Carroll, D. 1997. Recollection-based recognition: Word frequency effects. *J. Mem. Lang.* **37:** 502–516.

Heit, E., Brockdorff, N., and Lamberts, K. 2003. Adaptive changes of response criterion in recognition memory. *Psychon. Bull. Rev.* **10:** 718–723.

Hirshman, E. 1995. Decision processes in recognition memory: Criterion shifts and the list-strength paradigm. *J. Exp. Psychol. Learn. Mem. Cogn.* **21:** 302–313.

Hirshman, E., Fisher, J., Henthorn, T., Arndt, J., and Passannante, A. 2002. Midazolam amnesia and dual-process models of the word-frequency mirror effect. *J. Mem. Lang.* **47:** 499–516.

Hopfield, J.J. 1982. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci.* **84:** 8429–8433.

Howard, M.W. and Kahana, M.J. 2002. A distributed representation of temporal context. *J. Math. Psychol.* **46:** 269–299.

Humphreys, M.S., Pike, R., Bain, J.D., and Tehan, G. 1989. Global matching: A comparison of the SAM, Minerva II, Matrix, and TODAM models. *J. Math. Psychol.* **33:** 36–67.

Hwang, G.M., Jacobs, J., Geller, A., Danker, J., Sekuler, R., and Kahana, M.J. 2005. EEG correlates of verbal and nonverbal stimuli in working memory. *Behav. Brain Funct.* **1:** 20 [Epub].

Kahana, M.J. and Sekuler, R. 2002. Recognizing spatial patterns: A noisy exemplar approach. *Vision Res.* **42:** 2177–2192.

Kahana, M.J., Rizzuto, D.S., and Schneider, A. 2005. An analysis of the recognition-recall relation in four distributed memory models. *J. Exp. Psychol. Learn. Mem. Cogn.* **31:** 933–953.

Kesner, R.P., Gilbert, P.E., and Wallenstein, G.V. 2000. Testing neural models of memory with behavioral experiments. *Curr. Opin. Neurobiol.* **10:** 260–265.

Kučera, H. and Francis, W.N. 1967 . *Computational analysis of present-day American English.* Brown University Press, Providence, RI.

Lacroix, J.P.W., Murre, J.M.J., Postma, E.O., and van den Herik, H.J. 2006. Modeling recognition memory using the similarity structure of natural input. *Cogn. Sci.* **30:** 121–145.

Levy, D.A., Bayley, P.J., and Squire, L.R. 2004. The anatomy of semantic knowledge: Medial vs. lateral temporal lobe. *Proc. Natl. Acad. Sci.* **101:** 6710–6715.

Li, L., Miller, E.K., and Desimone, R. 1993. The representation of stimulus familiarity in anterior inferior temporal cortex. *J. Neurophysiol.* **68:** 1918–1929.

Malmberg, K.J. and Nelson, T.O. 2003. The word frequency effect for recognition memory and the elevated-attention hypothesis. *Mem. Cognit.* **31:** 35–43.

Malmberg, K.J., Steyvers, M., Stephens, J.D., and Shiffrin, R.M. 2002. Feature frequency effects in recognition memory. *Mem. Cognit.* **30:** 607–613.

Malmberg, K.J., Zeelenberg, R., and Shiffrin, R.M. 2004. Turning up the noise or turning down the volume? On the nature of the impairment of episodic recognition memory by midazolam. *J. Exp. Psychol. Learn. Mem. Cogn.* **30:** 540–549.

McClelland, J.L. and Chappell, M. 1998. Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychol. Rev.* **1-5:** 724–760.

Miller, E.K. 2000. Organization through experience. *Nat. Neurosci.* **3:** 1066–1068.

Miller, E.K., Li, L., and Desimone, R. 1991. A neural mechanism for working and recognition memory in inferior temporal cortex. *Science* **254:** 1377–1379.

Miller, E.K., Nieder, A., Freedman, D.J., and Wallis, J.D. 2003. Neural correlates of categories and concepts. *Curr. Opin. Neurobiol.* **13:** 198–203.

Miller, M.B. and Wolford, G.L. 1999. Theoretical commentary: The role of criterion shift in false memory. *Psychol. Rev.* **106:** 398–405.

Mintzner, M.Z. 2003. Triazolam-induced amnesia and the word-frequency effect in recognition memory: Support for a dual process account. *J. Mem. Lang.* **48:** 596–602.

Murdock, B.B. 1982. A theory for the storage and retrieval of item and associative information. *Psychol. Rev.* **89:** 609–626.

Murdock, B.B. 1998. The mirror effect and attention-likelihood theory: A reflective analysis. *J. Exp. Psychol. Learn. Mem. Cogn.* **24:** 524–534.

Murray, E.A. and Bussey, T.J. 1999. Perceptual-mnemonic functions of the perirhinal cortex. *Trends Cogn. Sci.* **3:** 142–151.

Nelson, D., McEvoy, C., and Schreiber, T. 2004. The University of South Florida free association, rhyme, and word fragment norms. *Behav. Res. Methods Instrum. Comput.* **36:** 402–407.

Norman, K.A. and O'Reilly, R.C. 2003. Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychol. Rev.* **110:** 611–646.

Norman, K.A., Newman, E.L., and Perotte, A.J. 2005. Methods for

reducing interference in the complementary learning systems model: Oscillating inhibition and autonomous memory rehearsal. *Neural Netw.* **18:** 1212–1228.
O'Reilly, R.C. and McClelland, J.L. 1994. Hippocampal conjunctive encoding, storage, and recall: Avoiding a trade-off. *Hippocampus* **4:** 661–682.
Reder, L.M., Nhoujvanisvong, A., Schunn, C.D., Ayers, M.S., Angstadt, P., and Hiraki, K. 2000. A mechanistic account of the mirror effect for word frequency: A computational model of remember-know judgments in a continuous recognition paradigm. *J. Exp. Psychol. Learn. Mem. Cogn.* **26:** 294–320.
Roediger, H.L. and McDermott, K.B. 1999. False alarms about false memories. *Psychol. Rev.* **106:** 406–410.
Rugg, M.D. and Yonelinas, A.P. 2003. Human recognition memory: A cognitive neuroscience perspective. *Trends Cogn. Sci.* **7:** 313–319.
Schulman, A.I. 1967. Word length and rarity in recognition memory. *Psychon. Soc.* **9:** 211–212.
Schwartz, G., Howard, M.W., Jing, B., and Kahana, M.J. 2005. Shadows of the past: Temporal retrieval effects in recognition memory. *Psychol. Sci.* **16:** 898–904.
Shepard, R.N. 1967. Recognition memory for words, sentences, and pictures. *J. Verb. Learn. Verb. Be* **6:** 156.
Shiffrin, R.M. and Steyvers, M. 1997. A model for recognition memory: REM retrieving effectively from memory. *Psychon. Bull. Rev.* **4:** 145–166.
Shiffrin, R.M., Ratcliff, R., and Clark, S. 1990. The list-strength effect: II. Theoretical mechanisms. *J. Exp. Psychol. Learn. Mem. Cogn.* **16:** 179–195.
Sobotka, S. and Ringo, J.L. 1993. Investigation of long term recognition and association memory in unit responsnses from inferotemporal cortex. *Exp. Brain Res.* **96:** 28–38.
Sohal, V.S. and Hasselmo, M.E. 2000. A model for experience-dependent changes in the responses of inferotemporal neurons. *Network Comp. Neural.* **11:** 169–190.
Standing, L. 1973. Learning 10,000 pictures. *Q. J. Exp. Psychol.* **25:** 207–222.
Steyvers, M. and Malmberg, K.J. 2003. The effect of normative context variability on recognition memory. *J. Exp. Psychol. Learn. Mem. Cogn.* **29:** 760–766.
Steyvers, M. and Tenenbaum, J.B. 2005. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cogn. Sci.* **29:** 41–78.
Steyvers, M., Shiffrin, R.M., and Nelson, D.L. 2004. Word association spaces for predicting semantic similarity effects in episodic memory. In *Cognitive psychology and its applications: Festschrift in honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer.* (ed. A.F. Healy), American Psychological Association, Washington, DC.
Strack, F., Förster, J., and Werth, L. 2005. "Know thyself!" The role of idiosyncratic self-knowledge in recognition memory. *J. Mem. Lang.* **52:** 628–638.
Stretch, V. and Wixted, J.T. 1998. On the difference between strength-based and frequency-based mirror effects in recognition memory. *J. Exp. Psychol. Learn. Mem. Cogn.* **24:** 1379–1396.
Watts, D.J. and Strogatz, S.H. 1998. Collective dynamics of 'small-world' networks. *Nature* **393:** 440–442.
Wickens, T.D. 2002. *Elementary signal detection theory.* Oxford University Press, New York.
Wickens, T.D. and Hirshman, E. 2000. False memories and statistical decision theory: Comment on Miller and Wolford (1999) and Roediger and McDermott (1999). *Psychol. Rev.* **107:** 377–383.
Wixted, J.T. and Stretch, V. 2000. The case against a criterion-shift account of false memory. *Psychol. Rev.* **107:** 368–376.
Xiang, J.-Z. and Brown, M.W. 1998. Differential neuronal encoding of novelty, familiarity and recency in regions of the anterior temporal lobe. *Neuropharmacology* **37:** 657–676.
Yonelinas, A.P. 2002. The nature of recollection and familiarity: A review of 30 years of research. *J. Mem. Lang.* **46:** 441–517.
Zaki, S. and Nosofsky, R. 2001. Exemplar accounts of blending and distinctiveness effects in perceptual old-new recognition. *J. Exp. Psychol. Learn. Mem. Cogn.* **27:** 1022–1041.

Received July 5, 2006; accepted in revised form January 4, 2007.