Cognitive Science 29 (2005) 41–78

The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth

Mark Steyvers[a], Joshua B. Tenenbaum[b]

[a]*Department of Cognitive Sciences, University of California, Irvine*
[b]*Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology*

Received 15 July 2003; received in revised form 19 June 2004; accepted 22 July 2004

Abstract

We present statistical analyses of the large-scale structure of 3 types of semantic networks: word associations, WordNet, and Roget's Thesaurus. We show that they have a small-world structure, characterized by sparse connectivity, short average path lengths between words, and strong local clustering. In addition, the distributions of the number of connections follow power laws that indicate a scale-free pattern of connectivity, with most nodes having relatively few connections joined together through a small number of hubs with many connections. These regularities have also been found in certain other complex natural networks, such as the World Wide Web, but they are not consistent with many conventional models of semantic organization, based on inheritance hierarchies, arbitrarily structured networks, or high-dimensional vector spaces. We propose that these structures reflect the mechanisms by which semantic networks grow. We describe a simple model for semantic growth, in which each new word or concept is connected to an existing network by differentiating the connectivity pattern of an existing node. This model generates appropriate small-world statistics and power-law connectivity distributions, and it also suggests one possible mechanistic basis for the effects of learning history variables (age of acquisition, usage frequency) on behavioral performance in semantic processing tasks.

Keywords: Semantic networks; Small worlds; Semantic representation; Growing network models

1. Introduction

Network structures provide intuitive and useful representations for modeling semantic knowledge and inference. Within the paradigm of semantic network models, we can ask at least three distinct kinds of questions. The first type of question concerns structure and knowl-

Requests for reprints should be sent to Mark Steyvers, Department of Cognitive Sciences, 3151 Social Sciences Plaza, University of California–Irvine, Irvine, CA 92697–5100; E-mail: msteyver@uci.edu or Joshua B. Tenenbaum, Department of Brain and Cognitive Sciences, 77 Massachusetts Avenue, Cambridge, MA 02139; E-mail: jbt@mit.edu

edge: To what extent can the organization of human semantic knowledge be explained in terms of general structural principles that characterize the connectivity of semantic networks? The second type of question concerns process and performance: To what extent can human performance in semantic processing tasks be explained in terms of general processes operating on semantic networks? A third type of question concerns the interactions of structure and process: To what extent do the processes of semantic retrieval and search exploit the general structural features of semantic networks, and to what extent do those structural features reflect general processes of semantic acquisition or development?

The earliest work on semantic networks attempted to confront these questions in an integrated fashion. Collins and Quillian (1969) suggested that concepts are represented as nodes in a tree-structured hierarchy, with connections determined by class-inclusion relations (Fig. 1). Additional nodes for characteristic attributes or predicates are linked to the most general level of the hierarchy to which they apply. A tree-structured hierarchy provides a particularly economical system for representing default knowledge about categories, but it places strong constraints on the possible extensions of predicates—essentially, on the kinds of knowledge that are possible (Keil, 1979; Sommers, 1971). Collins and Quillian proposed algorithms for efficiently searching these inheritance hierarchies to retrieve or verify facts such as "Robins have wings," and they showed that reaction times of human subjects often seemed to match the qualitative predictions of this model. However, notwithstanding the elegance of this picture, it has severe limitations as a general model of semantic structure. Inheritance hierarchies are clearly appropriate only for certain taxonomically organized concepts, such as classes of animals or other natural kinds. Even in those ideal cases, a strict inheritance structure seems not to apply except for the most typical members of the hierarchy (Carey, 1985; Collins & Quillian, 1969; Rips, Shoben, & Smith, 1973; Sloman, 1998).

Subsequent work on semantic networks put aside the search for general structural principles of knowledge organization and instead focused on elucidating the mechanisms of semantic processing in arbitrarily structured networks. The network models of Collins and Loftus (1975), for instance, are not characterized by any kind of large-scale structure such as a treelike hierarchy. In terms of their large-scale patterns of connectivity, these models are essentially unstructured, with each word or concept corresponding to a node and links between any two

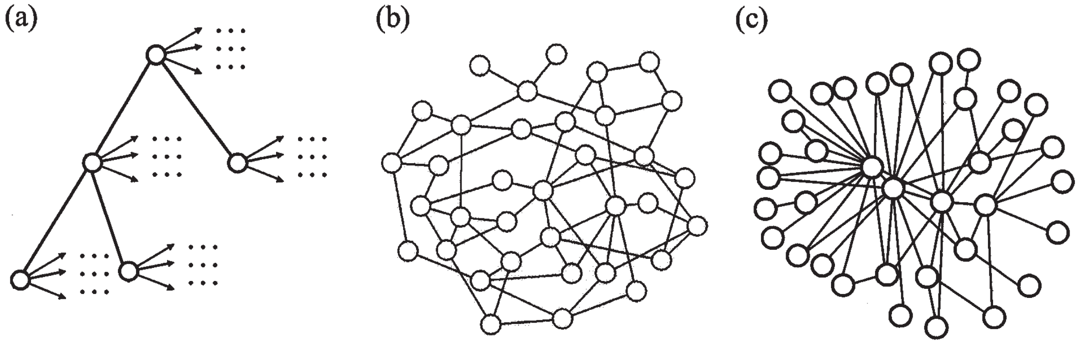

Fig. 1. Proposed large-scale structures for semantic networks: (a), a tree-structured hierarchy (e.g., Collins & Quillian, 1969); (b), an arbitrary, unstructured graph (e.g., Collins & Loftus, 1975); (c), a scale-free, small-world graph.

nodes that are directly associated in some way (Fig. 1B). Quantitative models of generic associative networks, often equipped with some kind of spreading-activation process, have been used to predict performance in a range of experimental memory retrieval tasks and to explain various priming and interference phenomena (Anderson, 2000; Collins & Loftus, 1975; Deese, 1965; Nelson, McKinney, Gee, & Janczura, 1998).

As a result of research in this tradition, there is now a fair consensus about the general character of at least some of the processes involved in the formation and search of semantic memory (Anderson, 2000). By contrast, there is relatively less agreement about general principles governing the large-scale structure of semantic memory, or how that structure interacts with processes of memory search or knowledge acquisition. Typical textbook pictures of semantic memory still depict essentially arbitrary networks, such as Fig. 1B, with no distinctive large-scale structures. The implications for semantic network theories of meaning are not good. Under the semantic net view, meaning is inseparable from structure: The meaning of a concept is, at least in part, constituted by its connections to other concepts. Thus, without any general structural principles, the semantic net paradigm offers little or no general insights into the nature of semantics.

In this article, we argue that there are in fact compelling general principles governing the structure of network representations for natural language semantics and that these structural principles have potentially significant implications for the processes of semantic growth and memory search. We stress from the outset that these principles are not meant to provide a genuine theory of semantics, nor do we believe that networks of word–word relations necessarily reflect all of the most important or deepest aspects of semantic structure. We do expect that semantic networks will play some role in any mature account of word meaning. Our goal here is to study some of the general structural properties of semantic networks that may ultimately form part of the groundwork for any semantic theory.

The principles we propose are not based on any fixed structural motif such as the tree-structured hierarchy of Collins and Quillian (1969). Rather, they are based on statistical regularities that we have uncovered via graph-theoretic analyses of previously described semantic networks. We look at the distributions of several statistics calculated over nodes, pairs of nodes, or triples of nodes in a semantic network: The number of connections per word, the length of the shortest path between two words, and the percentage of a node's neighbors that are themselves neighbors. We show that semantic networks, like many other natural networks (Watts & Strogatz, 1998), possess a *small-world* structure characterized by the combination of highly clustered neighborhoods and a short average path length. Moreover, this small-world structure seems to arise from a *scale-free* organization, also found in many other systems (Barabási & Albert, 1999; Strogatz, 2001), in which a relatively small number of well-connected nodes serve as hubs, and the distribution of node connectivities follows a power function.

These statistical principles of semantic network structure are quite general in scope. They appear to hold for semantic network representations constructed in very different ways, whether from the word associations of naive subjects (Nelson, McEvoy, & Schreiber, 1999) or the considered analyses of linguists (Miller, 1995; Roget, 1911). At the same time, these regularities do not hold for many popular models of semantic structure, including both hierarchical or arbitrarily (unstructured) connected networks (Figures 1A and 1B), as well as high-dimensional vector space models such as Latent Semantic Analysis (LSA; Landauer & Dumais, 1997). These principles may thus suggest directions for new modeling approaches, or for extending or revising existing models. Ultimately, they may help to determine which classes of models most faithfully capture the structure of natural language semantics.

As in studies of scale-free or small-world structures in other physical, biological, or social networks (Albert, Jeong, & Barabási, 2000; Barabási & Albert, 1999; Watts & Strogatz, 1998), we will emphasize the implications of these distinctive structures for some of the crucial processes that operate on semantic networks. We suggest that these structures may be consequences of the developmental mechanisms by which connections between words or concepts are formed—either in language evolution, language acquisition, or both. In particular, we show how simple models of network growth can produce close quantitative fits to the statistics of semantic networks derived from word association or linguistic databases, based only on plausible abstract principles with no free numerical parameters.

In our model, a network acquires new concepts over time and connects each new concept to a subset of the concepts within an existing neighborhood, with the probability of choosing a particular neighborhood proportional to its size. This growth process can be viewed as a kind of semantic differentiation, in which new concepts correspond to more specific variations on existing concepts, and highly complex concepts (those with many connections) are more likely to be differentiated than simpler ones. It naturally yields scale-free small-world networks, such as the one shown in Fig. 1C (see also Fig. 6).

Our models also make predictions about the time course of semantic acquisition, because the order in which meanings are acquired is crucial in determining their connectivity. Concepts that enter the network early are expected to show higher connectivity. We verify this relation experimentally with age-of-acquisition norms (Gilhooly & Logie, 1980; Morrison, Chappell, & Ellis, 1997) and explain how it could account for some puzzling behavioral effects of age of acquisition in lexical-decision and naming tasks, under plausible assumptions about search mechanisms in semantic memory.

Our growing network models are not intended as anything like a complete model of semantic development, or even a precise mechanistic model of some aspects of semantic development. Rather, they capture just one aspect of semantic development at a high level of abstraction: the origin of new meaning representations as differentiations of existing meaning representations. There are clearly many aspects of semantic development that these models do not address, such as the different routes to word learning, the relative roles of verbal and nonverbal experience, and the possibility that semantic relations may disappear as well as appear. Our model shows that one aspect of semantic development—growth of semantic networks by differentiations of existing nodes—is sufficient to explain the characteristic large-scale statistical structures we observed across different semantic networks. There are also many aspects of semantic structure not captured in our simplified semantic network models: the context sensitivity of meanings, the existence of different kinds of semantic relations, or the precise nature of the relations between word meanings and concepts. We leave these topics as questions for future research.

2. Basic concepts from graph theory

We begin by defining some terminology from graph theory and briefly introducing the statistical properties that we will use to describe the structure of semantic networks.[1] Underlying every semantic network is a *graph,* consisting of a set of *nodes* (also called *vertices*) and a set of *edges* or *arcs* that join pairs of nodes. The number of nodes in the network is denoted by *n.* An edge is an undirected link between two nodes, and a graph containing only edges is said to be *undirected.* An arc is a directed link between two nodes, and a graph containing only arcs is said to be *directed.* Every directed graph corresponds naturally to an undirected graph over the same nodes, obtained by replacing each arc with an edge between the same pair of nodes (see Table 1).

Two nodes that are connected by either an arc or edge are said to be *neighbors;* a *neighborhood* is a subset of nodes consisting of some node and all of its neighbors. When the network is directed, the *in-degree* and *out-degree* of a node refer to the numbers of arcs incoming to or outgoing from that node, respectively. The variables k_i^{in} and k_i^{out} denote the in- and out-degree of node *i,* respectively. When the network is undirected, the in-degree and out-degree are always equal, and we refer to either quantity as the *degree* of a node. We write the degree of node *i* as k_i. We will also use k_i with directed networks to denote the degree of node *i* in the corresponding undirected network (i.e., the total numbers of neighbors of node *i*).

In an undirected graph, a *path* is a sequence of edges that connects one node to another. In a directed graph, a (directed) path is a set of arcs that can be followed along the direction of the arcs from one node to another. We can also talk about undirected paths in a directed graph, referring to the paths along edges in the corresponding undirected graph, but by default any reference to paths in a directed network will assume the paths are directed. For a particular path from node *x* to node *y,* the *path length* is the number of edges (in an undirected graph) or arcs (in a directed graph) along that path. We refer to the *distance* between *x* and *y* as the length of the shortest path connecting them.[2] In a *connected* graph, there exists an undirected path between any pair of nodes. A directed graph is said to be *strongly connected* if between any pair of nodes there exists a path along the arcs. A (strongly) *connected component* is a subset of nodes that is (strongly) connected.

We characterize the structure of semantic networks primarily in terms of four statistical features defined using the previous terminology. These quantities are the average distance *L,* the diameter *D,* the clustering coefficient *C,* and the degree distribution *P*(*k*). *L* and *D* are closely related: *L* refers

Table 1
Definitions for terms and variables used throughout the paper

Term/variable	Definitions
n	number of nodes
L	the average length of shortest path between pairs of nodes
D	the diameter of the network
C	the clustering coefficient (see Equation 1 and accompanying text)
k, k^{in}, k^{out}	the degree, the in-degree, and out-degree (degree = number of connections)
P(k)	degree distribution
$<k>$	average degree
γ	power-law exponent for the degree distribution (see Equation 2)
random graph	network where each pair of nodes is joined by an edge with probability p
small-world structure	network with short average path lengths L and relatively high clustering coefficient C (by comparison with equally dense random graphs)
scale-free network	network with a degree distribution that is power-law distributed

to the average of the shortest path lengths between all pairs of nodes in a network, whereas D refers to the maximum of these distances over all pairs of nodes. In other words, at most D steps are needed to move from any node to any other, but on average only L steps are required.[3]

The clustering coefficient C, and the degree distribution $P(k)$, are two different probabilistic measures of the graph connectivity structure. C represents the probability that two neighbors of a randomly chosen node will themselves be neighbors, or alternatively, the extent to which the neighborhoods of neighboring nodes overlap. Following Watts and Strogatz (1998), we calculate C by taking the average over all nodes i of the quantity

$$C_i = T_i \Big/ \binom{k_i}{2} = 2T_i \big/ k_i(k_i-1) \tag{1}$$

where T_i denotes the number of connections between the neighbors of node i, and $k_i(k_i-1)/2$ is the number of connections that would be expected between i's neighbors if they formed a fully connected subgraph. Because T_i can never exceed $k_i(k_i-1)/2$, the clustering coefficient C is normalized to lie between 0 to 1, as required of a probability. When $C = 0$, no nodes have neighbors that are also each others' neighbors. In a fully connected network (i.e., every node is connected to all other nodes), $C = 1$. Although the clustering coefficient is sensitive to the number of connec-

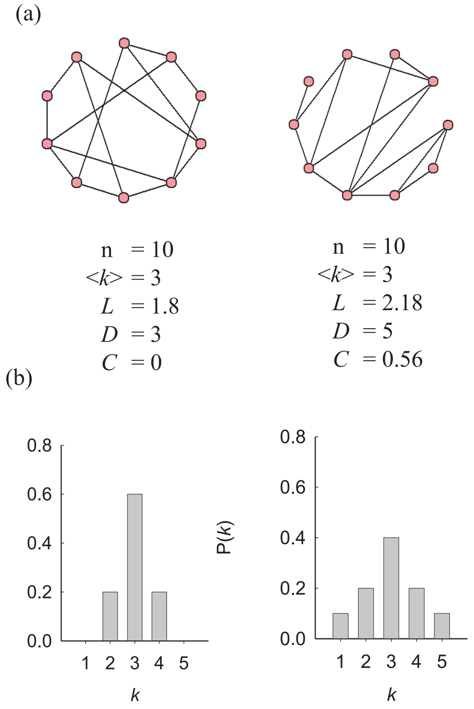

Fig. 2. An illustration of the graph-theoretic properties that we will apply to semantic networks. (a) Two networks with equal numbers of nodes and edges. For both networks, the variables n (number of nodes) and $<k>$ (average degree, i.e., average number of edges) are shown as well as the statistical properties L (average shortest path length), D (diameter) and C (clustering coefficient). Note that the two networks in have different clustering coefficients even though they have the same n and $<k>$. (b) The degree distributions corresponding to the two networks above. Both networks show the typical pattern for random graphs: approximately bell-shaped distributions with tails that decay exponentially as k increases.

tions in a network, it is possible for two networks to have the same number of connections but different clustering coefficients (see Fig. 2). Finally, note that because the definitions for T_i and k_i are independent of whether the connections are based on edges or arcs, the clustering coefficients for a directed network and the corresponding undirected network are equal.

The degree distribution $P(k)$ represents the probability that a randomly chosen node will have degree k (i.e., will have k neighbors). For directed networks, we will concentrate on the distribution of in-degrees, although one can also look at the out-degree distribution. We estimate these distributions based on the relative frequencies of node degrees found throughout the network. The most straightforward feature of $P(k)$ is the expected value $<k>$ under $P(k)$. This quantity, estimated by simply averaging the degree of all nodes over the network, ranges between 0 and n (for a fully connected network of n nodes) and represents the mean density of connections in the network. More information can be gleaned by plotting the full distribution $P(k)$ as a function of k, using either a bar histogram (for small networks, as in Fig. 2), a binned scatter plot (for large networks, as in Fig. 5), or a smooth curve (for theoretical models, as later illustrated in Fig. 3). As we explain in the next section, the shapes of these plots provide characteristic signatures for different kinds of network structure and different processes of network growth.

Fig. 2 shows these statistical measures for two different networks with 10 nodes and 15 edges. These examples illustrate how networks equal in size and density of connections may differ significantly in their other structural features. Fig. 2 also illustrates two general properties of random graphs—graphs that are generated by placing an edge between any two nodes with some constant probability p independent of the existence of any other edge (a model introduced by Erdös and Réyni, 1960). First, for fixed n and $<k>$, high values of C tend to imply high values of L and D. Second, the degree distribution $P(k)$ is approximately bell shaped, with an exponential tail for high values of k.[4] Although these two properties hold reliably for random graphs, they do not hold for many important natural networks, including semantic networks in natural language. We next turn to a detailed discussion of the small-world and scale-free structures that do characterize natural semantic networks. Both of these structures can be thought of in terms of how they contrast with random graphs: Small-world structures are essentially defined by the combination of high values of C together with low values of L and D, whereas scale-free structures are characterized by non-bell-shaped degree distributions, with power-law (rather than exponential) tails.

3. Small-world and scale-free network structures

Interest in the small-world phenomenon originated with the classic experiments of Milgram (1967) on social networks. Milgram's results suggested that any two people in the United States were, on average, separated by only a small number of acquaintances or friends (popularly known as "six degrees of separation"). Although the finding of very short distances between random pairs of nodes in a large, sparsely connected network may seem surprising, it does not necessarily indicate any interesting structure. This phenomenon occurs even in the random graphs described previously, where each pair of nodes is joined by an edge with probability p. When p is sufficiently high, the whole network becomes connected, and the average distance L grows logarithmically with n, the size of the network (Erdös & Réyni, 1960).

Watts and Strogatz (1998) sparked renewed interest in the mathematical basis of the small-world phenomenon with their study of several naturally occurring networks: the power grid of the western United States, the collaboration network of international film actors, and the neural network of the worm *Caenorhabditis elegans.* They showed that although random graphs with comparable size *n* and mean connectivity <*k*> describe very well the short path lengths found in these networks, they also exhibit clustering coefficients *C,* that are orders of magnitude smaller than those observed in the real networks. In other words, natural small-world networks somehow produce much shorter internode distances than would be expected in equally dense random graphs, given how likely it is that the neighbors of a node are also one another's neighbors (see Note 4). For the remainder of this article, we use the term *small-world structure* to refer to this combination of short average path lengths *L,* and relatively high clustering coefficients *C* (by comparison with equally dense random graphs).

Small-world structures have since been found in many other networks (reviewed in Strogatz, 2001), including the World Wide Web (WWW; Adamic, 1999; Albert, Jeong, & Barabási, 1999), networks of scientific collaborators (Newman, 2001), and metabolic networks in biology (Jeong, Tombor, Albert, Oltval, & Barabási, 2000). Watts and Strogatz (1998) proposed a simple abstract model for the formation of small-world structures, in which a small number of the connections in a low-dimensional regular lattice are replaced with connections between random pairs of nodes. The local neighborhood structure of the lattice leads to high clustering, whereas the long-range random connections lead to very short average path lengths.

Amaral, Scala, Barthélémy, and Stanley (2000) distinguished between different classes of small-world networks by measuring the degree distribution *P*(*k*). In one class of networks, such as *C. elegans* and the U. S. power grid, the degree distribution decays exponentially. This behavior is well described by random graph models or variants of the Watts and Strogatz (1998) model. In other systems, such as the WWW or metabolic networks, the degree distribution follows a power law (Barabási & Albert, 1999),

$$P(k) \approx k^{-\gamma} \tag{2}$$

for values of γ typically between 2 and 4. Fig. 3 illustrates the difference between power-law and exponential degree distributions. Intuitively, a power-law distribution implies that a small but significant number of nodes are connected to a very large number of other nodes, whereas

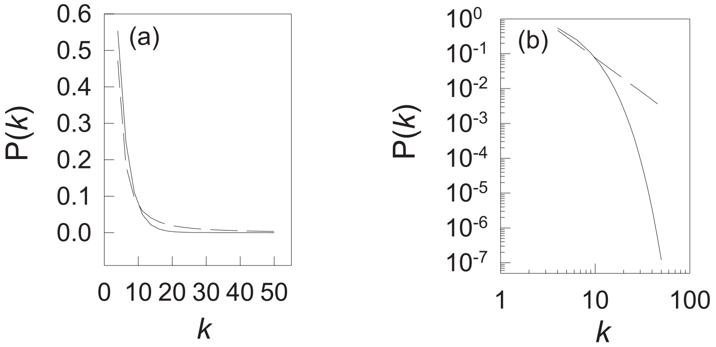

Fig. 3. Two different kinds of degree distributions that may be observed in complex networks. (a) A power-law distribution (dashed) and an exponential distribution (solid), shown with the same linear scales as the histograms in Figure 2b. (b) When plotted with log-log scaling, the same distributions are more clearly differentiated, particularly in their tails.

in an exponential distribution, such "hubs" are essentially nonexistent. Because networks with power-law degree distributions have no characteristic scale of node degree, but instead exhibit all scales of connectivity simultaneously, they are often referred to as *scale-free structures* (Barabási & Albert, 1999). Power-law and exponential distributions can be differentiated most easily by plotting them in log–log coordinates (as shown in Fig. 3B). Only a power-law distribution follows a straight line in log–log coordinates, with slope given by the parameter γ.

Barabási and Albert (1999) argued that the finding of power-law degree distributions places strong constraints on the process that generates a network's connectivity. They proposed an abstract model for scale-free network formation based on two principles that we explain in greater detail as follows: incremental growth and preferential attachment. This model yields power-law degree distributions, but it does not produce the strong neighborhood clustering characteristic of many small-world networks and the model of Watts and Strogatz (1998). In short, although the model of Watts and Strogatz naturally produces small-world structures, and the model of Barabási and Albert naturally produces scale-free structures, neither of these approaches explains the emergence of both scale-free and small-world structures as has been observed in some important complex networks such as the WWW. There is currently a great deal of interest within the theoretical physics community (see Albert & Barabási, 2002, for an overview) in developing models of network formation that can capture both of these kinds of structures.

In the next section, we show that semantic networks, such as the WWW, exhibit both small-world and scale-free structures. In the following section we introduce a model for network growth that is related to the Barabási and Albert (1999) model but which grows through a process of differentiation analogous to mechanisms of semantic development. This growth process allows our model to produce both small-world and scale-free structures naturally, with essentially no free parameters. The final section explores some of the psychological implications of this model and compares it to other frameworks for modeling semantic structure.

4. Graph-theoretic analyses of semantic networks

We constructed networks based on three sources of semantic knowledge: free association norms (Nelson et al., 1999), WordNet (Fellbaum, 1998; Miller, 1995), and Roget's thesaurus (Roget, 1911). Although the processes that generated these data surely differ in important ways, we see that the resulting semantic networks are similar in the statistics of their large-scale organization. To allow the application of conventional graph-theoretic analyses, we construct these networks with all arcs and edges unlabeled and weighted equally. More subtle analyses that recognize qualitative or quantitative differences between connections would be an important subject of future work.

4.1. Methods

4.1.1. Associative network

A large free-association database involving more than 6,000 participants was collected by Nelson et al. (1999). Over 5,000 words served as cues (e.g., "cat") for which participants had to write down the first word that came to mind (e.g., "dog"). We created two networks based on these norms. In the directed network, two word nodes x and y were joined by an arc (from x to *y*) if the cue *x* evoked *y* as an associative response for at least two of the participants in the database. In the undirected network, word nodes were joined by an edge if the words were associatively related regardless of associative direction. Although the directed network is clearly a more natural representation of word associations, our other networks were both undirected, and most of the literature on small-world and scale-free networks has focused on undirected networks. Hence the undirected network of word associations provides an important benchmark for comparison. Fig. 4 shows all shortest associative paths from VOLCANO to ACHE in the directed associative network.

4.1.2. Roget's thesaurus (Roget, 1911)

Based on the lifelong work of Dr. Peter Mark Roget (1779–1869), the 1911 edition includes over 29,000 words classified into 1,000 semantic categories (ignoring several levels of subcategories). Roget's thesaurus can be viewed as a *bipartite graph,* a graph in which there are two different kinds of nodes, word nodes and semantic category nodes, with connections allowed only between two nodes of different kinds. In this graph, a connection is made between a word and category node when the word falls into the semantic category. For Roget's thesaurus, the analysis of neighborhood clustering is more complex. Because the thesaurus is a bipartite graph, with connections between word and class nodes, but not between nodes of the same type, the neighbors of a word node can never themselves be neighbors. To define a meaningful measure of semantic neighborhood clustering in the thesaurus network, we converted the bipartite graph to a simple graph on the word nodes by connecting words if they shared at least one class in common.

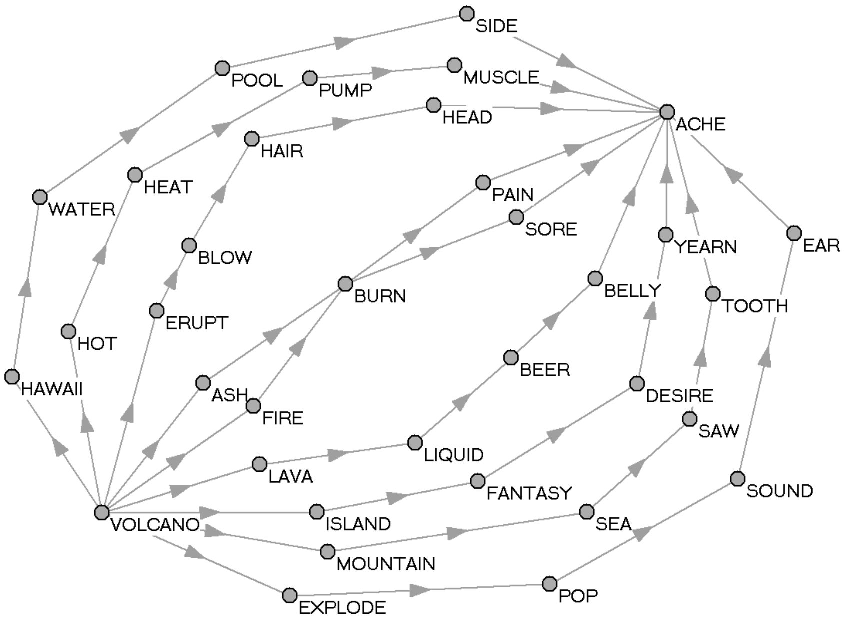

Fig. 4. Part of the semantic network formed by free association. Each directed edge illustrates an association between a cue and a response. A sample of associative directed paths from VOLCANO to ACHE is shown (the shortest path length is four).

4.1.3. WordNet

Somewhat analogous to Roget's thesaurus (Roget, 1911), but inspired by modern psycholinguistic theory, WordNet was developed by George Miller and colleagues (Fellbaum, 1998; Miller, 1995). The network contains 120,000+ word forms (single words and collocations) and 99,000+ word meanings. The basic links in the network are between word forms and word meanings. Multiple word forms are connected to a single word-meaning node if the word forms are synonymous. A word form is connected to multiple word-meaning nodes if it is polysemous. Word forms can be connected to each other through a variety of relations such as antonymy (e.g., BLACK and WHITE). Word-meaning nodes are connected by relations such as hypernymy (MAPLE is a TREE) and meronymy (BIRD has a BEAK). Although these relations, such as hypernymy and meronymy, are directed, they can be directed both ways depending on what relation is stressed. For example, the connection between BIRD and BEAK can be from bird to beak because birds have beaks but also from beak to bird because a beak is part of a bird. Because there are no inherently preferred directions for these relations, we treat WordNet as an undirected graph.

4.2. Results and analyses

Our analysis of these semantic networks focuses on five properties: sparsity, connectedness, short path lengths, high neighborhood clustering, and power-law degree distributions. The statistics related to these properties are shown in Table 2 (under the Data columns), and the estimated degree distributions for each network are plotted in Fig. 5. To provide a benchmark for small-world analyses, we also computed the average shortest-path lengths (L_{random}) and clustering coefficients (C_{random}) for ensembles of random networks with sizes and connection densities equal to those observed in the three semantic networks. These random graphs were created by randomly rearranging connections in the corresponding semantic networks.[5]

Table 2
Summary statistics for semantic networks

		Associative Network			
Variable	Type	Undirected	Directed	Roget	WordNet
n	words	5,018	5,018	29,381	122,005
	classes	—	—	1,000	99,642
$<k>$	words	22.0	12.7	1.7	1.6
	classes	—	—	49.6	4.0
L		3.04	4.27	5.60	10.56
D		5	10	10	27
C		.186	.186	.875	.0265
γ		3.01	1.79	3.19	3.11
L_{random}		3.03	4.26	5.43	10.61
C_{random}		4.35E-03	4.35E-03	.613	1.29E-04

Note. n = the number of nodes; $<k>$ = the average number of connections; L = the average shortest path length; D = the diameter of the network; C = clustering coefficient; γ = power law exponent for the distribution of the nunmber of edges in undirected netowrks and incoming connections in directed networks; L_{random} = the average shortest path length with random graph of same size and density; C_{random} = the clustering coefficient for a random graph of same size and density.

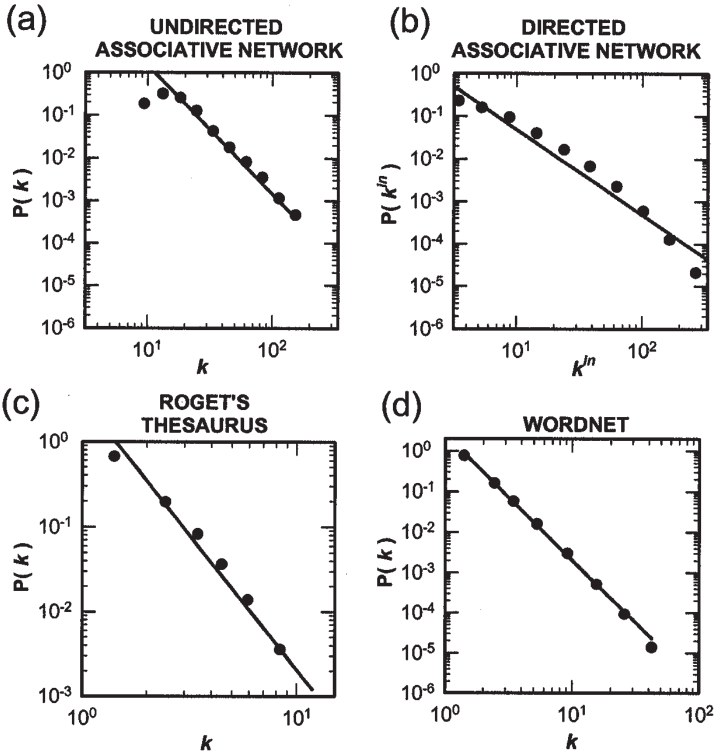

Fig. 5. The degree distributions in the undirected associative network (a), the directed associative network (b), Roget's thesaurus (c), and WordNet (d). All distributions are shown in log-log coordinates with the line showing the best fitting power law distribution. For Roget's thesaurus and WordNet, the degree distributions shown are for the word nodes only.

4.2.1. Sparsity

For WordNet (Fellbaum, 1998; Miller, 1995) and Roget (1911), the number of nodes can be separated into the number of word nodes and the number of class nodes (categories in Roget and word meanings in WordNet). For WordNet and Roget's thesaurus, Table 2 lists <*k*> (the average degree or average number of connections) separately for word and class nodes. Given the size of the networks and the number of connections, it can be observed that all three semantic networks are sparse: On average, a node is connected to only a very small percentage of other nodes. In the undirected associative network, a word is connected on average to only 22 (.44%) of the 5,018 total number of words. The semantic networks of WordNet and Roget's thesaurus exhibit even sparser connectivity patterns.

4.2.2. Connectedness

Despite their sparsity, each of these semantic networks contains a single large connected component that includes the vast majority of nodes. In the directed associative network, the largest strongly connected component consists of 96% of all words (i.e., for this set of words, there is an associative path from any word to any other word when the direction of association is taken into account). In the undirected associative network, the whole network is connected.

For both WordNet (Fellbaum, 1998; Miller, 1995) and Roget's (Roget, 1911) thesaurus, the largest connected component consists of more than 99% of all words. We restricted all further analyses to these components.

4.2.3. Short Path Lengths

All three networks display very short average path lengths and diameters relative to the sizes of the networks[6]. For instance, in the undirected associative network, the average path length (L) is only 3, whereas the maximum path length (D) is only 5. That is, at most, 5 associative steps (independent of direction) separate any two words in the 5,000+ word lexicon. These short path lengths and small diameters are well described by random graphs of equivalent size and density, consistent with Watts and Strogatz's (1998) findings for other small-world networks.

4.2.4. Neighborhood clustering

The clustering coefficient C for the associative network is well above zero, implying that the associates of a word tend also to be directly associated with a significant fraction (approximately 1/6) of time. The absolute value of C appears much lower for WordNet (Fellbaum, 1998; Miller, 1995), but that is primarily because the graph is much sparser (there are fewer possibilities to form overlapping neighborhoods of nodes). For both the associative network and WordNet, C is several orders of magnitude larger than would be expected in a random graph of equivalent size and density (C_{random}). For Roget's (Roget, 1911) thesaurus, the analysis of neighborhood clustering is more complex. Because the thesaurus is a bipartite graph, with connections between word and class nodes, but not between nodes of the same type, the neighbors of a word node can never themselves be neighbors. To define a meaningful measure of semantic neighborhood clustering in the thesaurus network, we converted the bipartite graph to a simple graph on the word nodes by connecting words if they shared at least one class in common. The clustering coefficient C was then computed on this word graph and compared with the mean clustering coefficient C_{random} computed in an analogous manner for an ensemble of random bipartite graphs with the same size and density as the original thesaurus network. As in the other two semantic networks, C was substantially higher for the thesaurus than for the comparable random graphs.

4.2.5. Power-law degree distribution

Fig. 5 plots the degree distributions for the word nodes of each network in log–log coordinates, together with the best fitting power functions (which appear as straight lines under the log–log scaling). For the directed associative network, the in-degree distribution is shown. As in conventional histograms, these distributions were estimated by grouping all values of k into bins of consecutive values and computing the mean value of k for each bin. The mean value of each bin corresponds to one point in Fig. 5. The boundaries between bins were spaced logarithmically to ensure approximately equal numbers of observations per bin.

For the three undirected networks, power functions fit the degree distributions almost perfectly. The exponents γ of the best fitting power laws (corresponding to the slopes of the lines in Fig. 5) were quite similar in all three cases, varying between 3.01 and 3.19 (see Table 2). The high-connectivity words at the tail of the power-law distribution can be thought of as the "hubs" of the semantic network. For example, in word association, the five words with the highest degrees are FOOD, MONEY, WATER, CAR, and GOOD. In WordNet (Fellbaum, 1998; Miller, 1995), they are BREAK, CUT, RUN, MAKE, CLEAR, and in Roget's (Roget, 1911) thesaurus they are LIGHT, CUT, HOLD, SET, and TURN. This analysis also reveals that the sets of high-degree words have no words in common among the three semantic networks. Similarly, when the analysis is expanded to 50 words with the highest degree, there are just a few overlapping words (e.g., Roget's thesaurus and WordNet overlap in BREAK, CHARGE, CLEAR, CLOSE, and CUT, but word association and WordNet overlap in CLEAN and GOOD). This shows that, whereas the hubs in these networks correspond to important semantic categories or highly polysemous verbs and nouns or both, they do differ in the exact set of words that make up these hubs.

For the directed associative network, the in-degree distribution shows a slight deviation from the power-law form, and the best fitting power law has an exponent γ somewhat lower than 2. The out-degree of words in the directed associative network (not shown in Fig. 5) is not power-law distributed but, instead, has a single peak near its mean and exponential tails, similar to a normal or ξ^2 distribution. We focus on the in-degree distribution as opposed to the out-degree distribution, because the out-degree of a node in the word associative network is strongly dependent on specific details of how the word association experiment was conducted: the number of subjects that gave associative responses to that cue and the number of different associates that each participant was asked to generate for that cue. We will discuss the differences between the in- and out-degree distributions in word association in more detail when we describe the growing network model that can explain these differences.

4.3. Discussion

All of the semantic networks studied shared the distinctive statistical features of both small-world and scale-free structures: a high degree of sparsity, a single connected component containing the vast majority of nodes, very short average distances among nodes, high local clustering, and a power-law degree distribution (with exponent near 3 for the undirected networks). The fact that these properties held for all networks despite their origins in very different kinds of data demands some explanation. It is unlikely that the commonalities are simply artifacts of our analysis, because they are not found in random graphs or even in many complex networks from other scientific domains that have been subjected to the same kinds of analyses (Amaral et al., 2000; Watts & Strogatz, 1998). It is more reasonable to suppose that they reflect, at least in part, some abstract features of semantic organization. This structure must be sufficiently deep and pervasive to be accessed by processes as diverse as rapid free association by naive subjects (Nelson et al., 1999) and considered analysis by linguistic experts (Miller, 1995; Roget, 1911).

The similarities in network structure may also depend in part on the coarse grain of our analysis, which treats all words and all connections among words equally. Surely these simplifications make our analysis insensitive to a great deal of interesting linguistic structure, but they may also enable us to see the forest for the trees—to pick up on general properties of meaning in language that might be missed in more fine-grained but small-scale studies of particular semantic domains. A promising direction for future work would be to refine our analyses based on some minimal linguistic constraints. For instance, words or connections could first be segregated into broad syntactic or semantic classes, and then the same statistical analyses could be applied to each class separately. Many authors have suggested that there are systematic differences in the semantics of nouns and verbs (Gentner, 1981) or nouns and adjectives (Gasser & Smith, 1998) or different kinds verbs (Levin, 1993), and it would be interesting to see if those differences correspond to different large-scale statistical patterns. It would also be interesting to apply the same analyses to the semantic networks of different languages. We expect that the general principles of small-world and scale-free structures would be universal, but perhaps we would find quantitative variations in the clustering coefficients or power-law exponents resulting from different language histories.

Power laws in human language were made famous by Zipf (1965), but were in fact discussed earlier by Skinner (1937). After conducting the previous analyses, we discovered some intriguing parallels with those classic studies. Zipf's best known finding concerns the distribution of word frequencies, but he also found a power-law distribution for the number of word meanings (as given by the Thorndike-Century dictionary). That is, most words have relatively few distinct meanings, but a small number of words have many meanings. If we assume that a word's degree of connectivity is proportional to the number of its distinct meanings, then Zipf's "law of meaning" is highly consistent with our results here, including a power-law exponent of approximately 3 that best characterizes his distribution.[7] In our analysis of the in-degree distribution for the directed associative network, the best fitting power-law exponent was somewhat lower than 2. This result was anticipated by Skinner (1937), who measured the distribution of the number of different associative responses to a much smaller set of cues than did Nelson et al. (1999). His plots show power-law distributions with a slope somewhat lower than 2, which is quite consistent with our findings for the in-degree distribution of the word association network.

Given the limited significance attributed to the work of Zipf (1965) and Skinner (1937) in most contemporary accounts of linguistic structure and processing, it is quite reasonable to ask whether the statistical regularities we have uncovered will prove to be any more important. Skinner's work on the associative basis of language has been discounted primarily on the grounds that it looked only at the surface forms of language, ignoring the unobservable representational structures that cause those observable forms to behave as they do (Chomsky, 1957, 1959). Our analysis, in contrast, examines both simple associations between surface forms (Nelson et al., 1999) and more complex relations between words mediated by unobservable classes (WordNet; Fellbaum, 1998; Miller, 1995) and finds similar patterns in them. The unified theory that Zipf proposed to account for his power-law findings, based on principles of least effort, has fared somewhat better than Skinner's theories of language. Yet it does not play a large role in contemporary computational studies of language, perhaps because its formulation is vague in many places and because many simple mathematical models have been subsequently proposed that can reproduce Zipf's most famous result—the power-law distribution of word frequencies—without offering any deep insights into how language works (Manning & Schutze, 1999). In contrast, the statistical properties we have identified are not predicted by the currently most popular accounts of semantic structure, or by many other mathematical models of network formation. We have also attempted to develop mathematically precise and psychologically motivated models for the formation of semantic networks that are consistent with all of these constraints. These models are the subject of the following section.

5. Growing network model

It has been argued by a number of researchers (Barabási & Albert, 1999; Kleinberg, Kumar, Raghavan, Rajagopalan, & Tomkins, 1999; Simon, 1955) that power-law distributions are a consequence of the characteristic ways that systems grow or develop over time. In particular, power laws in network degree distributions have been taken as the signature of a particular kind of network growth process known as *preferential attachment* (Barabási & Albert, 1999; see also, Simon, 1955): Nodes are added to the network successively, by connecting them to a small sample of existing nodes selected with probabilities proportional to their degrees. In other words, the more highly connected a node is, the more likely it is to acquire new connections.

Barabási and Albert (1999) proposed a simple model that directly instantiates the principle of preferential attachment: Each target for a new node's connections is sampled independently from the set of all existing nodes, with probability proportional to its current degree. This model produces power-law degree distributions with an exponent of 3, much like those we observed for semantic networks. However, it does not yield clustering coefficients that are nearly as high as those we observed. For instance, when the network size and density are comparable to the word association network, the model of Barabási and Albert yields values of C around .02, much lower than the observed value of .186. Asymptotically, as network size grows to infinity, C approaches 0 for this model, making it inappropriate for modeling small-world structures. Conversely, the classic model of small-world network formation, due to Watts and Strogatz (1998), does not capture the scale-free structures we have observed. In this section, we present an alternative model of preferential attachment that draws its inspiration from mechanisms of semantic development and that naturally produces both scale-free structures—with appropriate power-law exponents—and small-world structures—with appropriate clustering coefficients.

We should stress that in modeling the growth of semantic networks, our aim is to capture at an abstract level the relations between the statistics reported in the previous section and the dynamics of how semantic structures might grow. Our model's most natural domain of applicability is to semantic growth within an individual—the process of lexical development—but it may also be applicable to the growth of semantic structures shared across different speakers of a language or even different generations of speakers—the process of language evolution.

We frame our model abstractly in terms of nodes—which may be thought of as words or concepts—and connections between nodes—which may be thought of as semantic associations or relations. We do not attempt to model the development of particular words or concepts, based on particular experiences. Rather, our goal is to explain the statistics of semantic networks as the products of a general family of psychologically plausible developmental processes.[8] Over time, new nodes are added to the network and probabilistically attached to existing nodes on the basis of three principles. First, following the suggestions of many researchers in language and conceptual development (Brown, 1958a,b; Carey, 1978, 1985; Clark, 1993, 2001; Macnamara, 1982; Slobin, 1973), we assume that semantic structures grow primarily through a process of differentiation: The meaning of a new word or concept typically consists of some kind of variation on the meaning of an existing word or concept. Specifically, we assume that when a new node is added to the network, it differentiates an existing node by acquiring a pattern of connections that corresponds to a subset of the existing node's connections.

Second, we assume that the probability of differentiating a particular node at each time step is proportional to its current complexity—how many connections it has. Finally, we allow nodes to vary in a "utility" variable, which modulates the probability that they will be the targets of new connections. Utility variation is not necessary to produce any of the statistical features described in the previous section; it merely allows us to explore interactions between those features and aspects of word utility, such as usage frequency.

By focusing on the process of semantic differentiation, we do not mean to preclude a role for other growth processes. We have chosen to base our model on this single process strictly in the interests of parsimony. Incorporating additional processes would surely make the model more realistic but would also entail adding more free parameters, corresponding to the relative weights of those mechanisms. Given that the data we wish to account for consist of only the few summary statistics in Table 2 and the degree distributions in Fig. 5, it is essential to keep the number of free parameters to an absolute minimum. Our model for undirected networks (Model A) has no free numerical parameters, whereas our model for directed networks (Model B) has just one free parameter.

5.1. Model A: The undirected growing network model

Let n be the size of the network that we wish to grow, and $n(t)$ denote the number of nodes at time t. Following Barabási and Albert (1999), we start with a small fully connected network of M nodes ($M < n$). At each time step, a new node with M links is added to the network by randomly choosing some existing node i for differentiation and then connecting the new node to M randomly chosen nodes in the semantic neighborhood of node i. (Recall that the neighborhood H_i of node i consists of i and all the nodes connected to it.) Under this growth process, every neighborhood always contains at least M nodes; thus a new node always attaches to the network by connecting to a subset of the neighborhood of one existing node. In this sense, the new node can be thought of as differentiating the existing node, by acquiring a similar but slightly more specific pattern of connectivity.

To complete the model, we must specify two probability distributions. First, we take the probability $P_i(t)$ of choosing node i to be differentiated at time t to be proportional to the complexity of the corresponding word or concept, as measured by its number of connections:

$$P_i(t) = \frac{k_i(t)}{\sum_{l=1}^{n(t)} k_l(t)} \tag{3}$$

where $k_i(t)$ is the degree (number of connections) of node i at time t. The indexes in the denominator range over all existing $n(t)$ nodes in the network. Second, given that node i has been selected for differentiation, we take the probability $P_{ij}(t)$ of connecting to a particular node j in the neighborhood of node i to be proportional to the utility of the corresponding word or concept:

$$P_{ij}(t) = \frac{u_j}{\sum_{l \in H_i} u_l} \tag{4}$$

where the indexes in the denominator range over all nodes in the neighborhood H_i. To explore the interaction between word frequencies and network structure, we may equate a word's utility with its usage frequency (e.g., Kucera & Francis, 1967). For simplicity, we may also take all utilities to be equal, in which case the connection probabilities are simply distributed uniformly over the neighborhood of node *i:*

$$P_{ij}\left(t\right)=\frac{1}{k_i\left(t\right)} \tag{5}$$

For each new node added to the network, we sample repeatedly from the distribution in (4) or (5) until *M* unique nodes within the neighborhood of *i* have been chosen. The new node is then connected to those *M* chosen nodes. We continue adding nodes to the network until the desired network size *n* is reached. The growth process of the model and a small resulting network with $n = 150$ and $M = 2$ is illustrated in Fig. 6.

In our applications, *M* and *n* are not free to vary but are determined uniquely by the goal of producing a synthetic network comparable in size and mean density of connections to some real target network that we seek to model. The size *n* is simply set equal to the size of the target network. The parameter *M* is set equal to one half of the target network's mean connectivity $<k>$, based on the following rationale. Each new node in the synthetic network is linked to *M* other nodes, and the network starts with a small, fully connected subgraph of *M* nodes. Hence the average number of connections per node in the synthetic network is $<k> = 2M + M\,(M - 1)/n,$ which approaches $2M$ as *n* becomes large.

5.2. Model B: The directed growing network model

Our model for growing directed networks is practically identical to the model for undirected networks, with the addition of a mechanism for orienting the connections. The initial fully con-

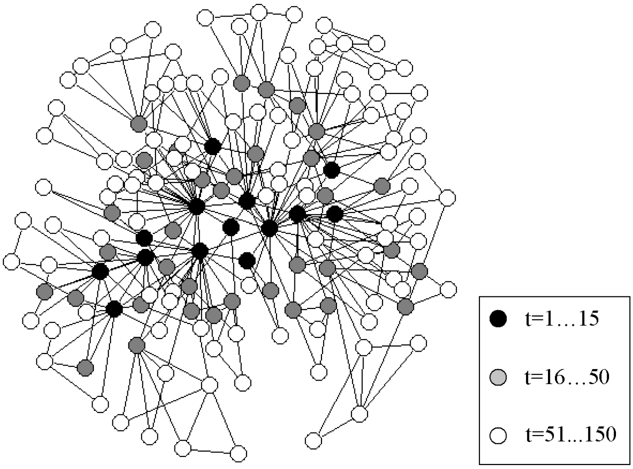

Fig. 6. Illustration of the undirected growing network model with n=150 and M=2. The shade of the nodes indicates the time step at which the nodes were first inserted.

nected network with M nodes now contains an arc from each node to every other node. The probability of choosing node i to be differentiated at a particular time t is still given by Equation (3), with the degree of a node in the directed graph defined as the sum of its in- and out-degrees, $k_i = k_i^{in} + k_i^{out}$. Each new node still makes connections to M existing nodes within the neighborhood of the node it differentiates, and those nodes are still sampled randomly according to Equations (4) or (5). The main novelty of Model B is that now each new connection may point in one of two possible directions, toward the existing node or toward the new node. We make the simplifying assumption that the direction of each arc is chosen randomly and independently of the other arcs, pointing toward the older node with probability α and toward the new node with probability $(1 - \alpha)$. The value of α is a free parameter of the model. Whether connections represent superficial associations or deep semantic dependencies, it seems plausible that they should typically point from the new node toward the previously existing node, rather than the other way around. Hence we expect the best fitting value of α to be significantly greater than .5, and perhaps just slightly less than 1.0.

5.3. Results and analyses

In principle, we could compare the products of these models with any of the real semantic networks analyzed previously. However, the computational complexity of the simulation increases dramatically as the network grows, and it has not been practical for us to conduct systematic comparisons of the models with networks as large as WordNet (Fellbaum, 1998; Miller, 1995) or Roget's (Roget, 1911) thesaurus. We have carried out a thorough comparison of Models A and B with the undirected and directed word association networks, respectively, and we report those results as follows. In all of these simulations, we set $n = 5{,}018$ to match the number of words in the free-association database. We set $M = 11$ in Model A and $M = 13$ in Model B to ensure that the resulting synthetic networks would end up with approximately the same density as the corresponding word association networks. We explored two different settings for the node utilities, one with all utilities equal and the other with utilities determined by word frequency, according to $u_i = \log(f_i+1)$. The word frequencies f_i were estimated from the Kucera and Francis (1967) counts of words in a large sample of written text. Because M and n were set to match the size and density of the associative networks, and the utilities were set uniformly or by observed word frequencies, there were no free numerical parameters in these mechanisms. The only free parameter occurred in the directed model: We varied α and obtained best results near $\alpha = 0.95$, corresponding to the reasonable assumption that on average, 19 out of 20 new directed connections point from a new node toward an existing node.

We evaluated the models by calculating the same statistical properties (see Table 2) and degree distributions (see Fig. 5) discussed previously for the real semantic networks. Because the growing models are stochastic, results vary from simulation to simulation. All the results we describe are averages over 50 simulations. These results are shown in Table 3 and Fig. 7.

For all of the statistics summarized in Table 3, both models achieved close fits to the corresponding real semantic networks. The mean statistics from 50 model runs were always within 10% of the corresponding observed values, and usually substantially closer. Fig. 5 shows that the degree distributions of the models matched those of the real semantic networks, both qualitatively and quantitatively. Model A, corresponding to the undirected associative network,

Table 3
Results of model simulations

	Undirected Associative Network			Directed Associative Network	
Variable	Data	Model A	Model B	Data	Model B
n	5,018	5,018	5,018	5,018	5,018
$<k>$	22.0	22.0	22.0	12.7	13.0
L	3.04	3.00 (.012)	3.00 (.009)	4.27	4.28 (.030)
D	5	5.00 (.000)	5.00 (.000)	10	10.56 (.917)
C	.186	.174 (.004)	.173 (.005)	.186	.157 (.003)
g	3.01	2.95 (.054)	2.97 (.046)	1.79	1.90 (.021)
L_{random}	3.03	—	—	4.26	—
C_{random}	4.35E–03	—	—	4.35E–03	—

Note. Standard deviations of 50 simulations given between parentheses. n = the number of nodes; $<k>$ = the average number of connections; L = the average shortest path length; D = the diameter of the network; C = clustering coefficient; γ = power law exponent for the distribution of the nunmber of edges in undirected netowrks and incoming connections in directed networks; L_{random} = the average shortest path length with random graph of same size and density; C_{random} = the clustering coefficient for a random graph of same size and density.

produced a degree distribution with a perfect power-law tail and exponent near 3. Model B, corresponding to the directed associative network, produced an approximate power-law distribution, with a slight downward inflection and an exponent somewhat less than 2. All of these results were very similar regardless of whether utilities were taken to be equal (as shown in Table 1 and Fig. 5) or variable according to the Kucera–Francis (Kucera & Francis, 1967) frequency distribution.

We also checked that the directed network model would reproduce the results of the undirected model when all directed links were converted to undirected links, which corresponds

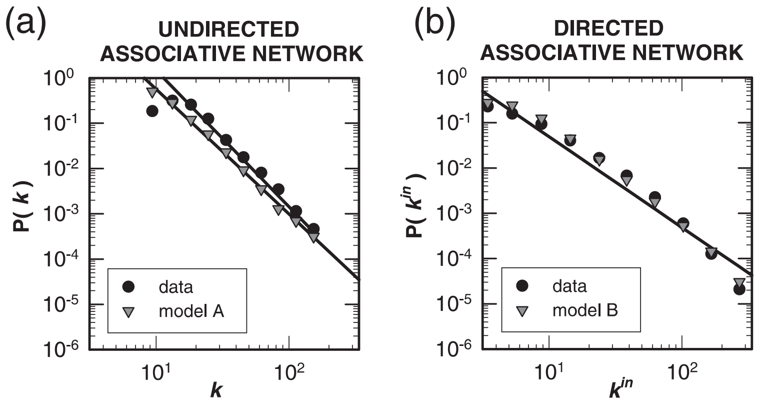

Fig. 7. The degree distributions of the word associative networks and model simulations with best fitting power law functions. (a) the undirected associative network is shown with the fit of model A: the undirected growing network model. (b), the in-degree distribution of the directed associative network is shown along with the fit of model B: the directed growing network model.

more accurately to the process by which the real undirected semantic network was constructed. We simulated Model B with $M = 11$ and $\alpha = 0.95$ and converted all arcs to edges at the end of each simulation. The results (shown in Table 3) were almost identical to those produced by Model A. This is to be expected, because of the particularly simple mechanism for choosing the directionality of new connections in Model B. The choice of directionality is made independently for each link and does not influence the other processes in the model. Removing the edge directionalities from Model B at the output stage thus renders it mechanistically equivalent to Model A.

5.4. Discussion

The qualitative features of the simulated networks make sense in light of the differentiation mechanism we have hypothesized for semantic network growth. Short average path lengths occur because of the presence of hubs in the network; many shortest paths between arbitrary nodes *x* and *y* involve one step from *x* to a hub and from the hub to node *y*. These hubs, and more generally, the power-law degree distributions, result from a version of the preferential attachment principle. At any given time step, nodes with more connections are more likely to receive new connections through the process of semantic differentiation, because they belong to more semantic neighborhoods, and those neighborhoods on average will be more complex (making them more likely to be chosen for differentiation, by Equation (3). The models produce high clustering coefficients because new connections are made only into existing semantic neighborhoods. This ensures high overlap in the neighborhoods of neighboring nodes.

It is not so easy to explain a priori the close quantitative match between the statistics and degree distributions of the word association networks and those of our simulated networks with comparable size and connection density. The fact that these results were obtained with no free parameters, in Model A, or one free parameter set to a reasonable value, in Model B, gives us reason to believe that something like the growth mechanisms instantiated in the models may also be responsible for producing the small-world and scale-free structures observed in natural-language semantic networks.

Our models are, at best, highly simplified abstractions of the real processes of semantic growth. This simplicity is necessary, given how little we know about the large-scale structures of real semantic networks—essentially, no more than is summarized in Table 2 and Figure 5. By pairing down the details of semantic growth to a simple mechanism based on differentiating existing words or concepts, we have been able to provide a unifying explanation for several nontrivial statistical structures.

There are clearly many ways in which our models could be made more realistic. Currently, new nodes always form their connections within a single semantic neighborhood, and new connections are added only when new nodes are added—never between two existing nodes. It would not be hard to remove these constraints, but it would necessitate additional free parameters governing the probability of making connections outside of a neighborhood (or in two or more neighborhoods) and the probability of adding a new connection between existing nodes. Removing these constraints would also make the models more flexible in terms of the kind of data they can fit; currently, the clustering coefficient and the shape and slope of the degree distribution are uniquely determined by the size and density of the network. It would also be possible to build models with different kinds of nodes and different kinds of connections, perhaps governed by different principles of growth. This complexity could be particularly important for modeling the network structure of WordNet (Fellbaum, 1998; Miller, 1995) or Roget's (Roget, 1911) thesaurus, which are based on a distinction between word nodes and class nodes. A thorough and rigorous exploration of the many possible models for semantic network growth should probably be deferred until we acquire more and better data about different kinds of semantic structures, and the computational resources to model the larger networks that we have already analyzed.

Finally, we acknowledge that we have been deliberately ambiguous about whether our model of semantic growth is meant to correspond to the process of language development within an individual's life span, or the process of language evolution across generations of speakers, or both. Although the mechanism of our model was primarily inspired by the language development literature (Brown, 1958a, 1958b; Carey, 1978, 1985; Clark, 1993, 2001; Macnamara, 1982; Slobin, 1973), we think that some kind of semantic differentiation process is also a plausible candidate for how new word meanings are formed between individuals. Clearly these two processes are coupled, as the critical period of language acquisition is a principal locus of cross-generational linguistic change (Pinker, 1994). In future work, we hope to relate our modeling efforts more closely to the knowledge base of historical linguistics, as well as recent mathematical models of language evolution (Niyogi & Berwick, 1997; Nowak & Krakauer 1999).

6. Psychological implications of semantic growth

Our proposal that the large-scale structure of semantic networks arises from the mechanisms of semantic growth carries more general implications for psycholinguistic theories beyond merely accounting for the graph-theoretic properties described previously. In this section, we focus on two issues: the viability of nongrowing semantic representations and the causal relations between a network's growth history, semantic complexity, and memory search behavior.

6.1. Power-law distributions and semantic growth

Conventional static views of semantic network organization—as either a hierarchical tree or an arbitrary graph—are consistent with the existence of short paths between any two nodes, but they do not predict either the small-world neighborhood clustering or the scale-free degree distributions that are found in real semantic networks. We have interpreted this constellation of features as the signature of a particular kind of network growth process, but we have not ruled out the possibility that some other kind of static semantic representation—perhaps not based on a network model at all—may give rise to these structures. A comprehensive survey of all previous semantic models is beyond the scope of this article, but we explore one currently popular alternative representation based on the analysis of co-occurrences of words in a large corpus of text.

LSA (e.g., Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998) has been proposed as a general theory for the representation and processing of semantic information. LSA is a machine learning model that induces a high-dimensional semantic space by a mathematical analysis of the way words are used in passages. The meaning of words is represented by vectors in the high-dimensional space, and the semantic similarity between words can then be determined by some measure of affinity in that space, such as the cosine of the angle between two vectors or the inner product. Landauer and Dumais showed that the local neighborhoods in semantic space successfully capture some subtle semantic relations. LSA differs from our network models in at least two important ways. First, it is a distributed representation; semantically similar words are represented by similar sets of features placing them in similar regions of a high-dimensional space. Second, it is a static model whose structure arises from a single-pass analysis of a large corpus, not a growth or learning process.

The question here is whether LSA captures some of the structural features that we observe in semantic networks, such as the presence of hubs or a power-law distribution of neighborhood sizes. If we can show that LSA does not naturally produce these same statistics, then it raises the possibility that there are nontrivial differences between growing network and spatial representations of semantic knowledge. To compare LSA representations with our semantic networks, we need some way of discretizing the continuous LSA space so that we can talk about the set of neighbors of a word rather than just the distances between words. We have explored two methods for creating local neighborhoods in LSA that we will refer to as the ε-method and the *k*-nn method.

The ε-method is similar in spirit to previous LSA analyses (Landauer & Dumais, 1997; Landauer et al., 1998); local neighborhoods are created by thresholding a continuous measure of similarity, such as the cosine of the angle between two word vectors, or the inner product of two word vectors. We could have restricted our analyses to the cosine measure only, as this is the measure in most information retrieval applications (see Landauer et al., 1998), but our initial analyses revealed important differences between the cosine and inner product measure. Two words are treated as neighbors if the similarity measure between their vectors exceeds some threshold ε. By varying the threshold ε, the size *n* and mean connectivity $<k>$ of the LSA network can be made to correspond with those of the other networks we have studied. In particular, we examined the LSA vector representation based on the Texas Association of Schools of Art corpus (Landauer et al., 1998) for all words (n =4,956), which also occurred in the word association database. This allowed us to compare the LSA network directly with the word association network and the outputs of our growing network model. We used thresholds ε that led to approximately the same $<k>$ as in the undirected word association network and our Model A.

The *k*-nn method differs from the ε-method in two important ways. First, it creates a directed network by selecting outlinks to the *k* nearest neighbors for each word in LSA. Second, the number of outlinks (*k*) is yoked to the number of outlinks for corresponding words in the word association network. Therefore, the out-degree distribution of the LSA network was made identical to the out-degree distribution of the word association network.

For both the ε and *k*-nn method, we applied two common measures for vector similarity based on the cosine of the angle and inner product. We also varied the dimensionality of the LSA space, $d = 50$, 200, and 400 (typical psychological applications of LSA set *d* at approximately 300).[9] In many cases, the LSA networks did not result in a single connected component.

In that case, we noted the number of connected components (denoted by *m*) and the size of the largest connected component (denoted by *n**). Table 4 presents the results for the LSA networks obtained with the ε and *k*-nn method for two similarity measures and three different dimensionalities. The network statistics <*k*>, *L, D, C,* and γ were all based on the largest connected component.

For both the ε and *k*-nn method and the two similarity measures, the LSA networks showed path lengths and clustering coefficients that could be reasonably characterized as a small-world structure, with much higher clustering and shorter path lengths than would be expected in a random network of equal size and density.

The networks based on LSA networks do not reproduce the observed degree distributions. In Fig. 8, the degree distributions based on cosine similarity show a bend that is typical of exponential distributions, not power-law distributions. Because the distributions are curved in log–log coordinates, it becomes difficult to interpret the slope of the best fitting line (γ) in these plots. In Table 4, we nevertheless show the best fitting values of γ to highlight this key difference between the LSA models and the real semantic networks.

The distributions based on inner products do look vaguely powerlike, especially for the LSA networks created with the ε-method. However, γ is lower than observed in word association. Also, for the ε-method with inner products, there is a large number of connected components, whereas in word association, there is just a single connected component.

We have shown that power laws of semantic connectivity do not arise in LSA when the data are processed in the standard way, with semantic similarity computed according to the cosine metric. We have also explored other ways of representing semantic association based on LSA, to see if power laws could emerge from processing the data in some nonstandard way. The only positive results we have obtained occur in networks based on an inner product metric, where two words are connected if their inner product in LSA space exceeds some threshold. Although such networks do yield power laws of connectivity in some cases, we doubt that they capture the same phenomenon we are seeing in human word association data and in our model. The power laws we obtain from LSA vectors under an inner product metric always have a slope that is significantly lower than the power laws we report here (e.g., approximately 1 or slightly higher). This suggests they may arise from a different source, such as the effects of word frequency, which is well known to follow a power-law distribution with a different exponent (Zipf, 1965) from the semantic power laws we focus on here. The inner product metric in LSA is just the cosine metric times the norm of each vector, and the norms of vectors in LSA are highly correlated with word frequency (approximately $r = .96$; Griffiths, personal communication, July 10, 2003). This correlation in LSA is a straightforward effect of the way that SVD operates on the (transformed) word-document co-occurrence matrix. Thus we have yet to find any strong evidence of power-law distributions in the semantic structures induced by LSA.

There are several reasons to think that scale-free connectivity is difficult to reproduce for many vector-space models of semantics, regardless of whether they are constructed through the machinery of LSA or some other procedure. Just as arbitrary graphs do not produce a power-law distribution of neighborhood sizes, neither do generic configurations of points in Euclidean space. For instance, if a large number of points are randomly distributed in a hypercube, and pairs of points are connected as neighbors whenever they are within some small Eu-

Table 4
Statistical properties of networks constructed from LSA semantic spaces

	Cosine similarity						Inner product similarity					
	ε-method			*k*-nn method			ε-method			*k*-nn method		
	d			*d*			*d*			*d*		
Variable	50	200	400	50	200	400	50	200	400	50	200	400
m	297	90	21	1	1	1	2,334	1,916	1,779	1	1	1
n^*	4,653	4,866	4,935	4,956	4,956	4,956	2,623	3,041	3,178	4,956	4,956	4,956
$<k>$	22.3	22.3	22.3	12.7	12.7	12.7	22.3	22.3	22.3	12.7	12.7	12.7
L	4.83	4.02	3.70	5.60	5.03	4.65	2.63	2.73	2.83	3.74	3.67	3.64
D	12	9	8	14	12	10	4	5	5	8	8	8
C	.456	.391	.298	.200	.219	.182	.601	.430	.340	.293	.358	.360
g	1.07	1.08	1.03	1.05	1.36	1.28	1.25	1.21	1.18	1.20	1.42	1.34

Note. LSA = Latent Semantic Analyses; d = dimensionality of the vector space; m = number of connected components, n^* = size of largest connected component; $<k>$ = the average number of connections; L = the average shortest path length; D = the diameter of the network; C = clustering coefficient; γ = power law exponent for the distribution of the nunmber of edges in undirected networks and incoming connections in directed networks.

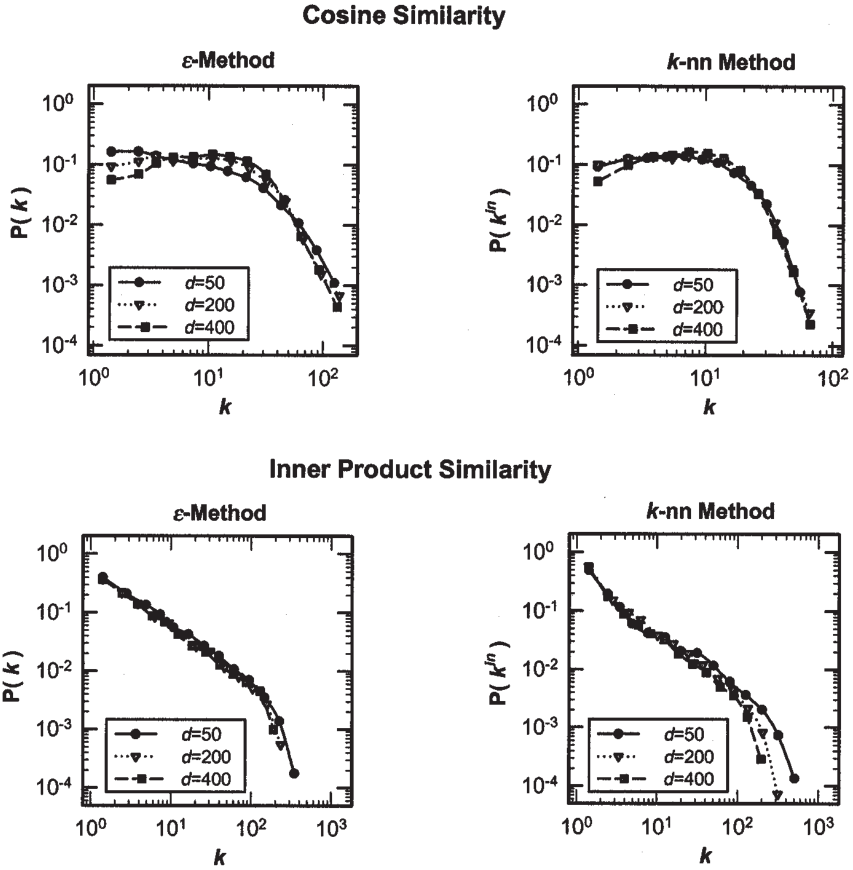

Fig. 8. The degree distributions for networks based on thresholded LSA spaces. For the ε-method, degree distributions of undirected networks are shown. For the *k*-nn method, the in-degree distributions are shown.

clidean distance ε of each other, the number of neighbors per point will follow a Poisson distribution (Stoyan, Kendall, & Mecke, 1995). This distribution is just a limiting case of the binomial distribution that describes neighborhood sizes in a random graph; both distributions have exponential tails that show up as nonlinear in log–log coordinates. In a similar vein, Tversky and Hutchinson (1986) identified certain geometric properties of Euclidean-space semantic representations that are not consistent with human similarity judgments. In particular, they argued that Euclidean geometry—particularly in low dimensions—places strong constraints on the maximum number of nearest neighbors that any point may have, and that these constraints are not satisfied in conceptual domains with even very elementary forms of non-Euclidean structure, such as a taxonomic hierarchy.

6.2. Age of acquisition, word frequency, and centrality

The core assumption of our model, that semantic structures derive from a growth process in which connections are introduced primarily between new nodes and existing nodes, predicts a causal relation between the history of a network's growth and its ultimate pattern of connectivity. This relation in turn motivates a unifying explanation for some previously observed behavioral phenomena of memory search, under the hypothesis that search processes exploit the structure of semantic networks in something like the way that search algorithms for information on the Web can exploit link structure of the Web (Brin & Page, 1998; Kleinberg, 2001).

Most generally, our growth model predicts a correlation between the time at which a node first joins the network and the number of connections that it ultimately acquires. More precisely, at any given time, older nodes should possess more connections than younger nodes, and this effect should interact with any variation in utility (e.g., word frequency) that influences the probability of connecting new nodes to particular existing nodes. Fig. 9 illustrates both the basic age effect and the nonlinear interaction with utility, using our undirected model of the word association data (Model A) and utility distributed according to the log transformed Kucera– Francis (Kucera & Francis, 1967) frequencies. (The directed network model shows similar results.) The age effect is strongest for nodes with highest utility because they acquire new connections at the highest rate.

To test these predictions of the model, we consulted age-of-acquisition norms that are available for small sets of words. Gilhooly and Logie (1980) asked adults to estimate, using an arbitrary rating scale, the age at which they first learned particular words. These ratings were converted to scores between 100 and 700, with a score of 700 corresponding to a word acquired very late in life. We took the average score for each word as a crude measure of the time at which that word typically enters an individual's semantic network. We also consulted the age-of-acquisition norms of Morrison et al. (1997), who in a cross-sectional study estimated the age at which 75% of children could successfully name the object depicted by a picture. Although these norms provide a more objective measurement for the age of acquisition, they were only available for a very small set of words.

Fig. 10 shows the relations between the number of connections that a word possesses, as measured in each of the three real semantic networks analyzed earlier, and its age of acquisi-

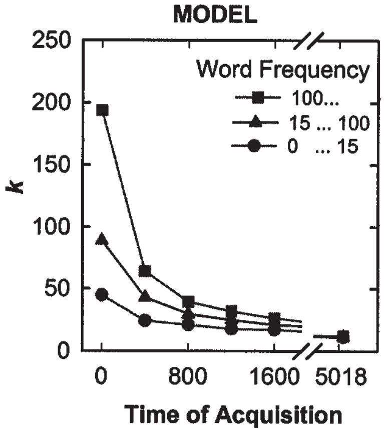

Fig. 9. In the growing network model, the degree of a node decreases nonlinearly as a function of the time since it was first connected to the network. This age effect holds regardless of variations in node utility (e.g., based on word frequency) and is greatest for the highest utility nodes. Here the nodes are binned into three sets corresponding to low, medium and high frequency in the norms of Kucera and Francis (1967) .

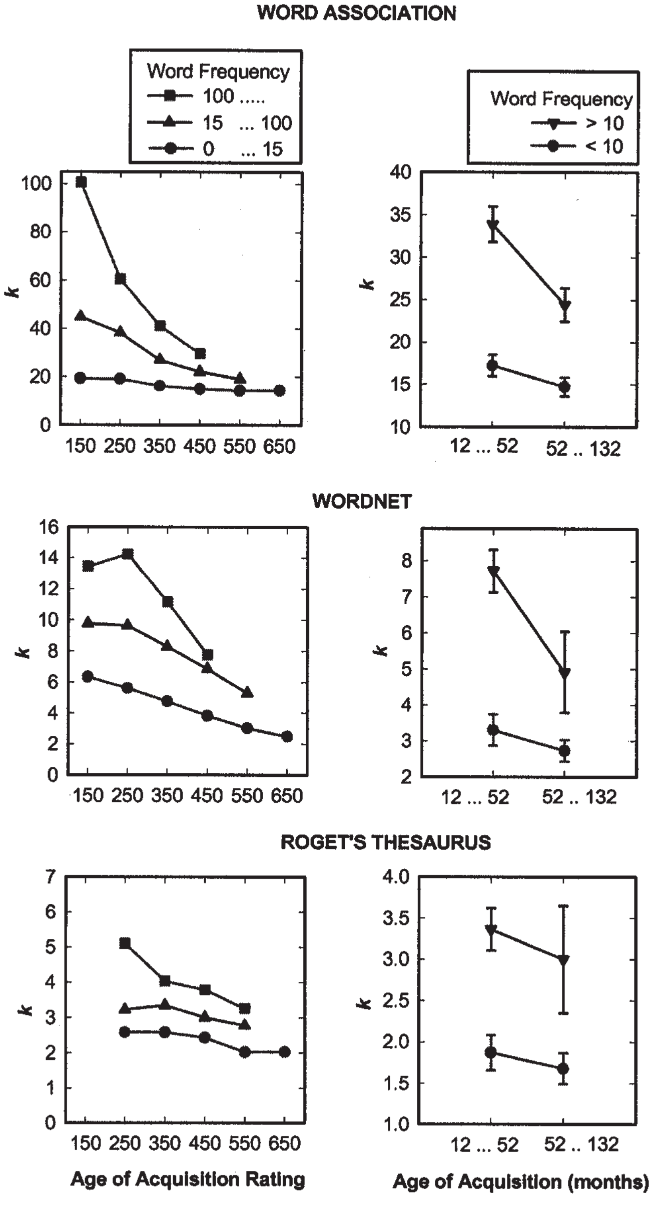

Fig. 10. The relation between degree and age of acquisition as measured by adult ratings (left panels) and the average age at which children can name pictures (right panels). Right panels include standard error bars around the means.

tion, as measured by both the adult rating and child picture-naming norms. We separated the words into three different frequency bins to show interactions between age of acquisition and word frequency. The data for all three networks are similar, but the word association network is most directly comparable to the model predictions shown in Fig. 9 (because the model was simulated to have the same size and density as this network). For both norms, early acquired words have more connections than late-acquired words. Also as predicted by the model, high-frequency words show higher connectivities, and the effect of age of acquisition on connectivity is most dramatic for high-frequency words.

Beyond the implication that semantic structure arises through some kind of temporally extended growth process, these relations between connectivity, frequency, and age of acquisition do not in themselves place strong constraints on the mechanisms of semantic network formation. In contrast with the small-world and scale-free structures discussed earlier, many different growth models would make similar predictions here. More significant is the possibility that some model of semantic structure based on growth processes may offer a unifying explanation for the behavioral effects of all of these variables.

Both high-frequency and early age of acquisition have been shown to correlate with short reaction-time latencies in naming (e.g., Carroll & White, 1973) and lexical-decision tasks (e.g., Turner, Valentine, & Ellis, 1998). Age of acquisition affects lexical tasks, such as naming, and lexical decision and age of acquisition has also been shown to affect semantic tasks, such as word association and semantic categorization (e.g., Brysbaert, Van Wijnendaele, & De Deyne, 2000; Ellis & Lambon Ralph, 2000). There has also been some debate on whether age of acquisition influences behavior independently of word frequency (e.g., Turner et al., 1998), or instead merely embodies cumulative frequency (Lewis, Gerhand, & Ellis, 2001), because high-frequency words are likely to be acquired earlier than low-frequency words.

In debating the causal status of variables such as frequency and age of acquisition, there are both functional (“why”) questions and mechanistic (“how”) questions. Functionally, it is clear that a bias toward high-frequency words or concepts would be useful in many cases, but it is not so clear what direct purpose an age-of-acquisition bias would serve. Mechanistically, there is good reason to doubt that either frequency or age of acquisition really is a direct cause of people’s behavior; it seems impossible for the history of learning to influence present behavior unless it has somehow left a structural trace in memory.

Our model of network growth suggests one possible structural substrate that could mediate the influences of both high frequency and early age of acquisition: the size of a node’s neighborhood, or the number of connections it makes to other nodes, which we have shown correlates with both of these history factors. In general, we can imagine various kinds of search processes that might activate nodes with higher connectivity more quickly, much as some connectionist modelers have suggested that retrieving traces with lower error could be faster for a number of reasons (e.g., Plaut, McClelland, Seidenberg, & Patterson, 1996). Mechanistically, a bias for retrieving high-connectivity nodes first could arise naturally if memory search is implemented by some kind of serial or parallel Markov process operating on a semantic network. Under many different probabilistic dynamics, a Markov search process would tend to find first those nodes with highest degrees (or in-degrees, for directed networks).

In addition to providing one clear mechanism for how the history of a word’s usage could affect present behavior, a bias to access words with high connectivity also has an intriguing functional basis. In Google, a state-of-the-art algorithm for Web searching (Brin & Page, 1998), sites are ranked in part based on a measure of their centrality in the WWW. The centrality of a node reflects both its authority—the extent to which other sites point to it—as well as the probability that it will be encountered in random walks on the network. Google measures the centrality of a node according to its projection onto the principal eigenvector of the normalized Web adjacency matrix. Essentially the same computation applied to a graph of feature dependencies, rather than WWW connections, has been used to assess the conceptual centrality of different object features (Sloman, Love, & Ahn, 1998). The (in-)degree of a node provides a simpler alternative measure of centrality (because it just involves counts of the number of incoming connections), which typically correlates highly with eigenvector measures of centrality and thus authority as well. Just as a Google search orders Web sites based on their centrality, so might the cognitive search processes involved in word production or lexical retrieval be functionally biased toward accessing more central, highly connected nodes first, as a means to direct processing toward the most authoritative and causally deep concepts that could be employed in a particular situation.

As a preliminary investigation of the behavioral significance of degree centrality, we undertook a correlational analysis of word frequency, age of acquisition, and node degree factors in predicting reaction times for several psycholinguistic tasks. Specifically, we looked at naming and lexical decision in two databases—a new naming latency database by Spieler and Brand (Brand, Rey, Peereman, & Spieler, 2001) for 796 multisyllabic words and a large lexical-decision latency database from Balota, Cortese, and Pilotti (1999) for 2,905 words. Node degrees were logarithmically transformed to avoid the extreme skew inherent in the degree distribution. Table 5 shows the correlations between latencies in naming and lexical-decision tasks and the three factors of degree centrality, age of acquisition, and word frequency. The results con-

Table 5
Correlations between naming and lexical decision latencies and three potential causal factors: word frequency, age of acquisition and degree centrality (in the word association, WordNet, and Roget's networks)

	Naming		Lexical Decision	
	n	*R*	*n*	*R*
Log(k), word association	466	–.330*	1676	–.463*
Log(k), WordNet	790	–.298*	2665	–.464*
Log(k), Roget	647	–.164*	2343	–.253*
Log (word frequency)	713	–.333*	2625	–.511*
AoA (rating)	199	.378*	566	.551*
AoA (picture naming)	44	.258	137	.346*
After partialing out log (word frequency)				
Log(k), word association	433	–.194*	1634	–.258*
Log(k), WordNet	706	–.171*	2503	–.274*
Log(k), Roget	602	–.110*	2243	–.136*
AoA (rating)	196	.337*	546	.450*
AoA (picture naming)	39	.208	131	.239*
After partialing out AoA (picture naming)				
Log(k), word association	33	–.279	107	–.414*
Log(k), WordNet	36	–.246	111	–.394*
Log(k), Roget	29	–.141	105	–.195*
Log (word frequency)	34	–.280	109	–.463*
After partialing out log (word frequency) & AoA (picture naming)				
Log(k), Word Association	32	–.171	106	–.234*
Log(k), WordNet	33	–.145	108	–.242*
Log(k), Roget	33	–.101	104	–.104

Note. R = correlation; *n* = number of observations; AoA = age of acquisition.
**p* < .05.

firm well-known findings that age of acquisition correlates positively with naming and lexical-decision latencies and that word frequency (using the norms of Kucera & Francis, 1967) correlates negatively with the same latencies. In other words, high-frequency or early acquired words are named faster and are quicker to be identified as words than low-frequency or late-acquired words. We also see that centrality is negatively correlated with these latencies: Words that are semantically more central have quicker reaction times. All of these correlations are relatively weak, but most are statistically significant (those with n >100). When the effects of word frequency or age of acquisition are partialed out of the analysis, the correlations between centrality and reaction-time latencies become weaker but remain significant for the lexical-decision task.

For the most part, these correlations are not surprising, given that we have already shown a correlation—and, in our growth model, a causal dependence—between degree of connectedness and the two history factors of age of acquisition and usage frequency. However, they suggest that in addition to perhaps being one mediator of learning history, degree centrality may also exert an independent effect on reaction times. This bolsters our hypothesis that memory search processes are dependent either functionally or mechanistically on the connectivity structure of semantic nets.

We are not suggesting that degree centrality is the only structural locus for learning history effects, nor that the causal relations between these factors only point in the one direction captured by our growth model. In particular, our model explains why more frequent, early acquired words are more densely connected, but it does not explain why certain words are acquired earlier or used more frequently than other words. There could well be a causal influence in the other direction, with age of acquisition and usage frequency dependent on the degree of connectivity in the semantic networks of competent adult speakers in a language community.

Our growing network model is also not the only model that attempts to capture the structural effects of age of acquisition and frequency. In several connectionist models (Ellis & Lambon Ralph, 2000; Smith, Cottrell, & Anderson, 2001), it has been found that the final accuracy for a particular training pattern depends on both its frequency of presentation and its age of acquisition (manipulated by hand in Ellis & Lambon Ralph, 2000, and estimated by Smith et al., 2001, to be the first time at which the training error drops below some threshold). The connectionist explanations depend on distributed representations: Early learned items induce a distributed representation that later learned items cannot easily change, and thus over time, the model loses the ability to encode new patterns accurately. Our growing network model offers an alternative view in which each word or concept is represented as a distinct entity, together with explicit connections to other entities. In contrast with the distributed representations account, this view has several distinctive features. Age-of-acquisition effects do not imply that recently learned meanings are encoded less accurately than older ones, only that they are less richly connected to other items in memory. If memory search processes are sensitive to connectivity, as we conjecture, then new items may be retrieved less effectively, but for other purposes they may be processed just as (or more) effectively or accurately than old items. Our model also provides a clear route for how age of acquisition could influence the dynamics of search behavior, via the structural feature of degree centrality that is both functionally and mechanistically relevant for effective network search procedures, and which seems to exert an independent effect on reaction times. It is not so clear in connectionist accounts how or why the magnitude of a pattern's training error should determine its accessibility or reaction time in a search process.

7. General discussion

In analyzing semantic networks, we are not endorsing traditional network models of semantic structure such as Collins and Quillian (1969) or Collins and Loftus (1975). Rather, we are analyzing one aspect of abstract semantic knowledge that could be captured in many different models defined at a more detailed level. We define a semantic network to be an abstract representation of one aspect of semantic knowledge—the pairwise relations between words. Semantic networks may directly form the basis for representing semantic knowledge, as in classic semantic net models, or they may be an abstraction computed from some other representation.

Both our semantic network analyses and our growing network model of semantic development are intended to capture abstract constraints on semantic representation and development that more precise mechanistic models may need to obey. As an example, consider vector-space models of semantic representation, such as LSA. LSA includes knowledge about the pairwise relations between words, which can be modeled as a semantic network by associating a node with each word and connecting two nodes if their corresponding points in LSA space are sufficiently close (e.g., within some small radius epsilon). We have shown that semantic networks constructed from LSA representations do not typically exhibit the characteristic large-scale statistical regularities found in our real-world semantic networks. To the extent that the statistical regularities turn out to be important constraints on semantic representation, this suggests a limitation in existing approaches to semantic representation, such as LSA, to model human knowledge. Such limitations can be overcome in a variety of ways. For example, Griffiths and Steyvers (2002, 2004) proposed the "topics model," a probabilistic approach that can capture most of the statistical regularities of semantic network structure that we have reported. By representing words as probability distributions over topics as opposed to spatial locations in a high-dimensional space, the topics model is able to avoid the inherent constraints associated with spatial models (see also, Tversky & Hutchinson, 1986) that make it difficult to produce small-world structures. Alternatively, one can model the observed statistical regularities by models of semantic development that embody something like the process of differentiation that drives our growing network model, or an alternative kind of "rich get richer" process.

We have not attempted to make a precise connection between our semantic network models and neural networks in the brain. There is nothing in our growth model that necessarily requires semantic concepts to be represented in a local or distributed fashion in the brain. However, our empirical findings of scale-free structure in semantic networks do place some constraints on how semantic networks can (or cannot) be implemented in neural hardware. Under a simple (localist) mapping of nodes and connections in our semantic networks onto neurons and synapses, the statistical structures we describe at the semantic level are not compatible with the general features of neuroanatomical structure. The neural network of the worm *C. elegans* has been shown to have a small-world structure (Watts & Strogatz, 1998), but it is definitely not scale-free it its connectivity. The degree distribution falls off according to a very clear exponential law with a single characteristic scale (Amaral et al., 2000). Likewise, the connectivity of neurons within a cortical area may have some small-world features, due to the existence of excitatory connections within a local two-dimensional neighborhood and long-range inhibitory connections between neighborhoods (Kleinberg, 2000), but it is almost certainly not scale-free. There are typically only one or a few scales of connectivity in any cortical area, with each neuron making a number of connections that is fairly similar to other neurons of the same type (Kandel, Schwartz, & Jessell, 1991). Any direct mapping from nodes in a semantic network onto single neurons or columns of neurons, and from the connections between nodes onto the synapses between those neurons, would therefore not preserve the essential scale-free structure of semantic relations. Thus our findings seem most compatible with relatively distributed neural representations, and they caution against drawing any simplistic analogies between nodes and neurons or connections and synapses. More likely, the correspondence between semantic nets and neural nets takes the form of a functional mapping, implemented in the physiology of neural populations, rather than a structural mapping implemented in their microanatomy. When the time comes for a serious study of these mind–brain correspondences, the quantitative principles that we have presented here may provide one source of guidance in identifying the physiological basis of semantic representations.

Our models appear to assume a *localist* representation for semantic concepts: Each concept is associated with a single distinct node in the semantic network. Some theorists have argued that semantic representations are better characterized as distributed across a network (e.g., O’Reilly & Munakata, 2000; Rogers & McClelland, 2004), in part based on the fact that with aging and the onset of dementia, semantic concepts typically degrade gradually rather than in an all-or-none fashion (Hodges, Graham, & Patterson, 1995; Warrington, 1975). However, the notion of distributed representation is not in fact in conflict with our proposal. We have assumed that the meaning of a concept is represented not by a single node in the semantic network, but by the pattern of connectivity between the node associated with that concept and nodes associated with related concepts. In terms of contemporary philosophy of mind, this theory is more like a version of conceptual role semantics (Block, 1998) than conceptual atomism (Fodor, 1998).

Gradual degradation of semantic concepts may be accounted for in our framework as the decay—either sudden or gradual—of the links between nodes that constitute meanings. All of our semantic networks have a giant connected component containing the vast majority of the nodes in the network. We might speculate that a concept becomes inaccessible when it becomes disconnected from this giant connected component. This proposal would match the observation that concepts acquired earliest disappear latest (e.g., Hodgson & Ellis, 1998; Lambon Ralph, Graham, Ellis, & Hodges, 1998), because they have the most connections and thus would take longest to become disconnected from the giant connected component of concepts in the semantic network.

8. Conclusion

We have found that several semantic networks constructed by quite different means all show similar large-scale structural features: high sparsity, very short average path lengths, strong local clustering, and a power-law distribution for the number of semantic neighbors, indicating a hublike structure for knowledge organization. The fact that we observe such similar and nontrivial statistical regularities across three different semantic networks constructed from very different kinds of data suggest that these abstract representations do in fact lock onto a useful level of analysis and that the structures we have uncovered could place important constraints on more detailed models of semantic structure.

These statistical principles of large-scale semantic structure may be valuable beyond offering the potential to link the organization of mature semantic representations to the processes of language development or evolution. In the most general terms, robust quantitative empirical relations typically provide some of the strongest constraints on computational models of cognition, and until now there have been few such laws for semantic structure. Any computational model that seeks to explain how semantic structures form and what shape they ultimately take will have to reckon with these results.

Models of semantic processing may also have to be sensitive to these structures because, in the words of one prominent network theorist, "structure always affects function." (Strogatz, 2001, p. 268). Since the seminal work of Collins and Quillian (1969), which explored the interaction between one simple kind of structure for semantic networks and its complementary processes, researchers have thought mainly in terms of general processes such as spreading activation operating on arbitrary structures. However, the finding of small-world and scale-free structures in semantic networks might cause us to rethink how search and retrieval could work in these networks. We have already suggested how retrieval processes in lexical decision and naming might be attuned to one aspect of semantic network connectivity—namely, node centrality—which leads as a natural consequence to effects of frequency and age of acquisition on reaction time. More generally, the processes involved in searching for relevant and useful knowledge in semantic networks might be adapted to their large-scale structure in any number of ways, and might be very different from search processes adapted to arbitrarily structured graphs, inheritance hierarchies, or high-dimensional Euclidean vector spaces. Developing search and retrieval algorithms that exploit the large-scale structures of semantic networks we have described here is likely to be a project of great promise, for both cognitive science research and information retrieval applications.

Notes

1. For reasons of space, we provide only heuristic definitions for some of these terms. See Watts (1999) for a more in-depth treatment of graph-theoretic concepts in connection to small-world networks.
2. Note that distances in an undirected graph always satisfy the three metric axioms of minimality, symmetry, and the triangle inequality (see Tversky, 1977), but distances in an undirected graph do not in general satisfy the latter two.
3. We implemented Dijkstra's algorithm with Fibonacci heaps (Cormen, Leiserson, & Rivest, 1990) as an efficient means to find the shortest paths between a given node and all other nodes. (Matlab code for this algorithm is available from the first author.) For very large networks, it is often computationally infeasible to calculate the shortest paths

between all pairs of nodes. In such cases, we can estimate *L* and *D* based on the shortest paths for a large sample of nodes (see Note 6).

4. The second property holds simply because *P*(*k*) in a random graph is a binomial distribution (Bollobás, 1985), and all binomial distributions have this shape. The first trend occurs because a lower value of *C*—lower overlap in the neighborhoods of neighboring nodes—implies that, on average, more distinct nodes can be reached in two steps from any given node. At the extreme value of $C = 0$, not one of the nodes that can be reached by taking two steps from node *i* is also a neighbor of *i,* and thus the number of nodes within a given distance of each node grows exponentially with distance. As *C* increases, more of the paths radiating out from a node become redundant and more steps are required on average to connect any two nodes.
5. For WordNet, there were connections between word and meaning nodes, between word and word nodes, and between meaning and meaning nodes; these connections were rearranged separately when constructing the random graphs.
6. For the word associative networks, *L* and *D* were calculated on the basis of the path lengths between all word pairs in the large (strongly) connected component. For the much larger networks of WordNet and Roget's thesaurus, *L* and *D* were based on the path lengths between all pairs of a sample of 10,000 words from the large connected component (see Note 3).
7. Zipf (1965) plotted the number of meanings of a word versus its rank of its word frequency in log–log coordinates and observed a slope $b = .466$. Adamic (2000) provided some simple analytic tools by which the slope $b = .466$ in this Zipf plot can be converted to $\gamma = 3.15$, the slope of the corresponding probability distribution in log–log coordinates.
8. These processes of semantic development may include verbal processes, as well as preverbal and nonverbal processes. Although the empirical data on which we evaluate our models of semantic growth come from statistical analyses of large linguistic databases, we do not mean to suggest that the mechanisms of semantic development in an individual are necessarily driven by statistical analyses of linguistic experience (e.g., Landauer & Dumais, 1997). Many important processes of conceptual development are preverbal or nonverbal, and these processes may be most likely to generate the early stages of the network structures in our model. As such, it would be desirable to evaluate our models of semantic growth on some form of nonverbal data, but currently the only available data directly assessing the large-scale structure of semantic networks are the linguistic or verbal data we consider here.
9. Both the path lengths and the clustering coefficients of the LSA networks show a decreasing trend as the dimensionality *d* is increased. It is possible that at dimensionalities higher than 400, these statistics will come closer to the values observed for the word association network. We were not able to investigate this possibility, as only 400 LSA dimensions were available to us. However, it is unlikely that increasing the dimensionality would threaten our main argument, because the lack of scale-free degree distributions is the primary feature distinguishing the LSA networks from naturally occurring semantic networks and our growing network models. Based on Fig. 7, it seems doubtful that these distributions would match significantly better at higher dimensionalities (unless perhaps *d* was increased far beyond the typical value of 300).

Acknowledgments

We are grateful to Tom Landauer and Darrell Laham for providing us with the TASA corpus. We also thank Eve Clark, Tom Griffiths, Doug Nelson, and Vin de Silva for discussions that helped shape this research.

References

Adamic, L. A. (1999). The small-world web. In S. Abiteboul & A-M. Vercoustre (Eds.), *Research and advanced technology for digital libraries: Third European conference, ECDL '99* (pp. 443–452). New York: Springer.

Adamic, L. A. (2000). *Zipf, power-laws, and pareto—A ranking tutorial.* Retrieved July 1, 2003, from http://www.hpl.hp.com/research/idl/papers/ranking/ranking.html

Albert, R., & Barabási, A. L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics, 74*(1), 47.

Albert, R., Jeong, H., & Barabási, A. L. (1999, September 9). Diameter of the world wide web, *Nature, 401,* 130–131.

Albert, R., Jeong, H., & Barabási, A. L. (2000, July 27). Error and attack tolerance of complex networks. *Nature, 406,* 378–382.

Amaral, L. A. N., Scala, A., Barthélémy, M., & Stanley, H. E. (2000). Classes of small-world networks. *Proceedings of the National Academy of Sciences, 97,* 11149–11152.

Anderson, J. R. (2000). *Learning and memory: An integrated approach* (2nd ed.). New York: Wiley.

Balota, D. A., Cortese, M. J., & Pilotti, M. (1999). Item-level analyses of lexical decision performance: Results from a mega-study. In *Abstracts of the 40th Annual Meeting of the Psychonomics Society* (p. 44). Los Angeles, CA: Psychonomic Society.

Barabási, A. L., & Albert, R. (1999, October 15). Emergence of scaling in random networks. *Science, 286,* 509–512.

Block, N. (1998). Semantics, conceptual role. *Routledge Encyclopedia of Philosophy Online.* Retrieved from http://www.rep.routledge.com.

Bollobás, B. (1985). *Random graphs.* London: Academic.

Brand, M., Rey A., Peereman, R., & Spieler, D. (2001). Naming bisyllabic words: *A large scale study.* Paper presented at the 12th Conference of the European Society for Cognitive Psychology (ESCOP 2001), Edinburgh, Scotland.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *WWW7/Computer Networks, 30,* 107–117.

Brown, R. (1958a). How shall a thing be called? *Psychological Review, 65,* 14–21.

Brown, R. (1958b). *Words and things.* Glencoe, IL: Free Press.

Brysbaert, M., Van Wijnendaele, I., & De Deyne, S. (2000). Age-of-acquisition effects in semantic processing tasks. *Acta Psychologica, 104,* 215–226.

Carey, S. (1978). The child as word learner. In J. Bresnan, G. Miller, & M. Halle (Eds.), *Linguistic theory and psychological reality* (pp. 264–293). Cambridge, MA: MIT Press.

Carey, S. (1985). Constraints on semantic development. In J. Mehler (Ed.), *Neonate cognition* (pp. 381–398). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Carroll, J. B., & White, M. N. (1973). Word frequency and age-of-acquisition as determiners of picture naming latency. *Quarterly Journal of Experimental Psychology, 25,* 85–95.

Chomsky, N. (1957). *Syntactic structures.* The Hague, The Netherlands: Mouton.

Chomsky, N. (1959). A review of B. F. Skinner's *Verbal behavior. Language, 35*(1), 26–58.

Clark, E. V. (1993). *The lexicon in acquisition.* New York: Cambridge University Press.

Clark, E. V. (2001). Making use of pragmatic inferences in the acquisition of meaning. In D. Beaver, S. Kaufman, B. Clark, & L. Cazillas (Eds.), *Stanford papers in semantics* (pp. 23–36). Stanford, CA: CSLI Publications.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82,* 407–428.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior, 8,* 240–248.

Cormen, T., Leiserson, C., & Rivest, R. (1990). *Introduction to algorithms.* Cambridge, MA: MIT Press.

Deese, J. (1965). *The structure of associations in language and thought.* Baltimore: Johns Hopkins University Press.

Ellis, A. W., & Lambon Ralph, M. A. (2000). Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: Insights from connectionist networks. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 26,* 1103–1123.

Erdös, P., & Réyni, A. (1960). On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences, 5,* 17–61.

Fellbaum, C. (Ed.). (1998). *WordNet, an electronic lexical database.* Cambridge, MA: MIT Press.

Fodor, J. A. (1998). *Concepts: Where cognitive science went wrong.* New York: Oxford University Press.

Gasser, M., & Smith, L. B. (1998). Learning nouns and adjectives: A connectionist account. *Language & Cognitive Processes, 13,* 269–306.

Gentner, D. (1981). Some interesting differences between nouns and verbs. *Cognition and Brain Theory, 4,* 161–178.

Gilhooly, K. J., & Logie, R. H. (1980). Age of acquisition, imagery, concreteness, familiarity and ambiguity measures for 1944 words. *Behaviour Research Methods and Instrumentation, 12,* 395–427.

Griffiths, T., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*(Suppl. 1), 5228–5235.

Griffiths, T. L., & Steyvers, M. (2002). A probabilistic approach to semantic representation. In W. Gray & C. Schunn (Eds.), *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society* (pp. 381–386). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Hodges, J. R., Graham, N., & Patterson, K. (1995). Charting the progression in semantic dementia: Implications for the organisation of semantic memory. *Memory, 3,* 463–495.

Hodgson, C., & Ellis, A. W. (1998). Last in, first to go: Age of acquisition and naming in the elderly. *Brain & Language, 64,* 146–163.

Jeong, H., Tombor, B., Albert, R., Oltval, Z. N., & Barabási, A. L. (2000, October 5). The large-scale organization of metabolic networks. *Nature, 407,* 651–654.

Kandel, E. R., Schwartz, J. H., & Jessell, T. M. (Eds.). (1991). *Principles of neural science.* Norwalk, CT: Appleton & Lange.

Keil, F. C. (1979). *Semantic and conceptual development: An ontological perspective.* Cambridge, MA: Harvard University Press.

Kleinberg, J. M. (2000). Navigation in a small world. *Nature, 406,* 845.

Kleinberg, J. M. (2002). Small-world phenomena and the dynamics of information. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems, 14,* (pp. 431–438). Cambridge, MA: MIT Press.

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English.* Providence, RI: Brown University Press.

Lambon Ralph, M. A., Graham, K. S., Ellis, A. W., & Hodges, J. R. (1998). Naming in semantic dementia—What matters? *Neuropsychologia, 36,* 775–784.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 211–240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes, 25,* 259–284.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation.* Chicago: University of Chicago Press.

Lewis, M. B., Gerhand, S., & Ellis, H. D. (2001). Re-evaluating age-of-acquisition effects: Are they simply cumulative frequency effects? *Cognition, 78,* 189–205.

Macnamara, J. (1982). *Names for things: A study in human learning.* Cambridge, MA: MIT Press.

Manning, C. D., & Schutze (1999). *Foundations of statistical natural language processing.* Cambridge, MA: MIT Press.

Milgram, S. (1967, May). The small-world problem. *Psychology Today, 2,* 60–67.

Miller, G. A. (1995). WordNet: An on-line lexical database [Special issue]. *International Journal of Lexicography, 3*(4).

Morrison, C. M., Chappell, T. D., & Ellis, A. W. (1997). Age of acquisition norms for a large set of object names and their relation to adult estimates and other variables. *Quarterly Journal of Experimental Psychology, 50A,* 528–559.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1999). *The University of South Florida word association norms.* Retrieved from http://w3.usf.edu/FreeAssociation

Nelson, D. L., McKinney, V. M., Gee, N. R., & Janczura, G. A. (1998). Interpreting the influence of implicitly activated memories on recall and recognition. *Psychological Review, 105,* 299–324.

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences, 98,* 404–409.

Niyogi, P., & Berwick, R. C. (1997). Evolutionary consequences of language learning. *Journal of Linguistics and Philosophy, 17,* 697–719.

Nowak, M. A., & Krakauer, D. C. (1999). The evolution of language. *Proceedings of the National Academy of Sciences, 96,* 8028–8033.

O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain.* Cambridge, MA: MIT Press.

Pinker, S. (1994). *The language instinct.* New York: Morrow.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103,* 56–115.

Rips, L. J., Shoben, E. J., & Smith, E. E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior, 12,* 1–20.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic knowledge: A parallel distributed processing approach.* Cambridge, MA: MIT Press.

Roget, P. M. (1911). *Roget's Thesaurus of English Words and Phrases* (1911 ed.). Retrieved October 28, 2004, from http://www.gutenberg.org/etext/10681

Simon, H. (1955). On a class of skew distribution functions. *Biometrika, 42,* 425–440.

Skinner, B. F. (1937). The distribution of associated words. *Psychological Record, 1,* 71–76.

Slobin, D. I. (1973). Cognitive prerequisites for the acquisition of grammar. In C. A. Ferguson & D. I. Slobin (Eds.), *Studies of child language development* (pp. 173–208). New York: Holt, Rinehart & Winston.

Sloman, S. A. (1998). Categorical inference is not a tree: The myth of inheritance hierarchies. *Cognitive Psychology, 35,* 1–33.

Sloman, S. A., Love, B. C., & Ahn, W (1998). Feature centrality and conceptual coherence. *Cognitive Science, 22,* 189–228.

Smith, M. A., Cottrell, G. W., & Anderson, K. L. (2001). The early word catches the weights. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems 13* (pp. 52–58). Cambridge MA: MIT Press.

Sommers, F. (1971). Structural ontology. *Philosophia, 1*(1), 21–42.

Stoyan, D., Kendall, W. S., & Mecke, J. (1995). *Stochastic geometry and its applications* (2nd ed.). New York: Wiley.

Strogatz, S. H. (2001, March 8). Exploring complex networks. *Nature, 410,* 268–276.

Turner, J. E., Valentine, T., & Ellis, A. W. (1998). Contrasting effects of age of acquisition and word frequency on auditory and visual lexical decision. *Memory & Cognition, 26,* 1282–1291.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84,* 327–352.

Tversky, A., & Hutchinson, J. W. (1986). Nearest neighbor analysis of psychological spaces. *Psychological Review, 93,* 3–22.

Warrington, E. K. (1975). The selective impairment of semantic memory. *Quarterly Journal of Experimental Psychology, 27,* 635–657.

Watts, D. J. (1999). *Small worlds: The dynamics of networks between order and randomness.* Princeton, NJ: Princeton University Press.

Watts, D. J., & Strogatz, S. H. (1998, June 4). Collective dynamics of 'small-world' networks. *Nature, 393,* 440–442.

Zipf, G. K. (1965). *Human behavior and the principle of least effort.* Hafner: New York.