Semantic Network Connectivity is Related to Vocabulary Growth Rate in Children

Nicole M. Beckage (nbeckage@indiana.edu)
Linda B. Smith (smith4@indiana.edu)
Department of Cognitive Science, Indiana University, 819 Eigenmann, 1910 E. Tenth Street
Bloomington, IN 47406 USA

Thomas Hills (thomas.hills@unibas.ch)
Department of Psychology, University of Basel, Missionsstrasse 64A
4055 Basel, Switzerland

Abstract

Adult semantic networks show small-world structural properties that are believed to support language processing and word retrieval. The focus of this paper is to understand when these properties emerge in lexical development. We believe that they relate to the rate of word acquisition and vocabulary size. To address this, we examine the connectivity patterns of semantic networks of individual children and compare children on faster and slower vocabulary growth trajectories. The results show that small-world properties emerge early. However, children on slower growth trajectories, who are at risk for significant language delay, do not show these properties. The differences between typical and these so-called "late-talkers" persist, even when vocabulary size is equated. Late talkers' vocabularies are not only acquired later, but also less cohesively, a fact that may relate to future language processing difficulties for these children. In brief, the results suggest that properties of network connectivity may play a role in early lexical development.

Keywords: semantic networks, language acquisition, corpus analyses, late talkers

Words connected to other words

Words exist in a sea of other words. The semantic relations among these words play an explanatory role in language comprehension and processing (e.g., Lund & Burgess, 1996; Jones & Mewhort, 2007). These relations are often studied in terms of semantic networks (Collins & Quillian, 1969; Steyvers & Tenenbaum, 2005). Recent advances in graph theory reveal that adult semantic networks have properties that may be important to language processing, and potentially also to word learning.

Graph theory, or network analysis, can be applied to any structure that consists of nodes connected to each other through links or edges. For example, nodes might be cities and links might be roads; or nodes might be proteins and links might be the molecules that bind with and activate them; or, nodes might be words and the links indices of semantic connectedness such as association strength or co-occurrence.

The semantic networks may be built from various sources, including corpora collected from written or spoken language, free association data, and hand-coded collections of words (e.g., Steyvers & Tenenbaum, 2005; Hills et al., 2009b). As such, they describe the typical mature language user. These mature semantic networks exhibit what is known as small world properties (see Steyvers & Tenenbaum, 2005; Hills et al., 2009a). Small world characteristics allow for local structure but global access. In a network with small world characteristics, there are often clusters of densely connected nodes. The connections between the nodes of a cluster tend to connect to nodes in the same cluster. This contributes to the high local structure. However, there are also a few nodes in these dense clusters that have connections to nodes in other potentially distant clusters. This is the global access that allows easy movement and transition from one cluster to another. Quantitatively, these features are apparent in a high clustering coefficient (a measure of local connectivity) and an average geodesic distance (the shortest path between two nodes) on par with a random network of similar size and connection density. To aid in exposition, these properties are illustrated in Figure 1. Small-world properties are believed to support efficient processing, word retrieval, categorization and robustness to damage and deletion (Hills et al., 2009a; Griffiths, Steyvers & Firl, 2007, Steyvers & Tenenbaum, 2005).

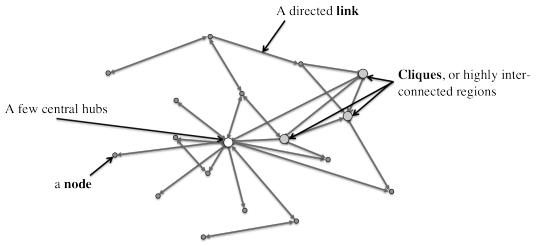


Figure 1: Characteristics of small world structure.

Although it is known that adult semantic networks have small world characteristics, only a few studies have addressed their development and the role of network structure in language acquisition (e.g. Vitevitch 2008, Hills et al., 2009a, Hills et al., 2009b). Here, for the first time, we examine the network structures of the vocabularies of individual children at different points in development. We ask whether small-world properties are dependent on acquiring some number of English words and whether, for any vocabulary size, some children's networks might show more robust connectivity patterns than other children's

networks. Is network connectivity a general fact about the structure of language, or can we show that it is a relevant property at the scale of an individual? Finally, is the connectivity pattern for individual children related to rate of vocabulary growth?

To these ends, we examine the connectivity within the semantic networks of individual children who –by normative standards –are on a path of typical development and children who are on a slower path and one that past research shows is predictive of later language difficulties (e.g., Thal et al., 1997; Bishop & Leonard, 2000; Heilmann, et al, 2005).

Trajectories of Early Vocabulary Growth

Early word learning is first slow and then accelerates (Bloom 2000; Dale & Fenson, 1996), a fact that suggests that already learned words help new word acquisition (see Mitchell & McMurray, 2009). Vocabulary size at any point in development is thus a predictor of future vocabulary growth rates (Dupuy, 1974; Raven, 1948; Bates et al., 1992; Fenson et al., 1993; Thal et al., 1997). Figure 2 illustrates the normative vocabulary size as a function of age for children at the 50th percentile and the 20th percentile (Fenson et al, 1993; see also Dale & Fenson, 1996). Percentile is calculated by considering a child's age, number of words in their productive vocabulary and gender.

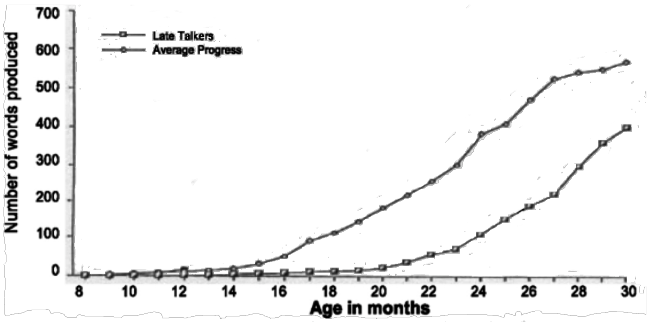


Figure 2:Trajectories of early vocabulary growth, representing children in the 50th percentile and children in the 20th percentile (drawn from Fenson et al, 1993).

The trajectory at the bottom–for children whose vocabulary size falls at or below the 20th percentile of children their same age–has attracted considerable attention in the study of early word learning. Many of these children not only stay on this slower trajectory, but about half go on to have serious deficiencies in language processing and even those who might seem to "catch up" often have measurable difficulties in language tasks (including reading) when they reach school age (e.g Anderson & Freebody, 1981; Bishop & Leonard, 2000; Thal et al., 1997; Moyle et al, 2007). Moreover, early as well as later in development, these children show retrieval errors and word-finding difficulties (Bishop & Leonard, 2000).

Accordingly, we ask how vocabulary size in young children relates to the structure of semantic relations within those vocabularies and whether this structure is related to individual children's rate of vocabulary growth. We examine a broad sample of children and specifically compare vocabularies of children not at risk for language deficits with children whose vocabulary size for their age puts them at risk for language difficulties. In the literature, these at-risk children are often called "late talkers;" we will also use that term although it is somewhat of a misnomer because they are not simply "late" but rather on a slower path of vocabulary growth. If small-world properties are important to the efficiency of language use–and perhaps also to new word acquisitions–then vocabulary structure and not just vocabulary size may be different for these children. Does the connectivity of words in the emerging semantic networks of late talkers differ from the network structure of children whose vocabulary has grown at a more typical pace?

Rationale for the Approach

We analyzed vocabularies from a broad sample of children who differed in age and vocabulary size but whose vocabulary size for age was above the 20th percentile and also from a sample of children, also varying in age and vocabulary size, whose vocabulary size fell below the 20th percentile for age at the time the vocabulary was collected. A semantic network was built for each vocabulary yielding a large set of individual networks that could be ordered by age and separately by vocabulary size.

To build individual networks, we connected the words in an individual's vocabulary, using co-occurrence in a large corpus of child-directed speech as the index of semantic relatedness. The co-occurrences in this corpus of child directed speech is presumed to index the relatedness of the individual words in the language (and that part of the language relevant to children) and in the learning environment in general. This measure of semantic relatedness is *not* the co-occurrences in the specific learning environments of individual children, a key point we will consider in the general discussion. Co-occurrences of words within the corpus formed the edges or links of a semantic network and the nodes were based on the words in each individual child's productive vocabulary.

In sum, the key question is whether and how semantic network connectivity changes as children's vocabularies grow and whether this differs for children whose vocabulary growth rate is sufficiently slow that they are considered at risk for language disorders.

Methods

Vocabularies.

Vocabularies from 73 children ranging in age from 16.2 to 34.6 months were selected for this study. These vocabularies derive from one-time visits of children to the Cognitive Development Laboratory at Indiana University and are measures of productive vocabulary via the Bates-MacArthur Communicative Developmental Inventory (Toddler or Infant form as appropriate to the child's age, Fenson et al, 1993). This is a parent checklist and parents were asked to indicate

which words on the checklist their child produced (Fenson et al, 1993). Total vocabulary as indicated by the parent was used to determine the percentile of the vocabulary size for the child's age. From this repository of child vocabularies we selected a random sample of vocabularies of children whose vocabulary size for their age fell above the 20th percentile and as large a sample as possible of children whose vocabulary size for their age fell below the 20th percentile (see Fenson et al, 1993). Table 1 provides the number of children in each group, means and ranges of their vocabulary size, age, and percentile.

Table 1: Age and percentile of children in study

	# children	Age range in months (mean)	Percentile range (mean)
All children	73	16.2-34.6 (22.1)	5-99 (25.6)
Late talkers	38	16.3-34.6 (24.3)	5-20 (12)
Typical talkers	35	16.2-26.6 (19.8)	25-99 (40.4)

Words.

For the network analysis, only the 291 words that are on both the Toddler and Infant forms were used. This allowed for a more accurate comparison across ages. Of the included words, 204 are nouns, 51 are verbs and the remaining 36 are adjectives, adverbs and function words.

Networks.

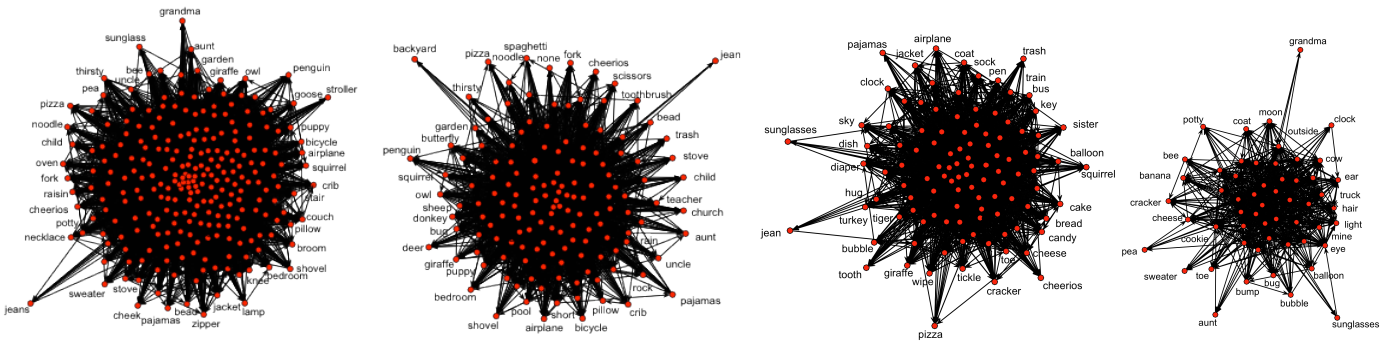
To build the networks, links between words were defined in terms of co-occurrences in the CHILDES database (MacWhinney, 2000). The co-occurrence method was taken from prior analyses by Riordan and Jones (2007) and related lemmas (cat, cats, hit, hitting) were counted as instances of the same lexeme. The matrix of co-occurrences was built using a process similar to the Hyperspace Analogue to Language (HAL) (Lund & Burgess, 1996) and the word co-occurrence detector (Li, Farkas & MacWhinney, 2004). For the 291 unique words, we formed a *291x 291 matrix*, where each cell, *ij*, is filled according to the following rule: a moving window of size 15 moves word-wise through the corpus, with each cell *ij*, changed to a value of 1 if word *j* occurs both downstream and together in the same window with word *i*. This produces a directed network where each word is connected to another word by a directed link if it co-occurs downstream of that word in child directed speech. Frequency counts were taken as the number of occurrences of a given word in the corpus.

Results

The analyses reported here use four network statistics: *median in-degree*, *global clustering coefficient*, *redundancy*, and *geodesic distance*. Each provides a means of assessing connectivity within networks. Figure 3 shows four networks for four typically developing children and the index of connectivity for each of these networks. The four individual networks show considerable small-world structure with as few as 106 (or even 55) words. This suggests that these properties–characteristic of mature semantic networks–are evident even from the earliest stages of lexical development. This could merely reflect the structure of language such that any learner (or random sample of words from early vocabularies) would show these properties. Or, these properties could be more fundamentally related to how individual children build semantic structures for efficient language learning and processing. The comparison of typically-developing and late-talking children provides the relevant evidence.

In-degree.

In-degree is a measure that captures how many connections each node has directed towards it from other nodes. In the present case, the in-degree of the target word or node is the number of distinct words that occurred 15 or fewer words after the target in the CHILDES corpus. The median in-degree provides an overall picture of how sparse or dense a network is. In a sparse network, the words in the



# of words	256	149	106	55
Median in-degree	53.5	29	36	22
Clustering coefficient	.472	.494	.555	.606
Geodesic dist.	1.720	1.737	1.608	1.566
Redundancy	16.04	14.15	9.23	4.84

Figure 3: These semantic networks of typically developing children show that children develop small world structure even with relatively few words. Throughout development the semantic networks of children show high clustering coefficients and low average geodesic distance. The networks also quickly develop a high number of connections and multiple traversable pathways.

vocabulary are not as related to each other and so there can be more words than connections. In a dense network, many words are connected to each other; e.g., the median in-degree is nearly equal to the total number of words or nodes, many words in the network are semantically related and co-occur frequently in speech.

Regression analysis, with median in-degree as the independent variable and the child's MCDI percentile as the dependent, yielded a significant relation between in-degree and percentile with lower median values characterizing late-talkers even when age ($p<.001$) and vocabulary size ($p=.0162$) were controlled. The relation between in-degree and vocabulary size for the two groups is shown in Figure 4.

This indicates that there are more links in a typical talker's network than in a late talker's network even when the networks have the same number of nodes. Typical talkers learn words that are semantically connected to each other but late talkers are less likely to do so, as if perhaps, they learn words as individual islands, as if the *next* word learned is somehow independent of the prior learned words.

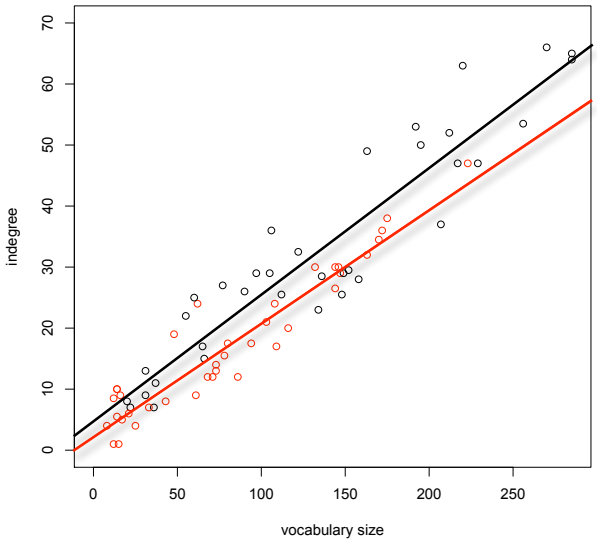


Figure 4: A graph of the median in-degree as a function of vocabulary size. The black line indicates typical talkers and the lighter line, the late talkers. ($p=0.016$).

Global clustering coefficient.

The clustering coefficient provides a measure of how well connected a node's neighbors are to each other. Small-world networks have high clustering coefficients, relative to networks of the same size (number of nodes) and density (ratio of observed links to possible links). A clustering coefficient of 1 indicates that all of a node's neighbors are themselves connected. A clustering coefficient of 0 indicates that none of a node's neighbors are connected to one another. This provides a measure of local clustering, as opposed to more global measure of density assessed with in-degree above. The late-talkers in the present study show a lower average clustering coefficient than late talkers when age is controlled ($ß=-54.6$, $SE=21.991$, $p=0.0154$) and a near significant effect when vocabulary size is controlled ($ß=34.53$, $SE=18.33$, $p=0.0638$). Figure 5 shows the clustering co-efficient as a function of vocabulary size for the two groups. As is apparent from the data points, there is both more variability by this measure among the youngest later-talkers than typically developing children and typically developing children appear to move toward a stable clustering coefficient earlier than do late talkers. The lower average clustering coefficient of late talkers suggests that they are less likely to learn words that fill out categories of closely related words that they already know, a result that again suggests that there may be fewer dependencies between new acquisitions and already learned words. Being unable to fill out categories of closely related words, these late talkers may have trouble reorganizing their current semantic understanding to create new categories and concepts.

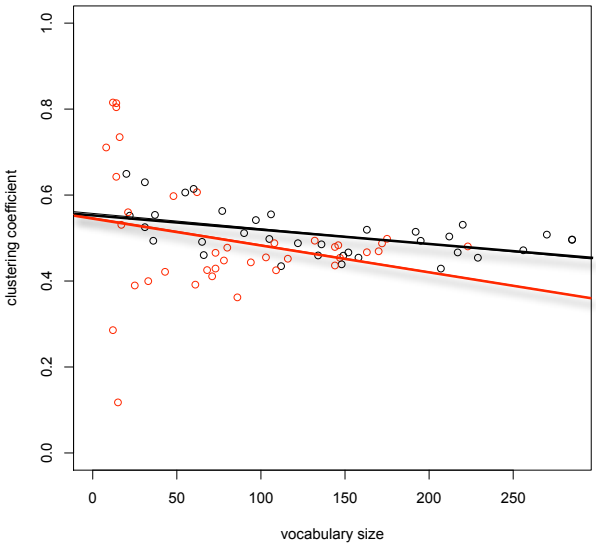


Figure 5: A graph of the clustering coefficient as a function of vocabulary size. The black line indicates typical talkers and the lighter line, late talkers (lm, $p=0.064$).

Redundancy.

Redundancy captures the robustness of the network: in a highly redundant network, if a random connection is deleted, the deleted link will not alter the likelihood of a connected path between two words. For example, with a road network, if there are multiple ways to get between two places, then a road closure is not an insurmountable problem. However, if only one road connects two locations, then a closure of that road makes the two locations inaccessible to one another. Higher redundancy means more possible paths. As opposed to clustering coefficient and in-degree, redundancy provides a measure of the ease of accessibility in the network (from one node or word to another). Compared with the clustering coefficient, this provides a more global measure of cohesion across the network.

Regression analyses yielded significant differences between the two groups, with late talkers having less redundant networks when controlling for age ($p<.001$) and

vocabulary size (p=0.015). To quantify this, for a given network of 200 words, a late talker would have on average 11 possible pathways compared to 13 possible paths in a network for a typical talker (t-test, p= 0.016, comparing all late talkers/typical talkers). Though this difference is small, the actual implication of this difference is that late talkers have many words that have only one or two connections, whereas typical talkers have fewer words with low redundancy suggesting a network more robust to change.

The difference in the number of possible pathways between nodes across the two groups suggests that the robustness of the two groups is also different. The typical talkers, building more redundant networks, are less likely to have trouble transitioning from one area of the network to another. The fluidity of their productive speech also would be less hampered by the forgetting of a few words. By having multiple ways of getting from one word to another, the typical talkers may more easily access one word following another. These differences may relate importantly to the word-finding and word-retrieval difficulties of late talkers, an important question for future work.

Average geodesic distance.

Geodesic distance represents shortest path length between two nodes. We computed the geodesic distance between all nodes excluding isolates or unconnected nodes. We then averaged the geodesic distance for all nodes, further excluding all cases in which there was no traversable path between two nodes. As networks grow larger, more connections are possible and the geodesic distance, or the shortest distance between two nodes, will often trend toward less than 2. This happens when a word that connects to all other words, such as "you", is added to the semantic network. If word A is not directly connected to word B, word A is connected to word B through "you", resulting in an average geodesic distance of approximately 2.

Late talkers have significantly different geodesic distances from typical talkers. When considering networks of similar size (i.e. words known), we see that typical talkers having a mean geodesic distance of 1.82 and late talkers having a mean geodesic distance of 2.55 (t-test, p=0.0276).

Another indication that these at-risk children are building networks with less global structure is the number of components in a network. Components are isolated clusters or words of a network that do not connect to other components in a network. Early on in vocabulary learning, it is possible to learn a word, or words, in complete isolation that is not semantically related to any other word or cluster. For example a child might learn a bunch of animals and a bunch of food words but be missing words like milk that would link the two clusters. Of the children in this study only 17 children showed networks that had more than one component, 14 of which are classified as late talkers.

The difference in geodesic distance and number of components suggests that late talkers are not building networks that allow for the same level of global access.

Discussion

The present study is the first analysis of the network structures of early vocabularies for individual children and the first to reveal potentially meaningful individual differences in the structures of these emerging networks. As such, there are still open questions and limitations that will need to be addressed. These include comparisons to randomly selected vocabularies of different sizes, linking of these differences in vocabulary structure to performance (such as word retrieval), and following individual children's vocabulary growth. Nonetheless, the results provide three new insights: (1) Small-world properties are evident in the network structure of even very small and early vocabularies; (2) these properties are not the consequence of just learning any subset of early English words since–at any vocabulary size–there are individual children with more robustly connected networks than other children; and (3) the structure of these individual differences in network connectivity appears related not just to vocabulary size but to the rate of vocabulary development with children at risk for serious language deficiencies (by normative standards) showing less cohesive and less efficiently structured networks.

The broad sample of typically-developing children, children above the 20th percentile and who are not at risk for language deficiencies, show less variance in network structure, specifically clustering coefficient in our analysis, than do the late-talking children, a remarkable fact in its own right. These typically-developing children seem to be building semantic networks with many of the small-world properties found in adult semantic networks, showing higher in-degree, clustering coefficient, and redundancy, indicating that typical talkers are learning words more cohesively, with more semantic connectivity between learned words—both globally and locally—than do the networks of late talkers. Late talkers are not only learning more slowly but appear to be learning differently. One possibility consistent with the present pattern is that typically developing children build their vocabularies in ways such that learning itself is dependant on the semantic relations among already learned words or the semantic relations in the learning environment (Hills et al, 2009b) whereas late talkers just learn words, adding words as individual and unrelated items, not picking up on the semantic relations in the learning environment.

Because the semantic relations in these networks are themselves normative–reflecting the structure of the general learning environment and not the child's specific learning environment-it is also, in principle, possible that these children's learning environments present less semantic connectivity. Previous research has shown that learning environments, in terms of the kind and number of words that are spoken to children, do influence the kinds and number of words that children learn (e.g., Hurtodo, Marchman & Fernald, 2008; Rowe, 2008; Hoff & Naigles, 2002: Huttenlocher et al, 1991). However, contemporary understanding of language-delayed children suggests that this may not be the sole factor in these delays (see Bishop &

Leonard, 2000). Still, a more detailed examination of individual language learning environments is in order.

Our evidence suggests that typical talkers are more likely to acquire words that share semantic associations with words they already know. This may be a consequence of the fact that they are more sensitive to semantic associations in the environment (what has been called *preferential acquisition*), or that they are more likely to use known words to direct the acquisition of new words (called the *lure of the associates*). Previous work has shown that both of these processes are predictive of word acquisition (Hills et al., 2009b), but these processes may also represent individual strategies for learning. This suggests an interesting direction for future research in individual differences in language acquisition.

Acknowledgments

This research was supported by NICCHD grant HD028675 to Linda Smith.

References

Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. T.Guthrie (Ed.), *Reading comprehension and education* (pp. 77–117). Newark, DE: International Reading Association.

Bates, E., Marchman, V., Thal, D., Fenson, L., Dale, P., Reznick, J. S., Reilly, J. & Hartung, J. (1992). Developmental and stylistic variation in the composition of early vocabulary. CRL Technical Report, UCSD.

Bishop, Dorothy V. M. (Ed) (1), & Leonard, L. B. (. (Eds.). (2000). *Speech and language impairments in children: Causes, characteristics, intervention and outcome*. New York, NY, US: Psychology Press.

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

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

Dale, P.S., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, & Computers*, *28*, 125–127.

Dupuy, H. P. (1974). *The rationale, development and standardization of a basic word vocabulary test (DHEW Publication No. HRA 74-1334)*. Washington, DC: U.S. Government Printing Office.

Fenson, L., Dale, P. S., Reznick, J. S., Thal, D., Bates, E. and Hartung, J. P., et al. 1993, The MacArthur Communicative Development Inventories: User's Guide and Technical Manual (San Diego: Singular)

Griffiths, T.L., Steyvers, M., & Firl, A. (2007). Google and the mind: Predicting fluency with PageRank. *Psychological Science, 18*, 1069–1076.

Heilmann, J., Weismer, S. E., Evans, J., & Hollar, C. (2005). Utility of the MacArthur-bates communicative development inventory in identifying language abilities of late-talking and typically developing toddlers. *American Journal of Speech-Language Pathology, 14*(1), 40-51.

Hills, T., Maouene, J., Sheya, A., Maouene, M., and Smith, L. (2009a). Emergent categories in the feature structure of early-learned nouns. *Cognition*, 112, 381-396.

Hills, T., Maouene, J., Sheya, A., Maouene, M., and Smith, L. (2009b). Longitudinal analysis of early semantic networks: preferential attachment or preferential acquisition? *Psychological Science*, 20, 729-739.

Hoff, E., & Naigles, L. (2002). How children use input to acquire a lexicon. *Child Development*, 73, 418-433

Hurtado, N., Marchman, V. A., & Fernald, A. (2008). Does input influence uptake? links between maternal talk, processing speed and vocabulary size in spanish-learning children. *Developmental Science,* 11, F31-F39.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input & gender. *Developmental Psychology,* 27, 236-248.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*(1), 1–37

Li, P., Farkas, I., MacWhinney, B. (2004). Early lexical development in a self-organizing neural network. *Neural Networks* 17, 1345-1362.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavioral Research Methods, Instruments, and Computers, 28*(2), 203-208.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Erlbaum.

Mitchell, C., & McMurray, B. (2009). On leveraged learning in lexical acquisition and its relationship to acceleration. *Cognitive Science*, 33, 1503-1523.

Moyle, M. J., Weismer, S. E., Evans, J. L., & Lindstrom, M. J. (2007). Longitudinal relationships between lexical and grammatical development in typical and late-talking children. *Journal of Speech, Language, and Hearing Research, 50*(2), 508-528.

Raven, J. C. (1948). The comparative assessment of intel-lectual ability. *British Journal of Psychology*, *29*, 12–19.

Riordan, B., and Michael N. J.. (2007). Comparing semantic space models using child-directed speech. In D. S MacNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Conference of the CogSci*. 599-604.

Rowe, M. L. (2008). Child-directed speech: Relation to socioeconomic status, knowledge of child development & child vocabulary skill. *Journal of Child Language,* 35, 185-205.

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cognitive Science, 29*, 41-78

Thal, D., Bates, E., Goodman, J., & Jahn-Samilo, J. (1997). Continuity of language abilities: An exploratory study of late and early-talking toddlers. *Developmental Neuro-psychology, 13*, 239–273.

Vitevitch, M.S. (2008). What can graph theory tell us about word learning and lexical retrieval? *Journal of Speech, Language, and Hearing Research, 51*, 408–422.