What if child vocabulary development followed network acquisition models exactly?

Christopher R. Cox[1], Stanley H. West[1], Thomas Hills[2], and **Eileen K. Haebig[3]**

[1]Department of Psychology, Louisiana State University, Baton Rouge, LA 70803 USA
[2]Department of Psychology, Warwick University, Coventy, UK
[3]Department of Communication Sciences and Disorders, Louisiana State University, Baton Rouge, LA 70803 USA

Abstract

The network science perspective on vocabulary development emphasizes the structured relationships between words and how they can influence learning. Words that appear in many contexts and thus develop associations with many other words tend to be learned earlier—a growth model called preferential acquisition. Likewise, children appear to have a bias towards learning new words associated with many known words—a growth model called lure of the associates. Although both are statistically related to age of acquisition estimates, much variance remains unexplained, and it is unknown what structures these models promote within the developing vocabulary. We simulated vocabulary growth strictly adhering to preferential acquisition and lure of the associates and found that they promote similar structures: more connectivity, more clustering, and much shorter path lengths than random growth would achieve. They clearly promote small-world structure, consistent with that seen in young children's vocabularies.

Keywords: language development; network modeling; vocabulary growth;

Introduction

Children begin producing their first words near the end of their first year of life with accelerating vocabulary growth throughout their second year. The impressive rate of vocabulary growth in young toddlers has prompted an extensive body of research spanning a myriad of methodological approaches. One productive research method utilizes computational modeling techniques that allow hypothesized drivers of early vocabulary learning to be explored and compared. Network growth modeling approaches explore the ways that children may be influenced by the relationships among words in their environments and within their existing vocabularies while acquiring new words. Although this line of research has provided novel insights into early word learning, substantial variance in the typical progression of vocabulary growth remains unexplained (Bilson et al., 2015; Cox & Haebig, 2023; Hills et al., 2009; Jiménez & Hills, 2022). This leaves open the possibility that the models of network growth currently understood to be mechanistic contributors to early language acquisition do not actually fit the data very well. The current study explores what early vocabulary growth would look like if the three leading vocabulary-learning models were to fully dictate the order of acquisition of 675 words from the MacArthur-Bates Communicative Development Inventory.

The three network growth models that have received the most attention as mechanisms of early vocabulary growth are *preferential acquisition*, *preferential attachment*, and *lure of the associates*. The preferential acquisition mechanism (Hills et al., 2009) prioritizes relationships among words as they exist in the external environment, rather than structure internal to the child's mind constrained by words they have already acquired. Words associated with many other words in the child's environment are more likely to be learned than those that are relatively specialized or conceptually isolated, regardless of the size or composition of the child's vocabulary. In contrast, the preferential attachment account (Steyvers & Tenenbaum, 2005) prioritizes the internal associative structure between words that the child already knows. Therefore, an unknown word that is associated with many known words that are themselves associated with many other known words has an increased chance of being learned relative to an unknown word associated with fewer known words or associated with known words that are each conceptually isolated within the vocabulary. The lure of the associates mechanism (Hills et al., 2009) is similar to preferential attachment but disregards the structure among known words. On this account, an unknown word that is associated with several known words is more likely to be acquired, no matter which known words it is associated with. Previous work has found evidence that each of these models explains a significant portion of early vocabulary growth (e.g., Bilson et al., 2015; Cox and Haebig, 2023; Hills et al., 2009; Jiménez and Hills, 2022; Sailor, 2013).

Although previous studies have found that each mechanistic model of vocabulary growth predicts significant variance in the age of acquisition of words, the models often include other psycholinguistic variables that are associated with word learning (e.g., word frequency, phonotactic probability, phonological neighborhood density). In fact, a multitude of other word features are known to impact word learning, such as concreteness (Braginsky et al., 2019; Verhagen et al., 2022), imageability (Lin et al., 2022; Ma et al., 2009; Masterson et al., 2008), iconicity and baby-ness (Perry et al., 2015), and child body-object interaction ratings (Muraki et al., 2022). Children are exposed to and implicitly learn from multiple cues within their environment, word features, and experiences. Thus, the small-world semantic network structure observed in children's vocabularies (a subset of highly-connected nodes facilitating short paths between most nodes), often attributed to the network growth mod-

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

els described above, could arise primarily from other growth drivers. We explore the evolving semantic network structures of vocabularies that grow exactly according to preferential acquisition, lure of the associates, and preferential attachment and compare them to a cross section of real child CDI data. This will improve our understanding of what structures each mechanism promotes and whether they are consistent with profiles of child vocabulary learning.

Method

Vocabulary data

WordBank (Frank et al., 2017) is a freely available archive of data acquired using the MacArthur-Bates Communicative Development Inventory (CDI; Fenson et al., 2007). The CDI is collection of parent-report questionnaires which include vocabulary checklists, where parents can indicate whether each word is in their child's expressive vocabulary. We analyzed 1,416 unique CDI data entries from WordBank's American English corpus. We attempted to ensure that the children in our sample are all developing language typically by excluding children who scored below the 15th percentile according to the CDI normative data (Fenson et al., 2007). Our analyses focused on 675 of the 680 words and short phrases assessed by the CDI, which excluded "give me five!", "gonna get you!", "this little piggy", "babysitter's name", "child's own name" and "pet's name". The short phrase "so big!" was included, and "pet's name" was substituted with "pet (noun)" in the word association study reference below.

Network definitions

Lexical network structure was derived from child-oriented word associations cued by 675 CDI items, excluding the short phrases mentioned above and replacing "pet's name" with the word "pet (noun)"; (Cox & Haebig, 2023). The network nodes correspond to the 675 cue words, and an unweighted, directed edge was drawn between each pair of nodes if the *receiving* word was generated in response to the *sending* (cue) word at least once. Responses that were not among the set of cue words are ignored for this analysis. This produced a fully connected sparse network ($\sim 4\%$ of possible directed connections exist).

Growth models

We estimated network growth according to three growth models: preferential acquisition, lure of the associates, and preferential attachment. Each can be implemented as a function that takes two arguments: 1) a network relating words that can be learned and 2) a list containing a subset of those words that have been acquired already. Each function returns a *growth value* for every word not in the vocabulary. An unknown word's growth value is defined by each growth model as:

- *Preferential acquisition*: the unknown word's own indegree in the context of the full network, irrespective of what words are known.
- *Lure of the associates*: the number of known words that are associated with the unknown word.
- *Preferential attachment*: the average indegree of the known words the unknown word is associated with.

Indegree refers to the number of edges that are directed toward a node. Code for computing growth values is provided in an R package available at github.com/crcox/netgrowr.

Generating seed vocabularies

For each of the 1,416 children for whom we have CDI data, we counted the number of words each child was reported to produce. This count is used as a proxy for the child's productive vocabulary size. The probability of each word existing in a child's vocabulary as a function of their vocabulary size was estimated via logistic regression. Using these fitted models, we estimated the probability of each word existing in a 60-word vocabulary. One hundred 60-word seed vocabularies were sampled at random, with the likelihood of inclusion in a seed vocabulary proportional to the estimated probability for each word. Therefore, through this process, the 60 seed words were early acquired words that have very young ages of acquisition.

Simulating model-prescribed vocabulary growth

Beginning with a seed vocabulary of 60 early-acquired words, a network growth model is used to assign growth values to all unknown words. The words associated with the top 20 growth values from each network growth model are then assigned as newly acquired words, with ties broken at random. This is repeated, acquiring 20 words at a time, until the vocabulary has grown to 600 words. This process is repeated with the same seed vocabulary for the two remaining network growth models and then repeated for the remaining 99 seed vocabularies. The order of word acquisition is recorded at each step. Any single growth trajectory is influenced exclusively by a single network growth model (e.g., by a word's associations to words in the learning environment for preferential acquisition) .

Network descriptive statistics

By extracting subsets of nodes from the 675-node network of child-oriented word associations, a network can be defined to correspond to any state of vocabulary growth over this set of CDI words. We calculated the median indegree, the global clustering coefficient, and the average shortest path length for each network statistic arrived at by each growth model based on each seed vocabulary (i.e., at each vocabulary size along each growth trajectory) in addition to the 1416 real vocabularies indicated by the CDI data. In a network with directed connections, indegree is the number of connections pointing towards a node. As the total number of connections (the network density) increases, the central tendency of the (in)degree distribution will increase. We report the median because indegree distributions can be heavily skewed by a

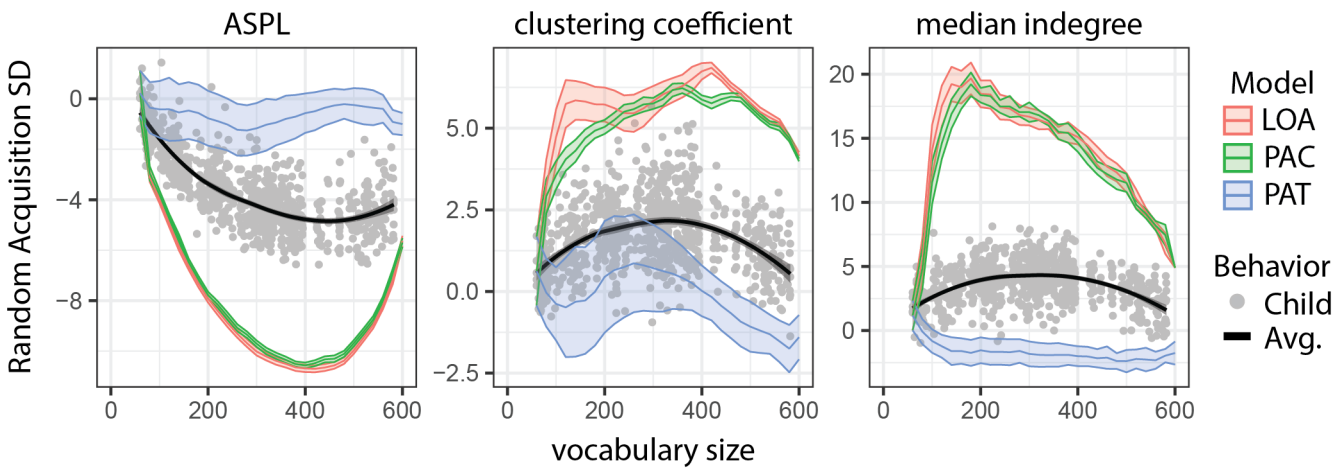

Figure 1: Random acquisition standardized network descriptive statistics for lure of the associates (LOA), preferential acquisition (PAC), and preferential attachment (PAT) growth models. CDI data for individual children are show as grey dots and summarized with a black line. Error bands are ± 1 SD of the mean.

few nodes having much higher indegree than the rest, particularly in networks with small world structure. With respect to networks where nodes are words and connections are made from word association data, the median indegree is an estimate of a child's overall level of semantic network density. We emphasize indegree because word association behavior is directional (the cue elicits the response).

In a network with a high global clustering coefficient, it is likely that the neighbors of a node will also be neighbors. A cluster is defined as a transitive triad: For example, "pig", "bunny", and "mouse" form a transitive triad because all three animals are directly associated with each other. Meanwhile, "pig", "mud", and "rain" form an intransitive triad: pigs are known to roll in mud, rain can cause mud to form, but "pig" and "rain" are not associated. The global clustering coefficient is the proportion of triads in a network that are transitive. Densely connected networks will have a larger clustering coefficient, but small world networks are more clustered than their density would imply.

The average shortest path length (ASPL) of a network is the average of the shortest paths between all pairs of nodes. The shortest path between two nodes is an indication of how much distance would be involved in traveling between them. In semantic networks, words separated by short paths are more likely to prime one another (De Deyne et al., 2016; Kenett et al., 2017). Relatedly, there are data suggesting that participants with shorter ASPL score higher on tasks designed to measure creative thinking (Benedek et al., 2017; Kenett et al., 2014) the hypothesis being that a more diverse array of words are more accessible from any given starting point within the network.

These metrics were chosen because they are commonly used in network modeling research and are part of the definition of what constitutes small world structure. In small world networks, most pairs of nodes are not directly connected (they are not neighbors), yet the number of connections that must be traversed to get from any word to another will tend to be small (Watts & Strogatz, 1998). A network with a large clustering coefficient, short ASPL, and heavily skewed degree distribution have small world structure.

Normalization by random acquisition

Network descriptive statistics tend to correlate with network size. Furthermore, the network structure within a vocabulary will necessarily reflect the network structure of the environment within which the words are being learned. To help eliminate these factors from the network descriptive statistics, we generated a set of 10,000 random growth trajectories to standardize the model-based growth trajectories. Starting with each of the same 60-word seed vocabularies, an additional 540 words are sampled without replacement 20 at a time to incrementally grow a 600-word vocabulary at random. This was repeated 100 times for each seed.

As described above, median indegree, global clustering coefficient, and ASPL were computed at each vocabulary size along each random growth trajectory. At each vocabulary size, we computed the mean and standard deviation. These were used to standardize the "true" descriptive statistics for each seed, growth model, and vocabulary size.

Results

Figure 1 presents the mean standardized network statistics tracking simulated vocabulary growth for each model. Error bands represent the standard deviation of the mean over seeds. Figure 2 presents the same data without standardization along with the mean of the randomly acquired vocabularies as a visual reference (orange line); error bands around the random acquisition means are too narrow to display. Also shown are the analogous network descriptive statistics for the

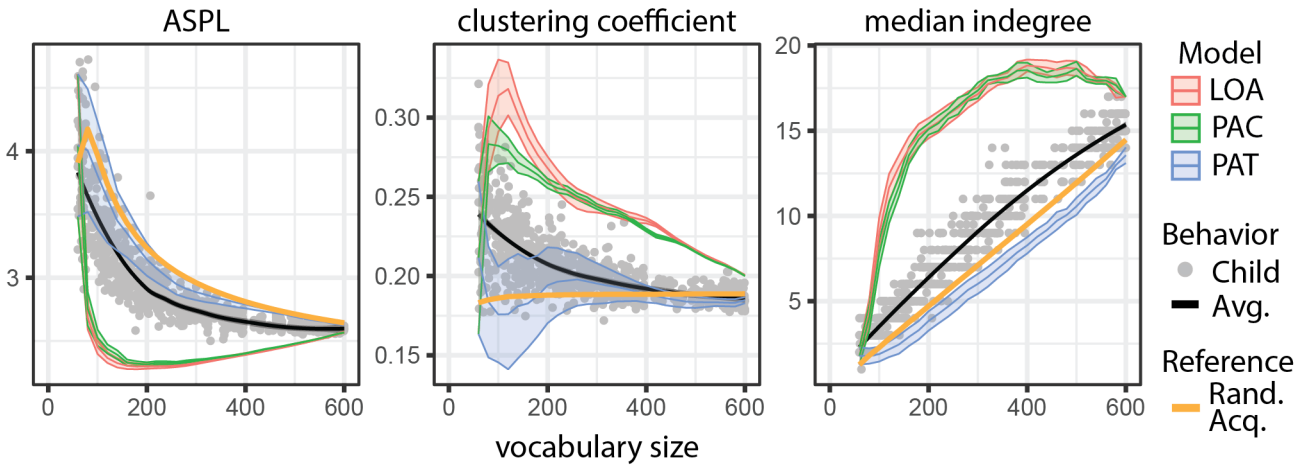

Figure 2: Network descriptive statistics for lure of the associates (LOA), preferential acquisition (PAC), and preferential attachment (PAT) growth models presented without standardization. CDI data for individual children are show as grey dots and summarized with a black line. The orange line is the average of the random acquisition networks. Error bands are ±1 SD of the mean.

subset of 867 children in our sample with vocabularies containing between 60 and 600 words (grey dots). A nonlinear best fit line is overlaid to emphasize the smoothed average over these measured vocabularies (black line).

The lure of the associates and preferential acquisition foster similar semantic network structures across this developmental range. They promote high median indegree, clustering, and short paths, which taken together are indicative of small world structure. Their developmental trajectories are very similar, with indegree and clustering rising rapidly within the first sets of words acquired. While these structures were extremely pronounced in the vocabularies simulated purely with preferential acquisition or lure of the associates, we note that the smoothed average of cross-sectional child CDI data followed a similar but less extreme trajectory of semantic network development as measured by all three network statistics. The structures observed in child CDI data and vocabularies acquired via preferential acquisition and lure of the associates are extremely unlikely to develop at random.

Preferential attachment, in contrast, yields vocabularies that are far less distinct from random acquisition. The structure fostered is far less pronounced than preferential acquisition and lure of the associates and appears to promote a developmental trajectory quite unlike that estimated in the cross-sectional child CDI data—indegree and clustering (at some vocabulary sizes) are lower and path lengths are somewhat longer.

The simulations reported in Figures 1 and 2 are the result of explicitly selecting the 20 words ranked most highly by each growth model. In Figure 3, we instead sampled probabilistically at each step, where the sampling distribution over unknown words was defined as in Bilson et al (2015):

$$P(i) = \frac{(\text{growthvalue}_i + 1)^{\beta}}{\sum_{i \in W} (\text{growthvalue}_i + 1)^{\beta}} \quad (1)$$

Where i indexes the set of unknown words W. When $\beta > 1$ the difference between the highest and lowest probability is exaggerated and when $\beta < 1$ it is reduced; when $\beta = 0$ the difference between the highest and lowest probability goes to zero (i.e., a uniform sampling distribution) and the growth model is completely ignored. In other words, β controls how much of a "say" the growth model has in which words will be acquired, and allows for more or less randomness in the sampling process.

For each growth model, 1,000 probabilistic trajectories were simulated for six values of β equally spaced between zero and one. Figure 3 presents the average trajectory for each model for each beta. By sampling probabilistically, rather than deterministically selecting the words with largest growth values, the models are less clustered and dense overall and are slower to become clustered and dense. By allowing more randomness (i.e., allowing for influences on word acquisition other than a single growth model), the simulated trajectories become more similar to the mean of the child CDI data (shown in black).

In short, semantic network models built by pairing the child-oriented word association norms with the CDI data indicate the same small-word network qualities observed in prior work. The preferential acquisition and lure of the associates growth models strongly encourage this structure over this sample of early-acquired words, while preferential attachment does not.

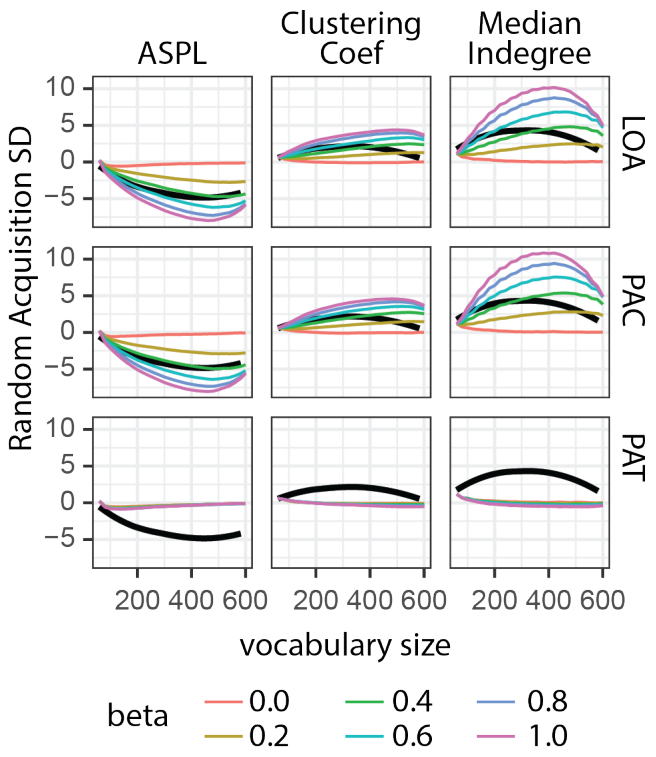

Figure 3: Random acquisition standardized descriptive statistics for networks simulated via probabilistically sampling according to the distribution of growth values at each step. Colored lines correspond to different values of the exponent β, which manipulates the importance of the growth model in driving vocabulary growth relative to random acquisition (see Eq. 1).

Discussion

Network science offers a rich set of tools for studying language development and has redirected attention from individual words to the relationships between words, within the environment and within the child's mind. Computational modeling of large collections of vocabulary data is a valuable tool for exploring the mechanisms driving learning and early word acquisition. However, showing that a hypothesized mechanism accounts for significant variance in a measured outcome does not imply that no other undiscovered mechanism might fit the data better.

Our results demonstrate that preferential acquisition and the lure of the associates tend toward developing dense and highly clustered (semantic) networks. Although the hypothesis that either of these growth models exclusively dictates early vocabulary development is obviously wrong, our simulations also demonstrate that both of these growth models are consistent with the kind of structure seen in child lexicosemantic networks.

This is not the first attempt to simulate vocabulary growth via preferential acquisition. Bilson et al. (2015) sampled probabilistically based on the distribution of growth values at each growth step and tuned β to best fit the child CDI data. They found that the optimal value was $\ll 1$, suggesting that the contribution of preferential acquisition to child language development may be small. Their analysis did not aim to explore what kind of structure preferential acquisition would develop in the limit, which is helpful for discerning how this mechanism would shape the vocabulary.

The current work is limited by focusing on a relatively small set of words and only one way of defining the relationships among them. Aside from the other ways that semantic similarity could have been estimated (e.g., co-occurrence statistics in the CHILDES corpus of child directed speech; MacWhinney, 2000), phonological and syntactic similarity have also been shown to be predictive of language growth (e.g., Ciaglia et al., 2023; Laing, 2024). When modeling vocabulary development, the contributions of the network growth model are often more apparent if psycholinguistic variables like frequency and phonological neighborhood density are incorporated into the model. Simulations of this kind could also be applied to study the influences of variables such as these.

Preferential attachment may appear more consistent with language development within appropriate context or after achieving a sufficiently large vocabulary. Steyvers and Tenenbaum (2005) demonstrated that larger semantic networks with better coverage of the English language have structure consistent with growth by preferential attachment. Perhaps a child's internal structure of their lexicosemantic environment must be sufficiently developed before that structure begins to meaningfully drive the learning process. Network growth analyses spanning longer developmental periods and exploring dynamics between larger vocabularies and richer environments will be necessary to investigate hypotheses such as these.

Another point of interest in our results is how similar preferential acquisition and lure of the associates performed. From a cognitive scientific perspective, a model that incorporates information about the internal state and what has been previously learned is distinctly different from one that does not. Theoretically and conceptually, lure of the associates and preferential acquisition are importantly different accounts of the learning process relying on distinctly different learning mechanisms. Our results are in line with prior work indicating that these two models tend to fit similarly well and explain substantially overlapping components of the variance in vocabulary growth (Beckage & Colunga, 2019; Cox & Haebig, 2023; Hills et al., 2009, 2010). However, additional work will be required to determine the generality of this similarity: will similar results obtain from a different set of words or a different similarity structure? For noww, the similar growth trajectories fostered by preferential acquisition and the lure of the associates remains an intriguing puzzle.

While more work remains to be done, the current work affirms the utility of preferential acquisition and the lure of the associates as models of early vocabulary learning and devel-

opment.

References

Beckage, N. M., & Colunga, E. (2019). Network growth modeling to capture individual lexical learning. *Complexity*, *2019*(1), 7690869.

Benedek, M., Kenett, Y. N., Umdasch, K., Anaki, D., Faust, M., & Neubauer, A. C. (2017). How semantic memory structure and intelligence contribute to creative thought: A network science approach. *Thinking & Reasoning*, *23*(2), 158–183.

Bilson, S., Yoshida, H., Tran, C. D., Woods, E. A., & Hills, T. T. (2015). Semantic facilitation in bilingual first language acquisition. *Cognition*, *140*, 122–134.

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2019). Consistency and variability in children's word learning across languages. *Open Mind*, *3*, 52–67.

Ciaglia, F., Stella, M., & Kennington, C. (2023). Investigating preferential acquisition and attachment in early word learning through cognitive, visual and latent multiplex lexical networks. *Physica A: Statistical Mechanics and its Applications*, *612*, 128468.

Cox, C. R., & Haebig, E. (2023). Child-oriented word associations improve models of early word learning. *Behavior research methods*, *55*(1), 16–37.

De Deyne, S., Navarro, D. J., Perfors, A., & Storms, G. (2016). Structure at every scale: A semantic network account of the similarities between unrelated concepts. *Journal of Experimental Psychology: General*, *145*(9), 1228.

Fenson, L., Marchman, V., Thal, D., Dale, P., Reznick, J. S., & Bates, E. (2007). *MacArthur-Bates communicative development inventories: User's guide and technical manual* (2nd ed.). Paul H. Brookes Publishing Company.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, *44*(3), 677–694.

Hills, T. T., Maouene, J., Riordan, B., & Smith, L. B. (2010). The associative structure of language: Contextual diversity in early word learning. *Journal of memory and language*, *63*(3), 259–273.

Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009). Longitudinal analysis of early semantic networks: Preferential attachment or preferential acquisition? *Psychological Science*, *20*(6), 729–739.

Jiménez, E., & Hills, T. T. (2022). Semantic maturation during the comprehension-expression gap in late and typical talkers. *Child Development*, *93*(6), 1727–1743.

Kenett, Y. N., Anaki, D., & Faust, M. (2014). Investigating the structure of semantic networks in low and high creative persons. *Frontiers in human neuroscience*, *8*, 407.

Kenett, Y. N., Levi, E., Anaki, D., & Faust, M. (2017). The semantic distance task: Quantifying semantic distance with semantic network path length. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(9), 1470.

Laing, C. E. (2024). Phonological networks and systematicity in early lexical acquisition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Lin, K. R., Wisman Weil, L., Thurm, A., Lord, C., & Luyster, R. J. (2022). Word imageability is associated with expressive vocabulary in children with autism spectrum disorder. *Autism & Developmental Language Impairments*, *7*, 23969415221085827.

Ma, W., Golinkoff, R. M., Hirsh-Pasek, K., McDonough, C., & Tardif, T. (2009). Imageability predicts the age of acquisition of verbs in chinese children. *Journal of child language*, *36*(2), 405–423.

MacWhinney, B. (2000). *The childes project: Tools for analyzing talk: Transcription format and programs*. Lawrence Erlbaum Associates Publishers.

Masterson, J., Druks, J., & Gallienne, D. (2008). Object and action picture naming in three-and five-year-old children. *Journal of Child Language*, *35*(2), 373–402.

Muraki, E. J., Siddiqui, I. A., & Pexman, P. M. (2022). Quantifying children's sensorimotor experience: Child body–object interaction ratings for 3359 english words. *Behavior Research Methods*, *54*(6), 2864–2877.

Perry, L. K., Perlman, M., & Lupyan, G. (2015). Iconicity in english and spanish and its relation to lexical category and age of acquisition. *PloS one*, *10*(9), e0137147.

Sailor, K. M. (2013). Is vocabulary growth influenced by the relations among words in a language learner's vocabulary? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(5), 1657.

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, *29*(1), 41–78.

Verhagen, J., Van Stiphout, M., & Elma, B. (2022). Determinants of early lexical acquisition: Effects of word-and child-level factors on dutch children's acquisition of words. *Journal of Child Language*, *49*(6), 1193–1213.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *nature*, *393*(6684), 440–442.