A Diffusive-Particle Theory of Free Recall

Francesco Fumarola

Center for Theoretical Neuroscience, Department of Neuroscience, Columbia University, New York, NY, USA

ABSTRACT

Diffusive models of free recall have been recently introduced in the memory literature, but their potential remains largely unexplored. In this paper, a diffusive model of short-term verbal memory is considered, in which the psychological state of the subject is encoded as the instantaneous position of a particle diffusing over a semantic graph. The model is particularly suitable for studying the dependence of free-recall observables on the semantic properties of the words to be recalled. Besides predicting some well-known experimental features (forward asymmetry, semantic clustering, word-length effect), a novel prediction is obtained on the relationship between the contiguity effect and the syllabic length of words; shorter words, by way of their wider semantic range, are predicted to be characterized by stronger forward contiguity. A fresh analysis of archival free-recall data allows to confirm this prediction.

KEYWORDS

cognition, neurosemantics, experimental semantics, episodic/short-term memory, free recall, psycho-linguistics, neurolinguistics

FREE RECALL: MATRIX MODELS AND GRAPH MODELS

Free-recall experiments are a key tool for the controlled investigation of episodic memory. A typical free-recall experiment takes place in two stages: During the "presentation stage", subjects are shown a list of words; during the "memory test", they are requested to recall them in any order.
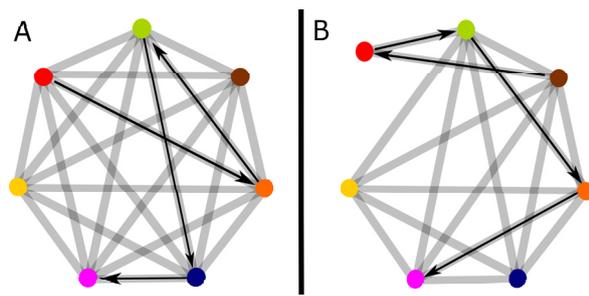
Some of the main effects reported are:

1. *Power-law scaling*: The number of retrieved items scales like a power law of the number of items in the list (Murray, Pye, & Hockley, 1976).

2. *Primacy* and *recency effects*: The first and last words in the list are recalled better than the rest (Murdock, 1962).

3. *Contiguity effect*: Items contiguous within the list tend to be recalled contiguously (Kahana, 1996).

4. *Forward asymmetry*: The tendency to recall items in forward order (already reported in Ebbinghaus, 1913).

5. *Semantic clustering*: Semantically related words tend to be recalled successively (Bousfield & Sedgewick, 1944).

6. The *word-length effect*: Lists of shorter words are recalled better than lists of longer words (Baddeley, Thomson, & Buchanan, 1975).

The contiguity effect, the recency effect, and several other phenomena, are now well understood by means of retrieved-context theories of episodic memory, such as the *temporal context model* of Howard and Kahana (2002). In these theories, the recovery of a memory is mediated by the recovery of its "temporal context," and temporal contexts are modeled through a matrix representation that undergoes a linear evolution in time.

While the effectiveness of these theories is undisputed, recently Romani, Pinkoviezky, Rubin, and Tsodyks (2013) have introduced a somewhat different approach to the modeling of free recall. After studying the process of memory retrieval on a mechanistic neural model, they introduced the idea of an "average graph" of attractors, and modeled free recall as diffusion on that graph (Romani et al., 2013, Appendix A2).

A "graph" is a mathematical object usually depicted as a set of dots (called *nodes*) joined by lines (called *edges*, see Figure 1, Panel A). In

Corresponding author: Francesco Fumarola, Center for Theoretical Neuroscience, Department of Neuroscience, Columbia University, New York, New York 10032, USA. E-mail: ff2154@columbia.edu

FIGURE 1.

Panel A: free recall as diffusion through a complete graph; the gray lines are the edges of the graph, the colored spots are the nodes, and a possible trajectory is shown as a sequence of black arrows. Panel B: free recall as diffusion through a noncomplete graph; the word depicted as a red node is now linked only to the *green* and *brown* words; as a consequence, *red* must be recalled contiguously with *green* or *brown*, whatever their serial position in the list (semantic clustering).

the approach of Romani et al. (2013), the psychological state corresponding to each word is modeled as a node in a graph. The number N of nodes in the graph is thus the number of words in the list.

Retrieval is effected by a *diffusive particle* moving over the graph. At each moment in time, the particle's position is at one of the nodes in the graph, and the subject's psychological state is encoded as the current position of the particle. The particle moves from node to node by travelling along the available edges of the graph. If the currently occupied node is an endpoint of multiple edges, one edge will be chosen at random amongst them, and will be travelled along by the particle (see Figure 1, Panel A). One says that the particle is *diffusing* over the graph, this type of motion being known as diffusion.

For example, if there are three edges departing from the currently occupied node, each will have a 1/3 probability of being chosen, and each choice will lead the diffusive particle to move on to a different node. Whenever the particle moves on to a certain node, the word associated to that node is recalled. Diffusion is terminated when the path self-intersects.

Romani et al. (2013) introduced this theory as a toy version of their neural-network model, and used it to compute explicitly the power-law scaling of retrieval. The calculation of this power law (as done in Romani et al., 2013, Appendix A2) assumes that the average graph over which diffusion takes place is complete—that is, every pair of distinct nodes is connected by an edge, as in Panel A of Figure 1 (for a simple introduction to graphs, see Frieze & Karonski, 2016). As a result, the power-law exponent is found to be ½, which is indeed close to experimental values.

This is a substantial result that may not have been as easy to obtain through more conventional theories, and, as such, it encourages further exploration of graph methods in the study of free recall. This motivated the present paper. While the argument of Romani et al. (2013) is sufficient to extract the power-law exponent, it is far from

providing a general understanding of free recall. In this paper, a more versatile graph-based theory is proposed, which proves able to provide an explanation for several known effects and to predict a new effect emerging from experimental data.

I begin by introducing, in the next section, a more realistic family of graph models, allowing for both missing edges and multiple meanings, and I proceed to demonstrate that the resulting theory exhibits both semantic clustering and forward asymmetry. I then recall some well-established results from linguistics concerning the correlation between meaning and word-length. Applying these to the diffusive-particle model yields a whole new prediction on the correlation between word-length and the contiguity effect. This prediction is tested through an original analysis of archival free-recall data. I then show that the underlying mechanism can easily explain another well-known feature of free recall, the word-length effect. To conclude, I discuss the application of the diffusive-particle approach to some further aspects of free recall.

A SEMANTIC GRAPH WITH RANDOM EDGES

When a pair of semantically related words (e.g., *pear* and *apple*) is embedded in the list to be recalled, the related words are often recalled contiguously. This tendency to successively recall semantically related words is known as semantic clustering (Bousfield & Sedgewick, 1944).

The toy model that Romani et al. (2013) described is unable to reproduce such an effect or any other phenomenon strictly dependent on semantics. This is no longer true, however, if we relax the assumption that the graph is complete—that is, if we remove some of the edges. The pairs of words linked by edges can then be interpreted as being semantically connected; we may thus refer to the graph as a *semantic graph*.

If the recall process is modeled in terms of diffusion on a semantic graph, semantic clustering is inevitable; two nodes that are more closely connected are more likely to be visited successively by any diffusive process. This holds true independently of the serial positions of the words whose meaning was found at the nodes.

A simple example of this is shown in Panel B of Figure 1. The *red* word is connected only to the *green* and *brown* words; by necessity, *red* will be recalled contiguously to *green* and/or to *brown*, even when those two words are located far from *red* within the list.

Since missing edges are now being allowed, the question arises of which edges should be assumed to be missing, and which should survive. In principle, a cluster analysis of textual corpora may help with this estimate; for example, words that appear mostly at close distance from each other may be assumed to be semantically related and the corresponding nodes to be linked by an edge. The criteria for such an analysis, however, involve an inevitable degree of arbitrariness. Moreover, because semantic associations are built through individual experience, they vary from subject to subject over any population.

Uncertainty and variability may both be taken into account by assuming the edges to be chosen probabilistically. The semantic graph is then a probabilistic graph with a fixed number (N) of nodes but a random choice of the edges.

In principle, this means that any graph with N nodes (including the complete graph) has some probability of being the semantic graph. Call $P(G)$ the probability that a specific graph G is the semantic graph. The distribution $P(G)$ encodes the probabilistic structure of semantics. The quantities we would like to predict (recall probabilities) can be computed by simulating trajectories on each possible graph G; results must then be averaged over all such graphs, and the averaging weighed with the factor $P(G)$.

Graph models of free recall may thus become helpful, among other things, as part of an endeavor to elucidate the semantic graph empirically. The connections between various semantic contexts are encoded in the distribution $P(G)$. If we compute recall probabilities for various choices of the distribution $P(G)$ and compare them with experimental values, the true structure of the semantic graph will emerge as the choice of $P(G)$ that yields the best agreement with the data.

In this paper, I try out the simplest possible trial distribution $P(G)$, which relies on no lexicographic knowledge and depends on a single parameter. This is done by assuming that all edges of the complete graph are kept or removed independently of each other, and each has a probability α of being removed. Otherwise said, the parameter α is the probability that any two nodes are *not* connected.

If α = 0, the semantic graph is (with probability equal to 1) the complete graph; for arbitrary values of α the probability associated to a specific graph with n edges is found to be $P(n) = (1 - \alpha)^n \alpha^{\frac{N(N-1)}{2} - n}$.

INTRODUCING POLYSEMY

Before computing measurable quantities—that is, recall probabilities, we must notice a second limitation to the model used in Romani et al. (2013). The "average attractor graph" considered therein represents every word in the vocabulary as a single node. Yet, fMRI measurements have convincingly shown that the neural response to free-recall tests exhibits a strong statistical dependence on the semantic variability of words (Musz & Thompson-Schill, 2015).

In linguistics, the degree of dependence of a word's meaning on context is called *polysemy* (Nerlich, Todd, Herman, & Clarke, 2003). Of course, since meaning is inevitably affected by context, no word is perfectly *monosemic* (i.e., having a single nuance of meaning); a word with comparatively little semantic variability is called *oligosemic* (Fernando, 1996). To graft polysemy into the graph model, we must identify the nodes of the semantic graph with meanings (or semantic nuances) rather than with words, allowing each word to label multiple nodes.

A word W will then have a degree of polysemy $k(W)$, defined as the number of nodes corresponding to word W. In the simplest scenario, the degree of polysemy will have a constant value K, the same for all words (see Figure 2, Panel B).

If the semantic graph is complete, each node will be linked to $K - 1$ nodes corresponding to the same word, and to K nodes corresponding to every other word in the vocabulary. If the semantic graph is random and its probability distribution characterized by a disconnectedness parameter α, a node corresponding to any given word will be linked on
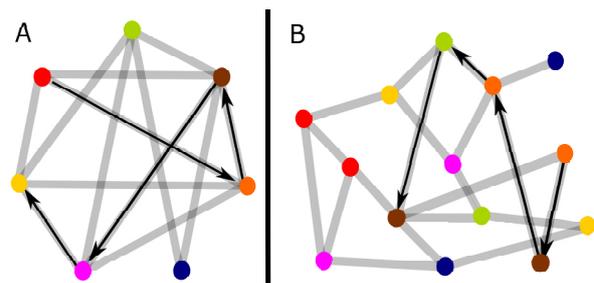

FIGURE 2.

Panel A: diffusion through a noncomplete graph; some of the edges are missing—that is, some pairs of nodes are not directly connected, and the particle can only travel along the available edges. Panel B: diffusion on a noncomplete graph with the inclusion of polysemy; in this particular example, each word has two semantic nuances, or meanings, represented by as many nodes. Nodes of the same color represent different meanings of the same word; edges (i.e., available connections between meanings) are again shown.

average to $(1 - \alpha)K$ nodes corresponding to every other word, as well as $(1 - \alpha) \times (K - 1)$ same-word nodes.

Given that each word corresponds to multiple nodes, a question arises concerning the retrieval process. Will a word be recalled when the diffusive particle touches any of the nodes corresponding to it? Or will each memory be encoded in a given node?

The literature on context-retrieved theories strongly suggests that the latter option holds true. Indeed, it has been proven that memories are anchored to the *contextual* region where they have been created during the presentation of the list (Howard & Kahana, 2002). Hence, if a word has multiple meanings, its recall will require retrieving the specific meaning that was attributed to that word during presentation.

In order to know which node corresponds to a given memory, we need to formalize the dynamics during presentation, which can be simply modeled as another diffusive process on the semantic graph. At every instant during the presentation stage, the diffusive particle lies on a definite node of the graph; once a word is presented, the particle diffuses until it recognizes that word—that is, until it stumbles on one of the nodes corresponding to it.

This process has an interpretive function: The system interprets each word through the meaning of that word on which the diffusing particle stumbles first, and that particular node becomes the location of the memory corresponding to the word.

Notice that, however, this recognition may never occur, as the graph has a finite probability of being composed of several noncommunicating subgraphs; if there is no path leading from the current position of the particle to any of the word's nodes, the particle is allowed to jump on to a node randomly chosen amongst them.

This interpretive process takes place for each word in succession: once a word has been interpreted, the next word in the list is presented, and the diffusion goes on. Thus, memories are created.

The model includes, therefore, two diffusive trajectories—one effecting the interpretation of words and one effecting their retrieval (see Figure 3). These two trajectories are meant to model processes rooted in different cognitive abilities, so it would be more correct to speak of two different particles, one employed for interpretation and one for retrieval.

In practice, we must only be cognizant that the two diffusive trajectories may develop over different time scales. The model itself is of course too abstract to provide an independent estimate of the two time scales.
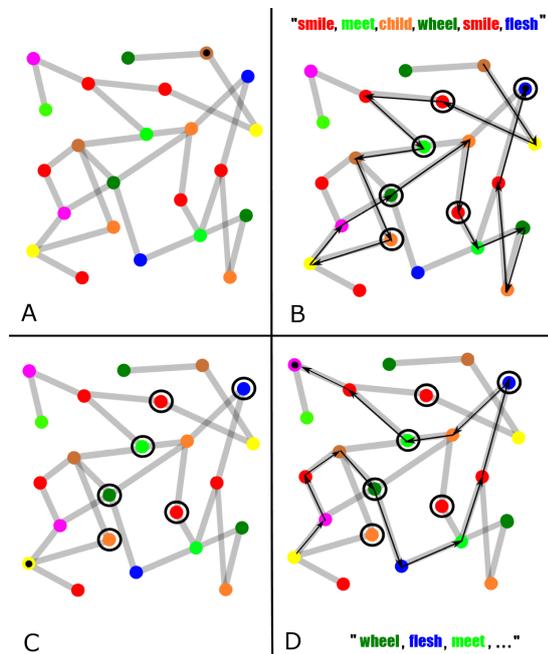
FORWARD ASYMMETRY

FIGURE 3.

Diffusive-particle model of a free-recall experiment. Panel A: a semantic graph, shown with a specific choice of its edge structure among the many such structures over which final results must be averaged; meanings corresponding to the same word are shown in the same color; the current position of the particle is indicated by a black dot. Panel B: presentation stage; each time a new word is presented, the particle keeps diffusing until it lands on any of the meanings described by that word; the resulting trajectory is shown as a sequence of arrows. Panel C: The nodes where meaning has been found during presentation have become transient memories (circled nodes); in the interval between presentation and memory test, the diffusive particle's position is reset to a random point indicated by the black dot. Panel D: During the memory test, a new diffusive process takes place, similar to the one described in Romani et al. (2013). The diffusive particle has to locate the circled nodes for the corresponding words to be recalled.

Let us call any sequence of two consecutively recalled words *transition*. Obviously, neither the first word retrieved in the recall stage nor an intrusive word or a word recalled after an intrusion is retrieved as part of a transition. We will call the difference between the serial positions of two words in a given transition *lag*; for example, if the fifth word in the list is recalled right after the eighth, the corresponding lag is $L = -3$.

In addition, let us call $p(L)$ the lag probability distribution—that is, the probability that an arbitrary transition will have a lag L. *Forward asymmetry* is the empirical fact that $\sum_{L>0} p(L) > \sum_{L<0} p(L)$—that is, lags are more often positive than negative, meaning that forward transitions are preferred; as we will see, this fact is due almost entirely to the contribution from contiguous transitions ($L = \pm 1$).
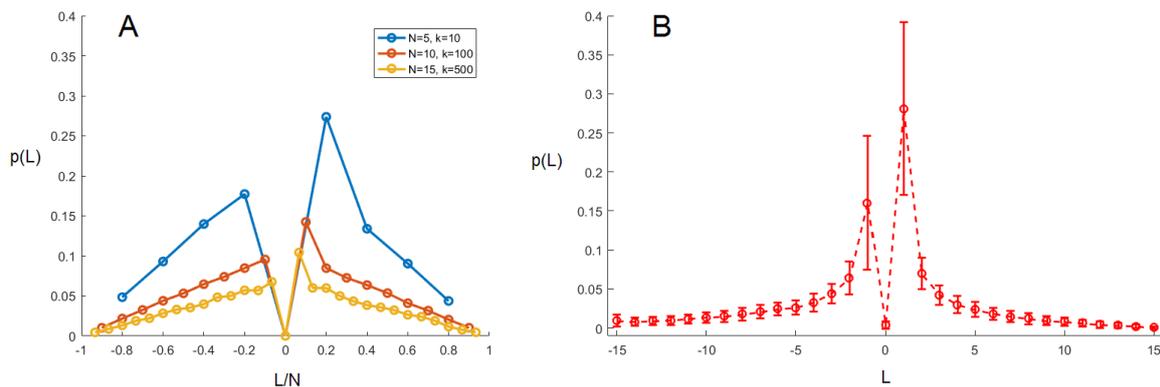
To compute $p(L)$, we proceed to simulate the diffusive-particle model. All simulations presented in this paper consist of the following steps:

1. A function $N(\kappa)$ is defined, describing the number of words with polysemy κ in the vocabulary; hence, the vocabulary has size $N_V = \sum_\kappa N(\kappa)$ and the graph contains $N_G = \sum_\kappa \kappa N(\kappa)$ nodes.

2. The semantic graph for a given subject is created by picking a matrix $N_G \times N_G$ whose elements are 0 with probability α (corresponding to two unconnected nodes) and 1 with probability $1 - \alpha$ corresponding to connected nodes).

3. A list of words to be recalled is generated by picking a random permutation of the vocabulary (i.e., a permutation of the first N_V integers).

4. Submission/interpretation of words in the list is simulated as diffusion through the semantic graph; whenever a node corresponding to the currently submitted word is met, a memory is recorded at that node, and the next word is presented.

5. The retrieval of memories is simulated as a second diffusion process starting from a random node; each memory met along the way is recorded as a new recall event, and the trajectory ends when it self-intersects.

6. Steps 3-5 are repeated a sufficient number of times to ensure the convergence of recall probabilities; this amounts to presenting multiple lists to a given subject.

7. Steps 2-6 are repeated on a large number of subjects—that is, for many different semantic graphs.

The dataset thus generated has the structure of realistic free-recall data; in particular, the number $n(L)$ of recall events with lag L can be divided by the total number of transition events to yield an estimate of the lag probability $p(L)$.

The results in Panel A of Figure 4 refer to graphs with $N(\kappa) = N \delta_{K,\kappa}$ (a Kröneger delta)—that is, all N words have the same degree of polysemy K. Thus simplified, the theory depends on only three parameters: the vocabulary size N, the polysemy level K, and the semantic disconnectedness α.

In the figure, the frequency of transitions has been plotted for various choices of these three parameters. As we are not considering repetitions, by construction, the curve vanishes at $L = 0$. The main features of the curve, as can be seen, are analogous for various combinations of

FIGURE 4.

Panel A: results of simulations of the diffusive-particle model for three choices of the vocabulary size *N* and polysemy *K* (see legend) and for α =1–1/*K*. Lists presented to the model were permutations of the whole vocabulary. The y-axis shows transition frequencies, the x-axis - the serial-position lag normalized by the size of the lists. Panel B: transition frequency as a function of lag, as computed from Penn Electrophysiology of Encoding and Retrieval Study (PEERS) data.

parameter values. There are two maxima at $L = \pm 1$, and the transition probability is a decreasing function of $|L|$, the absolute value of the lag.

Moreover, the curve is not symmetric around $L = 0$: The forward branch sums up to a larger cumulative, although it lies higher up only insofar as the peak at $L = 1$ is concerned. I will refer to this peak as the *sequential* peak, and to forward contiguous transitions as *sequential transitions*. The sequential peak is always considerably higher than the backward contiguous peak—a phenomenon widely documented in experiments (see Kahana, 2012).

To provide an example of how these features emerge in empirical results, Panel B of Figure 3 displays the curve of transition frequencies for archival data from Penn Electrophysiology of Encoding and Retrieval Study (PEERS), a large study conducted at the University of Pennsylvania. The data are those described in Lohnas, Polyn, and Kahana (2015), summing up to a total of 7,360 free-recall trials on 92 subjects, all performed with lists of 16 words. Participants consented according to the University of Pennsylvania's institutional review board (IRB) protocol and were compensated for their participation. Intrusions have been discarded from these data, and no availability correction has been introduced; repetitions, which are comparatively rare, have been counted in under the lag $L = 0$.

In the dataset corresponding to each subject, transition events with the same lag have been grouped, counted, and normalized by the total number of transition events to yield the subject's curve of transition frequencies. The averages of these curves over all subjects and the *SD*s of the corresponding distributions are shown respectively as the solid curve and the error bars of Panel B of Figure 4.

The empirical curve thus obtained and the curves obtained from simulations are not identical. Nonetheless, the features we have outlined above are prominent in both. In particular, the difference between the backward and the forward branch of the curve is concentrated in both cases at contiguous transitions, and the maximum at $L = 1$ is always the global maximum of the distribution. This is a substantial feature non-trivially displayed by the model, and the mechanism behind it should become clearer in the next two sections.

WORD LENGTH AND POLYSEMY

In the previous section, we simulated the model under the assumption that all words have the same degree K of polysemy—that is, the same number of semantic nuances. This is not the case in real-life experiments, and we may wonder how the recall probability of a word varies as a function of the word's degree of polysemy.

Polysemy is unfortunately a somewhat elusive variable, subtle to measure (Nerlich et al., 2003). Consider for instance the two words *lion* and *lioness* (a classic example); does the meaning of *lioness* vary with context? Surely less than the meaning of *lion*, because, aside from finer distinctions, the word *lion* has at least two potential meanings (a male lion, or a lion of unspecified gender) while *lioness* has, by comparison, just one (a female lion). Nonetheless, a typical dictionary may only mention gender in connection with *lioness* and not provide distinct definitions for the two meanings of the word *lion*.

Linguists have been studying this type of problem in depth for decades (Greenberg, 1966; Pomorska & Rudy, 1987). One of their most useful conclusions is that the syllabic length of words may be employed as a reliable, and easily measurable, statistical indicator for oligosemy. Said otherwise, longer words have proven to be robustly less polysemic than shorter ones, and (as in Rensinghoff & Nemcová, 2010) a Waring distribution seems to fit this dependence best. For numerical details on the correlation, see the statistical studies in the literature, in particular Zipf (1949), Guiter (1974), Sambor (1984), and Rothe (1994).

Hereinafter, by *word-length* I will always mean the number of syllables in a word. In the experiments of Lohnas et al. (2015), whose data I employed above, word lists were assembled from a pool consisting of 1,638 words with up to six syllables. However, only four 5-syllable words were present, and a single 6-syllable word (*encyclopedia*); hence, the statistics for these two lengths may not be representative.

An interesting feature that emerges from these data concerns the sequential peak of the lag probability distribution (the forward contiguous transition frequency). Suppose that the distribution is computed only over transitions to words of syllabic length M, so that it can be written as $p_M(L)$. It appears that the height of the sequential peak,

$p_M(+1)$, exhibits a nontrivial dependence on the length M of the word recalled—that is, the probability of sequential recall varies significantly over words of different lengths.

To estimate the value of the probability $p_M(+1)$, we must extract the relative frequency of sequential transitions from the data. This may be done at least in two separate ways, through a word-by-word statistics or through a subject-by-subject statistics. The results from both approaches are shown in Figure 5.

Panel A of Figure 5 shows results obtained by regarding every transition (from one recall to the next) as an independent event. Let us call $n(S, W, L)$ the number of observed transitions to word W with lag L in trials on subject S. The number of transitions in the dataset having a given word W as their word of arrival is $N(W) = \sum_S \sum_L n(S, W, L)$. Amongst them, $C(W) = \sum_S n(S, W, 1)$ are sequential—that is, have lag $L = +1$. The y-coordinate of each blue dot in Panel A of Figure 5 is the ratio $R(W) = \frac{C(W)}{N(W)}$ computed for a particular word—that is, the frequency with which the word is recalled sequentially.

The histogram of this quantity over all words with the same length has been plotted vertically for each number of syllables (black curves); red circles show the arithmetic means of these values over all words with M syllables: $m_1(M) = \frac{1}{|V(M)|} \sum_{W \in V(M)} R(W)$, where $V(M)$ is the set of all words with M syllables used in the database and $|V(M)|$ their number. The widths of the histograms serve as error bars to these mean values.

The trend of the resulting curve is decreasing. Extracting the correlation coefficient yields $r = -.12$, with a negligible p value $p < 10^{-5}$. This signifies that the longer a word, the smaller its chance of being recalled through a forward contiguous transition.

While this is an intriguing result, it relies on the assumption that all transition events could be treated independently. On the other hand,

transition events within the same trial are statistically correlated, and the same may be true for transition events within different trials performed on the same subject.

In Panel B of Figure 5, a different analysis is displayed. Instead of computing the recall statistics for each individual word, we characterize every transition event solely by the length of the word of arrival. Information on the particular word involved is ignored—that is, assumed to be averaged out.

For each subject S, let $N(S, M)$, be the number of transitions whose word-length of arrival is M (transitions to a word with M syllables); explicitly, we have $N(S, M) = \sum_L \sum_{W \in V(M)} n(S, W, L)$. Call $C(S, M)$ the number of sequential transitions among them—that is, $C(S, M) = \sum_{W \in V(M)} n(S, W, 1)$. The ratio $R(S, M) = \frac{C(S, M)}{N(S, M)}$ has been computed for each individual subject, and its values are shown as the y-coordinates of the blue points in Panel B of Figure 5.

Again, histograms of these quantities are shown in black. The mean values $m_2(M) = \frac{1}{N_S} \sum_S R(S, M)$ (where N_S is the number of subjects) are shown as red circles; the widths of the histograms serve as error bars to the means.

Notice that if the normalization factors depended solely on word length—that is, in the case where $\sum_L n(S, W, L) = n(M)$ for all S and all $W \in V(M)$, we would have $m_1(M) = m_2(M) = \frac{1}{|V(M)| N_S} \sum_{W \in V(M)} \sum_S \frac{n(S, W, 1)}{n(M)}$ for all M. This is the case, in particular, if the samples are identical over all subjects and over all words of the same length, which is of course not true in any realistic dataset. Nonetheless, the mean values we have obtained from the subject-by-subject statistics (see Figure 5, Panel B) appear to be fairly close to those obtained in the word-by-word statistics.
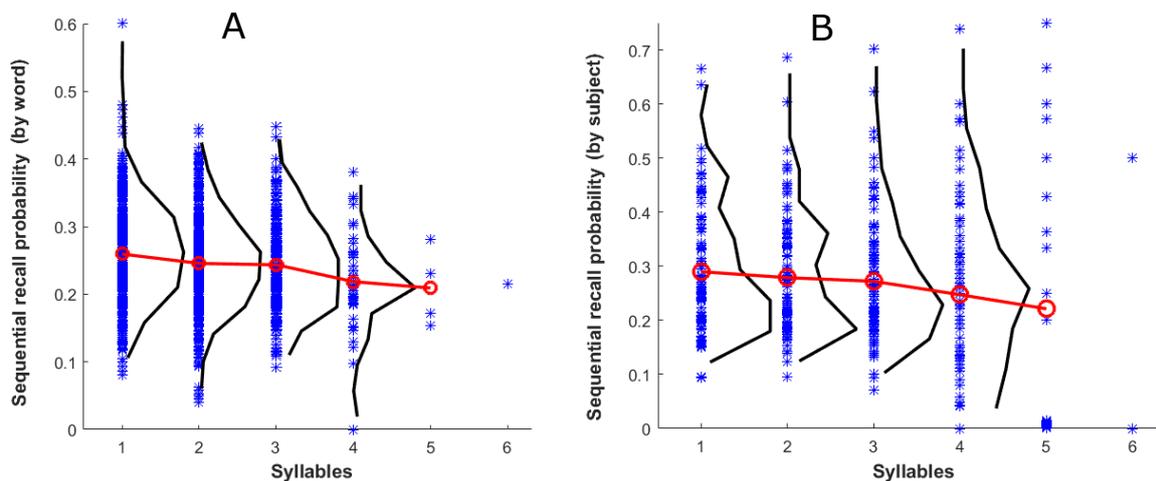

FIGURE 5.

Panel A: probability that a word, if recalled, will be recalled sequentially, computed from Penn Electrophysiology of Encoding and Retrieval Study (PEERS) data by regarding all recall events as independent. Each blue dot corresponds to a different word; for example, the high-lying one-syllable outlier is the word *belt*. The black curves are histograms of these probabilities over all words of a given length, as indicated on the x-axis; the red circles indicate their means, and the widths of the histograms serve as error bars. Panel B: probability that an individual subject will recall a word of a given length sequentially, obtained from PEERS data by regarding all words of the same length as equivalent. Each blue dot corresponds to a different subject; points overlapping at zero have been jittered for display; histograms over all words of the same length are shown as black curves, their means as red circles.

Moreover, we find once again that the mean probabilities for sequential transitions are monotonously decreasing as functions of word length. As for the correlation coefficient, it is also close to the value found above, $r = -.11$. The p value is higher, but still low enough to enable our correlation hypothesis ($p = .01$). All this provides substantial evidence that sequential transitions (with lag $L = +1$) are indeed more favored for shorter words.

We should also report that no significant correlation between transition probabilities and word-length has been found for transitions with lags other than $L = +1$. For example, suppose that the foregoing analysis is repeated for backward contiguous transitions, and that the dependence of $p_M(-1)$ on the word-length M is estimated from the data in an identical way—that is, by simply replacing $n(S, W, 1)$ with $n(S, W, -1)$ in the formulas. A p value of the order of $p \sim .2$ is thus obtained both from the word-by-word and from the subject-by-subject statistics—too high for any correlation to be considered relevant. We must conclude that the effect we are describing arises from mechanisms that concern exclusively sequential transitions.

To ascertain whether the effect is related to length per se or to polysemy, an independent measure of a word's polysemy would be helpful. As we argued above, measuring polysemy is an elusive task and counting the definitions of a word in a standard dictionary does not yield a measurement of its full semantic variability. Nonetheless, it can be interesting to compute correlations between a naïve definition count and the free-recall effect I have just reported.

Figure 6 shows results from the analysis of items from the PEERS wordpool within an up-to-date dictionary of contemporary American English (Dictionary.com, 2017) in which the definitions corresponding to each word are systematically numbered. The counting procedure

needs to follow criteria modeled on the experimental free-recall paradigm. In experiments, words are presented to subjects outside of any syntactic context, therefore, we must count together the definitions of a given word as any part of speech (e.g., both as a noun and as a verb). In PEERS experiments, words were shown visually, hence homographs with different pronunciations must be counted as one word. Moreover, because words were shown in upper-case, we must count homographs as one word also when they differ through capitalization (e.g., *China* and *china*). Finally, abbreviations and definitions corresponding to idiomatic usage have only been included if they were numbered separately within the source dictionary.

In Panel A of Figure 6, the histogram of definition counts is shown for PEERS words of each given length. Since longer words are rarer in the PEERS word-pool, the size of the histogram bins has been adjusted to the varying size of the sample. Medians are shown as vertical red lines. It can be seen that the histogram of definitions moves toward fewer definitions as word-length increases. The correlation coefficient between word-length and the definition count is found to be $r = -.43$, with a p value $p < 10^{-4}$.

Panel B of Figure 6 shows a scatter plot of the sequential recall probability versus the definition count. Each blue dot corresponds to a different word, while the least-square line is shown in red. The correlation coefficient is found to be $r = .16$ ($p < 10^{-4}$), of the same order of magnitude as the correlation coefficient obtained for word lengths, and indeed larger in magnitude.

This supports the notion that polysemy may be playing an important role in the phenomenon we have singled out. As will be shown in the next section, the diffusive-particle model provides a particularly simple explanation for this possibility.

A Histogram of definition counts for M-syllable words
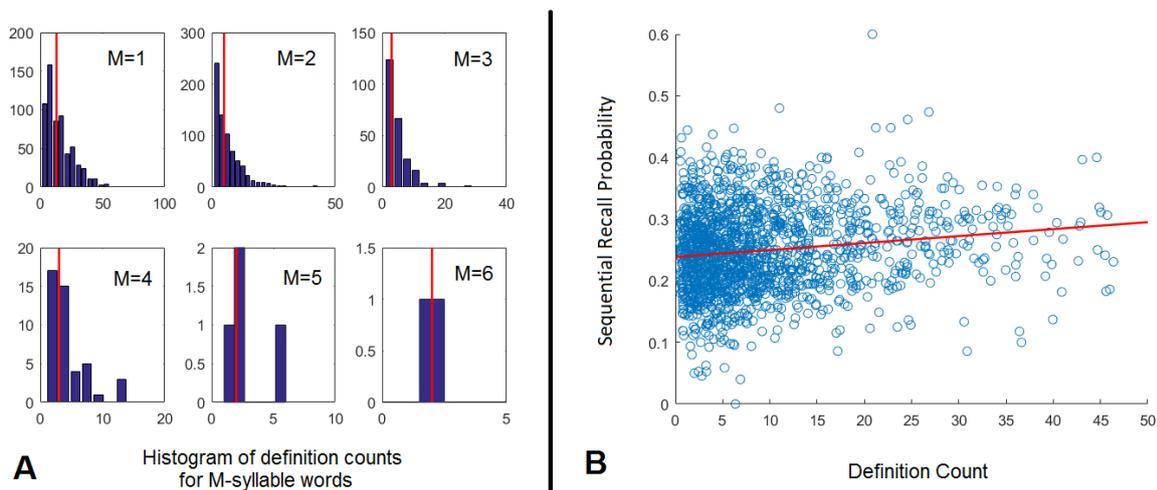
B Definition Count

FIGURE 6.

Panel A: histograms of the definition count in a contemporary dictionary (Dictionary.com, 2017) for words belonging to the Penn Electrophysiology of Encoding and Retrieval Study (PEERS) pool. Details of the counting procedure are provided in the main text. Each histogram refers to words containing the same number of syllables *M*; the size of the histogram bins has been adjusted to the varying size of each sample; medians are shown as vertical red lines. Panel B: scatter plot of the sequential recall probability in PEERS data versus the definition count. Each blue circle refers to a different word; the least-square line is shown in red; the correlation coefficient is $r = .16$ ($p < 10^{-4}$).

INTERPRETIVE CLUSTERING

We must now consider the semantic graph in the case where the polysemy $k(W)$ of word W varies over different words—that is, each word W has a different number $k(W)$ of semantic nuances (which, as we have seen, will be more numerous if the word is shorter).

The quantity we need to calculate is the lag probability distribution $p_k(L)$—that is, the conditional probability that a word with k semantic nuances, if recalled, will be recalled through a transition with lag L. If the effect we observed in the experimental data is indeed due to polysemy, we should expect the sequential transition probability $p_k(1)$ to be enhanced for more polysemic words. Moreover, because of the normalization constraint $p_k(1) + \sum_{L \neq 1} p_k(L) = 1$, this entails that the probability $p_k(L)$ for any $L \neq 1$ should be suppressed, on average, with more polysemic words.

Figure 7 shows the results of simulations on a semantic graph with disconnectedness parameter $\alpha = .9$. The lists presented to the system were permutations of the whole vocabulary. The conditional probability $p_W(L)$ that a word W, if recalled, will be recalled with a lag L, has been averaged over all words with the same degree of polysemy $k(W)$ and the means are displayed as bar plots of different colors.

Panels A, B, and C of Figure 7 refer to results for a vocabulary of $2N$ words, of which N are monosemic (i.e., have one meaning) and the

remaining N words are disemic (i.e., have two meanings). The values of N are respectively 2, 3, and 4, as shown over the plots, and all three yield qualitatively identical plots.

The most conspicuous feature of these plots is the sequential peak exhibited by the disemic word as opposed to the monosemic one. The sequential recall probability $p_k(L = 1)$ is a sharply increasing function of polysemy (hence, a decreasing function of word length, as we found in the data). Yet, the lag probability distribution for each word-type is normalized, so this gap should be made up for by nonsequential transitions. Indeed, we observe that nonsequential transitions are slightly more frequent for the monosemic words than for the disemic ones, the difference at $L = 1$ being redistributed over all nonsequential values of the lag.

We may ask now whether the correlation between sequentiality and polysemy holds also for words with more than two meanings. Simulations show that this is the case: Panel D of Figure 7 displays results of simulations for a vocabulary of five words, one for each degree of polysemy between $k = 1$ and $k = 5$.

The overall picture that emerges is a straightforward extension of what has been found in the case of only two word-types: Again, the sequential probability $p_k(1)$ is a sharply increasing function of a word's degree of polysemy k; again, all other values of $p_k(L)$ are faintly decreasing functions of polysemy.
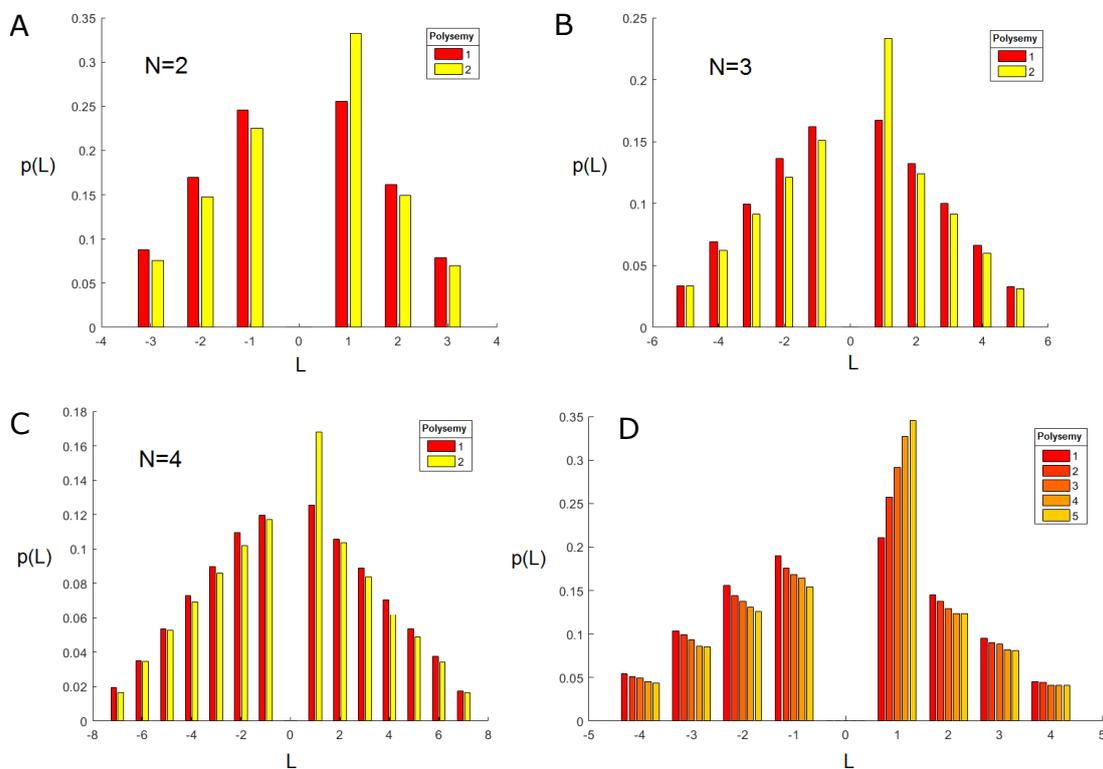

FIGURE 7.

Lag probability distribution $p(L)$ from simulations where the lists presented for recall are permutations of the vocabulary. The semantic graphs have disconnectedness $\alpha = .9$. Different bar colors refer to different degrees of polysemy k, shown in the legends. Panels A, B, and C: results for a vocabulary of $2N$ words of which N are monosemic (i.e., have one meaning) and N are disemic (two meanings); the values of N are shown over the plots. Panel D: results for a five-word vocabulary in which each word has a different degree of polysemy (from $k = 1$ to $k = 5$).

We conclude that the positive correlation between sequential recall and polysemy is a feature robustly displayed by this model. The more meanings a word has, the more easily it is recalled in the order in which it was presented. The remaining question is why this happens—that is, what is the ubiquitous mechanism at the root of this relationship.

To answer this question, we recall that, by introducing a degree of disconnectedness in the semantic graph, we have endowed it with a nontrivial geometry, in which some meanings are closer to each other while others lie further apart. A possible way to measure the distance between any two nodes on a graph is, for instance, by the length of the shortest path connecting them or by the time it takes to diffuse from one to the other.

It is in this spirit that one should regard Figure 8, where the distance between any two nodes represents the distance between them (i.e., length of the shortest path or time for first passage) within a wider semantic graph. Of the graph, only a few nodes are shown – those corresponding to three words (*red*, *green*, and *blue*).

Green and *blue* are monosemic words; *red* is monosemic in the semantic graph of Panels A, B, and C of Figure 8, and polysemic in the semantic graph of Panels D, E, and F (having two meanings). The arrays of colored squares over the drawings in Panels B and C and E and F represent lists of words presented to the system for a free-recall trial.

In Panels B and C of Figure 8, since all words are monosemic, memories of each word can only be created at a fixed node, and a different order of presentation does not generate different memories. Hence, *red* has the same probability of being recalled after *green* or after *blue*.

In Panels E and F of Figure 8, on the contrary, the memory created by presenting the word *red* tends to lie close to the memory created by the word that precedes it. This happens because *red* is polysemic, so the system can choose a meaning for it. If the graph is not too disconnected, the diffusive process that interprets words is continuous (jumps
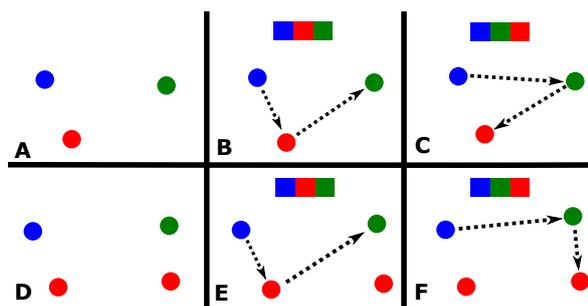
being rare), so a meaning close to the current position of the particle will be more likely to be hit first.

In Panel E of Figure 8, therefore, *red* is more likely to be recalled after *blue* than after *green*, while in Panel F, *red* is more likely to be recalled after *green* than after *blue*. In both cases, *red* is most likely to be recalled right after the word that precedes it in the list. Thus, the polysemy of *red* makes it more likely to be recalled sequentially.

We will refer to this phenomenon as *interpretive clustering*: Among the multiple meanings of an input, the cognitive system selects the one that fits best the content of the ongoing discourse. The more polysemic a word, the more numerous the meanings the system can choose from; hence, the more likely it is to find a meaning close by. This will logically translate, during the test stage, into an enhanced probability for sequential recall.

DISCUSSION: CHRONOLOGICAL STORAGE

It is well-known in the literature (Farrell, 2012) that a word-list presented for a free-recall test is effectively divided by the memory-storage process into *sequential chunks*, sections that tend to be recalled in sequential order. These chunks and their optimal length have been subjected to extensive studies (see, e.g., Cowan, 2001).

Indeed, if the peak at $p(L = +1)$ is large for a series of consecutive words, these are likely to be recalled in the order in which they were presented. With high probability, the peak will guide the recall process through a full sequential chunk, and the last word of the chunk will be the first after which the peak is suppressed; at that point, the recall process becomes more fully associative, that is, free association decides which chunk will be recalled next.

The probability value $p(+1)$ approaches unity only for rare subjects (Healey, Crutchley, & Kahana, 2014); the peak value is, on average, of the order of .3 (see Figure 4). Hence, even where information has been stored the most sequentially, the retrieval process has a finite probability of occurring in nonchronological orders.

The sequential peak, nonetheless, is regularly the global maximum of the probability distribution $p(L)$, and this fact makes it possible to retrieve the chronology of events with arbitrary accuracy, as one can easily argue in terms of diffusion.

If the chronological ordering is the most probable, a diffusive process has indeed a particularly simple way of singling it out with arbitrary accuracy; it is sufficient to re-explore the same contextual area a large number of times and to choose the ordering of memories that has been experienced most often during this re-exploration. The more strictly sequential the memory storage is (i.e., the larger the $p[+1]$), the less time it will take to perform the iterative sampling needed to establish a chronology with arbitrary accuracy.

It may then be conjectured that the value of $p(1)$ is optimized to compromise between two conflicting goals: (a) to allow for a fast-enough iterative sampling—as described—and (b) to keep the memories available nonetheless for use by free association.

FIGURE 8.

Nodes corresponding to three words (*red*, *green*, and *blue*) within a denser semantic graph; distances on the page are meant to represent roughly shortest-path distances within the graph. *Green* and *blue* are monosemic words; *red* is monosemic in the semantic graph of panels A, B, and C, polysemic in the semantic graph of panels D, E, and F, with two meanings. The arrays of colored squares over Panels B and C and E and F represent word-lists presented to the system. Dotted arrows depict diffusive motion through the semantic graph during presentation.

If the sequential peak is too low, the number of iterations needed to find the most probable ordering will become large, and the iterative sampling procedure slow; it may be impractical to devote more than a fraction of a second to ordering any sequence of past events.

If, on the contrary, the sequential peak is too high, associative retrieval of a given memory will be blocked, as follows from the normalization of probabilities; if we can only arrive at a memory from its chronological precedent, it cannot be accessed other than chronologically. It is, consequently, not available for associative tasks and becomes useless for most cognitive purposes.

Thus, sequentiality and retrievability are in conflict and a trade-off between the two requirements may be necessary. A memory must stay available for associative reasoning, and yet its chronology needs to be trackable through iterative sampling. From these two constraints, the optimal value of the $p(+1)$ may be determined.

This optimization process can further depend on the particular memory involved. In other words, what has been referred to as chunking may be a process based partly on a distinction between memories that need chronological storage and memories that do not.

The suggestion of this paper is that polysemy may be one of the criteria for this distinction. As long as words with adaptable meanings are being presented, the system may keep grafting them easily into the ongoing semantic chunk. But when a word with a highly specific meaning appears, there are few chances that the current discourse may accommodate it logically. Hence, a rift in the storage process may have to be introduced—and a new chunk will begin.

This may be conceptually understood as implementing a principle of least effort (Zipf, 1949). Polysemy compels the receiver of any verbal input to choose one of many possible understandings, and that can only be done on the basis of the chronology of events. Chronology is, therefore, a functional part of polysemic communication. This is not the case where the words being used are oligosemic; memorizing a chronology is arguably much less useful when it does not play a role in determining the meaning of the events.

WORD-LENGTH EFFECT

The empirical fact that lists of shorter words are easier to recall (word-length effect) is one of the early findings in the history of free recall (Baddeley et al., 1975). Theories of this effect may be classified as being either item-based or list-based—that is, they impute the effect either to an individual property of words or to a global property of a list.

Recently, item-based theories have been cast doubt upon by novel experiments; in particular, it appears that in experiments with mixed lists (composed of words of various lengths), the shorter words are not always easier to recall (Hulme, Suprenant, Bireta, Stuart, & Neath, 2004; Katkov, Romani, & Tsodyks, 2014; Xu & Li, 2009). This suggests that the word-length effect in pure lists may exist not because shorter words are more distinctive, but in spite of the fact that they are not, strongly pointing toward a list-based explanation for the effect.

In list-based theories, however, the global property on which the effect is made to depend is most frequently the total duration of the list

(Baddeley, 2007). But this explanation has been repeatedly called into question. Neath, Bireta and Suprenant (2003) have shown that with words having the same number of syllables but different pronunciation times, no unambiguous word-length effect arises. This suggests that the effect may depend on the number of syllables and not on the time it takes to pronounce them (Campoy, 2008). A review of the debate can be found in Jalbert, Neath, Bireta, and Surprenant (2011), where it is argued that "the word-length effect may be better explained by the differences in linguistic and lexical properties of short and long words rather than by length per se" (p. 338).

Could this elusive linguistic property be just polysemy? This hypothesis seems to not have been explored yet, and the diffusive-particle model may help to test it. To do so, I have simulated the model by presenting lists that contain words with a fixed degree of a polysemy, while keeping the semantic-graph structure unchanged. The results are shown in Figure 9.

For all choice of the graph-structure parameters, the relationship between recall probability and the degree of polysemy of the word list is monotonously increasing. The more polysemic the words in the list, the easier each will be to recall. Rephrased in terms of word-length, this is nothing but the word-length effect, as exhibited by the diffusive-particle model.

The reason for the word-length effect, within this model, is indeed a global or list-based mechanism: the fact that lists of shorter words, being more polysemic, produce a higher degree of interpretive clustering.

When a word has a higher degree of polysemy, it takes a smaller distance to reach one of its meanings from anywhere within the semantic graph. In other words, a diffusive particle will need to move less
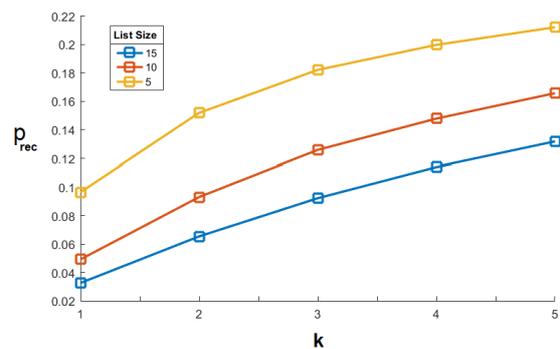
FIGURE 9.

Mean recall probability in the diffusive-particle model. The semantic graph employed for the simulations contains a vocabulary of 10 words, two for each degree of polysemy between $k = 1$ and $k = 5$, while the edges are distributed with a disconnectedness α = .7. The word-length effect was checked by simulating presentation of a large number of *pure* lists–that is, lists consisting entirely of words with the same degree of polysemy k. The recall probability was averaged over all trials with the same value of k and the results plotted as a function of k. The three curves refer to lists of three different sizes, shown in the legend.

far if it has to interpret shorter words. For shorter words, therefore, the semantic region within which memories are formed will be narrower and a smaller region will have to be explored during retrieval; thus, recall will be facilitated.

This is shown in Figure 10, where, again, distances on the page are meant to represent shortest-path distances in a denser semantic graph of which only a few nodes are shown. The nodes being shown refer to both some highly polysemic words (in shades of blue) and some highly oligosemic ones (in shades of red).

Panel A of Figure 10 shows the diffusive trajectory of the particle during the presentation of a list of polysemic words; Panel B - during the presentation of a list of oligosemic words. In the latter case, the desired meanings are less readily available, so longer distances have to be travelled and the memories will afterwards have to be sought over a larger area of the graph.

This is evidently not just an item-based effect. A comparatively long word, by causing a longer shift in the presentation trajectory, distances all the memories that will be created afterwards from the ones created before. Moving from memory to memory during the retrieval stage becomes, in principle, harder over the full scale of the list size.

OTHER FREE-RECALL EFFECTS

While we have shown that the model accounts satisfactorily for several free-recall effects, these are but a fraction of the wealth of phenomena studied over the last decades in the free-recall literature. Let us mention briefly some of them:
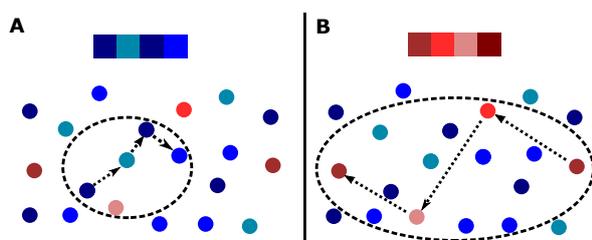

FIGURE 10.

Role of interpretive clustering in the word-length effect. Distances on the page are meant to represent roughly shortest-path distances within a denser semantic graph of which only a few nodes are shown. These nodes refer to three highly polysemic words (shown in shades of blue) and three highly oligosemic ones (in shades of red). Dotted arrows depict diffusive motion through the semantic graph. Panel A depicts the diffusive trajectory during the presentation of a list of polysemic words; Panel B—during the presentation of a list of oligosemic words, to the same system. In both panels, the list being presented is displayed over the drawing as a sequence of colored squares. In the oligosemic case, longer distances have to be travelled; therefore memories are distributed over a wider region (dashed ellipses), impairing recall.

1. *Power-Law scaling*: This was demonstrated to emerge from a limiting case of the present model (for α = 0) in Romani et al. (2013). By continuity, the effect is also bound to emerge for sufficiently small values of α. The exponent found for α = 0 ($\gamma[0] = \frac{1}{2}$) is somewhat larger than the experimentally measured value (Murray et al., 1976; Standing, 1973). The exponent for finite α can differ, of course, from the value computed in Romani et al. (2013) and will deserve further study.

2. *Recency effect*: If the interval between presentation and memory test is short enough, the initial position of the test-stage diffusion will be correlated to the point of arrival of the presentation-stage diffusion. Instead of choosing the initial position of retrieval at random (as done above), it may be realistic to choose it in the neighborhood of the last memory. As a corollary, the last memory will be more likely to be found first, and if the diffusive trajectory during presentation has been sufficiently continuous (jumps being rare), the last few words of the list are bound to be equally favored at the early stages of the recall process.

3. *Lag-recency effects*: The continuity of the diffusion process entails that the positive and negative branches of the lag probability curve $P(L)$ will be, on average, decreasing functions of $|L|$, just as in the empirical data. This would hold true, in principle, even for the case of infinite lists. The simple type of semantic graph ensemble we have considered yields only a qualitative agreement with the empirical curve (see Figure 4). In future work on the model, the observed form of the curve can serve as a key point of comparison for optimizing the semantic-graph distribution $P(G)$ over the data.

CONCLUSIONS

A diffusive approach to the modeling of free recall has been developed, in which the presentation of words and their recall are modeled as trajectories of a particle diffusing over a semantic graph (a graph whose edges are random and whose nodes represent meanings of potentially polysemic words).

The model has predicted correctly some well-known features of free recall (forward asymmetry, semantic clustering, the word-length effect) and has been argued to be a suitable model for others (power-law scaling, recency, and lag-recency effects). A novel prediction has also been obtained: Shorter words, being more polysemic, are characterized by a stronger sequentiality—that is, they are more likely to be recalled through forward contiguity—a prediction confirmed by a fresh analysis of archival data.

The mechanism behind the latter phenomenon (interpretive clustering) is the same that lies at heart of the word-length effect as predicted by this theory. The conversion of words into meaning involves interpretation, and our freedom of interpretation (which is larger for the more polysemic words) has the effect of turning temporal contiguity into semantic contiguity. Since we memorize each word through a meaning largely determined by its context, mixed temporal-semantic correlations are created amongst memories.

Future work on the theory may evolve in three directions: (a) comparing results from this model to additional features of available databases or to features well-documented in the literature (primacy,

intrusions, inter-response times, recall initiation probabilities), (b) trying out more realistic forms for the distribution *P*(*G*) of the probabilistic graph through which the particle moves and optimizing this distribution over the data, which may help in interpreting free-recall data as measurements of semantic connections within specific groups of words, and (c) studying the possible connections between the diffusive-particle model and more widely tested retrieved-context models, in order to ascertain to what extent they differ and in what respects they may correspond.

There are also several experiments that may help test the predictions made so far. In particular, it may be useful to perform ad hoc experiments with select pools of words for which the measurement of polysemy is not overly tricky. This could be done by using two pools, one composed of decidedly oligosemic words (such as *Parthenon*) and one of extremely polysemic words (such as *set*).

Experiments on such mixed lists would serve as a strict test of what we have claimed to be a polysemy effect in the sequential recall probabilities. Another task would be to test whether the word-length effect survives when each list harbors multiple word-lengths but is assembled entirely out of a single pool—either the highly polysemic or the highly oligosemic one. If recall probabilities would not depend on which pool has been used, that would disprove the explanation provided above, ruling out the role of interpretive clustering in the word-length effect.

Finally, the degree of importance of interpretive clustering may be quantified through experiments based on pseudowords. The meanings that a pseudoword evokes can affect its association value, playing a potentially important role in the recall process (Glaze, 1928); yet, the recall of pseudowords may be expected to be more phonetical than the recall of real words. If so, effects due to interpretive clustering will be reduced. Comparing data from experiments with words and from experiments with pseudowords may help ascertain how much semantics really matter in the emergence of the effects we have discussed.

ACKNOWLEDGEMENTS

REFERENCES

Baddeley, A. D. (2007). *Working memory, thought and action*. Oxford, England: Oxford University Press.

Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior, 14*, 575-589. doi: 10.1016/S0022-5371(75)80045-4

Bousfield, W. A., & Sedgewick, C. H. W. (1944). An analysis of sequences of restricted associative responses. *Journal of General Psychology, 30*, 149-165. doi: 10.1080/00221309.1944.10544467

Campoy, G. (2008). The effect of word length in short-term memory: Is rehearsal necessary? *Quarterly Journal of Experimental Psychology, 61*, 724-734. doi: 10.1080/17470210701402364

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral Brain Science, 24*, 87–114. doi: 10.1017/S0140525X01003922

Dictionary.com (2017). *American-language section of http://www.dictionary.com*. Retrieved from http://dictionary.com

Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology*. New York, NY: Teachers College, Columbia University.

Farrell, S. (2012). Temporal clustering and sequencing in short-term memory and episodic memory. *Psychological Review, 119*, 223–271. doi: 10.1037/a0027371

Frieze, A., & Karonski, M. (2016). *Introduction to random graphs*. Cambridge, England: Cambridge University Press.

Fernando, C. (1996). *Idioms and idiomaticity*. Oxford, England: Oxford University Press.

Glaze, J. A. (1928). The association value of non-sense syllables. *Pedagogical Seminary and Journal of Genetic Psychology, 35*, 255-269. doi: 10.1080/08856559.1928.10532156

Greenberg, J. (1966). *Universals of language*. Cambridge, MA: MIT Press.

Guiter, H. (1974). Les relations fréquence-longeur-sens des mots (langues Romaines et Anglais) [The frequence-length-meaning relationship among words (Romance languages and English]. *Atti del Congresso Internazionale di Linguistica* [Proceedings of the International Conference on Linguistics] (pp. 373-381). Amsterdam, the Netherlands: Benjamins.

Healey, M., Crutchley, P., & Kahana, M. J. (2014). Individual differences in memory search and their relation to intelligence. *Journal of Experimental Psychology, 143*, 1553–1569. doi: 10.1037/a0036306

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology, 46*, 269-299. doi: 10.1006/jmps.2001.1388

Hulme, C., Suprenant, A. M., Bireta, T. J., Stuart, G., & Neath, I. (2004). Abolishing the word-length effect. *Journal of Experimental Psychology, 30*, 98-106. doi: 10.1037/0278-7393.30.1.98

Jalbert, A., Neath, I., Bireta, T. J., & Surprenant, A. M. (2011). When does length cause the word length effect? *Journal of Experimental Psychology, 37*, 338-353. doi: 10.1037/a0021804

Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory and Cognition, 24*, 103-109. doi: 10.3758/BF03197276

Kahana, M. J. (2012). *Foundations of human memory*. Oxford, England: Oxford University Press.

Katkov, M., Romani, S., & Tsodyks, M. (2014). Word length effect in free recall of randomly assembled word lists. *Frontiers in Computational Neuroscience, 8*:129. doi: 10.3389/fncom.2014.00129

Lohnas, L. J., Polyn, S. M., & Kahana, M. J. (2015). Expanding the

scope of memory search: Modeling intralist and interlist effects in free recall. *Psychological Review, 122*, 337-363. doi: 10.1037/a0039036

Murdock, B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology, 64*, 482-488. doi: 10.1037/h0045106

Murray, D. J., Pye, C., & Hockley, W. E. (1976). Standing's power function in long-term memory. *Psychological Research, 38*, 319-331. doi: 10.1007/BF00309039

Musz, E., & Thompson-Schill, S. L. (2015). Semantic variability predicts neural variability of object concepts. *Neuropsychologia, 76*, 41-51. doi: 10.1016/j.neuropsychologia.2014.11.029

Neath, I., Bireta, T. J., & Surprenant, A. M. (2003). The time-based word length effect and stimulus set specificity. *Psychonomic Bulletin Review, 10*, 430-4. doi: 10.3758/BF03196502

Nerlich, B., Todd, Z., Herman, V., & Clarke, D. D. (2003). *Polysemy: Flexible patterns of meaning in mind and language*. Berlin, Germany: Mouton de Gruyter.

Pomorska, K., & Rudy, S. (1987). *Language in literature*. Cambridge, MA: Belknap Press.

Rensinghoff, S., & Nemcová, E. (2010). On word length and polysemy in French. *Glottotheory, 1*, 83-88. doi: 10.1515/glot-2010-0005

Romani, S., Pinkoviezky, I., Rubin, A., & Tsodyks, M. (2013). Scaling laws of associative memory retrieval. *Neural Computation, 25*, 2523-2526. doi: 10.1162/NECO_a_00499

Rothe, U. (1994). Wortlänge und Bedeutungsmenge: Eine Untersuchung zum Menzerathschen Gesetz an drei romanischen Sprachen [Word length and semantic density: An investigation of Menzerath's law in three Romance languages]. In R. Köhler & J. Boy (Eds.), *Glottometrika 5* (pp. 101-112). Bochum, Germany: Universitätsverlag Brockmeyer.

Sambor, J. (1984). Menzerath's law and the polysemy of words. In R. Köhler & J. Boy (Eds.), *Glottometrika 6* (pp. 94-114). Bochum, Germany: Universitätsverlag Brockmeyer.

Standing, L. (1973). Learning 10,000 pictures. *Quarterly Journal of Experimental Psychology, 25,* 207-222. doi: 10.1080/14640747308400340

Xu, Z., & Li, B. Q. (2009). The mechanism of reverse word length effect of Chinese in working memory. *Acta Psychologica Sinica, 41*, 802-811.

Zipf, G. K. (1949). *Human behaviour and the principle of least effort*. Cambridge, MA: Addison-Wesley.