Word Learning as Network Growth: A Cross-linguistic Analysis

Abdellah Fourtassi
afourtas@stanford.edu
Department of Psychology
Stanford University

Yuan Bian
ybian.uiuc@gmail.com
Department of Psychology
University of Illinois

Michael C. Frank
mcfrank@stanford.edu
Department of Psychology
Stanford University

Abstract

Children tend to produce words earlier when they are connected to a variety of other words along both the phonological and semantic dimensions. Though this connectivity effect has been extensively documented, little is known about the underlying developmental mechanism. One view suggests that learning is primarily driven by a network growth model where highly connected words in the child's early lexicon attract similar words. Another view suggests that learning is driven by highly connected word in the external learning environment instead of highly connected words in the early internal lexicon. The present study tests both scenarios systematically in both the phonological and the semantic domains, and across 8 languages. We show that external connectivity in the learning environment drives growth in both the semantic and the phonological networks, and that this pattern is consistent cross-linguistically. The findings suggest a word learning mechanism where children harness their statistical learning abilities to (indirectly) detect and learn highly connected words in the learning environment.

Keywords: semantic network, phonological network, network growth, mechanism of word learning

Introduction

What factors shape vocabulary learning over the course of early childhood? To investigate this question, scientists have adopted multiple research strategies, from conducting controlled laboratory experiments (e.g. Markman, 1990) to analysing dense corpora capturing language learning in context (e.g., B. C. Roy, Frank, DeCamp, Miller, & Roy, 2015). One strategy consists in documenting the timeline of words' acquisition, and studying the properties that make words easy or hard to learn. For example, within a lexical category, words that are more frequent in child-directed speech are acquired earlier (J. C. Goodman, Dale, & Li, 2008). Other factors include word length, the mean length of utterances in which the word occurs, and concreteness (see Braginsky, Yurovsky, Marchman, & Frank, 2016).

Besides these word-level properties, the lexical structure (that is, how words relate to each other) also influences the age of acquisition of words. The lexical structure is best characterized in terms of a network where each node represents a word in the vocabulary, and each link between two nodes represents a relationship between the corresponding pair of words. Previous studies investigated early vocabulary networks based on different word relations such as shared semantic features, target-cue relationships in free association norms, co-occurrence in child directed speech, and phonological similarity. These studies have found that children tend to produce earlier the words that have higher neighborhood density (i.e., high connectivity in the network) both at the phonological and the semantic level (Engelthaler & Hills, 2017; Hills, Maouene, Riordan, & Smith, 2010; Hills, Maouene, Maouene, Sheya, & Smith, 2009; Stella, Beckage, & Brede, 2017; Storkel, 2009).

While most studies focused on the static properties of the lexical network, a few have investigated the underlying developmental process. Steyvers & Tenenbaum (2005) suggested that the effect of connectivity is the consequence of how the lexical network gets constructed in the child's mind. According to this explanation, known as Preferential Attachment (PAT), highly connected words in the child's lexicon tend to "attract" more words over time, in a rich-get-richer scenario (Barabasi & Albert, 1999). In other words, what predicts word learning is the *internal* connectivity in the child's early lexicon. In contrast, Hills et al. (2009) suggested that what biases the learning is not the connectivity in the child's internal lexicon but, rather, *external* connectivity in the learning environment. They called this alternative explanation Preferential Acquisition (PAC). Figure 1 shows an illustration of both growth scenarios with the same simplified network. These two proposals represent two divergent ideas about the role of lexical networks in acquisition. On the PAT proposal, network structure is a causal factor in early word learning; in contrast, on the PAC approach, network structure is not internally represented and, therefore, might be an epiphenomenon of the statistics of the linguistic input (see discussion).

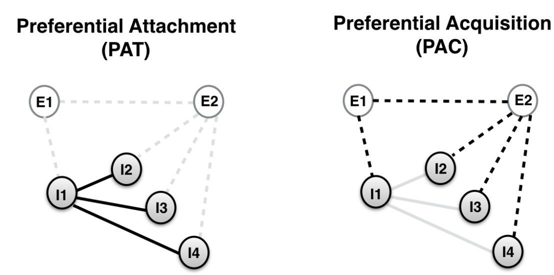

Figure 1: Illustration of the growth scenarios. Filled circles (I1-I4) represent known words (internal), and empty circles (E1 and E2) represent words that have not been learned yet (external). Black lines represent links that are relevant in each growth scenario, and gray lines represent links that are irrelevant. For PAT, the utility of a candidate, external node is the average degree (i.e., number of links) of the internal nodes that it would attach to. Thus, according to PAT, the node E1 is more likely to enter the lexicon first. For PAC, the utility of a candidate node is its degree in the entire network. According to PAC, the node E2 is more likely to enter the lexicon first.

Studies that investigate lexical network growth have focused on semantic networks using English data (Hills et al., 2010, 2009; Steyvers & Tenenbaum, 2005). The novelty of the current study is threefold: First, it investigates whether phonological networks, like semantic networks, grow by PAC, or if they rather grow by PAT. Second, it provides a systematic comparison of both network growth scenarios in the phonological and the semantic domains and assesses their relative contribution to the learning process. Third, it tests the generality of the findings across eight languages.

Networks

Data

We used data from Wordbank (Frank, Braginsky, Yurovsky, & Marchman, 2017), an open repository aggregating cross-linguistic language developmental data of the MacArthur-Bates Communicative Development Inventory (CDI), a parent report vocabulary checklist. Parent report is a reliable and valid measure of children's vocabulary that allows for the cost-effective collection of datasets large enough to test network-based models of acquisition (Fenson et al., 1994). We used the *Words and Sentence* version of the CDI which contains the productive vocabulary of toddlers (age varied between 16 to 36 months). Following previous studies (Hills et al., 2009; Storkel, 2009), we restricted our analysis to the category of nouns. We defined the age of acquisition of a given word by the month at which this word was produced by at least 50% of children (J. C. Goodman et al., 2008). We excluded nouns that have not been learned (according to this criterion) by the last month for which we have CDI data.

We obtained these nouns in eight languages: Croatian, Danish, English, Italian, Norwegian, Russian, Spanish, and Turkish. We used the subset of nouns that had entries in the Florida Association Norms (see below). Since these norms are available only in English, we used the hand-checked translation equivalents provided by Braginsky et al. (2016), allowing us to use the English association norms across languages. Table 1 gives an overview of the data used. Translation equivalents were originally constructed for a subset of words appearing on the toddler CDI form, and so not all words are currently available. Note, however, that all languages have at least 60% of nouns translated.

	language	total	translated	normed
1	Croatian	253	177	170
2	Danish	295	198	187
3	English	296	296	274
4	Italian	311	203	194
5	Norwegian	305	193	186
6	Russian	311	311	285
7	Spanish	240	173	163
8	Turkish	293	175	164

Table 1: Total number of nouns produced by toddlers in the CDI (left). We included in our study the subset of these nouns that had available English translations (middle). The final set consisted of nouns that had both available translations as well entries in the Free Association Norms (right).

Semantic networks

We constructed the semantic networks following the procedure outlined in Hills et al. (2009). We used as an index of semantic relatedness the Florida Free Association Norms (Nelson, McEvoy, & Schreiber, 1998). This dataset was collected by giving adult participants a word (the cue), and asking them to write the first word that comes to mind (the target). For example, when given the word "ball", they might answer with the word "game". A pair of nodes were connected by a directed link from the cue to the target if there was a cue-target relationship between these nodes in the association norms. The connectivity of a given node was characterized by its *indegree*: the number of links for which the word was the target. To model growth from month to month, we constructed a different network at each month, made of the words that have been acquired at that month.

Phonological networks

We generated approximate International Phonetic Alphabet (IPA) transcriptions from the orthographic transcription, across languages, using the open source text-to-speech software **Espeak.** We used the Levenshtein distance (also known as edit distance) as a measure of phonological relatedness between two nodes. The measure counts the minimum number of operations (insertions, deletions, substitutions) required to change one string into another.

In previous studies, two nodes were linked if they had an edit distance of 1 (e.g., Storkel, 2009). However, in these previous studies the network was built using an adult vocabulary. However, since the children's vocabulary contains very few word pairs with an edit distance of 1, the resulting networks were too sparse and uninformative. Thus, we increased the threshold from 1 to 2, that is, two nodes were related if their edit distance was equal to 1 or 2. The connectivity of a given node was characterized with its *degree*: the number of links it shares with other words.

Analysis

Static properties of the global network

We start by analysing word connectivity in the global (static) network. We constructed this network using nouns at the oldest age for which we have CDI data (e.g., in English this corresponds to the network by 30 months). This global network is the end-state towards which both PAT and PAC should converge by the last month of learning. Moreover, following Hills et al. (2009), we used this end-state network as a proxy for the external connectivity in the learning environment. Below we analyse properties of this global networks that are relevant to PAT and/or PAC.

Connectivity predicts the age of acquisition Connectivity in the global network is directly related to PAC as it represents the explicit criterion PAC uses to determine what words should be learned first (Figure 1). Therefore, a direct consequence of a PAC-like growth scenario is a correlation between connectivity in the global network and the age of acquisition.[1] Figure 2 shows how the age of acquisition for each word varies as a function of its degree (or indegree for the semantic network). For ease of visual comparison, the predictor (i.e., the degree) was centered and

[1]Note that this correlation is also compatible with PAT, even if this growth scenario does not rely explicitly on external connectivity. Indeed, from a PAT perspective, higher connectivity in the global network is caused by earlier learning, i.e., some words end up being highly connected in the global network precisely because they happen to be acquired earlier (and therefore have a higher chance of forming more links over time).

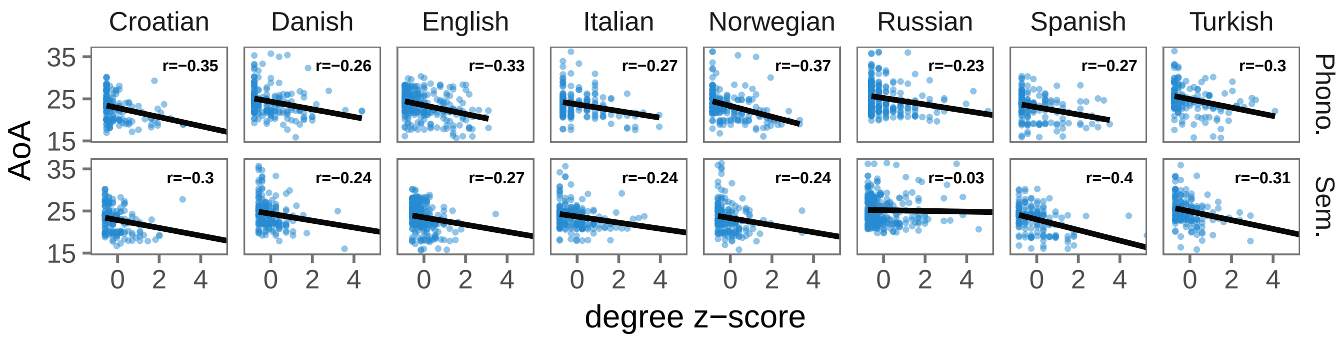

Figure 2: Age of acquisition in the global network as predicted by the degree in this network. Results are shown in each language for phonological and semantic networks. Each point is a word, with lines indicating linear model fits.

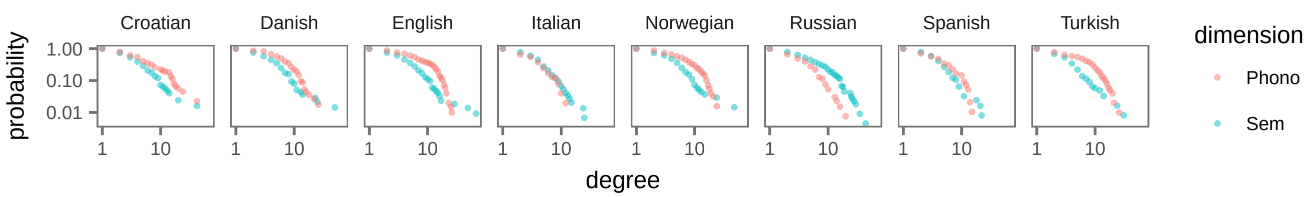

Figure 3: Log-log plot of the cumulative degree distribution function for the global phonological and semantic networks across languages. A perfect power-law distribution should appear as a straight line in this graph.

scaled across languages. The plots show, overall, a negative correlation between the month of acquisition and the degree, indicating that nouns with higher degrees are generally learned earlier.

Power-law degree distribution? We also analysed the global network's degree distribution. This property is particularly relevant to PAT as this growth scenario is known to generate networks with a power-law degree distribution (i.e., a distribution of the form $p(k) \propto \frac{1}{k^{\alpha}}$, Barabasi & Albert, 1999). If the network displays this property, this fact would suggests a PAT-like generative process. Conversely, if the degree distribution does not follow a power law, this fact would weaken the case for PAT. The log-log plots are shown in Figure 3. We fit a discrete power law to each empirical degree distribution following the procedure outlined in Clauset, Shalizi, & Newman (2009) and using the related R package (poweRlaw, Gillespie, 2015). In brief, the analysis consists in two steps. First, we derive the optimal cut-off, k_{min}, above which the distribution is more likely to follow a power law[2], and we estimate the corresponding scaling parameter α. Second we calculate the goodness-to-fit, which results in a p-value quantifying the plausibility of the model. Overall, we did not find strong evidence against a power-law distribution: the p-value was generally above 0.1, except for the Italian phonological network where we obtained $p < 0.05$ (suggesting that the power law can be ruled out in this particular case).

In sum, the static properties of the global network do not provide conclusive evidence in favor or against any of the growth scenarios in question. For a more direct test of the developmental process, we need to fit explicit growth models to the data.

[2]In natural phenomena, it is often the case that the power law applies only for values above a certain minimum.

Network growth models

How does each growth scenario predict noun development? To test the network growth scenarios, we fit different growth models to the data. We calculated the probability that a word w_i, with a growth value d_i would enter the lexicon at a given month, using a softmax function:

$$p(w_i) = \frac{e^{\beta d_i}}{\sum_j e^{\beta d_j}} \quad (1)$$

where β is a fitted parameter that captures the magnitude of the relationship between network parameters and growth (analogous to a regression coefficient). A positive value of β means that words with higher growth values d_i are acquired first, and a negative value means that words with lower growth values are acquired first (see Figure 1 for an illustration of how growth values d_i are defined in each growth scenario). The normalization includes all words that could be learned at that month.

We estimated the parameter β using a Bayesian approach. The inference was performed using the probabilistic programming language WebPPL (N. Goodman & Stuhlmller, 2014). We defined a uniform prior over β, and at each month, we computed the likelihood function over words that could possibly enter the lexicon at that month, fit to the words that have been learned at that month (using formula 1). Markov Chain Monte Carlo sampling resulted in a posterior distribution over β, which we summarized in Figure 4.

Besides fitting a growth model to the data, we conducted a separate evaluation. This second evaluation consists in determining, in each month, the growth value distribution of all words that could possibly be learned at this month, and then computing the z-score of each learned word with respect

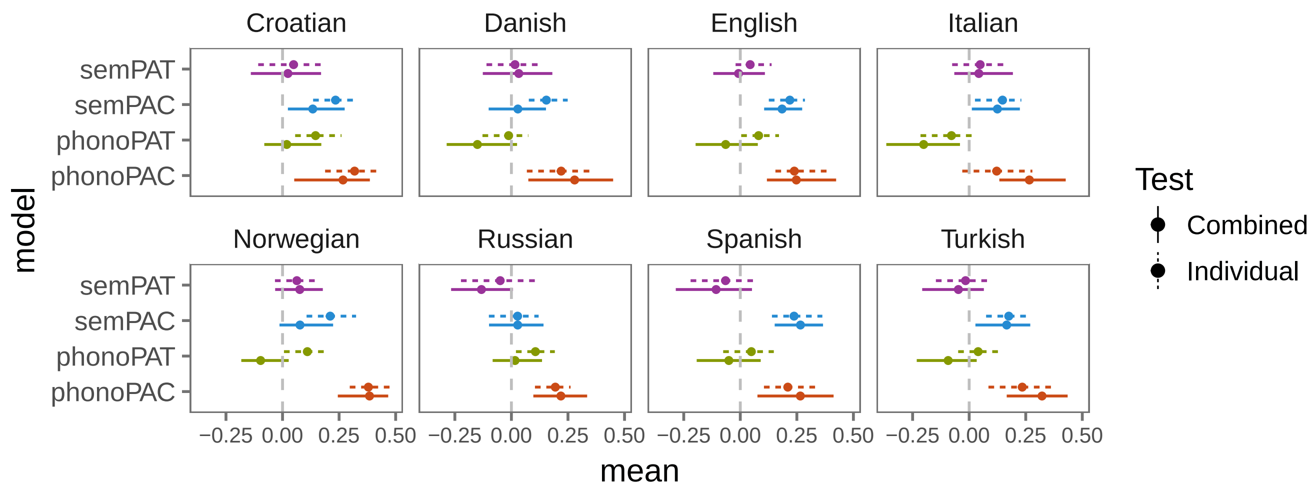

Figure 4: Evaluation of network growth scenarios both individually (dotted), and when combined in the same growth model (solid). Each dot represents the mean of the posterior distribution of the corresponding growth parameter, with ranges representing 95% credible intervals (computed using the highest density intervals). Positive values mean that learning proceeds according to the predictions of the growth scenario. Negative values mean that learning proceeds in opposition to the predictions of the growth scenario.

to this distribution. For each growth scenario, we tested if the distribution constituted by the z-scores of all learned words was different from zero, using a one-sample t-test.

The results from both evaluations were very similar and lead essentially to the same conclusions.[3] For the semantic networks, the results replicate Hills et al.'s finding in English, which was that the semantic network grows by PAC, not by PAT. Moreover, this finding held in seven of the eight languages we examined. The PAC model also fit better than PAT for phonological networks. We note however that PAT, though weaker, fares better for the phonological networks (where it predicts part of the growth process in some languages such as Croatian, English, Norwegian and Russian) than it does for the semantic networks (where it is rather universally unpredictive).

What is the relative contribution of each growth model? Above we evaluated the network growth scenarios individually. As a next step, we analysed their relative contribution to the learning process. This was done through adding more fitted parameters to the model, that is, by substituting βd_i in formula (1) with:

$$\beta_1 d_{i,1} + \beta_2 d_{i,2} + \beta_3 d_{i,3} + \beta_4 d_{i,4}$$

where the indices represent the 4 networks: semPAT, semPAC, phonoPAT and PhonoPAC. Using the same fitting technique, we obtained the values shown in Figure 4. PAC dominates the learning. Both phonological and semantic networks contribute to lexical growth, but the phonological network appears to be stronger and more consistent across languages. In summary, the findings show that both semantic and phonological networks contribute to the learning process, and that they both grow primarily by PAC, relying on the external connectivity in the learning environment, rather than the internal connectivity in the acquired lexicon.

[3]we do not show here the results of the second evaluation because they were redundant with the results of the first evaluation

Comparison to other known predictors of age of acquisition

We saw that the way semantic and phonological information is structured in the learning environment (i.e., PAC) contributes to noun learning across languages. However, we know that other factors influence learning as well (e.g., Braginsky et al., 2016). Next we investigated how semantic and phonological connectivity interact with two other factors. The first one is word frequency, a well studied factor shown to predict the age of acquisition in a reliable fashion (e.g. J. C. Goodman et al., 2008). The second factor is word length, which correlates with phonological connectivity.

Since PAT was uninformative, we dropped it from this analysis, keeping only PAC. This simplified the model because we no longer needed to fit growth month-by-month.[4] A more direct way to assess and compare the contribution of PAC in relation to other word-level factors is through conducting linear regressions, where connectivity in the learning environment, frequency and length predict the age of acquisition.

We used the frequency estimates from Braginsky et al. (2016) where unigram counts were derived based on CHILDES corpora in each language.[5] For each word, counts included words that shared the same stem (e.g., cats counts as cat), or words that were synonymous (e.g. father counts as daddy). For word length, we used our generated IPA transcription.

We conducted two analyses. We fit a linear regression for each language, and we fit a linear mixed-effect model to all the data pooled across languages, with language as a random effect. Figure 5 shows the coefficient estimate for each predictor in each language, and Figure 6 shows the coefficient estimates for all

[4]This was a requirement only for PAT whose utilities varied from month to month, depending on previously learned words.

[5]Note that these frequency counts are based on transcripts from independent sets of children and represent a general estimate of environmental frequency across children

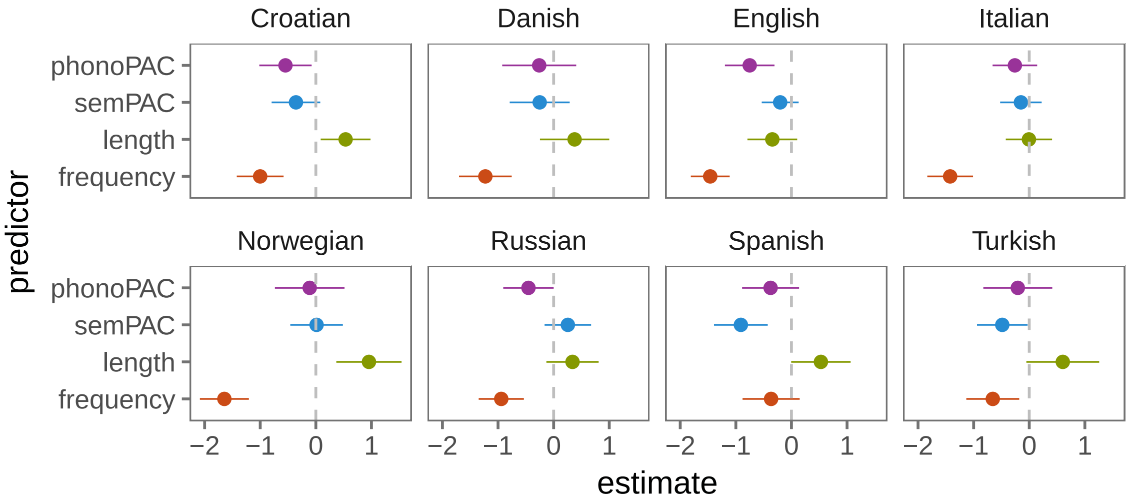

Figure 5: Estimates of predictor coefficients by language, with ranges indicating 95% confidence intervals. Positive values indicate a positive relationship (e.g. longer words tend to have a higher AoA), while negative values indicate a negative relationship (e.g. words with higher frequency tend to have a lower AoA).

languages combined (all predictors were centered and scaled). The findings were as follows. Overall, frequency is the largest and most consistent predictor of age of acquisition, replicating results for nouns across a variety of analyses (Braginsky et al., 2016; J. C. Goodman et al., 2008; B. C. Roy et al., 2015). Word length predicts learning in some languages such as Croatian and Norwegian, but not in others (including English). It remains, however, a significant predictor in the global model. As for the factors of interest, i.e., semantic and phonological connectivity, we also found cross-linguistic differences. Phonological connectivity contributes to learning in languages such as Croatian, English and Russian, whereas semantic connectivity contributes to learning in Turkish, Spanish and to some extent in Croatian, but not in English.[6] Despite this cross-linguistic variation, both phonological and semantic connectivity remain significant predictors in the global model.

Discussion

The present study provided a comprehensive analysis of how lexical connectivity influences the age of acquisition of nouns in toddlers. We compared two network growth scenarios and assessed their relative contributions across eight languages. One scenario, PAT, described a "rich get richer" network growth model in which the structure of the learner's internal network determines future growth; the other, PAC, described a model in which the external, global environmental network structure determines learners' growth patterns. Part of the findings largely replicates the results obtained in Hills et al. (2009), i.e., semantic networks (based on free associations) grow by preferential acquisition, not by preferential attachment. A novel finding was that phonological networks also grow primarily by preferential acquisition, especially when both scenarios (PAT and PAC) were pitted against each other in the same model. These findings generalize well across languages. Moreover, both semantic and phonological connectivity in the learning environment (i.e., PAC) predict growth in a consistent way across many languages. Nevertheless, when pitted against other known predictors of age of acquisition (word frequency and length), the effect of word connectivity shows a cross-linguistic variation, predicting learning in some languages, but not in others. Despite this cross-linguistic variation, both phonological and semantic connectivity contribute to the overall learning (when data is pooled across languages).

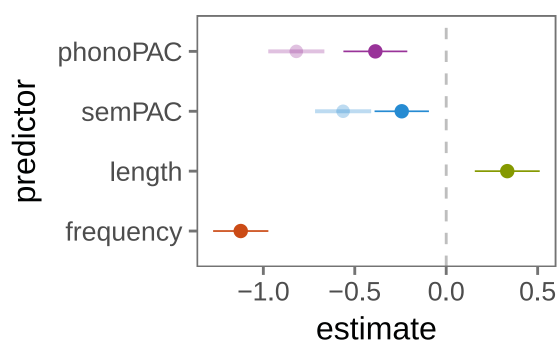

Figure 6: Estimates of predictor coefficients in the combined mixed-effect model with language as a random effect. Ranges indicate 95% confidence intervals. Lighter points indicate estimates of PAC predictors in a model that does not include frequency and length as covariates.

One important result of the study is that children start by learning words that have high phonological and semantic similarity to a variety of other words in the learning environment, not in the child's available lexicon. This suggests that children are sensitive to connectivity even without having first acquired the connected

[6] Semantic connectivity does not explain variance in English data beyond that explained by phonological connectivity, frequency and length. This contrasts with the original finding in Hills et al. 2009. However, in this previous study, semantic connectivity was not tested in a model that included frequency, length and phonological connectivity as covariates. Another important difference is the number of words tested: our study uses a larger set of nouns.

words. How can children indirectly detect highly connected words, and why would such words be more readily learned?

In the semantic case, free association can be predicted through the patterns of word co-occurrence (Griffiths, Steyvers, & Tenenbaum, 2007), meaning that highly connected words tend to be the words that co-occur with many other words in various contexts. One possibility, suggested by Hills et al. (2010), is that the referents of such words are more easily disambiguated from other potential referents because their presence in multiple contexts provides more cross-situational, disambiguating statistics about their true referents (Smith & Yu, 2008).

In the phonological case, connectivity is inherently correlated with phonotactic probability (Vitevitch, Luce, Pisoni, & Auer, 1999). That is, highly connected words tend to be made of frequent sound sequences. Even infants (whose vocabulary is still very rudimentary) show a sensitivity for high frequency sound sequences in the ambient language (Jusczyk, Luce, & Charles-Luce, 1994). Moreover, phonotactic probability facilitates learning and recognition (e.g., Storkel, 2001). In other words, children's sensitivity to local phonotactic regularities might lead them to learn higher-probability words more easily. This learning effect, in turn, would lead to an observed pattern of growth that would appear to follow the PAC growth model even though learners themselves would only be tracking local statistics.

Finally, while validating previous results using network growth models, our study suggests that these correlational patterns may emerge from the operation of simpler mechanisms in both the semantic and phonological domains. One question for future experimental work is whether such patterns of growth can be produced in controlled behavioral experiments.

All data and code for these analyses are available at `https://github.com/afourtassi/networks`

Acknowledgements

This work was supported by a post-doctoral grant from the Fyssen Foundation, NSF #1528526, and NSF #1659585.

References

Barabasi, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509–512.

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2016). From uh-oh to tomorrow: Predicting age of acquisition for early words across languages. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*.

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, *51*(4), 661–703.

Engelthaler, T., & Hills, T. T. (2017). Feature biases in early word learning: Network distinctiveness predicts age of acquisition. *Cognitive Science*, *41*, 120–140.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., ... Stiles, J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, *59*(5), i–185.

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.

Gillespie, C. S. (2015). Fitting heavy tailed distributions: The poweRlaw package. *Journal of Statistical Software*, *64*(2), 1–16. Retrieved from `http://www.jstatsoft.org/v64/i02/`

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language*, *35*(3), 515–531.

Goodman, N., & Stuhlmller, A. (2014). The Design and Implementation of Probabilistic Programming Languages. `http://dippl.org`.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 2007.

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.

Jusczyk, P. W., Luce, P. A., & Charles-Luce, J. (1994). Infant's sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, *33*(5), 630–645.

Markman, E. M. (1990). Constraints children place on word meanings. *Cognitive Science*, *14*(1), 57–77.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. Retrieved from `http://w3.usf.edu/FreeAssociation/`

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*, *112*(41), 12663–12668. `http://doi.org/10.1073/pnas.1419773112`

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568.

Stella, M., Beckage, N. M., & Brede, M. (2017). Multiplex lexical networks reveal patterns in early word acquisition in children. *Scientific Reports*, *7*.

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.

Storkel, H. L. (2001). Learning new words: Phonotactic probability in language development. *Journal of Speech, Language, and Hearing Research*, *44*(6), 1321–1337.

Storkel, H. L. (2009). Developmental differences in the effects of phonological, lexical and semantic variables on word learning by infants. *Journal of Child Language*, *36*(2), 29–321.

Vitevitch, M. S., Luce, P. A., Pisoni, D. B., & Auer, E. T. (1999). Phonotactics, neighborhood activation, and lexical access for spoken words. *Brain and Language*, *68*(1), 306–311.