Lexico-Semantic Structure and the Recognition Word Frequency Effect

Joseph D. Monaco† L. F. Abbott
Columbia University

Michael J. Kahana
University of Pennsylvania

The recognition word frequency effect (WFE) is the phenomenon that rarer words are better recognized than more common words. We demonstrate that an associative network model of familiarity discrimination operating on data from a semantic word association space yields a robust WFE. Also, capacity effects of the network produce a qualitatively correct list length effect. Further, when performance is determined with a particular multiple-criterion decision process, both word frequency and list length effects show both hit and false alarm components characteristic of recognition mirror effects. This model may reflect the contribution of inter-item semantic similarity to these effects. Finally, we suggest that word frequency is non-intuitively encoded in the semantic structure of language, and that this is a causal basis for the observed WFE.

Keywords: *recognition memory*; *word frequency*; *semantic space*; *perirhinal cortex*; *familiarity*; *Hopfield network*.

Old–new item recognition is the task of deciding whether items in a probe list were present (old) or not (new) in a study list. Performance is quantified as the probability of old responses to study items (hit rate or HR) and non-study probes (false alarm rate or FAR). One of the most prominent recognition memory effects 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 equal FAR effect. Mirror effects are considered one of the regular features of recognition memory (Glanzer, Adams, Iverson, and Kisok 1993). 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 these recognition effects. The former perform familiarity discrimination (FD) using similarity measures such as global feature matching. This typically requires some additional transformation, such as log-likelihood computation, to achieve the required symmetry between old and new familiarity distributions (Murdock 1998). To explain the WFE, various 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 output a unidimensional scalar value for the strength of familiarity for a given stimulus allowing 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; Wickens 2002). Dual-process models, however, rely on differential contributions of recollective and familiarity-based processes to explain the performance differences. Recollection, or controlled item retrieval, is characterized as less error-prone than a simple, automatic familiarity process (Guttentag and Carroll 1997; Reder et al. 2000). Thus, asymmetric process involvement can be used to explain mirror effects (Arndt and Reder 2002).

In humans, both types of retrieval processes appear to be seated in the medial temporal lobe (MTL; Levy, Bayley, and Squire 2004; Zubicaray, McMahon, Eastburn, Finnigan, and Humphreys 2005). Also, electrophysiological studies in monkey have shown perirhinal cortex (PRC), specifically, to have a substantial proportion of familiarity-sensitive neurons (Miller, Li, and Desimone 1991; Li, Miller, and Desimone 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, this computation may approximate the familiarity computation performed by perirhinal neurons (Bogacz, Brown, and Giraud-Carrier 2001a; Brown and Bashir 2002) and has been used to determine theoretical limits on recognition capacity (Bogacz, Brown, and Giraud-Carrier 2001b; Bogacz and Brown 2002). We hypothesized that if this is representative of perirhinal familiarity processing, then it may be able to demonstrate a robust human memory effect such as the WFE. For this purpose, we used input vectors from a word association space (WAS; Steyvers, Shiffrin, and Nelson 2004),

This work was supported by grants XXXXXX for M.J.K. and XXXXXX for L.F.A.

†Correspondence should be addressed to J.D.M., Center for Theoretical Neuroscience, Department of Neurobiology and Behavior, Columbia University, 1051 Riverside Drive Unit 87, New York, NY, 10032-2695 (e-mail: joe@neurotheory.columbia.edu).

which is an empirical model of semantic similarity based on normative data from free association (Nelson, McEvoy, and Schreiber 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 exhibits a WFE mirror effect. These results have novel implications for the role of distinct retrieval processes in recognition memory.

Model

Familiarity as Network Energy

In this FD model, item stimulus vectors are associatively encoded into a Hopfield network (Hopfield 1982). The familiarity signal is just the internal energy of the network (Amit 1989) when activated with a probe stimulus (Bogacz et al. 2001a). FD processes such as this are more efficient and have much higher capacities than models involving recall. Consider a Hopfield network of N units trained on random and unbiased input vectors. 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).

Hopfield networks are recurrent attractor networks of fully-connected binary units. The network weights are trained on an input set ξ_N^S of S N–dimensional stimulus vectors such that $\xi_i^\mu \in \{-1,+1\}$ for all $i \in \{1..N\}$ and $\mu \in \{1..S\}$. That is, each unit is either active $(+1)$ or inactive (-1) for a given input vector. All input sets considered here are unbiased so that, on average, an input vector will have as many active units as inactive. Letting $W = [w_{ij}]_{i,j=1}^N$ be the weight matrix, its elements are computed according to the Hebbian learning rule,

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

and $w_{ii} = 0$ 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 there are no network dynamics involved here. This 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 is computed as

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

Due to the negative prefactor, 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^S.

Semantically Structured Input

Previous applications of the above FD model have, for simplicity, assumed the input vectors to be random (Bogacz et al. 2001a; Bogacz and Brown 2002). Further, these applications have not attempted to model the benchmark behavioral phenomena of recognition memory. Given that words are the stimuli most frequently used in human recognition memory studies, it seems prudent to construct input vectors whose similarity relations approximate the semantic relations of words in the English language. Recent 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 dataset containing the probabilities with which subjects provided a given word as the first associate to another word (Nelson et al. 2004). These data can be taken as a measure of direct associative strength among over 5,000 words. Indirect, or second-order, associative strengths can also be calculated from the dataset. To create the WAS, singular-value decomposition (SVD) was applied on these direct and indirect associations so as to place words in a reduced 400-dimensional space. This was done so that the cosine between any two word vectors (i.e., normalized inner product, $\cos(\theta) = (V \cdot U)/(\|V\|\|U\|)$) reflects the overall similarity of their associative patterns. Words with similar associative patterns have $\cos(\theta)$ values approaching 1, whereas semantically unrelated words have values approaching 0. The dimensional reduction reveals latent, high-order semantic relations within the dataset. Importantly, 400 dimensions was found to be lowest dimensionality that remains highly predictive of experimental data such as semantic similarity ratings in recognition, free recall intrusion rates, and extralist cued recall (Steyvers et al. 2004). The resultant WAS shares gross structural characteristics with other types of semantic networks (Steyvers and Tenenbaum 2005).

Now, we can ask whether common and rare words differ in their similarity structure. Here, we use a set of 1,748 WAS vectors for which we have the associated Kučera-Francis word frequency (WF; Kučera and Francis 1967). In the cosine similarity matrix for the 100 most common and the 100 rarest words in our WAS pool (Figure 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.

Finally, we binarized the WAS vectors in our word pool. This allows proper operation of the Hopfield learning rule (1) and energy computation (2). The elements of WAS vectors are nonzero but symmetrically distributed around zero. Thus, taking the sign of each vector element produces a set of binarized vectors, on average, with unbiased activity levels. The normalized Hamming distances between these vectors decrease monotonically with increases in the cosine similarities of the corresponding continuous WAS vectors (Figure 1b). This indicates that the binarization significantly preserves the similarity relations between word vectors.

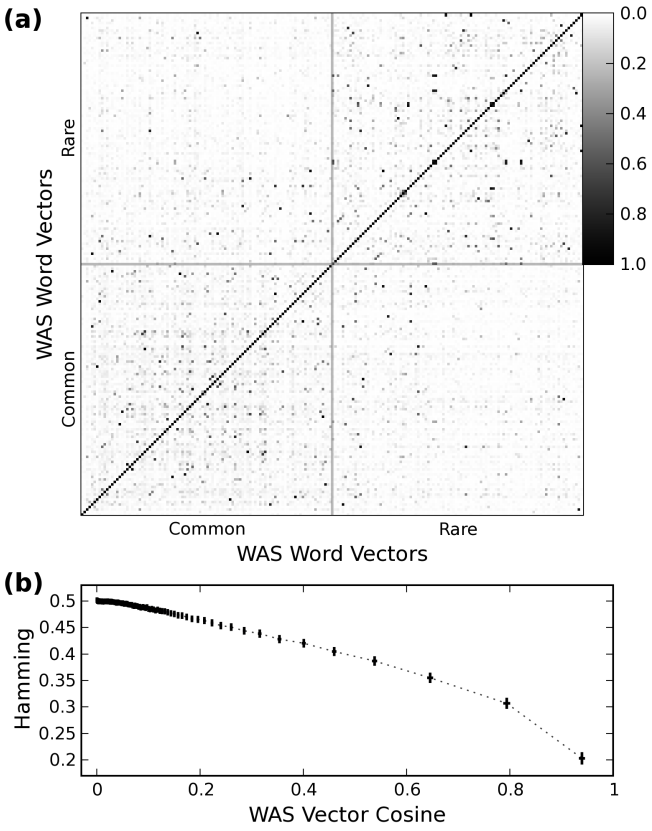

Figure 1. (a) Similarity matrix of the WAS vectors for the 100 highest and lowest frequency words in the set. The color of each pixel denotes the cosine of the angle between the *i*th and *j*th vectors. The lower-left and upper-right quadrants represent the cosine similarity among pairs whose members 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 here are the Hebbian learning of a Hopfield network and the dimensional reduction of a word association matrix. These serve, respectively, as the FD mechanism and the semantic input space of our recognition model. In considering this combination, we have to be able to interpret the necessary combination of the assumptions inherent to both. We must properly frame the limitations of the results and emphasize that they comprise, at most, a high-level explanation. However, we argue for the specificity and functional plausibility of the components as well as their composition.

First, we assume that employing the energy (2) of a trained Hopfield network (1) as a familiarity signal captures some salient characteristic of familiarity processing in the perirhinal cortex. Bogacz et al. (2001a, 2001b) provide support for this assumption by arguing from a standpoint of functionality and efficiency as well as from modeling results. It is easily verified that human recognition capacity is very high (Standing 1973). 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 ms), but neural responses for recency and novelty have also been found (Fahy, Riches, and Brown 1993; Xiang and Brown 1998; Brown and Bashir 2002). Despite this functional diversity, only familiarity-sensitive neurons are considered here. Further, 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; Murray and Bussey 1999; Aggleton and Brown 1999). This independence suggests that PRC is the site of the neural substrate for familiarity judgments (Rugg and Yonelinas 2003; Yonelinas 2002, for review). From its IT afferents, the PRC has access to high-order stimulus features, which are theoretically ideal inputs for an autoassociative network model.

Autoassociative neural networks, such as the binary Hopfield network (Hopfield 1982), use recurrent connectivity and Hebbian-like synaptic modification to produce stimulus-dependent attractors. Simply reading out the internal energy as a familiarity signal is much more efficient than additionally involving recollective processes (Bogacz et al. 2001a; Bogacz and Brown 2003), which typically involve "relaxing" the network to reconstruct an associated attractor state. For random vectors, this FD process has a very high storage capacity and constitutes a rapid network response as it does not require any activity cycling. It is also more robust than other network architectures; e.g., encoding via feedforward competitive synaptic processes can exhibit forgetting after only 105 subsequent stimuli (Sohal and Hasselmo 2000). Bogacz et al. (2001b) argue that perirhinal FD neurons must form an autoassociative network in order to exhibit these characteristics and recognition efficiency. Further, they show that a more realistic multilayer spike response model (Fuentes, Ritz, Gerstner, and Hemmen 1996) with a recurrent FD layer develops familiarity neurons that replicate perirhinal FD responses. From this, we posit that it is a reasonable assumption 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 in speakers of English. This amounts to the equivalence between behavioral associativity and the neural encoding of semantic similarity. The WAS consists of transformed statistical behavioral data from 6,000 subjects, and as such can only be an inference of the structure of semantic space for a given subject. That it serves well as a predictive model for known human memory effects supports it as a useful inference. Further, a graph theoretic study demonstrated that the WAS shares large-scale characteristics,

specifically small-world structure and scale-free connectivity, with other empirically derived semantic networks and models of semantic growth (Steyvers and Tenenbaum 2005). Thus, the WAS has many characteristics consistent with being isomorphic to real semantic representations. Even though the WAS vectors were binarized in order to be used as proper inputs to the Hopfield network, Figure 1b shows that this transformation preserves the gross structure of the space.

Finally, we make the assumption that the semantic stimulus vectors of the WAS serve as appropriate inputs to the FD model of PRC. This 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, Jagadeesh, and Desimone 2000; Miller 2000; Miller, Nieder, Freedman, and Wallis 2003). Such abstraction is a hallmark of semantic information processing and, at the least, indicates that the PRC has access to semantic features among its inputs. Presumably, the semantic features of words, and possibly pictures (Karlsen and Snodgrass 2004), presented to subjects as stimuli in recognition experiments are accessible from PRC. We will refer to this bipartite recognition model, derived from WAS and the equivalence of familiarity with energy, as *WAS–FE*.

Methods

Signal Detection Analysis

The noise in the weights of the network depends both on the number of stored items (S) and the variability among items. This synaptic noise is translated into randomness in the internal energy computation (2) and, therefore, the familiarity measurement of a given probe. For *WAS–FE* to serve in a recognition experiment, a binary old–new decision must be made from its noisy unidimensional output. These conditions satisfy the assumptions necessary to assess recognition performance using signal detection methods (Wickens 2002).

A decision threshold, or criterion, can be used to efficiently decide if a probe stimulus is familiar or not. 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 input vectors is distinct from that of probes not belonging to the input set. Consider a random and unbiased set ξ_N^S of stored vectors. The distribution of synaptic weights in W can be approximated by a Gaussian with $\mu_W = 0$ and $\sigma_W^2 = S/N^2$. The energy distributions for both untrained probes and stored input vectors have $\sigma_{\mathrm{E}}^2 = S/2$. However, 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, a logical decision criterion would be $\lambda = -N/4$, the midpoint between the old and new energy distributions. This is the criterion employed by Bogacz et al. in their signal–noise analysis of capacity (2001b, 2002). For the semantically structured inputs considered here, λ is chosen as the midpoint between the empirical means of the distributions.

The performance of the model is assessed by calculating HRs and FARs. The HR is the fraction of stored vectors with $\mathbf{E} < \lambda$, while the FAR is the probability for an untrained vector to have $\mathbf{E} < \lambda$. Also, the discriminability between the old ($\mathbf{E}(\xi^{\mathrm{o}})$) and new ($\mathbf{E}(\xi^{\mathrm{n}})$) energy distributions may be characterized by their distance in standard deviations,

$$d'(\mathbf{E}(\xi^{\mathrm{n}}), \mathbf{E}(\xi^{\mathrm{o}})) = \frac{\langle \mathbf{E}(\xi^{\mathrm{n}}) \rangle - \langle \mathbf{E}(\xi^{\mathrm{o}}) \rangle}{\sqrt{(\sigma_{\mathrm{n}}^2 + \sigma_{\mathrm{o}}^2)/2}}. \tag{3}$$

For the experimental data in Figure 4b, d' was calculated using an unbiased estimator assuming underlying Gaussian distributions: $\hat{d'} = z(\langle HR \rangle) - z(\langle FAR \rangle)$. An ROC curve is constructed by plotting HR against FAR for a range of possible decision criteria. It thus provides a criterion-independent assessment of how well the familiarity signal is discriminated that is especially informative in the case of non-Gaussian distributions. 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

We used the following procedure to simulate a standard old–new recognition experiment. In this type of experiment, the subject first studies a list of known items from a training set. Then, at test, the subject is shown a list of probe items, some of which had appeared in the training set (old items) and others which 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 (1). Each Θ vector is associated with a WF value via the word it represents. Using these frequencies to index the vectors, Θ is sorted and evenly partitioned into 6 bins with Θ_1 containing the highest and Θ_6 the lowest frequency words. The other four subsets contain word vectors of intermediate WF values. The Θ_i are equal-sized to within one vector due to rounding. 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. These WF study bins are equal-sized to within one vector due to rounding. The study list is presented to the model by retraining the network on all the study vectors. If ξ_Λ is a matrix containing the study vectors row-wise, then W is updated as

$$W \rightarrow W + \frac{1}{N}\, \xi_\Lambda^T \xi_\Lambda, \tag{4}$$

and then zeroing out diagonal terms. This is analogous to strengthening the pre-existing neural representation of the items in a study list attended to by a subject. Specifically, this 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 compose

the *reference pool*. Therefore, the size of the pool is the size of the training set Θ minus the length of the study list Λ. The items in the study list and the reference pool serve as the old and new probes during test, respectively. Finally, we calculate the internal energy (2) of each vector in Θ using the updated weight matrix. We then compare the resulting sample energy distributions by calculating d' distances (3) and ROC curves. To determine HRs and FARs, we use a WF-based multiple-criterion decision strategy (see Discussion) with means-based thresholds,

$$\lambda_i = \left[\langle \mathbf{E}(\Phi)\rangle + \langle \mathbf{E}(\Lambda_i)\rangle\right]/2, \qquad (5)$$

where $\mathbf{E}(\Phi)$ is the energy distribution for the reference pool. This process, starting with a new random Θ chosen from our binarized WAS, is repeated for 2,000 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 are created by collapsing the energy vectors together across trials, computing counts for 100 equally-spaced bins across the range of energies, and then scaling the bin frequencies.

Neighborhood Measures

We perform two simple neighborhood analyses. First, we computed the WF-dependence of the mean number of neighbors in cosine space. Consider the metric $d_{\cos} = 1 - \cos(\theta)$ so that close neighbors have a cosine approaching one and a $d_{\cos}$ approaching zero. For every WAS vector, our algorithm traversed the range of possible inter-item distances from zero to one while counting the number of vectors within that distance from the given vector. This is a cumulative measure, as it is a population count for neighborhoods of increasing radii in cosine space. The mean populations were computed for every $d_{\cos}$ radius for each of six WF classes (Figure 5a). Second, to address the question of the WF composition of neighbors, we employed a $\cos(\theta)$-weighted frequency measure. We calculated the quantity,

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

for all WAS vectors ξ^μ, where $f_{\mathrm{KF}}(\cdot)$ is the function mapping a vector to its associated Kučera-Francis WF value. This measure is a discrete pairwise convolution of $f_{\mathrm{KF}}(\cdot)$ in cosine space. 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.

Results

We first performed the item-recognition experiment with the Hopfield FD model using an input set of random and unbiased vectors. This condition serves primarily as a control for the semantic structure of the WAS-derived vectors. The energy distributions of old and new items are binomial with means of -399 and -199 (Figure 2a, left column). Assuming random inputs with $N = 400$, the theoretical means are $\mu_\Lambda = -400$ and $\mu_\Phi = -200$. Study lists here have length $L = 100$. Two reference pool sizes are used: $P = 400$ (top row) and $P = 1600$ (bottom row) conditions are shown in Figure 2a. We refer to these as the "low-loading" and "high-loading" training conditions, respectively. They demonstrate the effects of adding noise to the network in the form of additional stored vectors. In all cases, the old and new distributions have equal variance. The low- and high-loading conditions have standard deviations s_E of 20 and 31, respectively. Theoretically, σ_E is proportional to the square root of the number of stored vectors. This increase in spread decreases d' from 10.1 to 6.5 (3) in the high-loading condition, however both values are larger than necessary for effectively perfect recognition performance ($d' > 4.6$). Overlap can be increased by reducing the magnitude of the weight update for study list items (4), but this serves as a direct comparison for the semantic case.

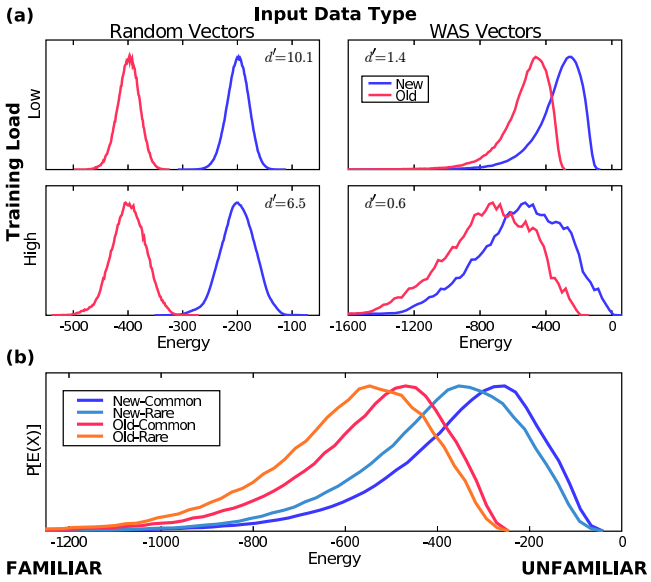

Figure 2. Effects of semantic input vectors and training load on resultant energy distributions. (a) Increasing training load for both random (left) and semantic (right) vectors increases overlap between energy distributions (100 study items, 400/1600 new items for low/high (top/bottom) training load). 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.

Semantic Inputs

The energy distributions resulting from the semantically structured input set (Figure 2a, right column) differ strongly from the random case. The distributions are not normal and

are instead negatively skewed (i.e., biased toward increasing familiarity). Furthermore, the statistics of the distributions have changed significantly. First, the distribution means are lower than the means for random vectors. In the low-loading condition, mean old and new energies are -524 and -326, respectively; however, the high-loading case shows -777 and -578, respectively. So, not only are the distributions exhibiting enhanced familiarity, the means are load-dependent. The more semantic vectors on which we train the network, the more negative energy distributions become. Second, only relative separation matters in signal detection, and these semantic distributions exhibit much lower d' distances than the random case. The discriminability as measured by d' decreases from 1.4 at $P = 400$ (top right) to 0.61 at $P = 1600$ (bottom right, Figure 2a). This is a 57% reduction in separation compared to a 36% decrease in the case of random inputs. Third, the distributions are much noisier as indicated by standard deviations of 146 and 329 exhibited by the low- and high-load conditions, respectively. These values are an order of magnitude larger than those observed for random inputs. Further, the high-loading condition produces energy distributions with noise-like irregularities that are not evident in the other cases. These were not investigated, but they could be a network capacity effect or a function of structural heterogeneity in the input space.

Statistical changes such as these could be expected for any sufficiently non-random input set. However, there are systematic differences in the energy distributions among word frequency classes. We found that vectors representing LF words tend to have lower energies, and thus enhanced familiarity, than those of HF words (Figure 2b). This is observed for frequency classes in both old and new energy distributions. Figure 2b shows the distributions for the highest and lowest frequency thirds of the study list and reference pool. This effect of increasing familiarity with decreasing WF is observed robustly across the full range of possible list and reference pool lengths. For the data shown here, using $L = 150$ and $P = 600$, all four distributions exhibit standard deviation $s_E = 192$ and both WF-dependent effects are discriminable at $d' = 0.23$. This effect is present in the new distribution and, as such, does not depend on item study (4).

ROCs computed from the WAS-based inputs and resultant familiarity distributions in Figure 2a are presented in Figure 3. The WF-dependence of these operating characteristics is shown for both the low-loading (Figure 3a) and the high-loading (Figure 3b) conditions. That is, for each training condition, the "common" and "rare" ROCs compare Λ_1 and Λ_6, respectively, to the reference pool. These are these same types of 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 also two loading effects. First, the low-loading ROCs (Figure 3a) indicate better overall performance, evident as higher HRs and lower FARs, than the high-loading ROCs (Figure 3b). Second, the WF-dependence of the ROC is greater in the high-loading than in the low-loading condition. That is, the ROCs in Figure 3b are farther separated than those in Figure 3a. Both loading effects are a result of the increase in energy variance and decrease in d' distances evident in Figure 2a. In the low-loading condition, the d' distances are 1.2 and 1.5 for common and rare words, respectively. For high-loading, 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, 1.7×10^3 stored semantic vectors is severely detrimental to FD performance, indicating that the correlations inherent the semantic input set reduce network capacity, which is analogous to the effect of repeating patterns (Bogacz and Brown 2002).

Word Frequency Effect

Calculating HRs and FARs requires a decision criterion. Here, we employ a multiple-criterion decision paradigm that compares study list frequency classes with the non-differentiated reference pool. The reasons for and possible implications of this are addressed in the Discussion. For each frequency comparison, the decision criterion λ_i is chosen as the midpoint between the empirical means of the Λ_i and reference pool distributions (5). We observed the WFE mirror effect across the possible range of the list length parameters for the study list and reference pool. Mean HR and FAR trends for the low-loading condition are shown in Figure 4a along 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. This is a symmetric 4% WFE mirror effect. Human recognition data collected by Schwartz, Howard, Jing, and Kahana (2005) are shown for comparison in Figure 4b. The experimental d' distances (Figure 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. This data approximately matches the 4% mirrored 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 (4), 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

In our experiments, we strictly used mixed-frequency lists of words chosen randomly from the WAS. The resultant energy distribution for LF words demonstrated lower energies (Figure 2b). From this and the learning mechanism of Hopfield networks, we hypothesized that LF words have more close neighbors in cosine space than common words and, further, that these neighbors tend to be other LF words. We refer to this hypothesis as the "tight clustering" of LF words. Now, consider a WF-sorted and evenly spaced 6-partition of our semantic space. Figure 5a addresses the first part of the

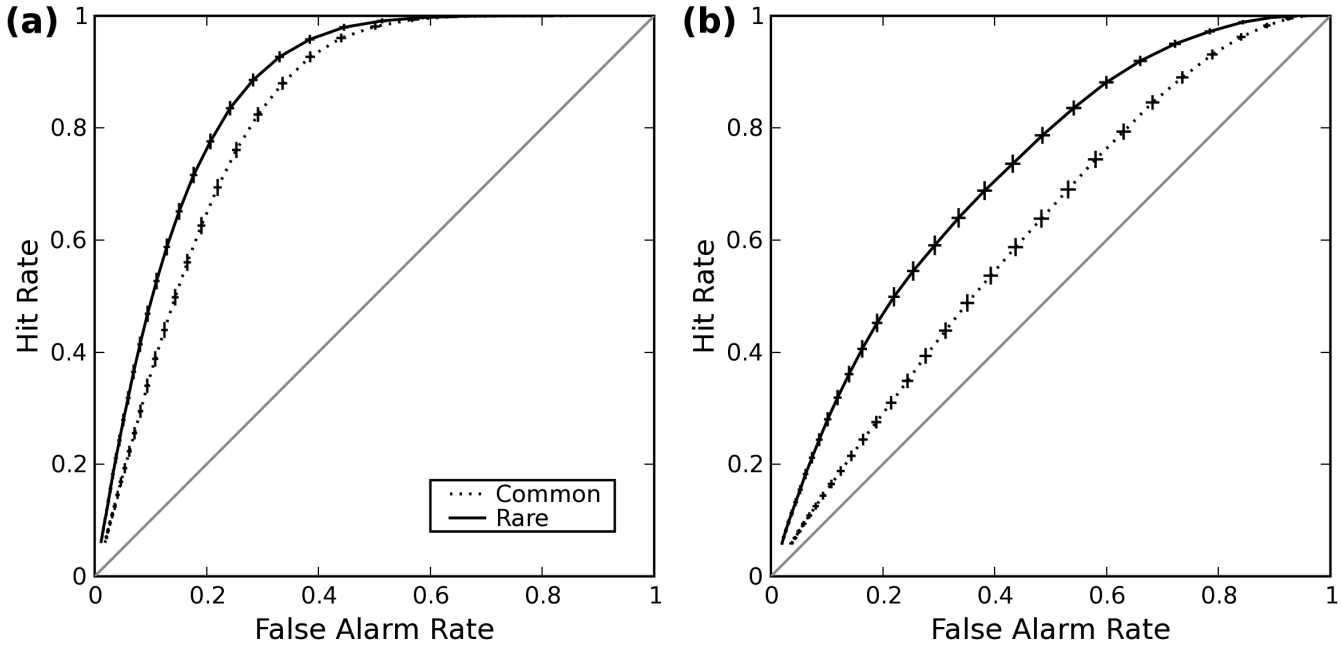

Figure 3. Operating characteristics for item-recognition performance under low (a; $P = 400$) and high (b; $P = 1600$) training load. The study list is composed of 100 items in both conditions. Performance across word frequency is assessed using a 6-partition of the study list indexed and sorted by Kučera-Francis 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).

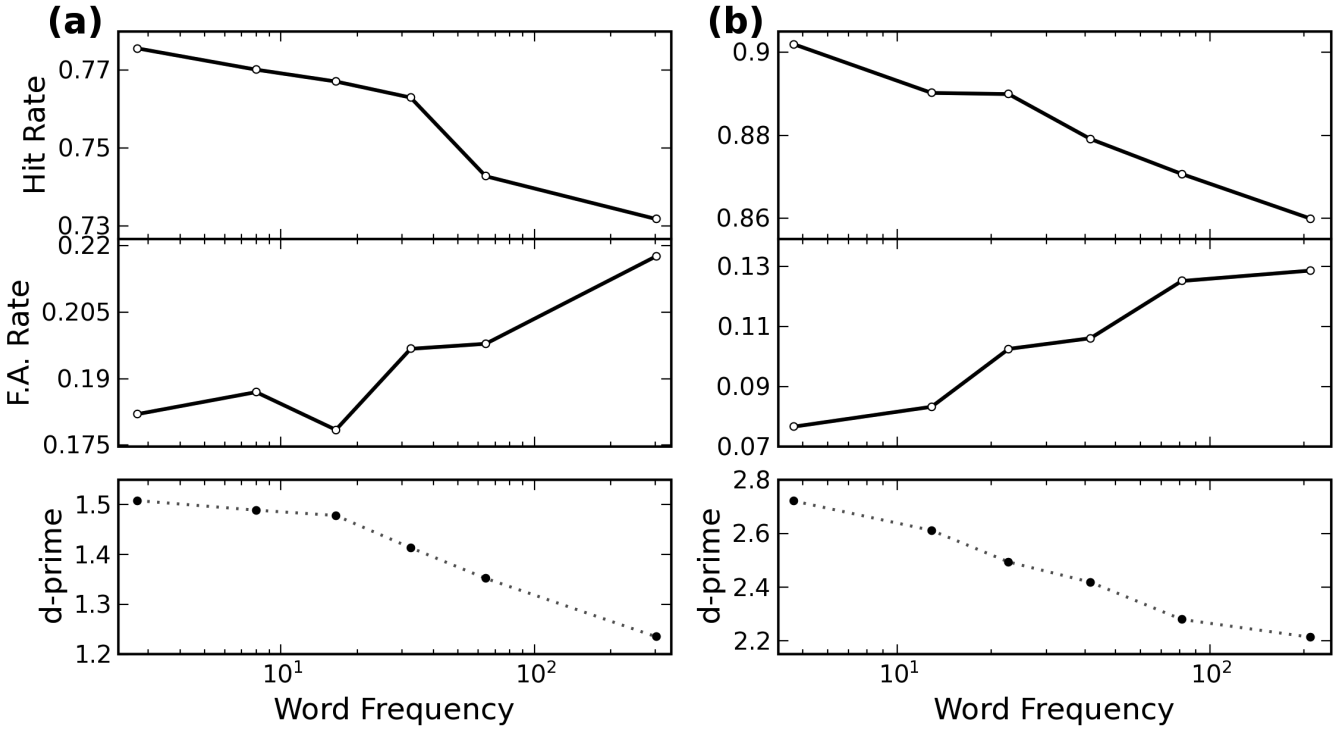

Figure 4. Word frequency mirror effect from 2,000 trials of item-recognition experiment simulations (a; $L = 100$, $P = 400$) and as seen in human memory experiments (b; data from Schwartz *et al.* 2005). The HR and FAR effects compose the mirror effect (top) and are due to changes in discriminability (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.

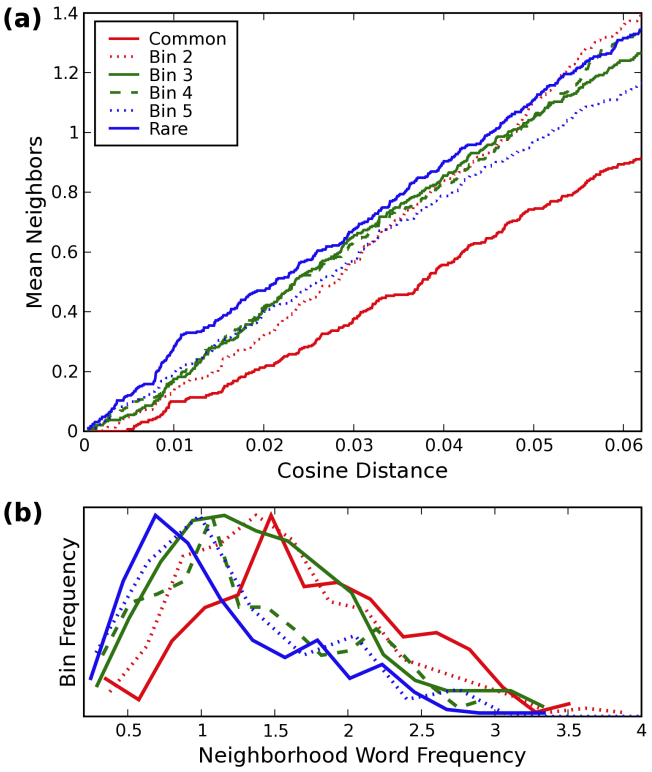

Figure 5. Word frequency-dependent clustering of word vectors in the WAS. Using a 6-partition of the word set, the mean population size for a neighborhood of a given radius in cosine space shows that rare words have more close neighbors than common words (a; the abscissa is $d_{\cos} = 1 - \cos(\theta)$).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

tight clustering hypothesis. It shows, for each of the WF bins, the mean population counts for word-centered neighborhoods of varying radii in cosine space. It counts all neighbors, irrespective of WF. Only radii up to $d_{\cos} = 0.062$ are shown, as that is sufficient to illustrate the WF-dependence of the number of close neighbors as $d_{\cos}$ approaches zero. The rarest words (blue solid line) show more mean neighbors than common words (red lines) for most of this range, and especially around $d_{\cos} = 0.01$ to 0.02. This indicates that LF words tend to have more close neighbors than HF words, which are, therefore, coded more diffusely throughout the semantic space. Figure 5b addresses the second property of tight clustering: the WF composition of those neighbors. The distributions of a discrete pairwise WF–$\cos(\theta)$ convolution (6), $\nu(\xi)$, are shown as 15-bin histograms in Figure 5b. Even though LF words have more high-$\cos(\theta)$ neighbors, the rare word convolutions are distributed to lower frequencies than those of common words. 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, we know that rare words tend to have closer neighbors than common words, so given only the $\cos(\theta)$ coefficients in (6) one would expect the distributions to be oppositely orientated from what is shown in Figure 5b. Thus, the dominant factor in these $\nu(\xi)$ values must be the WF measure, indicating that LF words are indeed tightly clustered with other LF words in WAS, whereas HF words are more diffusely self-clustered.

Discussion

Here, we bring together a simple model of familiarity-based recognition (Bogacz et al. 2001a, 2001b) and a recent model of semantic similarity among words (Steyvers et al. 2004) and demonstrate a word frequency effect. In this combination model of item-recognition, which we call *WAS–FE*, the familiarity of a probe stimulus is read out as the internal energy of a network trained on some subset of activity vectors corresponding to WAS word representations. In our experimental protocol emulating a typical word recognition experiment, the word lists are randomly chosen and words in the study list are simply presented twice to the network. We posit here that retraining is functionally analogous to having recently experienced a word in the context of an experimental study list. The WFE is dependent on the relative discriminability between study words. The only free parameters for the model are the sizes of the study and test lists, across which the WFE is robust, only differing in magnitude. These differences constitute a robust list length effect, which is due to the intrinsic capacity effects of the network-based familiarity mechanism. Notably, and as discussed below, the observed WFE is a mirror effect when decisions are determined using a stimulus-dependent criterion shift.

Input Structure Effects

If a complex object 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 (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. That is, synaptic weights are simply inner products of across-pattern activity vectors. Internal network energy (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. Because this is a measure only of independent pairwise correlations, a probe vector need not be highly similar to any given stored vector (or tight cluster of vectors) in order to yield a low network energy. This is in opposition to recognition models based on probe–stimulus summed similarity (Shiffrin and Steyvers 1997; Zaki and Nosofsky 2001), for which it has been noted that accounting for inter-item association is beneficial (Kahana and Sekuler 2002). For random and unbiased pattern sets and probe vectors, the statistics of the weights and energies can be determined *a priori* (see Methods); sample energy distributions resulting from the item-recognition experiment are shown in Figure 2a. The WAS-based semantic vectors are not activity biased, and the

lists in each trial are unbiased with respect to word frequency. Training and testing with the semantic inputs produce energy distributions which are negatively skewed and exhibit statistics different from the random case (Figure 2a). We suggest that this familiarity effect is attributable to the introduction of a structured input space.

It is possible to discern the mechanism underlying this structure-induced effect. Because the input space exhibits small-world structure, characterized by high local connectivity but short average path lengths (Watts and Strogatz 1998; Steyvers and Tenenbaum 2005), there must be subsets of input vectors that significantly share pairwise activation properties. Each of these groups, or clusters, will bias those synaptic weights corresponding to their respective set of shared features. Local groups of vectors will result in attractors with strength directly related to the size and connectivity of the corresponding cluster. A probe vector can yield a relatively low network energy only if its pairwise bit structure correlates signficantly with the features that form the correlations learned by the network. Further, these matching features may be characteristic of distinct attractors in the network: there is a combinatoric aspect to the diversity of such "spurious" attractors (Amit 1989). Here, this means that a probe vector judged as highly familiar need not match any single semantic input or cluster but instead have a pairwise bit structure with elements characteristic of different clusters.

From this network perspective and the relatively low means for the new distributions (Figure 2a), we can infer that the vectors in the semantic input space tend to be near learned or spurious attractors of the network. This is expected in a small-world network, because most vectors are part of the local clusters which give rise to the network attractors in the first place. The presence of spurious attractors further contributes to lower energies across the space. Additionally, the observed shape of the energy distributions is a function of the density and spatial distribution of vector clusters.

Vectors populating a space with a non-random large-scale structure carry redundant information in their correlations. This redundancy serves, similarly to repeated presentations (Bogacz and Brown 2002), to decrease the effective recognition capacity of the network. This is evident in the large drop in discriminability (3) between the random and semantic input spaces and between the low- and high-loading training conditions (see d' in Figure 2a). It is inefficient in terms of familiarity processing, but results in much more realistic (i.e., measurably worse than perfect) recognition performance. Also, orthogonalization of stimulus coding is thought to occur downstream of IT cortex in the dentate gyrus (O'Reilly and McClelland 1994; Kesner, Gilbert, and Wallenstein 2000), so presumably the raw structure of the semantic space is available to the FD neurons in PRC.

Frequency Effects via Structure

The WFE of recognition memory is one of the most robust and extensively studied human memory effects (e.g., Schulman 1967; Shepard 1967; Glanzer and Adams 1985; Guttentag and Carroll 1994; Karlsen and Snodgrass 2004). On the hypothesis that such an ubiquitous effect has a basis in semantic coding and so might be evident within *WAS–FE*, we sorted and partitioned the task lists into WF bins using a normative measure of frequency (Kučera and Francis 1967). These frequency-differentiated bins resulted in separable energy distributions for both old and new lists (Figure 2b). Further, this frequency effect is in the observationally correct direction of increased familiarity for rarer words: e.g., in Figure 2b, the energies for each LF ("Rare") bin are distributed more negatively than those of the corresponding HF ("Common") bin. This effect is robust across all free parameters in the model, which consist only of the list lengths.

Given this effect on familiarity judgements, there must be some structural or statistical characteristic of the WAS vectors that causes the effect: aside from the network mechanism, there simply is nothing else in *WAS–FE* to cause it. Everything besides semantic structure as represented by the WAS is controlled for in the random input case (Figure 2a). We performed two simple analyses of the original WAS vectors to support the hypothesis that LF words tend to be tightly clustered with other LF words. Cumulative population counts across distance in cosine space (Figure 5a) show that LF words have more close neighbors than HF words. The structural hypothesis only makes sense if those neighbors tend to be other LF words and that the sparser neighborhoods of HF words tend to consist of HF words. Using a discrete pairwise cosine space convolution of word frequencies (6), we can get an indication of the WF composition of the neighborhood of a given WAS vector. The distributions of these convolutions, sorted into their respective WF bins, do indeed demonstrate the required tendencies that words are encoded similar to other words of the same WF class (Figure 5b). This WF-based coding scheme is evident despite the tighter LF clustering shown in Figure 5a: the closest neighbors of LF words contribute larger $\cos(\theta)$ factors to the convolution (6) than HF words, so it follows that if clusters were heterogenous with respect to WF then the opposite tendency would be observed.

These two structural characteristics of the semantic input vectors are sufficient cause for the differentiation of relative stimulus familiarity observed in Figure 2b. Among single-process recognition models, there has not been a consistent approach for the structural representation of semantic stimuli, but the intuitive line of thinking seems to favor tighter clustering for HF words. 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. Notably, a normative measure of context variability has been shown to have a recognition mirror effect independent of WF (Steyvers and Malmberg 2003). 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 is based on the hypothesis that fewer features of HF words are attended to by the subject. This effectively imposes a reduction of semantic information in HF words, which is functionally similar to adding noise or placing them in tighter spatial clusters. Algorithmically, these previous models combine feature matching with the computation (or estimation) of log-likelihood ratios. Here, we use simple Hebbian learning and a physical energy computation.

Decision Process & Performance

The WFE is fundamentally a behavioral effect of recognition performance, so a decision-making process is needed. The human 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. Accordingly, our decision process must be able to demonstrate both the HR and FAR aspects of the LF enhancement if *WAS–FE* is to have any explanatory power.

As presented here, *WAS–FE* may be classified as a single-process signal detection model of recognition: LF and HF stimuli are not dissociated or processed any differently within the model, which only outputs a continuous-range energy value. This familiarity signal is noisy, as described by the distributions in Figure 2, and a decision must be made whether a given probe was studied (old) or not (new). Commensurate with the the simplicity of *WAS–FE*, we consider a familiarity threshold below which a probe is recognized as familiar. In the case of random vectors, Bogacz et al. (2001a, 2001b) simply used the midpoint of the theoretical means of the two binomial energy distributions as a decision criterion. With skewed distributions and empirically determined statistics, we can translate that here as the midpoint of the empirical means of the distributions. The question becomes that of which two distributions, in particular, are being compared in the decision process. The answer to this affects the interpretation of *WAS–FE* as a candidate explanation for the recognition WFE.

There are four possible decision comparisons: the WF-differentiated bins or entirety of the study list against the WF-differentiated bins or entirety of the reference pool. In the notation used above, these are Λ_i–Φ_i, Λ_i–Φ, Λ–Φ_i, and Λ–Φ, respectively, where i traverses WF bins. The Λ–Φ comparison is not WF-dependent and thus meaningless in terms of the WFE. Considering a means-based familiarity threshold, the Λ_i–Φ_i comparison will not produce either component of the mirror effect because the WF-dependence of both distributions is the same, resulting in no net performance change. A fixed, WF-independent criterion could be used, but it would reverse the FAR component of the mirror effect. 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 entirety of the other. The Λ–Φ_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 WF bins have more overlap with the Λ distribution. However, the remaining Λ_i–Φ comparison addresses both counts: cognitively and intuitively, it makes sense that the subject has information regarding the word frequency classes of recently studied stimuli; and performance increases with rarity because the old WF bins are farther from the Φ distribution. Also, this comparison requires WF-dependent decision criteria, as a fixed-criterion process would not yield a false alarm effect. There are different possible forms for a stimulus-dependent criterion shift, but most tend to be based on the hypothesis that the criterion increases in the signal direction "with the memorability of old items" (Hirshman 1995). Since we hypothesize *WAS–FE* to be capturing only gross characteristics of recognition processing, we choose simply to follow Bogacz et al. (2001b) and use a means-based midpoint criterion. Thus, supposing that the semantic familiarity process of *WAS–FE* can at least partially form a causal basis for the WFE mirror effect delimits certain requirements for both the discimination comparision (Λ_i–Φ) and the decision criterion (5). The resultant WFE has mirrored HR and FAR effects that match recognition data (Figure 4).

Criterion-independent performance results can be illustrated by old–new operating characteristics. For both small (Figure 3a) and large (Figure 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 correspond to the Λ_6–Φ and Λ_1–Φ comparisons, respectively. The shapes of these characteristic curves derive from the non-Gaussian form of the corresponding energy distributions (Figure 2a, right column). As such, they largely resemble those of other recognition models except that they are not symmetric around the negative diagonal. The corresponding z-space ROC curves (not shown) typically consist of a short linear region (slope less than unity) followed by a longer linear region (slope greater than unity). These two examples of ROCs also demonstrate two performance effects of the number of trained stimuli. The low-loading condition (Figure 3a) shows high absolute performance, but not a relatively small WFE; the high-loading condition (Figure 3b), however, shows worse overall performance, but a larger difference between the operating characteristics for LF and HF words. These are capacity effects of the attractor-based FD mechanism. Larger lists entail more synaptic noise and less recognition accuracy. Further, we can infer that such increases in synaptic noise affect the performance of HF words more than LF words. This WF-dependence may be a result of the relatively sparse encoding of HF words in the WAS: the weaker attractors are more sensitive to synaptic noise than the strong attractors associated with LF word clusters. Thus, *a fortiori*, for any given reference pool size, increases in study list length push the network closer to capacity decreasing HRs and increasing FARs regardless of WF class (data not shown). That is, *WAS–FE*

coupled with a means-based criterion shift exhibits a list-length mirror effect, which previous single-process models have also demonstrated (Shiffrin, Ratcliff, and Clark 1990; Shiffrin and Steyvers 1997; McClelland and Chappell 1998). Conversely, for constant study list length, this model predicts better overall recognition performance and a smaller WFE for subjects with relatively less exposure to semantic objects (e.g., children vs. adults).

Attaining a qualitatively correct false alarm effect was our criterion for choosing a decision process here. However, explaining FAR effects with a criterion shift remains controversial. Stretch and Wixted (1998) provide evidence that the strength-based but not the frequency-based recognition mirror effects depend on criterion shift. However, Miller and Wolford (1999) argued that a signal detection account of a simple false memory paradigm does support criterion shift as a mechanism for generating recognition false alarms. This was refuted (Roediger and McDermott 1999; Wixted and Stretch 2000) in part by asserting the compatibility of the other models with such an account (Wickens and Hirshman 2000). As such, the role of criterion shift in recognition false alarms has not been conclusively determined. It remains that if *WAS–FE* is to demonstrate WF-dependent recognition differences, then it must employ a criterion shift. Further, subjects performing recognition tasks are able to modulate their decision criteria on-line according to stimulus class in order to optimize performance (Hirshman 1995; Heit, Brockdorff, and Lamberts 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, Förster, and Werth 2005). The multiple-criterion decision process required here (5) by *WAS–FE* is plausibly in line with these observed decision strategies in human subjects.

Role of Contextual Information

Dual-process recognition theories employ 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, Fisher, Henthorn, Arndt, and Passannante 2002; Mintzner 2003), but not to the ultimate exclusion of single-process accounts (Malmberg, Zeelenberg, and Shiffrin 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. The process interaction implied here is different from the frequency tradeoff proposed in the source of activation confusion (SAC; Reder et al. 2000) model. In the decision comparison Λ_i–Φ, words from the study context are treated categorically as members of their respective WF bins. However, non-study probes are not likewise differentiated. This is a contextual distinction that is necessitated by *WAS–FE* as discussed above. This is not to say that some recollective process is making perfect old–new discriminations only to then involve an error-prone familiarity process. Instead, the contextual distinction consists of the subject having formed stimulus categories such as frequency class 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 seems plausible. Also, further episodic information could allow the discrimination between, for example, several study lists in a session. This could be modeled within the framework of *WAS–FE* by integrating a drifting context code (e.g., Howard and Kahana 2002). Regardless, the main point here is that the decision comparison requires some contextual distinction of this sort in order for *WAS–FE* to yield a proper WFE mirror effect.

Dual-process models like SAC predict differential recruitment of recollective processing. Given this, it would be expected to see a significant WF-dependent modulation of activation for regions involved in controlled item retrieval such as the hippocampal formation and other MTL structures. In contrast, *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, which we posit as the basis of the WFE, shows WF modulated activation. Using event-related fMRI at retrieval, 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; Zubicaray et al. 2005). Thus, LTC is well-positioned as a possible semantic input region for familarity processing in PRC. Zubicaray et al. (2005) suggest these results are consistent with context-noise models of the recognition WFE, but they are also consistent with our account here. Recently, electrophysiological measures such as EEG have been able to dissociate verbal from nonverbal retrieval (Hwang et al. 2005), indicating the possibility of investigations using higher temporal resolution techniques. More such studies are needed to complement the large body of behavioral data.

Conclusion

In the present work, we take advantage of a recent empirically determined model of semantic space to demonstrate a benchmark effect of human memory. Using this 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 non-intuitively 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.

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, U. K.: Cambridge University Press.

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: 830-42.

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 Sc* 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. *Phil Trans R Soc Lond B* 357: 1083-1095.

Erickson, C. A., Jagadeesh, B., and Desimone, R. (2000). Clustering of perirhinal neurons with similar properties following visual experience in adult monkeys. *Nature 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.

Fuentes, U., Ritz, R., Gerstner, W., and Hemmen, J. L. van. (1996). Vertical signal flow and oscillations in a three-layer model of the cortex. *J Comput Neurosci* 3: 125-136.

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 Nat Acad Sci, USA* 84: 8429-8433.

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

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.

Karlsen, P., and Snodgrass, J. (2004). The word-frequency paradox for recall/recognition occurs for pictures. *Psychol Res* 68: 271-276.

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.* Providence, RI: Brown University Press.

Levy, D. A., Bayley, P. J., and Squire, L. R. (2004). The anatomy of semantic knowledge: Medial vs. lateral temporal lobe. *Proc Nat Acad Sci USA* 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., 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. *Nature 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. (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 Meth Instr* 36: 402-407.

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. *Pscyhol 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., Ratcliff, R., and Clark, S. (1990). The list-strength effect: II. Theoretical mechanisms. *J Exp Psychol Learn Mem Cogn* 16: 179-195.

Shiffrin, R. M., and Steyvers, M. (1997). A model for recognition memory: REM - retrieving effectively from memory. *Psychon Bull Rev* 4: 145-166.

Sobotka, S., and Ringo, J. L. (1993). Investigation of long term recognition and association memory in unit responsses 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., Shiffrin, R. M., and Nelson, D. L. (2004). Word association spaces for predicting semantic similarity effects in episodic memory. In A. F. Healy (Ed.), *Cognitive psychology and its applications: Festschrift in honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer.* Washington, DC: American Psychological Association.

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.

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.* New York: Oxford University Press.

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.

Zubicaray, G. I. de, 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.