The Large-Scale Structure of Semantic Networks

Mark Steyvers (msteyver@psych.stanford.edu)

Josh Tenenbaum (jbt@psych.stanford.edu)

Department of Psychology, Stanford University,

Stanford, CA 94305-2130

Working draft, comments welcome

To be submitted to Cognitive Science

Abstract

We present graph-theoretic analyses of three 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 suggest a hub structure similar to that found in other natural networks, such as the world wide web. We propose a model for the growth of semantic networks, in which new words are preferentially attached to well-connected words and their neighbors. This model generates appropriate small-world statistics and power-law connectivity distribution and also predicts an observed correlation between age-of-acquisition and connectivity.

The Large-Scale Structure of Semantic Networks

Network structures provide intuitive and useful representations for semantic knowledge and inference systems. Semantic networks were introduced in the theory of Collins and Quillian (1969), which represented concepts as hierarchies of interconnected nodes with nodes linked to characteristic attributes. The structure of semantic networks such as networks formed by word association has proven to be useful to predict performance in a variety of experimental tasks such as recall and recognition (e.g., Deese, 1965, Nelson, ??) but the structure itself of semantic networks has not been a subject of investigation. It is important to understand the structure of semantic networks because it reflects the organization of meaning and language. In this paper, we aim to investigate the large-scale structure of several semantic networks constructed by different means by measuring a few statistical properties. These statistical properties can then be used to distinguish semantic networks from other networks such as random networks where concepts are linked by random connections. With these statistical properties, we can also compare the structure of semantic networks with the structure of other semantic representations such as high dimensional semantic spaces (e.g., Landuaer & Dumais, 1997) in which words are represented as points in a high dimensional space.

We will use some tools from graph theory to specify the large-scale organization of semantic networks by distributions over a few variables, such as the length of the shortest path between two words and the number of connections per word. We show that these distributions display similar, nontrivial patterns for several semantic networks. We then argue that these regularities place constraints on the mechanisms by which connections between words develop – either in language evolution or language acquisition, or both – and we propose a simple framework for modeling the growth of a semantic network consistent with these constraints.

In particular, we will show that the large-scale organization of semantic networks reveals a “small-world” structure, similar to that found in many other natural networks (Watts & Strogatz, 1998). In addition, we will propose a network model that mimics in several important ways the global organization of semantic networks. This network acquires new concepts over time and connects these concepts preferentially to existing concepts that are rich in connections to other concepts.

Two kinds of predictions follow from the model. First, because new concepts are preferentially attached to rich concepts, the distribution of the connectivity follows a power law: some concepts have a connectivity that is

orders of magnitude larger than the average concept. A related prediction is that semantic networks are scale-free: as the learner adds new concepts to the network, the distribution of the connectivity remains a power law with the same shape. Second, because the model builds the representation of new concepts on older concepts, the order in which concepts are learned is crucial in determining their connectivity. Concepts that are learned early in life should show higher connectivity. We will show how this growth model can predict some behavioral effects of age of acquisition in lexical decision and naming tasks.

Basic Concepts from Graph Theory

Before we provide an introduction into research involving small-world networks, we will first define some terminology from graph theory and introduce the statistical properties that we will use to describe the structure of semantic networks. Because of space limitations, we provide only heuristic definitions for these terms. The reader is referred to Watts (1999) for a more in-depth treatment of graph-theoretic concepts in connection to small-world networks.

A *graph* or *network* consists of a set of *nodes* (also called *vertices*) and a set of *edges* or *arcs* that join the nodes. The number of nodes in the network will be 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*. When the network is undirected, the *degree* of a node is the number of edges connected to the node. The degree of node *i* will be denoted by the variable k_i. When the network is directed, the *in*- and *out-degree* of a node is the number of arcs connected to the node that are incoming and outgoing respectively. The variables k_i^{in} and k_i^{out} will refer to the in- and out-degree of node *i* respectively. The average degree will be denoted by $\langle k \rangle$, which will be important to assess the density of connections in a network.

In an undirected graph, a *path* is a sequence of edges that connects one node to another. In a directed graph, a path is a set of arcs that can be followed along the direction of the arcs from one node to another. We will refer to the *path length* between *x* and *y* as the number of edges or arcs along the path from node *x* to *y*. In a *connected graph*, there exists between any pair of nodes a path, i.e., one can follow edges to from any node to any other node. A *strongly connected graph* is a graph in which there exists between any pair of nodes a path along the arcs. A (strongly) *connected component* is a subset of a nodes that is (strongly) connected.

We will now introduce four statistical features that are important indicators for the structure of a network: the *average shortest path length* (*L*), *diameter* (*D*), *clustering coefficient* (*C*), and the *distribution of node degrees*. The variable *L* refers to the average shortest path length between nodes while *D* refers to the diameter, the length of the longest of these shortest paths between nodes (i.e., at most *D* steps are needed to go from one node to another while on average *L* steps are needed)[1]. For very large networks, it may become computationally infeasible to calculate the shortest paths between all pairs of nodes. In such cases, *L* and *D* can be estimated on the basis of the shortest paths between a small sample of nodes.

Another important statistical feature that determines the structure of networks is the clustering coefficient. The clustering coefficient *C* measures the amount of local clustering or the tendency in the network for neighbors of nodes to be also each others neighbors. The clustering coefficient can be defined as follows (Watts, 1999):

$$C = \frac{3\,x\ \ number\ of\ \ triangles}{number\ of\ \ connected\ \ \ triples\ of\ \ nodes} \qquad (1)$$

A triangle is a set of nodes *x*, *y*, and *z* that are fully connected (i.e., there exist edges or arcs between node pairs *x*-*y*, *y*-*z*, and *x*-*z*). A connected triple of nodes is a set of nodes *x*, *y*, and *z* in which there is a path from any node to any other node (e.g., edges or arcs between node pairs *x*-*y*, and *y*-*z* would form a connected triple). By dividing the number of triangles in the network by the number of connected triples, *C* is normalized in the range of 0 to 1. When *C*=0, no nodes have neighbors that are also each others neighbors. In a fully connected network (all nodes are connected to all other nodes), *C*=1. While the clustering coefficient is sensitive to the number of connections in a network, it is possible for two networks to have the same number of connections but different clustering coefficients (see Figure 1). Finally, note that because the definitions for triangles and connected triples are independent of whether the connection are based on edges or arcs, the clustering coefficient for a directed network remains the same when the network is turned into an undirected network (by replacing all arcs by edges).

The last statistical feature of networks that we will consider is the distribution of node degrees which we refer to as the degree distribution. To create this distribution, the relative frequencies of nodes with degree 1, 2, 3, etc. are calculated so that a plot can be constructed of P(*k*) versus *k*. This visualizes the probability of observing a node with degree *k* for all possible values of *k*. As will become clear later, the shape of this plot is an important indicator for the structure of the network.

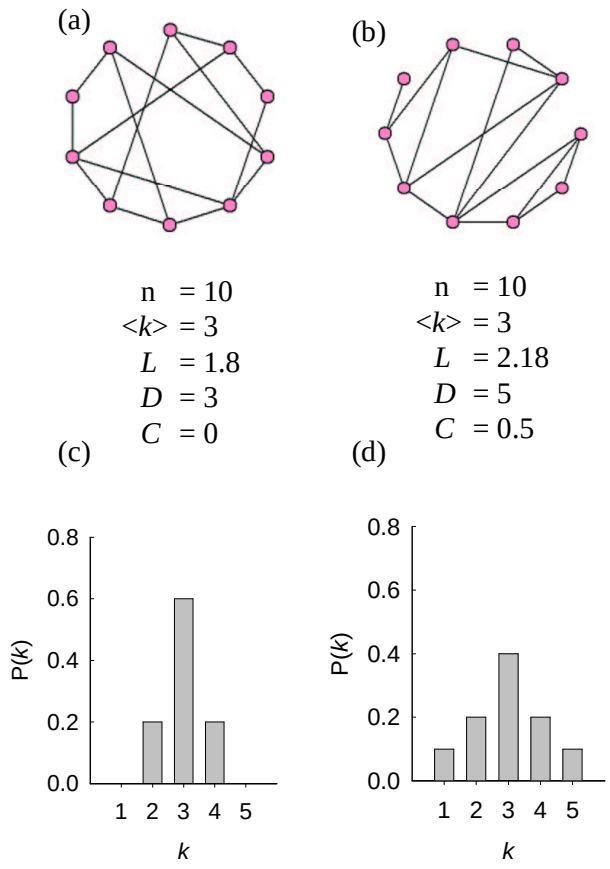

Figure 1. Two networks (a) and (b) with equal number of nodes and edges. For both networks, the variables *n* (number of nodes) and <*k*> (average degree, i.e., average number of edges) are shown as well as the statistical properties *L* (average shortest path length), *D* (diameter) and *C* (clustering coefficient). Note that the networks in (a) and (b) have different clustering coefficients even though they have the same *n* and <*k*>. In (c) and (d), the degree distributions are shown corresponding to the networks in (a) and (b).

Figure 1 shows an illustration of two networks with 10 nodes and 15 edges. The statistical properties of *L*, *D*, and *C* are given (the reader can verify the numbers with a little bit of work). The network on the left has *C*=0 (no nodes have neighbors that are also each others neighbors) and the other has C=18/36=0.5 (some nodes have neighbors that are each others neighbors). This shows that networks can be equal in size and density of connections but are different in the amount of clustering. Figure (c) and (d) also show the corresponding degree distributions of

the two networks in (a) and (b). From these distributions, several observations can be made such as the maximum number of connections (4 and 5 respectively) and the overall shape of the distribution (bell-shaped for both).

Small-World Networks

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 were, on average, separated by only a few acquaintances or friends (e.g., "six degrees of separation"). While the finding of very short path lengths between random pairs of nodes in a network may seem surprising, the phenomenon is well described by one of the simplest models of random graph theory such as the model by Erdös and Réyni (1960). In a Erdös and Réyni random graph with n nodes, any pair of nodes is connected by an edge with probability p. When p is sufficiently high, the whole network becomes connected and the average path-length L grows logarithmically with n, the size of the network.

Watts and Strogatz (1998) investigated several networks such as the United States power grid, the collaboration network of (international) film actors and the neural network of the worm C. Elegans. They showed that while random graphs describe very well the short path-lengths found in these networks, random graphs lack the strong local clustering (see previous section) observed in these networks: the neighbors of a node are often also each other's neighbors. Watts and Strogatz (1998) found that random graphs produce clustering coefficients orders of magnitude lower than those observed for the film actor network, the power grid and the neural network of C. Elegans. They proposed a model in which some of the connections in a lattice are randomly rewired. The local neighborhood of the lattice leads to high clustering while the long-range random connections lead to very short average path lengths.

Recently, the large-scale organization of the world-wide-web (WWW) has been analyzed with similar techniques. Based on an estimate of the whole WWW containing 8×10^8 sites, it was shown that random sites on the WWW are on average only 19 clicks away from each other (Albert, Jeong, & Barabasi, 1999). It has also been shown that the WWW shows strong local clustering (Adamic, 1999): a website typically refers to sites that also refer to each other.

Amaral, Scala, Barthélémy, and Stanley (2000) have distinguished between different classes of small-world networks by measuring the degree distribution. In one class of networks, such as C. Elegans and the

collaboration network of film actors, the degree distribution decays exponentially. This is well described by random graph theory and variants of the Watts and Strogatz model. In contrast, in the WWW, the number *k* of hyperlinks into and out of a site follows a power law distribution (Barabási & Albert, 1999):

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

In Figure 2a, a power-law and exponential degree distribution is shown (note that this plot is directly comparable to Figure 1c and d). Intuitively, a power-law distribution implies that a small but significant number of nodes are connected to a very large number of other nodes, while in an exponential distribution, such "hubs" are essentially nonexistent. The two kinds of distributions can be more easily differentiated by plotting in log-log coordinates as shown in Figure 2b). Only a power-law distribution follows a line in log-log coordinates, with slope given by the parameter γ.

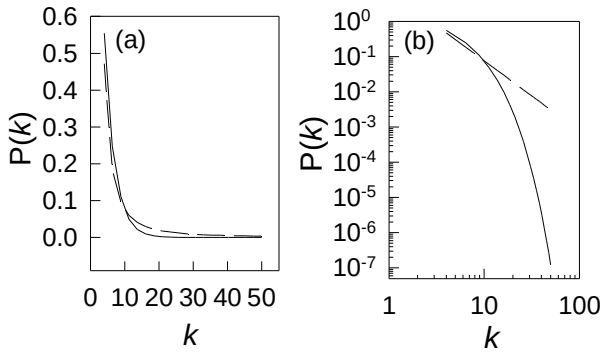

Figure 2. (a) the tails of a power-law distribution (dotted) and exponential distribution (solid). (b) log-log plot.

Barabási and Albert (1999) have argued that the finding of power laws in the degree distribution places strong constraints on the process that generates the underlying connectivity. They proposed a graph model based on two principles: (1) incremental growth and (2) preferential attachment. This leads to a scale-free distribution of degree, following a power law. Unfortunately, their model does not produce sufficiently strong clustering as observed in some real-life networks with power-law degree distributions, such as the WWW.

In the next section, we show that semantic networks have the small world properties of short path-lengths as well as strong clustering. In addition, we will show that the degree distributions closely follow power law

distributions. In a subsequent section, we will model the structure of semantic networks by a growth model inspired by the Barabási and Albert model but with specific principles for the development of semantic networks.

Analyses of Semantic Networks

We constructed semantic networks from three sources: free association, WordNet and Roget's thesaurus. Although the processes underlying these sources of semantic knowledge might be different, we will show that the resulting semantic networks are similar in their large-scale organization. For simplicity, we will construct these networks with all arcs and edges unlabeled and weighted equally.

Associative Network. A large free-association database involving more than 6000 participants was collected by Nelson, McEvoy, and Schreiber (1999). Over 5000 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 undirected network, word nodes were joined by an edge if the words were associatively related regardless of associative direction. In the directed network, any 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. Figure 3a shows a small part of the undirected semantic network highlighting one of the shortest associative path of length 4 from VOLCANO to ACHE (there are many different shortest paths between these words). In Figure 3b, all shortest associative paths from VOLCANO to ACHE are shown in the directed associative network.

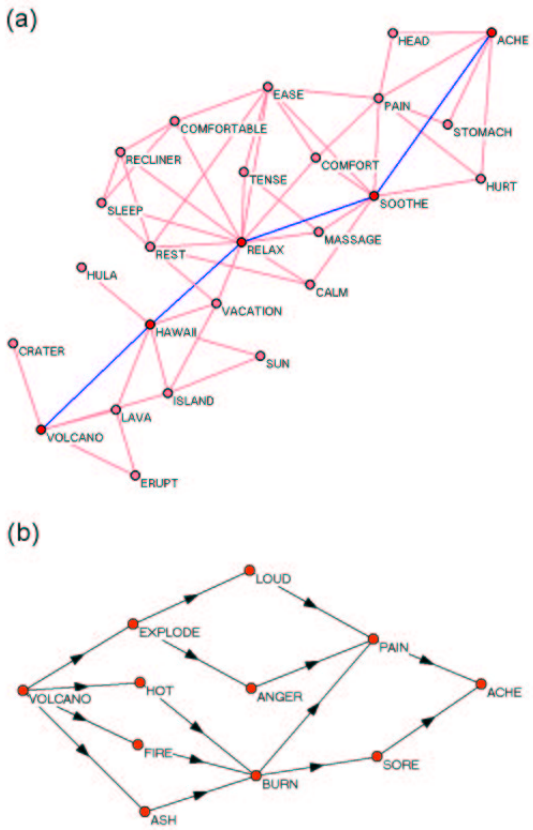

Figure 3. (a) Part of the semantic network formed by free association. The shortest path from VOLCANO to ACHE is highlighted. (b) all shortest directed paths from VOLCANO to ACHE.

Roget's Thesaurus (1911 edition). Based on the life long work of Dr. Peter Mark Roget (1779-1869), the 1911 edition includes over 29,000 words classified into 1000 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 kind of nodes, word nodes and semantic category nodes and connections are only allowed between two different kinds of nodes. In this graph, a connection is made between a word and category node when the word falls into the semantic category.

WordNet. Somewhat analogous to Roget, but inspired by modern psycholinguistic theory, WordNet was developed by George Miller and colleagues (see Fellbaum, 1998). 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. 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 relationship 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 in the connections, we will treat WordNet as an undirected graph.

Statistical Properties of Semantic Networks

We analyzed the three semantic networks for the following five properties: sparsity, connectedness, short-path lengths, local clustering, and power-law degree distributions. The statistics related to these properties are shown in Table 1, under the Data columns. We also compared the three semantic networks with random networks with the same number of nodes and edges by measuring the average path lengths (L_{random}) and amount of local clustering (C_{random}) in these random networks. In the next section, we will explain how two models (model A and model B) account for the data observed in the associative networks (as shown in the columns for model A and model B). We will now discuss each of these five properties in turn.

Sparsity. For WordNet and Roget, 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 1 lists $\langle k \rangle$ (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 5018 total number of words. The semantic networks of WordNet and Roget's thesaurus exhibit even sparser connectivity patterns.

Connectedness. Despite their sparsity, all of these semantic networks contain a single large connected components that includes the vast majority of nodes. In the undirected associative network, the whole network is connected (i.e., there is an associative path from any word to any other word when the direction of association is ignored). 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). For WordNet and Roget's thesaurus, the largest connected connected component consists of more than 99% of all words. We restricted all further analyses to these components.

Short Path-Lengths. All three networks display very short path-lengths and diameters relative to the sizes of the networks[2]. For the undirected associative network, the average path-length (L) is only 3 while the maximum path length (D) is only 5; at most 5 associative steps (regardless of direction) separate any two words in the 5,000+ word lexicon. In Table 1, L_{random} refers to the average shortest path lengths observed in random graphs with equivalent size and density. Such random graphs were created by randomly rearranging the connections in the semantic networks[3]. The short path lengths in the semantic networks are well described by random graphs with equivalent size and density, consistent with Watts & Strogatz's (1998) findings for other small-world networks.

Local Clustering. For the undirected associative network, the clustering coefficient is well above zero (see Table 1); the associates of a word tend to be also each others associates. Because the clustering coefficient does not take the direction of the connections between words into account, the clustering coefficients for the directed and undirected associative networks are identical. The amount of local clustering for the associative network and WordNet is orders of magnitude larger than can be expected from random graphs of equivalent size and density (denoted by C_{random}). Because Roget's thesaurus is a bipartite graph, the neighbors of a word node can never be each others neighbors by definition. In order to measure the local clustering in Roget's network, the bipartite graph was converted to a unipartite graph by joining word nodes by an edge if they belonged to the same semantic category. The clustering coefficient in this unipartite graphs is larger than observed in a random bipartite graph (with same the size and density as Roget's network) transformed in the same way to a unipartite graph.

Power-Law Degree Distribution. The degree distributions for the word nodes of the undirected and directed associative networks, WordNet and Roget's thesaurus are shown in Figure 4. For the directed associative network, the in-degree distribution is shown. All distributions are plotted in log-log coordinates and the best fitting power-law curves are shown with straight lines. Note that the power-law curve fits the tails of the observed degree distributions well. The high-connectivity words at the tail of the power-law distribution can be thought of as "hubs" of the semantic network. In word association, these hubs typically correspond to important general categories, such as GOOD, BAD, FOOD, LOVE, WORK, MONEY, and HOUSE. In WordNet, they correspond to polysemous verbs such as BREAK, RUN, and MAKE. The power law exponent γ (slope of the lines in Figure 4) is approximately 3 for the undirected associative network, WordNet and Roget's thesaurus (see Table 1). Similar exponents

characterize the degree distributions of many complex natural networks, and it has been argued that they reflect the processes by which these networks are formed (Barabási and Albert, 1999). For the directed associative network, the in-degree distribution shows a slight deviation from power-law and the best fitting power law exponent γ is somewhat lower than 2. The out-degree of words in the directed associative network (not shown in Figure 4) is not power-law distributed but is more similar to a normal distribution. We focused on the in-degree as opposed to out-degree because the out-degree in the word associative network is dependent on the specific details on how the word association experiment was conducted. The out-degree for a given word in word association is dependent on the number of subjects that gave associative responses to that cue as well as the number of different responses that were generated. We will discuss the differences between the in- and out-degree distributions in word association in more detail when discussing the growing network model that can explain these differences.

Table 1. Summary statistics**.**

		Undirected Associative Network			Directed Associative Network		WordNet	Roget
Variable[1]	Type	Data	Model A	Model B[2]	Data	Model B	Data	Data
n	words	5,018	5,018	5,018	5,018	5,018	122,005	29,381
	classes	-	-	-	-	-	99,642	1,000
<*k*>	words	22.0	22.0	22.0	12.7	12.0	1.6	1.7
	classes	-	-	-	-	-	4.0	49.6
L		3.04	2.82 (.032)	2.83 (.036)	4.27	4.13 (.046)	10.56	5.60
D		5	4.87 (.340)	4.86 (.360)	10	10.40 (.790)	27	10
C		.186	.186 (.005)	.183 (.006)	.186	.171 (.005)	.0265	.875
γ^{in}		3.01	2.83 (.070)	2.82 (.050)	1.79	1.88 (.062)	3.11	3.19
L_{random}		3.03	-	-	4.26	-	10.61	5.43
C_{random}		4.35E-03	-	-	4.35E-03	-	1.29E-04	.613

Note: Standard deviations of 50 simulations given between parentheses.

(1) The following notation was used: 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^{in} (power law exponent for the distribution of the number of incoming connections), L_{random} (the average shortest path length with random graph of same size and density), and C_{random} (the clustering coefficient for a random graph of same size and density)

(2) In these simulations, the directed networks from model B were converted to undirected networks.

To summarize, the word associative network, WordNet and Roget's thesaurus show five statistical features: sparsity, connectedness, short-path lengths, local clustering and power-law degree distributions. These statistical features of semantic networks reflect how the semantic connections among words are organized. The sparsity

reflects how on average words are related to only a few other words. The local clustering reflects the fact that those few connections among words are coherent and transitive: if *x* is related to *y* and *y* is related to *z*, then it is likely that *x* and *z* are related. The short path lengths and small diameters of semantic networks reflects the fact that language is very expressive and flexible; because some words have many different meanings (through polysemy or homonymy), it is easy to find a connection between any pair of words with only few degrees of separation. Finally, the power-law degree distributions show that semantic connections follow a hub structure: most words are related to only few other words while a non negligible minority of words are connected to a large number of other words.

Power laws in human language were made famous by Zipf (1965). 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 listed in 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 the power-law exponent of approximately 3 that best characterizes his distribution[4]. In our results for the in-degree distribution for the directed associative network, the power-law exponent was somewhat lower than 2. This result can be compared with the results by Skinner (1937) who measured the distribution of the number of different associative responses to a small set of cues. His plots show power-law distributions with a slope somewhat lower than 2, which is highly consistent with our results.

It has been shown that power law degree distributions can be produced by simple growth processes on networks (Barabási and Albert, 1999; see also Simon, 1955) where nodes are added to the network one at a time and connected to a small sample of existing nodes selected with probabilities proportional to their degrees. In the next section, we will show that the large-scale structures observed in semantic networks can be produced by a growth process in which a semantic network continually grows by adding new words and links between words. This model is inspired by a previously proposed growing network model of Barabási and Albert (1999), but our model is to our knowledge the only model capable of predicting all five small-world properties of the semantic networks listed in this section.

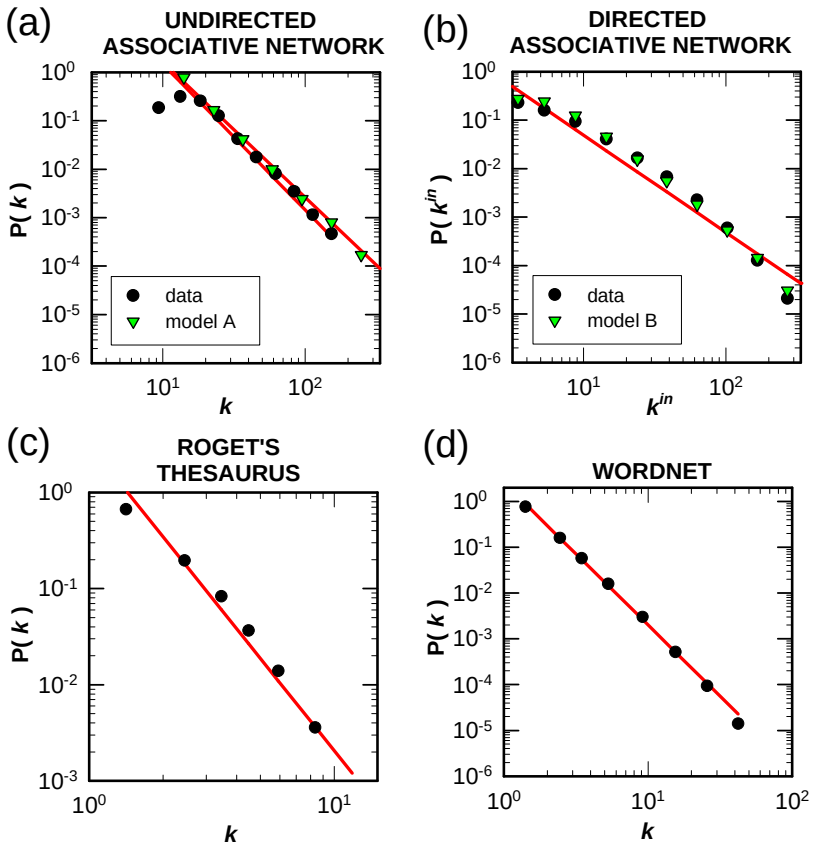

Figure 4. The degree distributions in three semantic networks, word association (a) and (b), Roget's thesaurus (c), and WordNet (d). All distributions are shown in log-log coordinates with the line showing the best fitting power law distribution. In panel (a), the undirected associative network is shown with the fit of model A: the undirected growing network model. In panel (b), the in-degree distribution of the directed associative network is shown along with the fit of model B: the directed growing network model. For Roget's thesaurus and WordNet, the degree distributions shown are for the word nodes only.

Growing Network Model

Here we introduce a simple model for the growth of semantic networks. Our aim is not to describe in detail any specific psychological mechanism, but 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 assume that semantic differentiation is the process by which semantic structures grow; when a new concept is acquired, it differentiates an existing concept by acquiring a meaning that is similar to the existing concept but different in some way that corresponds to a slightly different pattern of connectivity. We also assume that the more complex a concepts is, the more likely it is to be further differentiated. Finally, we assume that concepts that are used more frequently are more likely to be involved in the process of differentiation.

These cognitive assumptions of the growth process allows us to construct a model on the basis of a few abstract principles. 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. Nodes are also assumed to vary in utility which we assume is dependent on frequency of use. This is not an essential feature of the model, but it allows us to explore the effects of word frequency on network structure (assuming that higher frequency words have higher utility). Over time, new nodes are added to the network and probabilistically attached to existing nodes on the basis of three principles:

(1) Locality: a new node makes connections only into a local neighborhood.
(2) Size: when choosing neighborhoods, large neighborhoods are preferred over small neighborhoods.
(3) Utility: a new node connects preferentially to nodes in the neighborhood that have high utility.

The locality principle states that new links are created only into a local neighborhood. A local neighborhood is defined to be a set of nodes that have a common neighbor. By targeting connections into a local neighborhood, it is likely that the neighbors of the new node are also each others neighbors. The size of the neighborhood refers to the number of nodes in the neighborhood. The size principle states that when new words or concepts are learned, they will make connections preferably into neighborhoods that already have a large number of connections. By acquiring more connections, a neighborhood grows in size and is in turn better able to attract new connections. The utility principle states that within a neighborhood, connections are made preferentially to words or concepts that have high utility. This principle will relate the degree of nodes to utility: nodes with high utility will

tend to be nodes with high connectivity. While this principle will correlate degree with utility, we will show later how we predict independent effects of these variables.

We will now discuss two versions of a growing network model that implement these three principles. The first model (model A) produces undirected networks while in the second model (model B), the direction of connections between words or concepts is taken into account.

Model A: the Undirected Growing Network Model

Let us assume that we want grow a network with n nodes. The number of nodes at time t will be denoted by $n(t)$. 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 that targets its connections to some neighborhood i (in accordance with the locality principle). Let us define the neighborhood of a node i by the set of neighbors H_i of node i including the node i itself. Let us also define the neighborhood size of node i to be the number of neighbors of i (this equals the degree k_i).

The probability of choosing a neighborhood is based on neighborhood size:

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

where $k_i(t)$ is the degree of node i at time t. The indices in the denonimator range over all current $n(t)$ nodes in the network to normalize $P(i)$ properly. This equation implements the size principle by sampling neighborhoods with a probability proportional to their neighborhood size: large neighborhoods are preferred over small neighborhoods. The connections of the new node are targeted towards nodes within the chosen neighborhood H_i. The probability of connecting to a node j in the neighborhood of node i is based on its utility:

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

where the indices in the sum of the denominator range over all nodes in the neighborhood H_i. This choice rule implements the utility principle by favoring connections to nodes that have high utility. If all utilities are equal, then it follows that:

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

so that the connections to nodes in neighborhood *i* are chosen by a random sampling process without regard for the utilities.

The sampling from the distribution in (4) continues until *M* unique nodes within the neighborhood are chosen. The new node is then connected to the *M* chosen nodes. The process of adding nodes to the network stops when the desired number of nodes, *n* is reached. Given that each new node is linked to *M* other nodes, and that the model starts with a small fully connected network of *M* nodes, the average number of connections per node is $\langle k \rangle = 2M + M(M-1)/n \cong 2M$. Therefore, this relationship between <*k*> and *M* can be used to set *M* in order to achieve a desired average density of connections. The growth process of the model and a small resulting network with *n*=150 and *M*=2 is illustrated in Figure 5.

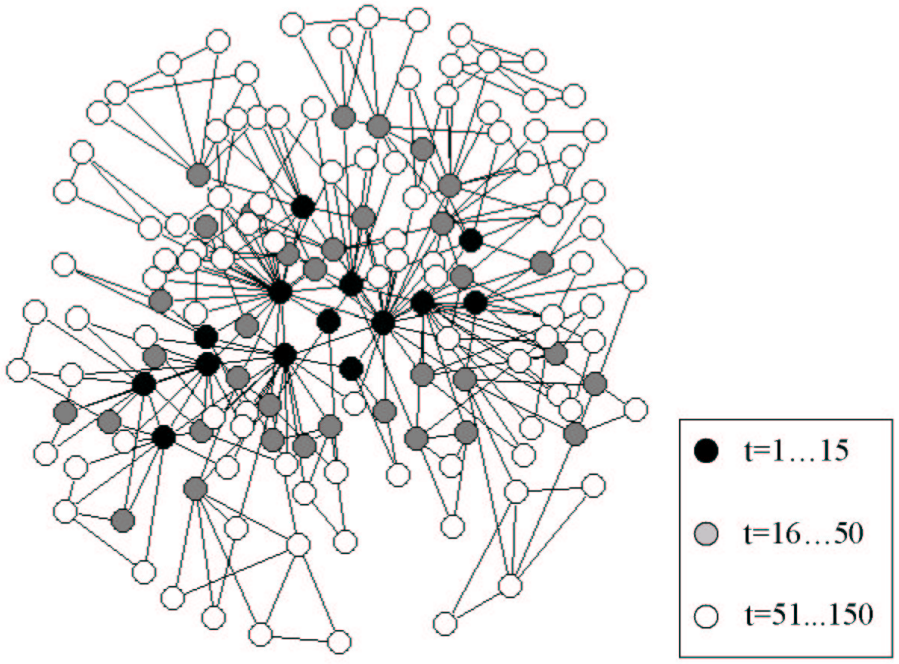

Figure 5. Illustration of the undirected growing network model with *n*=150 and *M*=2. The color of the nodes indicates the time at which the nodes were first inserted.

Model B: the Directed Growing Network Model

The growing network model for creating directed links between nodes is very similar to the model for creating undirected links. We start with a small fully connected network with *M* nodes so that each node has an arc pointed to each other node. Then, at each time step, we insert a new node. Each new node has *M* arcs that will be connected to or from nodes in a particular local neighborhood. If we define $k_i = k_i^{in} + k_i^{out}$, where k_i^{in} and k_i^{out} are the number of incoming and outcoming links for node *i* respectively, then the probability of choosing a neighborhood is also based on Equation (1): large neighborhoods are preferentially chosen over small neighborhoods. Then, the *M* nodes in the neighborhood are sampled from Equation (2): nodes with high utility in the local neighborhood are preferred over nodes with low utility. The main difference between model A and B is specifying how the links between nodes are directed. We followed a fourth principle to determine the direction of links:

4) Direction: the majority of arcs are pointed from new nodes to existing nodes in the network.

In order to implement this principle, the direction of each of the *M* arcs was determined probabilistically. For each arc, the probability that the arc will point away from the new node is α and we will assume that α>0.5 so that most arcs will point towards the existing nodes in the network.

Model Results

We applied the undirected and directed growing network model toward predicting the large-scale organization of the undirected and directed associative network (the model could also be applied to other networks, but it is computationally difficult to scale up the model to the size of networks such as WordNet and Roget's thesaurus). We set *n*=5018 to match the number of words in the network of free association. In the undirected and directed model, we set *M*=11 and *M*=12 respectively so that the resulting networks would end up with approximately the same density as the corresponding associative networks. The utility of nodes was determined by the word frequency according to: $u_i = \log(f_i + 1)$. The word frequencies f_i of the words were determined by the

Kucera & Francis (1967) word frequency count 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 according to observed word frequencies, there were no free parameters in the undirected model and there was one free parameter in the directed model. The parameter α for the directed model was set to 0.95 so that most arcs would go from a new node towards existing nodes in the network.

We evaluated the models by calculating the statistical properties of path lengths, diameter, amount og clustering and shape of degree distributions. Because both growing networks models are stochastic, results of the model vary from simulation to simulation. In Table 1, the results of the models averaged over 50 simulations are shown (standard deviations are shown between parentheses) under the columns of Model A and Model B. The networks produced by the models are characterized by path-lengths (*L*), diameters (*D*), and clustering coefficients (*C*) similar to that observed in the associative networks.

In Figure 4, panel a, it is shown that the degree distribution of the undirected growing model closely follows a power-law with an exponent of around 3, similar to that observed in the undirected associative network. In Figure 4, panel b, the in-degree distribution of the directed growing network model is shown. This distribution is very similar to the in-degree distribution observed in the directed associative network with an exponent somewhat lower than 2.

We were also interested in checking whether the directed growing network model would reproduce the results of the undirected growing network model when all directed links were converted to undirected links. We simulated model B with M=11 and a=0.95 and converted all arcs to edges at the end of each simulation. In Table 1, the results for this converted model B show that the small-world statistics are very comparable to those of model A. This result has to be expected because the process of choosing local neighborhoods and establishing connections between nodes is identical in the two models. The only difference between model A and B is that model B distinguishes between in- and out-links but the choice of direction is made only after it has already been established what connections to make. Therefore, when this distinction between in- and out-links is lost in model B, the model should exactly produce the results of model A (except for some minor fluctuations because of variability in the model).

Discussion

The growing network model produces short average path-lengths 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 the node *y*. Local clustering is produced by the model because of the locality principle: by targeting connections to a local neighborhood, the neighbors of a new node are likely to be also each others neighbors. Power-law degree distributions could be expected from the model because of the size principle: concepts that have a large number of connections are more likely to receive new connections when concepts are integrated into their neighborhoods. This is a rich-get-richer scheme leads to power-law degree distributions.

A priori, the principles behind the growing network model were expected to lead to the right qualitative results but how well the model would reproduce quantitatively the statistical properties of the word associative network was not a priori clear. While the qualitative results of short path lengths, local clustering, and power-law degree distributions are clearly the result of the two principles of locality and size, it was not clear how the principles would interact.

Power law distributions and semantic growth

Power-law degree distributions in growing network models have been observed earlier by Barabási and Albert (1999) and are related to observations on stochastic growth processes by Simon (1955). In this model, the power-law distributions are caused by the size principle: nodes that already have a high number of connections are more likely to be the target of new connections by which they gain even more connections. The result of this rich-get-richer process is the existence of a few hubs that connect to a large number of other nodes. The fact that the model produces power-law degree distributions that are similar to those observed in the three semantic networks raises the question to what extent power laws in the degree are diagnostic of growth processes. In order to increase our confidence that semantic growth is the explanatory principle for such findings, we need to show that other models without growth processes do not produce such power laws. We explore one alternative model here, based on an analysis of co-occurences of words in a large corpus of text.

Table 2. Statistical properties of networks constructed by LSA semantic spaces for different word sets and different dimensionalities.

	Words Association Words			Most Frequent Words			All Words		
	m			m			m		
Variable	50	200	400	50	200	400	50	200	400
n	4,956	4,956	4,956	4,956	4,956	4,956	92,408	92,408	92,408
ε	.614	.338	.225	.608	.332	.221	.614	.338	.225
<*k*>	22.3	22.3	22.3	22.3	22.3	22.3	219.4	201.1	209.7
L	4.83	4.02	3.70	4.77	3.86	3.62	-	-	-
D	12	9	8	11	8	7	-	-	-
C	.456	.391	.298	.454	.354	.274	-	-	-
γ	1.07	1.08	1.03	1.03	0.64	0.34	0.76	0.40	0.06

Note: empty cells in this table correspond to variables that could not be computed due to computational constraints; m = dimensionality of the vector space; ε = similarity threshold on the cosine of the angle for connecting two words

Recently, Latent Semantic Analysis (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. By analyzing the co-occurrence statistics of words across a large number of contexts in a corpus (where context is defined as a set of a few hundred words about a specific topic), the meaning of words can be represented by vectors in a high dimensional space. The semantic (dis)similarity between words can then be determined by the Euclidian distance, the inner product or the cosine of the angle between two vectors. Landauer and Dumais (1997) have shown that the local neighborhoods in semantic space successfully captures some subtle semantic relations. The question here is whether LSA captures other important semantic features such as the presence of hubs that we observe in semantic networks. To answer this, a measure for the number of local neighbors in the high dimensional space is needed to see if the degree distribution is governed by a power-law[5]. Local neighborhoods were created by thresholding the continuous measure for dissimilarity based on the angle between two vectors: two words were each others local neighbors if the cosine of the vector angle exceeded a threshold ε. By varying the threshold ε, the average number neighbors (<*k*>) could be varied.

We took the LSA vector representation based on the TASA corpus for three different word sets: (a) the words from the word associative networks (4956 of the 5018 words from word association were available in the LSA representation), (b) the 4956 most frequently occurring words in the TASA corpus, and (c) all words in the LSA vector representation (92,000+ words). The first set of words allowed us to compare the LSA network directly with the word associative network. The second set of words provided an important control for the first set to see if the results are dependent on the frequencies of the words as well as the particular set of words. The third set was important to see if some of the results with the smaller word-sets scale up. We also varied the dimensionality, m, of the LSA representation (i.e, the number of values in the LSA vectors). The dimensionality m was set at 50, 200 and 400.

Because word sets (a) and (b) contained almost the same words or the same number of words as the word associative network, we were able to compare the networks formed by the thresholding procedure directly with the word associative network. In order to do this, we found values of ε that lead to the same $<k>$ as observed in the word associative network (through a simple binary search). In Table 2, we show the size of these networks (n), the average number of connections ($<k>$), the thresholding value (ε) asi well as the values of the statistical properties. For both word-sets (a) and (b), and for the three different values of dimensionality (m), the networks showed somewhat higher path lengths (L) and diameters (D) and local clustering (C) than observed in the word associative networks (see Table 1). However, these results are qualitatively comparable to the observed properties of the word associative network. The networks are very different when the degree distributions are calculated. In Figure 6, panel (a) and (b), the degree distributions for the two word sets (a) and (b) are shown respectively for three different values of m. None of the observed distributions follow a power-law degree 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 1, we nevertheless show the values of γ to highlight how different the degree distributions for these LSA networks and semantic networks are. Finally, in order to check whether the results would hold up when more words were included in the analysis, we also repeated the simulation with word-set (c) that included all 92,000+ words of the LSA representation. We used thresholds ε that were identical to the thresholds used to construct the LSA network based on word set (a). Unfortunately, computational constraints prevented us from calculating the statistical properties of path lengths and local clustering but we were able to compute the degree distributions for these networks. Figure 6, panel (c) shows that for these networks, the distributions are not power-law distributed.

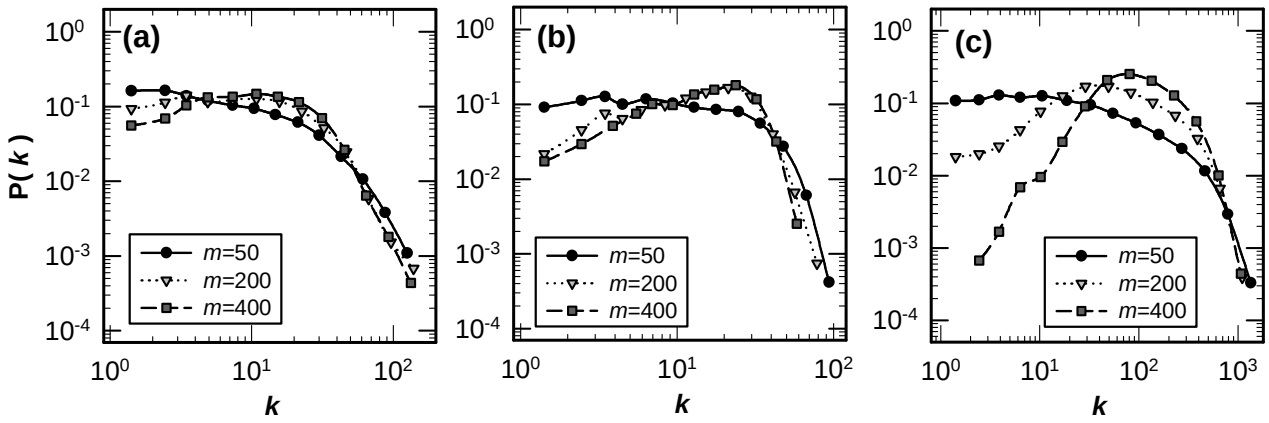

Figure 6. The degree distributions for networks based on thresholded LSA spaces for different dimensionalities. Panels (a), (b), and (c) correspond to simulations based on different word-sets: 4956 words from word association, 4956 most frequent words, and all 90,000+ words respectively.

In sum, these results suggest that the local neighborhood structure of semantic spaces in LSA does not follow power-law distributions in the number of neighbors or in other words, LSA does not produce hubs with words semantically connecting to a very large number of other words. While these results show that the particular distribution of points in LSA does not support hubs, that does not mean no static feature representation can produce these structures. For example, take a binary semantic representation (with a feature present or not present in a given word). If each feature is present in a number of words that is power law distributed and a connection between words is made whenever words share at least one feature, then the degree distribution of the resulting network will show a power law. This is one counterexample in which a non-growing semantic representation is capable of giving power law degree distributions. However, such representations also beg the question of where the power-law distribution of the number of words that have a given feature comes from. A growing network model, in contrast, provides a principled explanation for the origin of power-law degree distributions. Words that have a high connectivity tend to acquire even more connections over developmental time and with this rich-get-richer scheme, it is to be expected that some words will develop as hubs in the semantic network.

Age of acquisition, Word Frequency and Centrality

Because of the growth processes in the growing network model, the number of connections that a concept will acquire is related to the time at which the concept was first integrated into the network. The model predicts that concepts that are learned early acquire more connections over time than concepts learned late. Also, utility should interact with this effect. Concepts with high utility (i.e., high word frequency) should be better able to compete for links than concepts with low utility. This prediction is shown in Figure 7 for the simulation of the undirected growing network model reported in the last section (the directed growing network shows similar results). The degree of nodes is shown as a function of the time at which the node was inserted and word frequency (binned in three ranges). The simulation shows that early acquired nodes and nodes with higher utility end up with higher connectivity. Also, the effect of the time of acquisition is more pronounced for nodes with high utility than nodes with low utility.

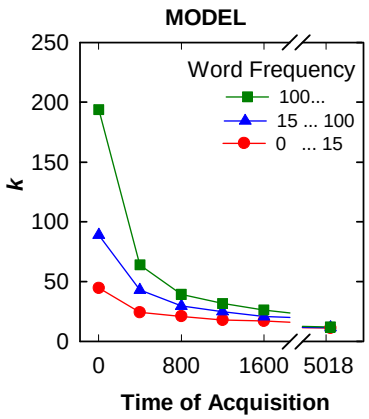

Figure 7. Degree of nodes as a function of the time of acquisition and word frequency in the model.

In order to test this prediction of the model, we consulted age of acquisition norms that are available for small sets of words. Gilhooly and Logie (1980) collected ratings in which adults estimated the age at which they thought they first learned the word on a rating scale (these ratings were converted to numbers between 100 to 700 with a rating of 700 corresponding to a word acquired very late in life). For these norms, we took the average ratings for each word as a crude measure for the time at which a word was acquired in a semantic network. We also

consulted age of acquisition norms from Morrison, Chappell, and Ellis (1997) who in a cross-sectional study estimated the age at which 75% of children could successfully name the object depicted by a picture. While these norms provide a more objective measurement for the age of acquisition, these norms were only available for a small set of words.

In Figure 8, for three semantic networks, the relation is shown between the degree of a word, and its mean acquisition rating/time. We separated the words into three different frequency bins to show interactions between age of acquisition and word frequency. For the associative network as well as WordNet and Roget's thesaurus, the results are qualititatively similar to that predicted by the model (that was simulated to have the same size and density as the associative network). For both the adult rating norms and the picture naming norms, early acquired words have more dense connections than late acquired words according to each of the three semantic networks. Also, high frequency words show higher connectivities than low frequency words and the effect of age of acquisition on degree is higher for high frequency words than low frequency words, an effect similar to that predicted by the model.

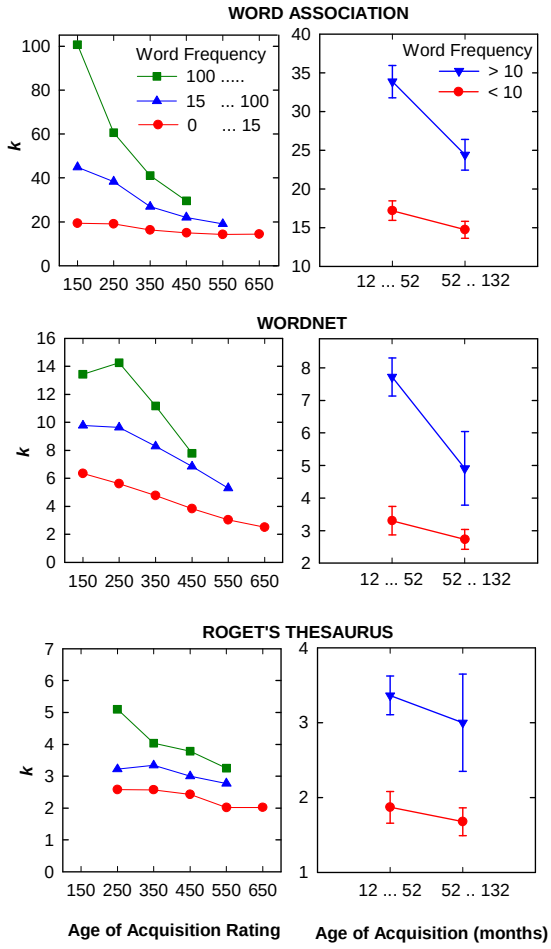

Figure 8. The relation between degree and age of acquisition as measured by adult ratings (left panels) and the average age at which children can name pictures (right panels). Right panels include standard error bars around the means.

The results relating number of semantic connections to age of acquisition and word frequency has potentially important consequences for research in which the behavioral effects of age of acquisition and word frequency are investigated. For example, early acquired words show short naming latencies (e.g., Carroll & White, 1973) and lexical decision latencies (e.g., Turner, Valentine, & Ellis). While it has been suggested that age of acquisition affects mainly the speech output system (Ellis & Lambon Ralph, in press), it has been shown that age of acquisition also effects non-phonological tasks involving face recognition and semantic tasks such as word association and semantic categorization (e.g., Brysbaert, Van Wijnedaele, & De Deyne, 2000). There has also been some debate on whether age of acquisition and word frequency exert their influence on behavioral tasks independently (e.g., Turner et al., 1998) or that age of acquisition is mere cumulative frequency (Lewis, Gerhand, & Ellis, 2001) because high frequency words are likely to be acquired earlier than low frequency words.

In our model, we relate the effects of age of acquisition and word frequency to the number of semantic connections developed over time. A different way of stating that early acquired words have more semantic connections is that early acquired words have higher *degree centrality*. The degree is not the only way to define centrality. Another way to measure centrality is by the computing the eigenvector of the adjacency matrix with the largest eigenvalue. Words with high eigenvector centrality would be words that are highly connected to other words that are themselves highly connected. The eigenvector centrality has been used to measure conceptual coherence (Sloman, Love, & Ahn, 1998) and to find authoritative websites on the WWW by the search engine Google (Brin & Lawrence, 1998). In Google, the idea is to rank websites partly on the basis of eigenvector centrality because information is more likely to be found in websites that are linked to by other websites (that are themselves linked to by other websites).

One reason why early acquired words might be named faster and lead to fast lexical decision times is because early acquired words are more central in an underlying semantic network. Just as the Google search engine orders websites based on their centrality, word production and lexical retrieval systems might be biased toward production or retrieval of words that are more central. Because the centrality of a word is a structural variable of the

model that is effected by environmental variables such as age of acquisition and word frequency, this variable might provide a principled explanations for the behavioral effects of age of acquisition and word frequency.

We applied correlational analyses on degree centrality, age of acquisition (both rating and picture naming norms) and word frequency on naming and lexical decision on two databases; a new naming latency database by Spieler and Brand (personal communication) for 796 multi-syllabic words and a large lexical decision latency database from Balota, Cortese, and Pilotti (1999) for 2905 words. Degree centrality was logarithmically transformed to avoid the extreme skew in the degree distribution. In Table 2, we show the correlations between degree centrality, age of acquisition (referred to as AoA) and word frequency for the three semantic networks and naming and lexical decision latencies. The results show that centrality is moderately negatively correlated with naming and lexical decision latencies; words that are semantically more central have lower naming and lexical decision latencies. It also confirms well known results that age of acquisition correlates positively with naming and lexical decision latencies and that word frequency (using the norms of Kucera and Francis) correlates negatively with naming and lexical decision latencies. In other words, high frequency words/ early acquired words are named faster and are quicker to be identified as words than low frequency words/ late acquired words. Interestingly, when the effects of word frequency or age of acquisition are partialled out of the correlational analyses, the correlations between centrality and lexical decision latencies remain largely significant. When both word frequency and age of acquisition are partialled out the correlational analyses, the correlations become very low but remain statistically significant for lexical decision.

Table 2. Correlations of word frequency, age of acquisition and degree centrality in word association, WordNet, and Roget's thesaurus with Naming and Lexical Decision Latencies.

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

Note: R=correlation; n=number of observations; * is placed next to significant correlations (p<.05)

General Discussion

The model explains the relationship between age of acquisition and degree of words as an *order effect*: when word A is learned before word B, A will (on average) acquire more connections than B. A weakness of this model is that it does not explain *why* certain words are acquired earlier than other words. In this model, degree of connectivity is causally dependent on the age of acquisition. Reversing this causal connection suggests an obvious alternative explanation: age of acquisition is causally dependent on the degree of connectivity. In such an explanation, words that have a high number of connections are learned earlier than words with a low number of connections (the number of connections of a word is an environmental variable that somehow the learner can utilize in a learning situation). We do not claim that the growing network model is capable of modeling all causal

connections between the variables of degree, age of acquisition and word frequency. One of the contributions of this research is showing how the degree might be causally dependent on age of acquisition which does not preclude a causal connection in the opposite direction.

The growing network model is not the only model that can explain the effects of age of acquisition. Ellis and Lambon Ralph (in press) have shown how a connectionist model in which words that are learned early are learned better than word that are learned late. The model was trained to develop a distributed representation for the input patterns and was initially only trained on patterns corresponding to early learned words. As training progressed, new words were interleaved with the old patterns already in the training set. The early trained items induced a distributed representation that later trained items could not easily change. In this explanation, age of acquisition effects occur because the model loses the ability to encode new patterns effectively over time. This model provides a simple explanation for age of acquisition effects in a general learning system. The similarity with the growing network model account for age of acquisition effects is that early learned patterns are more centrally represented in the distributed representation than late learned patterns. The difference between the connectionist model and the growing network model is that the latter is specifically tied to semantic growth and the large-scale structure of semantic networks. The growing network model also produces a power law degree distribution which is a signature of semantic growth.

We have tried to show that power law degree distributions can be understood by a semantic growth process and that non-growing semantic representations such as LSA do not seem to produce such distributions (or produce such distributions only when explicitly building in power laws in the distribution of feature values). Our analyses with the semantic space developed by LSA show that it is very unlikely for any word to be a hub (to have a very large number of local neighbors). Tversky and Hutchinson (1986) came to a very similar conclusion about the problems of semantic spaces to capture the local neighborhood structure of semantic relations (especially in low-dimensional spaces). They pointed out that many conceptual domains have a hierarchical structure that puts strong constraints on our perception of similarity between words. For example, in a similarity rating dataset involving many fruit words including the word fruit itself, the word fruit was rated as the most similar among 18 fruit words, making the word fruit the closest neighbor to 18 fruit words (this is not surprising because the word fruit is the category word). When a multidimensional scaling algorithm was applied to place the fruit words in a two dimensional semantic space, the word fruit was the closest neighbor to only 2 words (using Euclidian distance as a

measure for dissimilarity). Therefore, the nearest neighbor relations for the word fruit does not seem to be captured well by the two dimensional space. In theory, in a two dimensional space, a point can be the closest neighbor of at most 5 points (REF??), so even a specific placement of points in the two dimensional space could not have resulted in the nearest neighbor relationship observed for the word fruit. Of course, in some high dimensional space, it could be possible to achieve these nearest neighbor relationships for specific words. This requires a theory for how words can be placed in a high dimensional semantic space. LSA is one such a theory and we showed that the particular placement of points in the semantic space does not seem to capture the degree distributions observed in semantic networks.

Conclusion

We found that three semantic networks constructed by different means are sparse, exhibit very short average path-lengths and strong local clustering. As in the WWW, the number of neighbors follows a power law, suggesting a hub-like structure for knowledge organization. Similar power-law distributions were observed in a growing network model in which concepts are incrementally added and integrated into the existing network. The model's prediction that early acquired concepts end up with more rich connectivity was confirmed with age of acquisition norms.

Footnotes

1. We implemented "Dijkstra with heaps" (Cormen, Leiserson, & Rivest, 1990) as an efficient algorithm to find the shortest paths between a given node and all other nodes. The Matlab code for this algorithm is available from the authors.

2. Unlike for the word associative networks, where *L* and *D* were calculated on the basis of the path lengths between all word pairs, for the large networks of WordNet and Roget's thesaurus, *L* and *D* were based on the path lengths between all word pairs of a sample of 10,000 words.

3. 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.

4. Zipf 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) provides some simple analytic tools by which the slope b=.466 in this Zipf plot can be converted to γ=3.15, the slope of the corresponding probability distribution in log-log coordinates.

5. In general, when points are uniformly distributed in a hypercube and points connect when balls of size ε centered at the points overlap, the number of neighbors will be Poisson distributed; such a distribution has an exponential tail and will not show linearity in log-log coordinates. Because we do not know what the distribution of points is in LSA space and what exactly the difference is between Euclidian distance, inner product and cosine of the angle measures for (dis)similarity, we simulated different thresholding using these three (dis)similarity measures.

Acknowledgments

We are very thankful to Tom Landauer and Darrell Laham for providing us with the TASA corpus. We also thank Doug Nelson and Tom Griffiths for discussions that helped shape this research.

References

Adamic, L.A. (1999). The small-world web. Proceedings of ECDL'99, LNCS 1696, Springer.

Adamic, L.A. (2000). Zipf, Power-laws, and Pareto - a ranking tutorial.
http://www.parc.xerox.com/istl/groups/iea/papers/ranking/ranking.html

Albert, R., Jeong, H., & Barabasi, A.L. (1999). Diameter of the world wide web, *Nature, 401,* 130-131.

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.

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). Emergence of scaling in random networks. *Science, 286*, 509-512.

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

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

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

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, MD: The Johns Hopkins Press.

Ellis, A.W., & Lambon Ralph, M.A. (in press). 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*.

Erdös, P., & Réyni, A. (1960). On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences, 5*, 7-61.

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

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

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

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.

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.

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

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. *http://www.usf.edu/FreeAssociation*.

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*, 71-6.

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

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.

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

Watts, D.J. (1999). *Small Worlds: The Dynamics of Networks between Order and Randomness*. Princeton University Press.

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