Small Worlds in Semantic Networks

Mark Steyvers (msteyver@psych.stanford.edu)
Josh Tenenbaum (jbt@psych.stanford.edu)
Department of Psychology, Stanford University,
Stanford, CA 94305-2130

Abstract

We present graph-theoretic analyses of three types of semantic networks: associative networks, WordNet, and Roget's thesaurus. We show that they have a small-world structure: they are sparse, exhibit short average path-lengths between words as well as strong local clustering. In addition, the distributions of the number of connections follow power laws that suggests a hub structure similar to the WWW. We propose a network model that over time acquires new concepts and integrates them into the existing network. By preferentially attaching new concepts to well connected concepts and its neigbors, the model captures the small-world characteristics of semantic networks and also exhibits power-law distributions in the number of connections. With age of acquisition norms for adults and children, we confirm the model's prediction that concepts acquired early on are concepts with rich connectivity.

Introduction

Semantic networks are useful tools as representations for semantic knowledge and inference systems. Historically, semantic networks bring to mind the classic network theory of Collins and Quillian (1969) in which concepts are represented as hierarchies of interconnected nodes with nodes linked to certain attributes.

In this research, the goal is to understand the large-scale organization of semantic networks. By applying graph-theoretic analyses, the large scale structure of semantic networks can be specified 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 constructed by different means. We then argue that these regularities place strong constraints on the developmental principles by which connections between words are formed, and we propose a simple framework for modeling the acquisition and decay of semantic knowledge which is consistent with these constraints.

In particular, we will show that the large-scale organization of semantic networks reveals a small-world structure that is very similar to the structure of several other real-life networks such as the neural network of the worm C. elegans, the collaboration network of film actors and the WWW. In addition, we will propose a new network model that mimics 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 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 important*. The model predicts that concepts that are learned early in life should show higher connectivity and should be more resistant to memory disorders. We will show how this growth model can predict effects related to age of acquisition and how it might be utilized in models for semantic memory disorders such as semantic dementia.

Small-World Networks

Interest in the small-world phenomenon started by classic experiments in real life social networks Stanley Milgram (1967) that suggested that any random pair of people are separated by an average of only six degrees. The finding that random pairs of nodes in a network are separated by very short path-lengths is well described by random graph theory by Erdös and Réyni (1960). In a 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 (i.e., there is a path from any node to any other node) and the average path-length grows logarithmically with *n*, the size of the network.

Watts and Strogatz (1998) investigated several networks such as the 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 observed in these networks: the neighbors of a node are often also each other's neighbors. For each node *i*, they calculated the clustering coefficient, C_i by dividing the number of neighbors that were also each other's neighbors by the total possible number of neighbors' connections. They found that random graphs produce average 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 WWW has been analyzed with similar techniques. Based on an estimate of the whole WWW containing 8 x 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 of networks (the degree of a node is the number of neighbors a node has). In one class of networks, such as C. Elegans and the collaboration network of filmactors, 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 distribution of number of hyperlinks from and toward a site follows a power law (Barabási & Albert, 1999). In other words, a few sites refer to and are referred from a very large number of other sites. For the WWW, the probability of observing a degree *k* can be described by:

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

In Figure 1a, a power-law distribution shows a heavier tail than an exponential distribution. A power law can be more easily differentiated from an exponential distribution by plotting the distribution in log-log coordinates: in Figure 1b, only the power-law distribution follows a line in log-log coordinates. The parameter γ in the power-law distribution determines the slope of the line in log-log coordinates.

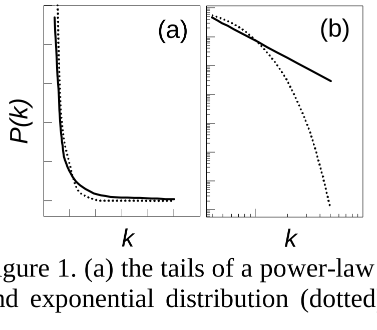

Figure 1. (a) the tails of a power-law distribution (solid) and exponential distribution (dotted). (b) log-log plot.

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 real-life networks with power-law degree distribitions, such as the WWW. In a later section, we will introduce a variant of the Barabási and Albert that does produce strong local clustering.

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 as undirected graphs with all 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"). The network was constructed by joining associatively related words by an edge. Figure 2 shows a small part of the semantic network highlighting the shortest associative path from VOLCANO to ACHE.

WordNet. Inspired by 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. Word meaning nodes are connected by relations such as hypernymy (MAPLE is a TREE) and meronymy (BIRD has a BEAK).

Roget's Thesaurus (1911 edition). Based on the life long work of Dr. Peter Mark Roget (1779-1869), the first system of verbal classification was developed. The 1911 edition includes over 29,000 words classified in

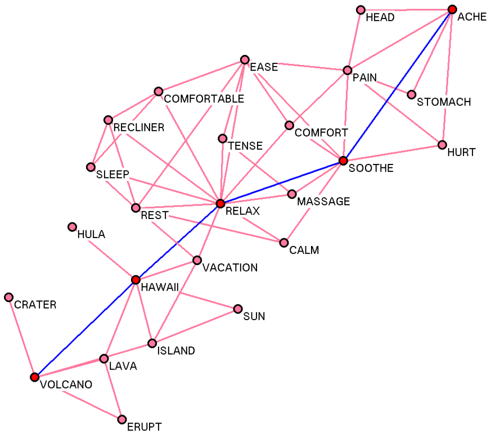

Figure 2. Part of the semantic network formed by free association. The shortest path from VOLCANO to ACHE is highlighted.

1000 semantic categories (ignoring various levels of classification). A bipartite graph was created by joining a word node and semantic category node by an edge if the word was part of the semantic category.

The summary statistics for the three semantic networks are shown in Table 1. The following notation was used: *n* (number of nodes), <*k*> (average of *k*, the degree of a node), *L* (average path-length between word nodes), L_{random} (average path-length between nodes of random graph with same size and density), *D* (diameter: maximum path-length between words), *C* (average clustering- coefficient), C_{random} (average clustering-coefficient for random graph of same size of density), and γ (slope in power-law distribution). The three semantic networks were analyzed for the following five properties:

Table 1. Summary statistics.

					Simulation of WA		
Variable	Type	WA	WordNet	Roget	α=1	α=2.5	α=4
n	word	5,018	122,005	29,381	5,018	5,018	5,018
	meaning	-	99,642	1,000	-	-	-
<*k*>	word	22.0	1.6	1.7	22.0	22.0	22.0
	meaning	-	4.0	49.6	-	-	-
L		3.04	10.56	5.60	2.98	2.90	2.86
L_{random}		3.03	10.61	5.43	-	-	-
D		5	27	10	5	5	5
C		0.186	0.029	0.875	0.018	0.187	0.281
C_{random}		0.004	0.000	0.613	-	-	-
γ		2.92	2.95	3.25	2.88	2.75	2.68

Note: WA=word association

Sparsity. All three semantic networks are sparse: on average, a node is connected to only a very small percentage of other nodes.

Connectedness. Despite the sparsity, the network based on free association forms one large connected component: from any word, any other word can be reached by some associative path. For WordNet and Roget's thesaurus, the largest connected components contained more than 99% of the words. The analyses were restricted to these components.

Path-Lengths. All three networks displayed very short path-lengths relative to the sizes of the networks. For word association for example, average path-length is only 3 while the maximum path-length is only 5; at most 5 associative steps separate any two words in the 5,000+ lexicon. The short path-lengths are well described by random graphs with equivalent size and density.

Local Clustering. For all three networks, the clustering coefficient[1] shows values well above zero. For the associative network and WordNet, the clustering is orders of magnitude larger than can be expected from random graphs of equivalent size and density.

Degree Distribution. The degree distributions for the word nodes are shown in Figure 3 with the best fitting power-law curves. Note that the power-law curve fits the tails of the observed degree distributions well (the front end of the distribution for the association network was not used in the estimation of γ). Therefore, some words have a very large connectivity and they could be described as the "hubs" of the semantic network. In word association, these hubs are words such as GOOD, BAD, FOOD, LOVE, WORK, MONEY, and HOUSE.

Growing Network Model

We introduce a growing network model in which knowledge is represented as a semantic network: the nodes represent concepts while the links between nodes represent different relationships between concepts. The model is based on the following three principles:

Growth: over time, the model acquires new concepts and links these concepts to existing concepts.

Preferential attachment to highly connected concepts: new concepts preferentially attach to highly connected concepts while preserving local neighborhoods.

[1] For each word node, a clustering coefficient C_i was calculated as the fraction of the number of neighbors of node *i* that were also each other's neighbors and the total possible number of neighbors' connections. Table 1 lists the average *C* over all word nodes where word nodes with only one neighbor were excluded to avoid artificial inflation of *C*. By definition, *C*=0 for a bipartite graph so for Roget's thesaurus, *C* and C_{random} were computed on a converted network in which words were joined by edges if they appeared in the same category.

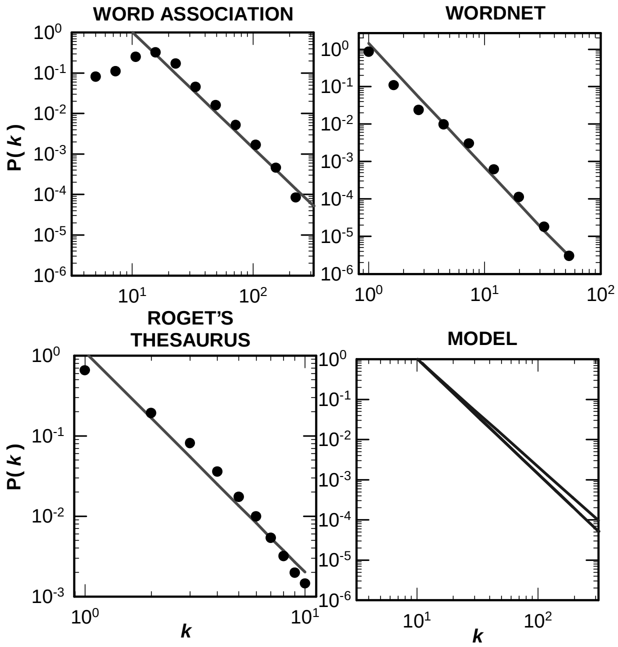

Figure 3. The log-log degree distributions for the word nodes in the three semantic networks and the model. For the model, the degree distribution is shown for α=1 and α=4 (top line).

Preferential attachment to concepts high in utility[2]: concepts vary in their utility and connections to high utility concepts should be preferred. We will assume that high frequency words have higher utility than low frequency words.

We start with a fully connected network of *M* nodes. At each timestep, a new node with *M* links is added to the network. The new node is attached to the network by a combination of degree of connectivity, utility, and local neighborhood structure. Let k_i and u_i be the degree and utility of node *i* respectively. The probability of attaching the first link of the new node to node *i* is:

$$P(i) = (u_i k_i) \Big/ \sum_j (u_j k_j)$$

This process favors choosing highly connected nodes and let *x* be the node chosen according to this process. The new node is then attached to *M*-1 other nodes, preferentially to the local neighborhood of *x* motivated by the assumption that a new concept should be related to other concepts that are themselves semantically related. Let *L*(*i*,*j*) be the path-length from *i* to *j*. We first calculate l_i which is inversely related to the path-length of *i* to *x*:

$$l_i = e^{-L(i,x)^{\alpha}}$$

The parameter α is a scale parameter. The probability of attaching each of the *M*-1 remaining links to node i is:

$$P(i) = (u_i l_i) \Big/ \sum_j (u_j l_j)$$

The process of adding nodes to the network stops when the desired number of nodes, *n* is reached. Because of the second linkage process, strong local clustering in the network can be obtained depending on the value of α. In Figure 4, the difference is illustrated between the Barabási and Albert (1999) model that incorporates only the first linkage process and our model.

We applied this model toward predicting the large-scale organization of the semantic network based on free association. We set *n*=5018 and *M*=11 so that the network would end up with the same size and density as the associative network. This leaves us with a single free parameter α. In order to check how much the results depend on α, three different values were explored: 1, 2.5, and 4. The utility variable was used to simulate differences in word frequency. For each new node, with probability 1/3, *u* was set to 1, 2, or 4. This arbitrarily divided the nodes into three levels of utility. The results of the model are shown in Table 1 and Figure 3. The network produced by the model is characterized by short path-lengths, and a power-law degree distribution similar to that observed in associative networks. These characteristics were relatively uninfluenced by different values of α. As expected, the parameter α did influence the amount of local clustering: with α=2.5, the amount of clustering in the model and the associative network was very similar.

Age of acquisition and Connectivity

An interesting prediction of the model is that concepts that are learned early in the network acquire more connections over time than concepts learned late. This prediction follows directly from the principles of incremental growth and preferential attachment. Also, utility should interact with this effect. Concepts with high utility (e.g., high word frequency) should be better able to compete for links than concepts with low utility. This prediction is shown in Figure 5 for the simulation reported in the last section. The degree is shown for words with different utilities acquired at different times: early acquired words and words with higher utility end up with higher connectivity. Also, differences in utility are more pronounced for words that are acquired early in the model. The prediction of the model was tested by consulting age of acquisition norms. Gilhooly and Logie (1980) and Bird, Franklin, and Howard (in press) collected ratings in which adults estimated the age at which they thought they first learned the word. We combined the ratings from these two databases. More

[2] In Bianconi and Barabási (submitted), a similar fitness variable was introduced so that late acquired nodes can compete successfully with earlier acquired nodes when they have sufficiently high fitness.

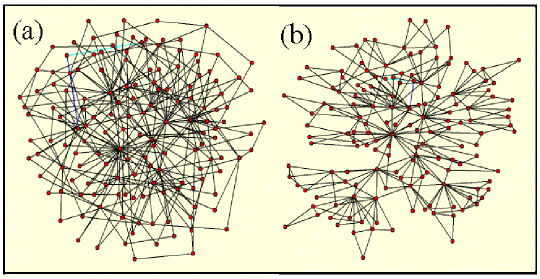

Figure 4. (a) the Barabási and Albert model with M=2, n=150 (b) our network model with M=2, n=150, α=2.

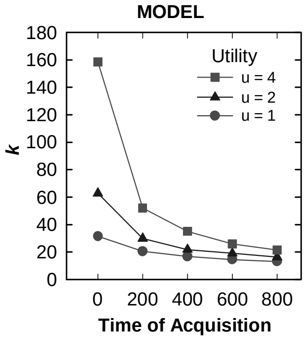

Figure 5. Degree of words as a function of the time of acquisition and the utility in the model.

objective norms are available from Morrison, Chappell, and Ellis (1997). In a cross-sectional study, they estimated the age at which 75% of children could successfully name the object depicted by a picture. In Figure 6, the relation is shown for the age of acquisition and the degree of words for different word frequencies. 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.

These results are potentially important to explain results in the literature because age of acquisition effects performance in a variety of tasks. It might be that the differences in the density of connectivity might provide greater explanatory power in describing effects of age of acquisition than age of acquisition itself. For example, early acquired words show short naming latencies (e.g., Carroll & White, 1973). While it has been suggested that age-of-acquisition effects mainly the speech output system (Lambon Ralph, Graham Ellis, & Hodges, 1998; Ellis & Lambon Ralph, in press), it has been shown that AoA 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). One factor that might explain this effect is the connectivity difference between words with different ages of acquisition. This simple explanation can be contrasted with connectionist accounts for age of acquisition effects. Ellis and Lambon Ralph (in press) have shown how a connectionist model can produce an advantage for early learned words. 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, they started interleaving new patterns with the old patterns already in the training set. They found that 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. We would like to propose that one component to the naming latency advantage for early acquired words could be that *early acquired words are more central* in an underlying semantic network where we can define centrality by the amount of connectivity. Another way to measure

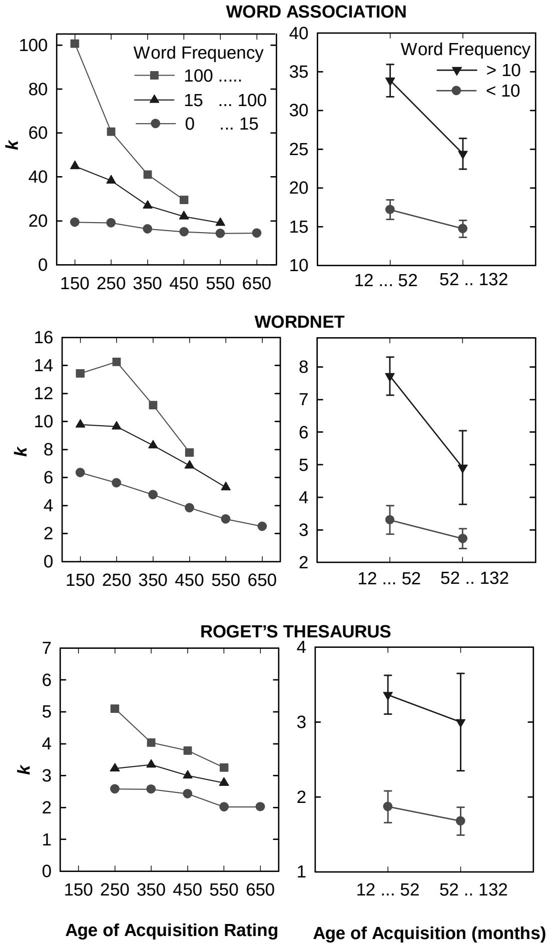

Figure 6. The relation between degree and age of acquisition as measured by adult ratings (left panels) and the average age at which children can name pictures (right panels). Right panels include standard error bars around the means.

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 highly connected. The eigenvector centrality has been used to find for example authoritative websites on the WWW by the search engine Google (Brin & Lawrence, 1998) and to measure conceptual coherence (Sloman, Love, & Ahn, 1998).

Another example where the relation between age of acquisition and centrality can be used to understand the effects of age of acquisition is semantic dementia. Patients with sematic dementia show a loss of conceptual knowledge while still retaining good short-term and episodic memory (e.g., Lambon Ralph, Graham, Ellis, & Hodges, 1998). As the disorder progresses, naming of common objects becomes increasingly less accurate. The results also suggest that early learned objects are more resistant to the naming deficit than later learned objects. In the connectionist model of Ellis and Lambon Ralph, early acquired patterns are better represented in the distributed representation so that lesions to the network tend to disrupt the representation of later acquired patterns more than early acquired patterns. Our model suggests an alternative explanation based on the underlying connectivity in a semantic network. Since early acquired concepts are more central in the semantic network (i.e., they are more highly connected), diffuse damage to the connections would tend to disrupt the representation for late acquired concepts more than for early acquired concepts.

Discussion

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 at least partially confirmed with age of acquisition norms. While the model suggests a causal direction in which *any* concept can end up with rich connectivity as long as it is learned early, an alternative is that concepts that have the potential for rich connectivity are exactly the concepts that are learned at an early age. We are currently investigating how to distinguish between these causal and non-causal interpretations of the relation between age of acquisition and connectivity.

Acknowledgments

We would like to acknowledge Doug Nelson who pointed out the small-world literature to the first author.

References

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

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.

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.

Bianconi, G., & Barabási, A.L. (submitted). Competition and multiscaling in evolving networks.

Bird, H., Franklin, S. & Howard, D. (in press). Age of acquisition and imageability ratings for a large set of words, including verbs and function words. *Behavior Research Methods, Instruments, and Computers*.

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.

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.

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

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*.

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

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