The Role of Early Semantic and Phonological Networks in Noun Learning: 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 learn first the words that have a high density neighborhood in the lexical network (i.e., words that are connected to a variety of other words). However, the reason why connectivity influences learning is still not very well understood. Previous studies suggested this effect could be the consequence of the underlying learning mechanism. However, these studies focused on semantic networks and English data only, thus severely limiting the generality of the findings. Here we provide a systematic analysis of network growth models in both the phonological and the the semantic domains, and across 8 languages. We replicate the results obtained by Hills et al. (2009) with English semantic networks, show that the same mechanism drives the growth of both semantic and phonological networks, and demonstrate that these effects hold cross-linguistically.

Keywords: keywords: semantic network, phonological network, word learning, semantic

Introduction

A crucial question in language development is to understand why children learn some words at an earlier age than others. Previous studies documented various word properties that influence the age of acquisition. For example, it was shown that, within a lexical category, words that are more frequent in child-directed speech are acquired earlier (Goodman, Dale, & Li, 2008). Other predictors 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, the network that specifies how words relate to each other) also influences the learning process. For instance, children tend to produce first the words that have higher neighborhood density (i.e., high connectivity in the network) both at the phonological and the semantic level (Stokes, 2010; Storkel, 2009). However, it is still not very well understood *why* connectivity in the lexical structure influences learning. Steyvers & Tenenbaum (2005) suggested that this effect is nothing but the consequence of how these structures get constructed in the child's mind. According to this explanation (known as Preferential Attachment), highly connected words in the child's lexicon tend to "attract" more words over time, in a rich-get-richer scenario. Nonetheless, Hills, Maouene, Maouene, Sheya, & Smith (2009) found that what biases the learning is not the connectivity in the child's internal lexicon but the connectivity in the learning environment.

These analyses, however, focused on growth in semantic networks only. A recent study by Stella, Beckage, & Brede (2017) analysed both semantic and phonological connectivity, but assumed only one growth model (preferential acquisition). More importantly, almost all available studies are limited in their generality because they focus almost exlusively on English data. The novely of this study is that it investigtes, for the first time, whether phonological networks grow by preferential attachement or preferential acquisition. It also provides a systematic comparision between network growth scenarios in both the phonological and the semantic domains and assess 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, and the network growth scenarios we tested. Next, we show the results of the various analyses we conducted comparing network growth scenarios across languages. Then we compare 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.

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). The latter consists of parent report vocabulary checklists. We used the *Words and Sentence* version which contains the productive vocabulary of toddlers (age varied between 16 to 35 months). We included words that were produced by at least 50% of children, and we considered that these words were learned at the month when this percentage was reached (Goodman et al., 2008). Moreover, following previous studies (Hills et al., 2009; e.g., Storkel, 2009), we restricted the analysis to the category of nouns, since there are known discrepancies between the learning of nouns, predicates, and function words (e.g., Braginsky et al., 2016). 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 mapped words onto translation equivalents across CDI forms. This allowed us to use the English as-

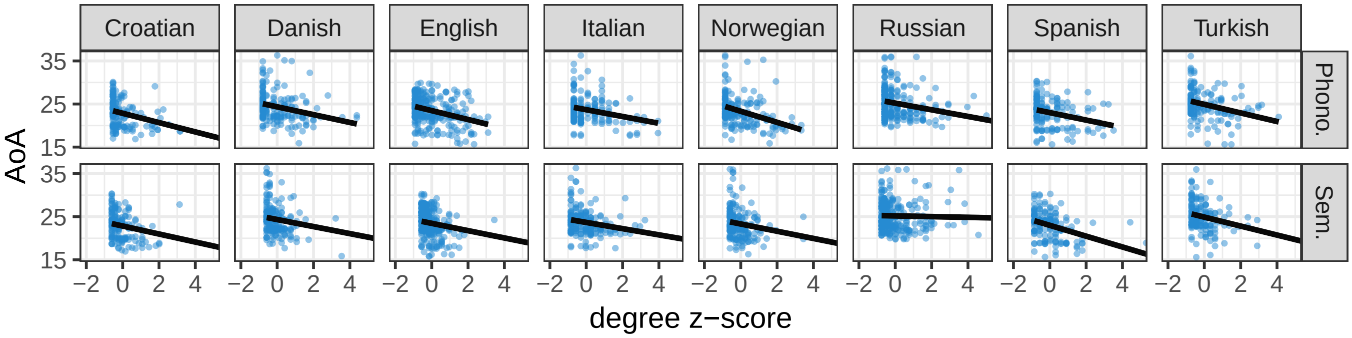


Figure 1: Age of acquisition in the end-state network as predicted by the degree in this network. Results are shown in each language for the phonological network (top) and the semantic network (bottom). Each point is a word, with lines indicating linear model fits.

sociation 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 have available English translations (middle), as well as entries in the Free Association Norms (right).

With this data, we constructed both semantic and phonological networks.

Semantic networks

We used as 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. Following Hills et al. (2009), we used as a proxy for the learning environment, the network constructed using the full 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 IPA forms 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 was characterized with its degree, i.e., the number of links it shares with other words.

Network growth

We tested two network growth scenarios discussed in Hills et al. (2009). The first one was *Preferential Attachment* (PAT). According to this mechanism, the network structure of known words predicts what words will be learned next. A word is more likely to be learned if it is linked to one of the highly connected nodes in this internal network. The second mechanism was *Preferential Acquisition* (PAC). In this model, what predicts what word will enter the lexicon is not connectivity in the internal network, but connectivity in the learning environment. Figure 3 shows an illustration of both growth scenarios with the same simplified network. For PAT, each candidate node was characterized with the average degree of the existing nodes that it would attach to. Thus, according to PAT, the node N1 is more likely to enter the lexicon first. For PAC, each candidate node was characterized with its degree in the entire network. According th PAC, the node N3 is more likely to enter the lexicon first.

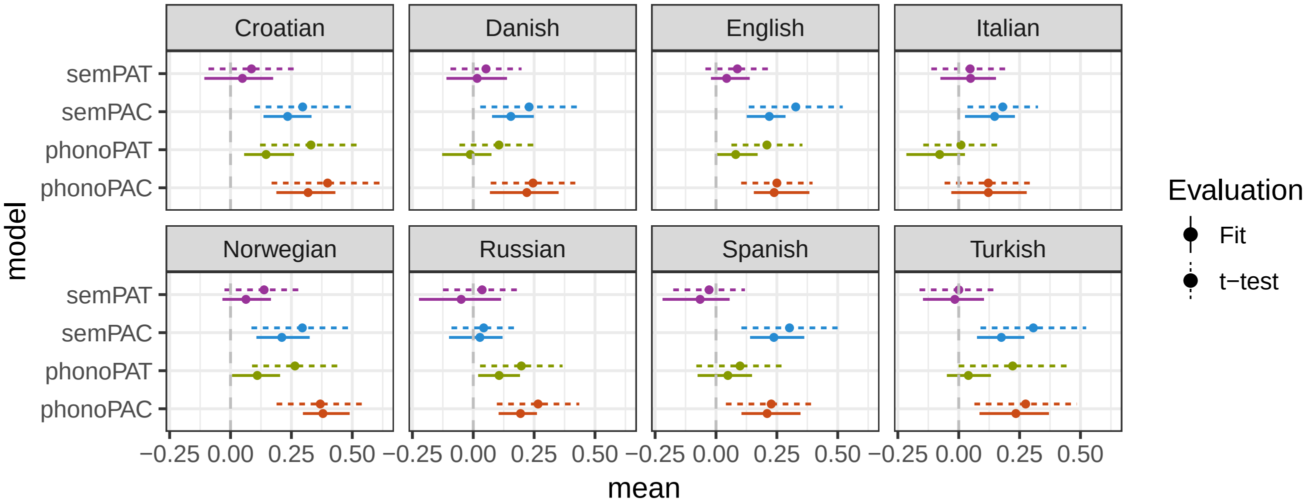
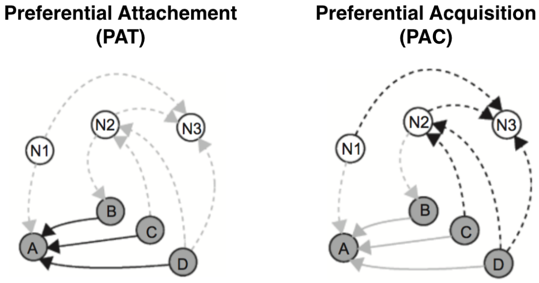


Figure 2: Evaluation of individual network growth models using both one-sample t-test over the z-score distribution of learned words (dotted), and model fitting (solid). Each dot represents, in the first case, the mean of the z-score distribution of the learned words with ranges representing 95% confidence interval, and in the second case, the mean of the posterior distribution of the model's parameter, with ranges reppresenting 95% credible intervals (computed using the highest density intervals).



Preferential Attachement (PAT)
Preferential Acquisition (PAC)
N1
N2
N3
A
B
C
D

Figure 3: Illustration of the two growth models with the same network. Filled circles (A to D) represent known words, and empty circles (N1 to N3) represent words that have not been learned yet. Back lines represent links that are relevant in each growth scenario, and gray lines represent links that are irrelevant (from Hills et al. 2009).

Analysis

Connectivity vs. Age of acquisition

Preferential attachment and preferential acquisition provide different generative scenarios for the same end-state. The latter is defined here as the network corresponding to the oldest age for which we have CDI data. However in both cases, it is predicted that nodes with the higher degrees in this end-state are learned earlier than nodes with the lower degrees. Figure 1 shows how the age of acquisition for each word varies as a function of its degree (or indegree for the semantic network) in the end-state network. In order to compare this correlation across languages, the predictor (i.e., the degree) was centered and scaled. The plot shows a negative correlation between the month of acquisition and the degree (with the exception of the Russian semantic network), indicating that nouns with higher degrees are generally learned earlier.

Network growth models

Following Hills et al. (2009), we evaluated the growth models in two separate ways. The first 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 are shown in Figure 2.

The second evaluation consists in fitting an explicit model to the data. The purpose of this second evaluation is, first, to validate the findings obtained with the one-sample t-test, and second, to determine (in a next step) the relative contribution of each network growth scenario (PAT vs. PAC) across domains (phonological vs. semantic).

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:

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

Where β is the fitting parameter. A positive value of β means that words with higher groth values d_i are acquired first, and a negative value means that words with lower growth values are acquired first. The normalization included words that were not yet learned at the month the word at hand was acquired. We estimated the parameter β using a Bayesian approach. 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

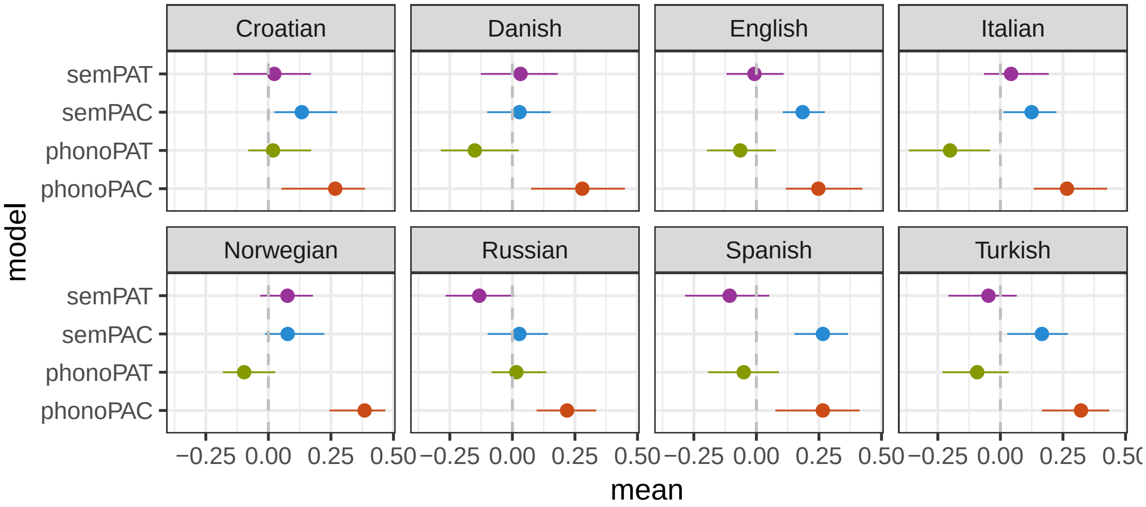


Figure 4: Estimates of the growth models' parameters in each language. Dots represent the means of the posteriori distribution of the parameters, and ranges represent 95% credible intervals (computed using the highest density intervals).

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 2.

The results from both evaluations were very similar and lead essentially to the same conclusions. 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, it generalizes this finding to all other languages (with the exception of Russian). For the phonological network, the results show that, generally speaking, PAC fits the developmental trajectory better than PAT in all languages. We note however that PAT, though weaker, fares better for the phonological networks than it does for the semantic networks. In the former, it predicts part of the growth process in some languages such as Croatian, English, Norwegian and Russian. In the latter, it is rather universally unpredictive.

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 equation (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 (phonological vs. semantic), 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 through preferential acquisition, and that, generally speaking, semantic and phonological networks both contribute to the learning process.

The big picture: 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. Goodman et al., 2008). The second factor is word length because it correlates with phonological connectivity (Pisoni, Nusbaum, Luce, & Slowiaczek, 1985).

Since PAT was uninformative, we dropped it from this analysis. Thus, we no longer needed to 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 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. This allowed us to get a more accurate estimate of

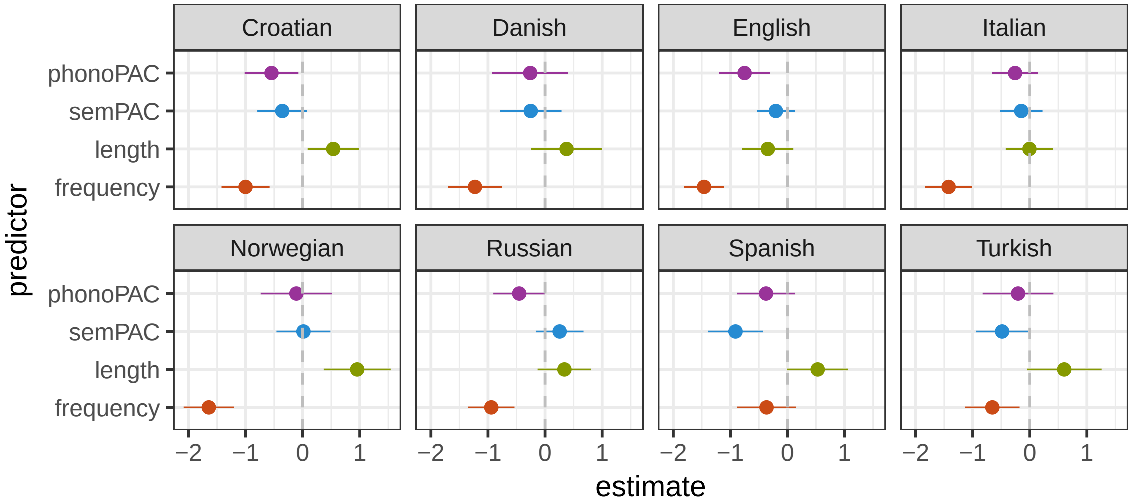


Figure 5: Estimates of predictor coefficients by language. Values above 0 indicate a positive relationship (e.g. longer words tend to have a higher AoA), while values below 0 indicate a negative relationship (e.g. words with higher frequency tend to have a lower AoA). Ranges indicate 95% confidence intervals.

word length across languages than the number of characters in the orthographic transcription.

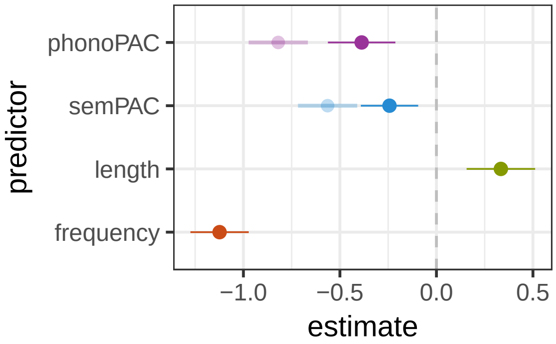


Figure 6: Estimates of predictor coefficients in the combined model. Ranges indicate 95% confidence intervals. Fade color indicates estimates of PAC predictors in a model that does not include frequency and length as covariates.

We conducted two analyses. On the one had, we fit a linear regression for each language, and on the other hand, 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 most reliable predictor of age of acquisition. Word length predicts part of 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 connectivities, 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[1]. 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. Moreoever, 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 part of learning in some languages, but not in others. Despite this coross-lingusitic variation, both phonological and semantic connectivity contribute to the overall learning (when data is pooled across languages).

[1] Semantic connectivity does not explain variance in English data beyond that explained by frequency and length. This contrasts with the original finding in Hills et al. 2009. However, in this previous studiy, semantic connectivity was not tested in a model that included both frequency and length as covariates. Morever, our sample size is bigger.

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 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, Maouene, Riordan, & Smith (2010), is that the referents of such words are more easily disambiguated as demonstrated, for instance, 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, and vice versa. We know even infant (whose vocabulary is still very rudimentary) develop a sensitivity for high frequency sound sequences in the ambient native 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 apparent sensitivity and preference for high connectivity in the lexical network.

Finally, this study shares a number of limitations with previous studies using similar datasets. For instance, we used normative age of acquisition. However, individual children may not follow a unique learning trajectory. The use of longitudinal data would allow us to assess how the aggregate behavior relates to individual trajectories (e.g., typical vs. late talkers, Beckage, Smith, & Hills, 2011). Moreover, the results obtained in this study provide correlational but not causal evidence. Thus, conclusion drawn from our network analysis would require parallel evidence, especially through controlled behavioral experiments.

Acknowledgements

This work was supported by a post-doctoral grant from the Fyssen Foundation.

References

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*.

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.

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/`

Pisoni, D. B., Nusbaum, H. C., Luce, P. A., & Slowiaczek, L. M. (1985). Speech perception, word recognition and the structure of the lexicon. *Speech Communication*, *4*(1), 75–95.

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–