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 first the words that are connected to a variety of other words along both the phonological and the 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 ealy lexicon attract similar words. Another view suggests that learning is driven, not by highly connected words in the early internal lexicon, but by highly connected word in the external learning environment. 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 controlled laboratory experiments to corpus analysis. 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 how concrete the word is (see Braginsky, Yurovsky, Marchman, & Frank, 2016).

Besides these word-level properties, researchers found that 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 relationship in free association norms, co-occurrence in child directed speech, and phonological similarity. These studies 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 (Beckage, Smith, & Hills, 2011; Engelthaler & Hills, 2017; Hills, Maouene, Riordan, & Smith, 2010; Hills, Maouene, Maouene, Sheya, & Smith, 2009; Stella, Beckage, & Brede, 2017; Stokes, 2010; Storkel, 2009).

A crucial question that naturally emerges from these analyses is how does network connectivity influence word learning? 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 *internal* connectivity in the child's early lexicon. In contrast, Hills et al. (2009) found 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.

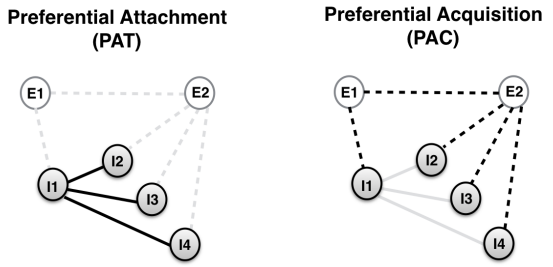

Figure 1: Illustration of the two growth models with the same network. Filled circles (I1-I4) represent known words (internal lexicon), and empty circles (E1 and E2) represent words that have not been learned yet (external lexicon). Black lines represent links that are relevant in each growth scenario,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

Studies that investigated lexical network growth focused on semantic networks using English data only (Hills et al., 2010, 2009; Steyvers & Tenenbaum, 2005). The novelty of the current study is that it investigates whether phonological networks, like semantic networks, grow by PAC, or if they rather grow by PAT. It also provides a systematic compari-

son of network growth scenarios in the phonological and the semantic domains and assesses their relative contribution to the learning process. Moreover, it tests the generality of the findings across eight languages.

The paper is organized as follows. First, we describe the dataset we used, the procedure we followed to construct both semantic and phonological networks. Next, we show the results of the various analyses we conducted comparing network growth scenarios across languages. Then we test the overall contribution of the networks in the learning process relative to other known predictors of age of acquisition (frequency and length). Finally we discuss the major results in relation to known experimental facts in language development, and we speculate about the potential learning mechanism that might drive early vocabulary learning.

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. 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; e.g., Storkel, 2009), we restricted our analysis to the category of nouns. The age of acquisition was defined by the month at which a word was produced by at least 50% of children (J. C. Goodman et al., 2008).

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 translation of non-English words provided by Braginsky et al. (2016). This allowed us to use the English association norms across languages. Table 1 gives an overview of the data used. The translation of non-English words is still an ongoing process. 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 productive nouns in the CDI (left). We used a 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 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. Each node was characterized by its "in-degree", which represents 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 words that have been learned at that month. Following Hills et al. (2009), we used as a proxy for the learning environment, the network constructed using the full set of nouns in the CDI data in each language (e.g., in English this corresponds to the network by 30 months).

Phonological networks

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

In previous studies, two nodes were linked if they had an edit distance of 1 (Stokes, 2010; Storkel, 2009). However, in these previous studies the network was built using an adult vocabulary. Here, similar to the approximation we made for the semantic network, we used the full set of nouns in the CDI data as a proxy for phonological connectivity in the learning environment. However, since the children's vocabulary contains very few word pairs with an edit distance of 1, the resulting network was too sparse and uninformative. Thus, we increased the threshold from 1 to 2, that is, two node were related if their edit distance was equal to 1 or 2. Each node

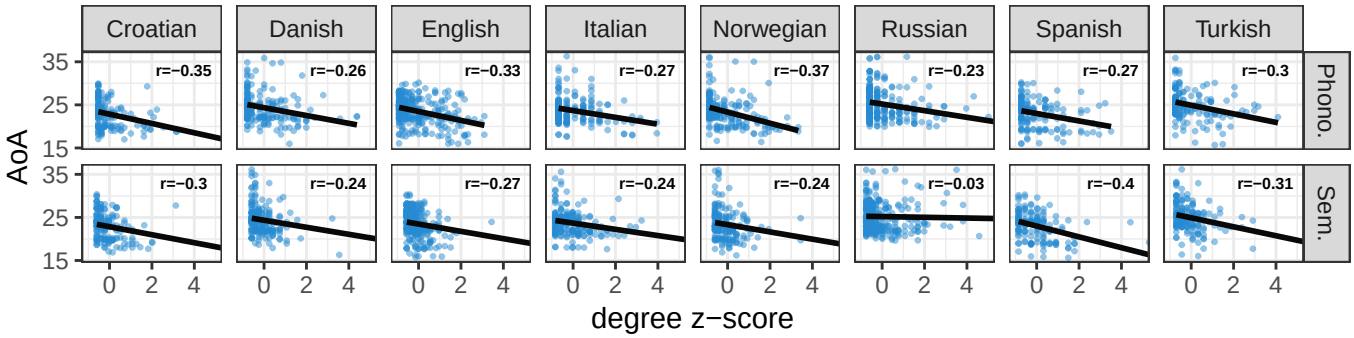

Figure 2: Age of acquisition in the end-state network as predicted by the degree in this network. Results are shown in each language for the phonological network (top) and the semantic network (bottom). Each point is a word, with lines indicating linear model fits.

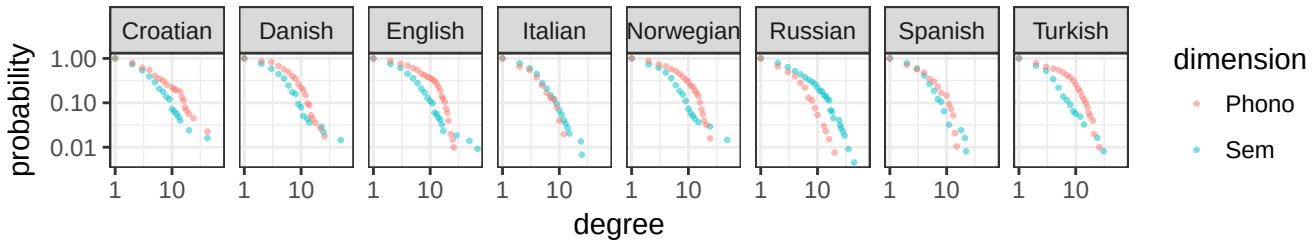

Figure 3: Log-log plot of the cumulative degree distribution function for the global phonological and semantic networks across languages. A perfect power law distribution should appear as a straight line in this graph.

was characterized with its degree, i.e., the number of links it shares with other words.

Analysis

Preliminary analysis

We start by examining how connectivity in the global network influences the age of acquisition. Regardless of the growth mechanism that might have given rise to this network (i.e., PAC or PAT), we expect nodes with the higher degrees in this network to be learned earlier than nodes with the lower degrees. The global network was constructed using nouns at the oldest age for which we have CDI data. 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 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.

We also analysed the global network degree distribution. PAT is known to generate networks with a power-law degree distribution (Barabasi & Albert, 1999). Thus, if the networks display this property, it means that PAT cannot be ruled out *a priori* as a potential growth mechanism. In contrast, if there is strong evidence against a power law distribution, then PAT (but not PAC) can be ruled out. The distributions are shown in Figure 3. Overall, the results do not provide strong evidence against a power-law distribution. Moreover, the degree to which the distribution approximates a power law varies across dimension and languages. For a more conclusive test, we fit explicit growth models to the data (next section).

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 proceeded as follows. 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 the fitting parameter. 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. As explained in Figure 1, each candidate word w_i is characterized with a growth values d_i that depends on the growth scenario tested. In PAT, the growth value of a given word is equal to the average degree of the existing nodes that it would attach to. In PAC, the growth value of a word is equal to its degree in the entire network. 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,

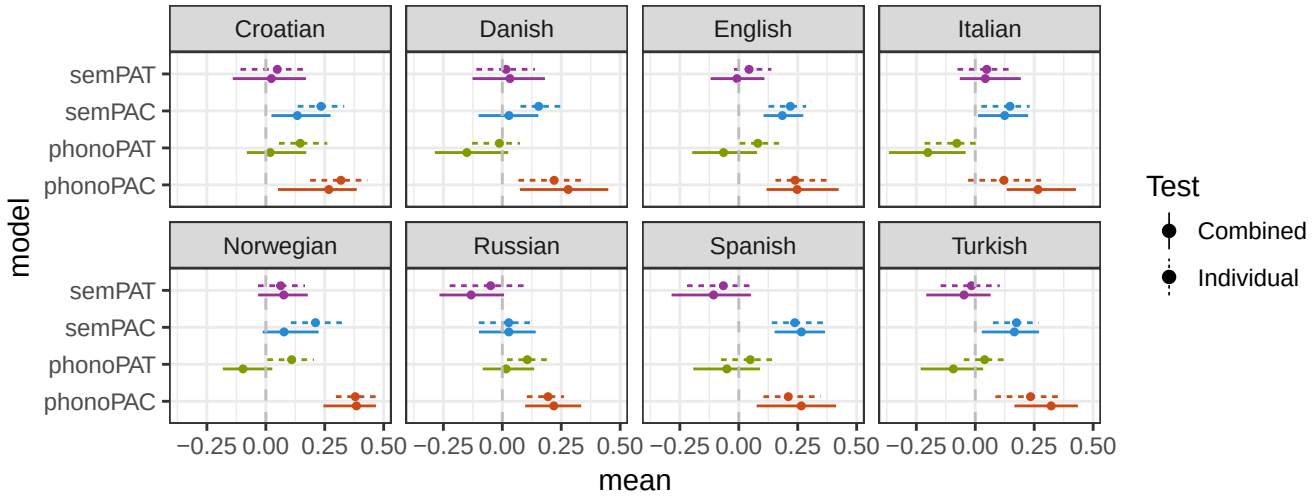

Figure 4: Evaluation of network growth scenarios both individually (dotted), and when combined in the same growth model (solid). Each dot represents the mean of the posterior distribution of the corresponding growth parameter, with ranges representing 95% credible intervals (computed using the highest density intervals).

2014). We started with a uniform distribution, 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 actually learned at that month (using formula 1). We ended up with a posterior distribution of β, which we summarized in Figure 4.

Besides fitting a growth model to the data, we conducted another evaluation. This second evaluation consists in determining, in each month, the model-dependent 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 to this distribution. For each model, 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[1]. For the semantic network, the results replicate Hills et al.'s finding in English, which was that the semantic network grows by PAC, not by PAT. Moreover, it generalizes this finding to all other languages (with the exception of Russian). For the phonological network, the results show that, generally speaking, PAC again fits the developmental trajectory better than PAT. 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 fitting 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. In term of growth scenarios, we found that PAC dominates the learning. In particular, though we found previously that phonological PAT predicted part of the learning in some languages when tested individually, here PAT appears to lose its predictive power when pitted against phonological PAC. In terms of domain, both phonological and semantic networks appear to contribute to learning, although the phonological network appears to be stronger and more reliable across languages. In summary, the findings show that both semantic and phonological networks grow primarily by PAC, and that, generally speaking, semantic and phonological networks both contribute to the learning process.

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) contribute to noun learning across languages. However, we know that other factors influence learning as well (e.g., Braginsky et al., 2016). In what follows, we will investigate 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. Thus, we no longer needed to fit the growth model month-by-month as in the previous section. In fact, word utilities in the case of PAC are fixed, they do not depend on

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

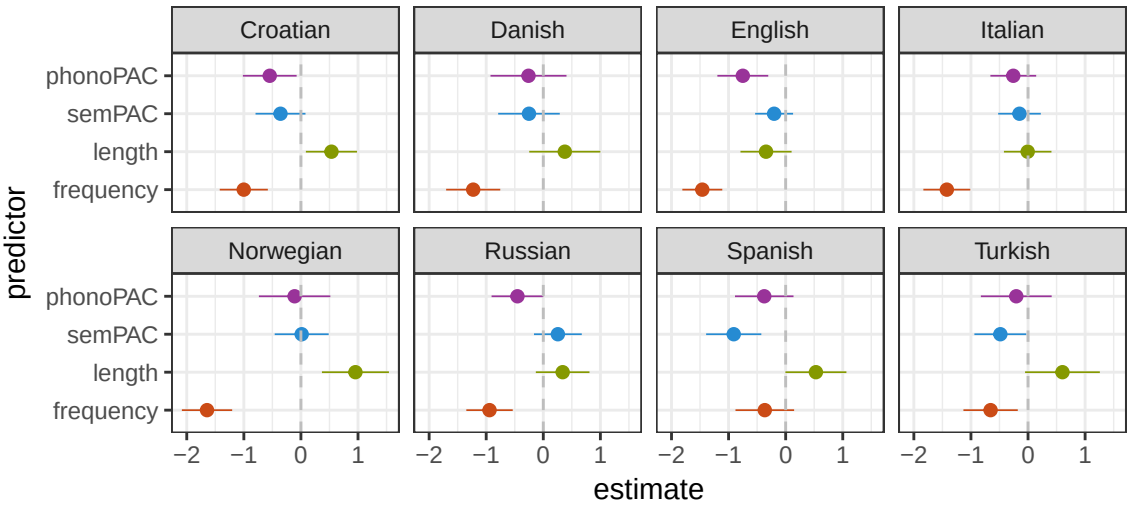

Figure 5: Estimates of predictor coefficients by language. Values above 0 indicate a positive relationship (e.g. longer words tend to have a higher AoA), while values below 0 indicate a negative relationship (e.g. words with higher frequency tend to have a lower AoA). Ranges indicate 95% confidence intervals.

previously learned words. A more direct way to assess and compare the contribution of PAC in relation to other factors is through conducting linear regressions, where connectivity in the learning environment (at both the phonological and semantic level), 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. 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 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 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. 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. The 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 interestingly not in English[2]. 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 the relative contribution of phonological and semantic information in 8 languages. Part of the findings largely replicate the results obtained in Hills et al. (2009), i.e., semantic networks (based on free associations) grow by preferential acquisition, not by preferential attachment. Another 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. However, 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).

[2]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 both frequency and length as covariates. Another important difference is the number of words tested. Our study uses a larger set of nouns.

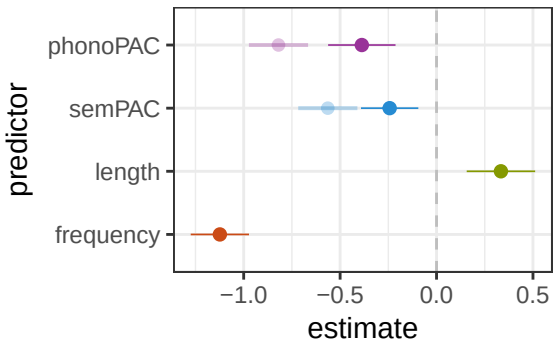

Figure 6: Estimates of predictor coefficients in the combined model pooling across languages . Ranges indicate 95% confidence intervals. Fade color indicates estimates of PAC predictors in a model that does not include frequency and length as covariates.

A major results of the study is that children start by learning words that have high phonological and semantic similarity with 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 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. Such effect was demonstrated by cross-situational learning experiments (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. We know even infant (whose vocabulary is still very rudimentary) develop a sensitivity for high frequency sound sequences in the ambient language (Jusczyk, Luce, & Charles-Luce, 1994). Interestingly, it was shown that phonotactic probability facilitates learning and recognition of novel words in toddlers and preschoolers (MacRoy-Higgins, Shafer, Schwartz, & Marton, 2014; Storkel, 2001). Thus, children's ability to keep track of co-occurrence statistics (both at the semantic and phonological level) might explain their sensitivity and preference for highly connected words in the learning environment.

Finally, this study shares a number of limitations with previous studies using similar datasets. In particular, the results provide correlational but not causal evidence. Thus, the conclusions of this study require parallel evidence, especially from 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.

References

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

Beckage, N. M., Smith, L., & Hills, T. T. (2011). Small worlds and semantic network growth in typical and late talkers. *PLOS ONE*, *6*(5), 1–6.

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*.

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

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.

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.

MacRoy-Higgins, M., Shafer, V. L., Schwartz, R. G., & Marton, K. (2014). The influence of phonotactic probability on word recognition in toddlers. *Child Language Teaching and Therapy*, *30*(1), 117–130.

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/`

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.

Stokes, S. F. (2010). Neighborhood density and word frequency predict vocabulary size in toddlers. *Journal of Speech, Language, and Hearing Research*, *53*(3), 670–683.

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.