Word Learning as Network Growth: A Cross-linguistic Analysis

Abdellah Fourtassi[1], Yuan Bian[2], & Michael C. Frank[1]

[1] Department of Psychology, Stanford University

[2] Department of Psychology, University of Illinois

Author Note

Abdellah Fourtassi

Department of Psychology

Stanford University

50 Serra Mall

Jordan Hall, Building 420

Stanford, CA 94301

Correspondence concerning this article should be addressed to Abdellah Fourtassi, Postal address. E-mail: afourtas@stanford.edu

Abstract

Children tend to produce words earlier when they are connected to a variety of other words along 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 predicted by a network growth model where highly connected words in the child's early lexicon enable learning of similar words. Another view suggests that learning is driven by highly connected words 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 semantic domains across 10 languages. We show that external connectivity in the learning environment drives growth in both production- and comprehension-based vocabularies, even controlling for word frequency and length. This pattern of findings suggests a word learning mechanism where children harness their statistical learning abilities to detect and learn highly connected words in the learning environment.

Keywords: Language understanding; audio-visual processing; word learning; speech perception; computational modeling.

Word Learning as Network Growth: A Cross-linguistic Analysis

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 analyzing dense corpora capturing language learning in context (e.g., B. C. Roy, Frank, DeCamp, Miller, & Roy, 2015). One prominent strategy consists in documenting the timeline of words' acquisition, and studying the properties that make words easy or hard to learn (e.g., J. C. Goodman, Dale, & Li, 2008; Huttenlocher, Haight, Bryk, Seltzer, & Lyons, 1991). For example, J. C. Goodman et al. (2008) found that, within a lexical category (e.g., nouns), higher parental frequency is associated with earlier learning. Researchers have studied the role of several other factors such as word length and the mean length of utterances in which the word occurs (e.g., Braginsky, Yurovsky, Marchman, & Frank, 2016; Swingley & Humphrey, 2018).

Besides word-level properties, the structure of the lexicon (that is, how words relate to one another) was also found to predict the Age of Acquisition (AoA) of words. The lexical structure can be 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 (e.g., Collins & Loftus, 1975; Luce & Pisoni, 1998). Previous studies have investigated early vocabulary structure by constructing networks using a variety of word-word relations including shared semantic features (McRae, Cree, Seidenberg, & McNorgan, 2005), target-cue relationships in free association norms (Nelson, McEvoy, & Schreiber, 1998), co-occurrence in child directed speech (MacWhinney, 2014), and phonological relatedness (Vitevitch, 2008). These studies have generally found that children tend to produce words that have higher neighborhood density (i.e., high connectivity in the network) earlier, both at the phonological and the semantic level (Carlson, Sonderegger, & Bane, 2014; Hills, Maouene, Riordan, & Smith, 2010; Hills, Maouene, Maouene, Sheya, &

Smith, 2009; Stella, Beckage, & Brede, 2017; Storkel, 2009).

While most studies have focused on the static properties of the lexical network, a few have investigated the underlying developmental process. In particular, Steyvers and Tenenbaum (2005) suggested that the observed effects of connectivity are the consequence of how the lexical network gets 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 (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. For clarity of reading, we will call preferential attachment the Internally-driven mechanism (INT), and preferential acquisition the Externally-driven mechanism (EXT). 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 INT proposal, learning is driven by known words with high connectivity to other known words (Figure 1, left). Thus, the network structure is a causal factor in early word learning, that is, children rely on the organization of their past knowledge to determine future learning (Altvater-Mackensen & Mani, 2013; Borovsky, Ellis, Evans, & Elman, 2016; Chi & Koeske, 1983; Storkel, 2009). In contrast, on the EXT approach, learning is driven by the connectivity of words that are not known yet (Figure 1, right). Thus, the relevant network structure is not internally represented by children, and the observed connectivity effect might be an epiphenomenon of some properties of the linguistic input. For example, highly connected words in the input could be more easily learned because of their contextual diversity, allowing for an easier meaning disambiguation (McMurray, Horst, & Samuelson, 2012; Smith & Yu, 2008; Yurovsky & Frank, 2015). Another reason could be that these words are emphasized by the cargivers in their child-directed speech (Clark, 2007; Hoff &

Naigles, 2002; Huttenlocher et al., 1991).

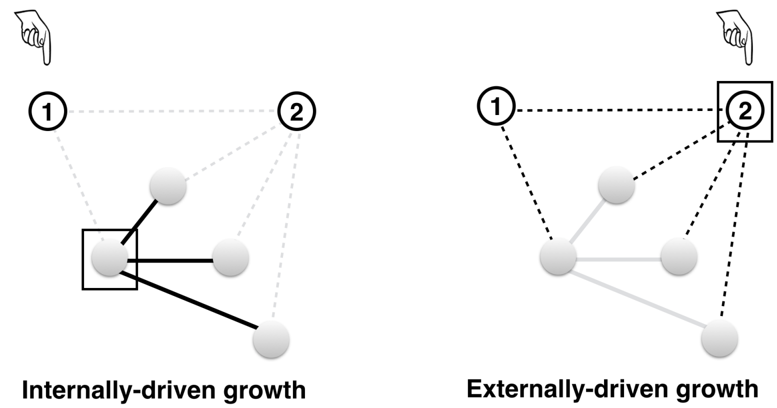

Figure 1. Illustration of the two growth scenarios. Filled grey circles represent known words (Internal) at a certain point in time. The empty, numbered circles represent words that have not yet been learned (External) and which are candidates to enter the lexicon next. The identity of the word that is going to be learned depends on the growth scenario. Here the squares indicate the node that drives growth in each scenario and the hand pointer indicates which word is likely to be learned. For INT, 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 INT, node 1 is more likely to enter the lexicon. For EXT, the utility of a candidate node is its degree in the entire network. According to EXT, node 2 is more likely to enter the lexicon next.

Hills et al. (2009) investigated the growth of lexico-semantic networks in toddlers and found that growth did not proceed according to INT as was originally hypothesised by Steyvers and Tenenbaum (2005), but rather according to EXT.[1] This is an important finding because it suggests that learning in the early stages is mostly driven by properties of the external input, regardless of how past knowledge is organized. However, this work has been

[1]Besides INT and EXT, the authors tested a third mechanism (called the lure of associates) which resembles EXT in that it is driven by the connectivity of external nodes, except that this connectivity is computed with respect to words that are known. However, EXT is the externally-driven scenario that best predicted the data in this previous work.

done in the special case of networks that are based on 1) semantic associations, 2) production-based vocabularies, and 3) data from English-learning children, only. The extent to which this result depends on the domain (e.g., semantic vs. phonological connectivity), the vocabulary measure (production vs. comprehension) and culture/language is still unclear. In this work, we test the generality of the finding along these three dimensions.

First, we study the phonological network in addition to the semantic network. These two networks represent different ways the mental lexicon is structured (Beckage & Colunga, 2016). In particular, words that are neighbors in the semantic network (e.g., "cat", "dog") are not necessarily neighbors in the phonological network, and vice versa. Does the phonological network also predict word learning? Previous work did find an effect of words' connectivity in the phonological network on their age of learning (Carlson et al., 2014; Stella et al., 2017; Storkel, 2009). In other words, words learned earlier in life tend to sound similar to many other words than a word learned later in life. However, this finding is *a priori* compatible with both INT and EXT, and previous studies did not explicitly compare these two mechanisms. Here, we investigate whether phonological networks, like semantic networks, grow through EXT, or if they rather grow via INT (Figure 1).

Second, we study vocabularies measured using both comprehension and production. Previous studies have found differences between these vocabularies in terms of their content and rate of acquisition (Benedict, 1979; Fenson et al., 1994). These differences may reflect the fact that comprehension and production do not share the same constraints. For instance, whereas comprehension depends on the ease with which words are stored and accessed, production depends, additionally, on the ease with which words are articulated, e.g., shorter words are produced earlier (Braginsky et al., 2016). By investigating comprehension-based vocabularies, we assess the extent to which the network growth mechanism captures general learning patterns beyond the specific constraints of production.

Finally, we use developmental data in 10 languages. Lexical networks can show more or less cross-linguistic variability along both the semantic and phonological domains (Arbesman, Strogatz, & Vitevitch, 2010; Youn et al., 2016). Besides, cultures might differ in the way caregivers talk to children (Cristia, Dupoux, Gurven, & Stieglitz, 2017; Kuhl et al., 1997), and this difference in the input could influence the way in which the children's networks grow. Thus, cross-linguistic comparison is crucial to test what growth mechanism is cognitively universal and is used by all children, and what mechanism is specific to some patterns of learning that emerge due to the particulars of a given language or culture (Bates & MacWhinney, 1987; Slobin, 2014).

We adopted the following research strategy. We used a cross-linguistic dataset of children's vocabulary development (Fenson et al., 1994; Frank, Braginsky, Yurovsky, & Marchman, 2017), and we studied the timeline of word learning using the normative/average age of acquisition (i.e., the age at which at least 50% of children know a given words). Our choice of studying the normative learning trajectory rather than the individual trajectories was motivated by the nature of the dataset used—which is primarily based on cross-sectional studies. Children may vary in their individual learning trajectories, but the aggregate data provide highly robust measures of the average learning patterns (Fenson et al., 1994). The use of such measures has lead to important insights on the mechanisms of word learning (Braginsky et al., 2016; J. C. Goodman et al., 2008; Hills et al., 2010, 2009; Stella et al., 2017; Storkel, 2009).

The paper is organized as follows. First, we describe the datasets we used and explain how we constructed the networks. Second, we analyze static properties of words' connectivity in these networks (correlation with age of acquisition and shape of the distribution) and we explain how these properties infom hypotheses about network growth. Next, we explicitly fit the two hypothesized growth mechanisms to the data. We investigate the extent to which the results obtained in Hills et al. (2009) generalize to phonological networks and comprehension-based vocabularies, and whether this generalization holds cross-linguistically.

Networks

Data

We used data from Wordbank (Frank et al., 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). When filling out a CDI form, caregivers are either invited to indicate whether their child "understands" (comprehension) or "understands and says" (production) each of about 400-700 words. For younger children (e.g., 8 to 18 months in the English data), both comprehension and production are queried, whereas for older children (16 to 36 months) only production is queried. Due to these limitations, we use data from younger children to test comprehension and data from older children to test production. Following previous studies (Hills et al., 2009; Storkel, 2009), we restrict our analysis to the category of nouns.[2]. Table 1 gives an overview of the data we used.

Age of acquisition

For each word in the CDI data, we compute the proportion of children who understand or produce the word at each month. Then we fit a logistic curve to these proportions and determined when the curve crosses 0.5, i.e., the age at which at least 50% of children know the word. We take this point in time to be each word's age of acquisition (Braginsky et al., 2016; J. C. Goodman et al., 2008).

[2]This choice was also forced by the fact that the great majority of association data produced by adults–and which we used to build semantic networks–were nouns.

Table 1

Statistics for the dataset we used.

	Comprehension		Production	
Language	Nouns	Ages	Nouns	Ages
Croatian	209	8-16	312	16-30
Danish	200	8-20	316	16-36
English	209	8-18	312	16-30
French	197	8-16	307	16-30
Italian	209	7-24	312	18-36
Norwegian	193	8-20	316	16-36
Russian	207	8-18	314	18-36
Spanish	208	8-18	312	16-30
Swedish	205	8-16	339	16-28
Turkish	180	8-16	297	16-36

Semantic networks

We constructed semantic networks for English data following the procedure outlined in Hills et al. (2009), as follows. We used as an index of semantic relatedness the Florida Free Association Norms (Nelson et al., 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.[3] To model growth from month to month, we constructed a different network at each month, based on the words that have been acquired by that month.

Since the free association norms are available only in English, we used the hand-checked translation equivalents available in Wordbank, which allowed us to use the English association norms across 10 languages. Using the same association data across languages does not necessarily lead to similar networks. Indeed, though this approximation assumes that the semantic similarity measure is universal–which is a reasonable approximation (e.g., Youn et al., 2016), the set of words acquired by children as well as the timeline of this acquisition can still vary from language to language leading to possibility different learning strategies.

Phonological networks

To construct phonological networks we first mapped the orthographic transcription of words to their International Phonetic Alphabet (IPA) transcriptions in each language, 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 (Carlson et al., 2014; Stella et al., 2017; Storkel, 2009). However, these studies reported a contribution of phonological connectivity to word learning when networks were built using a rich adult vocabulary. Since the focus of the current study is on the mechanism of growth, the networks should be based on children's early vocabulary. The latter, however, contains very few word pairs with an edit distance of 1. Thus, we increased the threshold from 1 to 2, that

[3]This choice was based the prior work by @hills2009 stating that analyses with both outdegrees (sum of the links where the word is the cue in a cue-target pair) and total degree (outdegree plus indegree) led to results weaker than those calculated for indegree.

is, two nodes were related if their edit distance was equal to 1 or 2.[4] 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 analyzing word connectivity in the global (static) network. We constructed this network using nouns learned by the oldest age for which we have CDI data (e.g., in English this corresponds, in comprehension, to the network by 18 months, and in production, to the network by 30 months). This global network is the end-state towards which both INT and EXT 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 analyze properties of this global networks that may a priori hint at an INT- or EXT-like growth. In order to compare various predictors on the same data, we restrict the analysis to the subset of nouns for which we had both semantic and phonological information in each language.

Connectivity predicts the age of acquisition. Connectivity in the global network is directly related to EXT as it represents the explicit criterion this growth scenario uses to determine what words should be learned first (Figure 1). Therefore, a direct consequence of an EXT-like growth scenario is a correlation between connectivity in the global network and the age of acquisition. This correlation is also necessary to INT, although the causality is reversed: higher connectivity in the global network is caused by earlier learning, not the other way around. 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 accumulating more links over time. Thus, the correlation between connectivity in

[4]We also considered the case of an edit distance of 1 as well as the continuous measure, i.e., the inverse edit distance without threshold. In both cases, the results were weaker than those obtained with a threshold of 2. We did not consider the case of a threshold larger than 2 since the pairs will become less and less phonologically related.

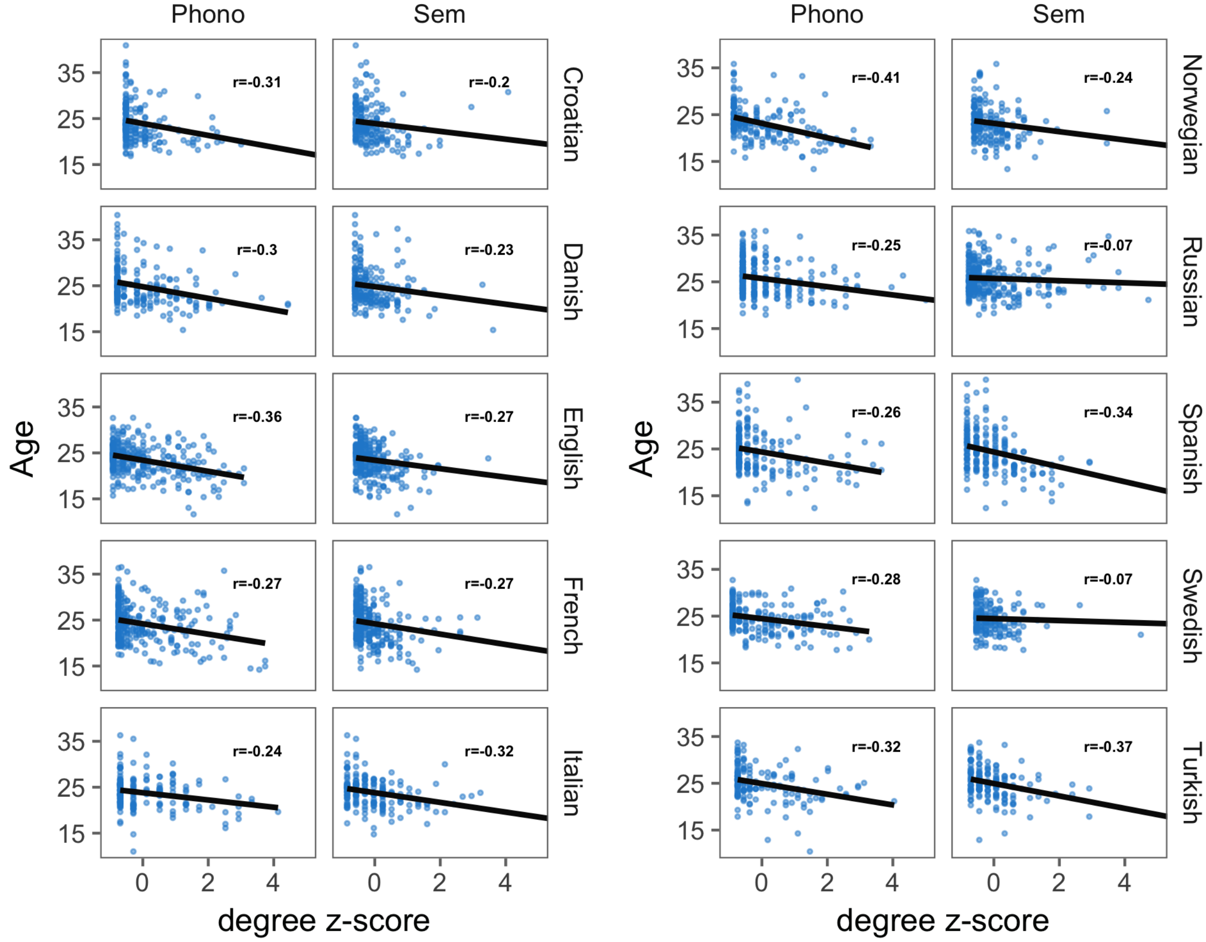

Figure 2. Age of production in the global network as predicted by the degree (i.e., connectivity) in this network. Results are shown in each language for phonological and semantic networks. Each point is a word, with lines indicating linear model fits, and numbers indicating the Pearson correlation coefficients.

the end-state network and AoA can result from both EXT and INT. If there is no such correlation, neither growth scenario can be posited as a possible learning mechanism.

Figures 2 and 3 show how the age of production and comprehension, respectively, correlates with the degree (or indegree for the semantic networks). 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. In production data, the average correlation across languages was -0.24 (SD=0.10) for the

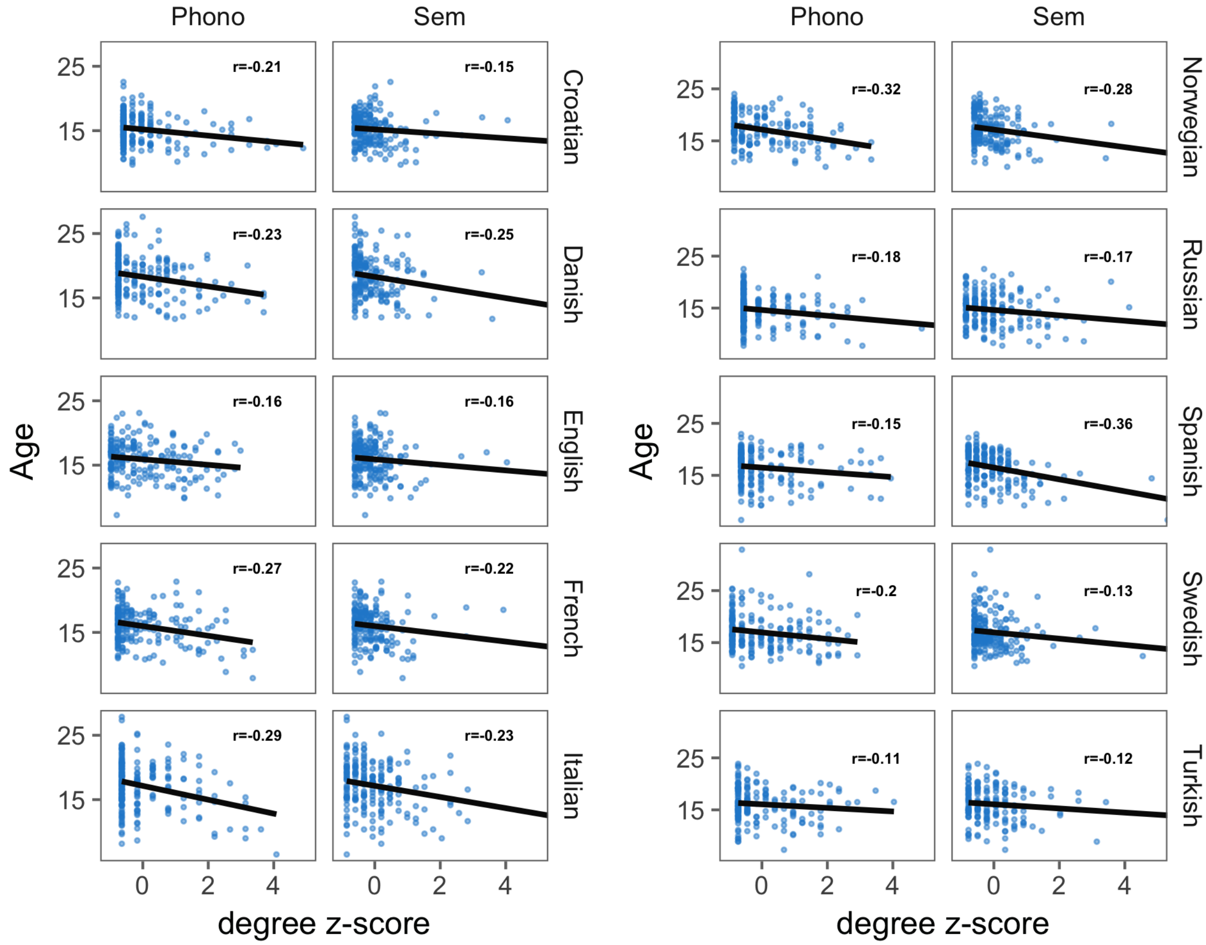

Figure 3. Age of comprehension in the global network as predicted by the degree (i.e., connectivity) in this network. Results are shown in each language for phonological and semantic networks. Each point is a word, with lines indicating linear model fits, and numbers indicating the Pearson correlation coefficients.

semantic networks and -0.30 (SD=0.05) for the phonological networks. In comprehension data, the average correlation was -0.21 (SD=0.08) for the semantic networks and -0.21 (SD=0.07) for the phonological networks. These results indicate that nouns with higher degrees are generally learned earlier, thus replicating previous findings in English (e.g., Storkel 2004, 2009; Hills et al. 2009) and extending these findings to 10 different languages, generally, in both production- and comprehension-based vocabularies.

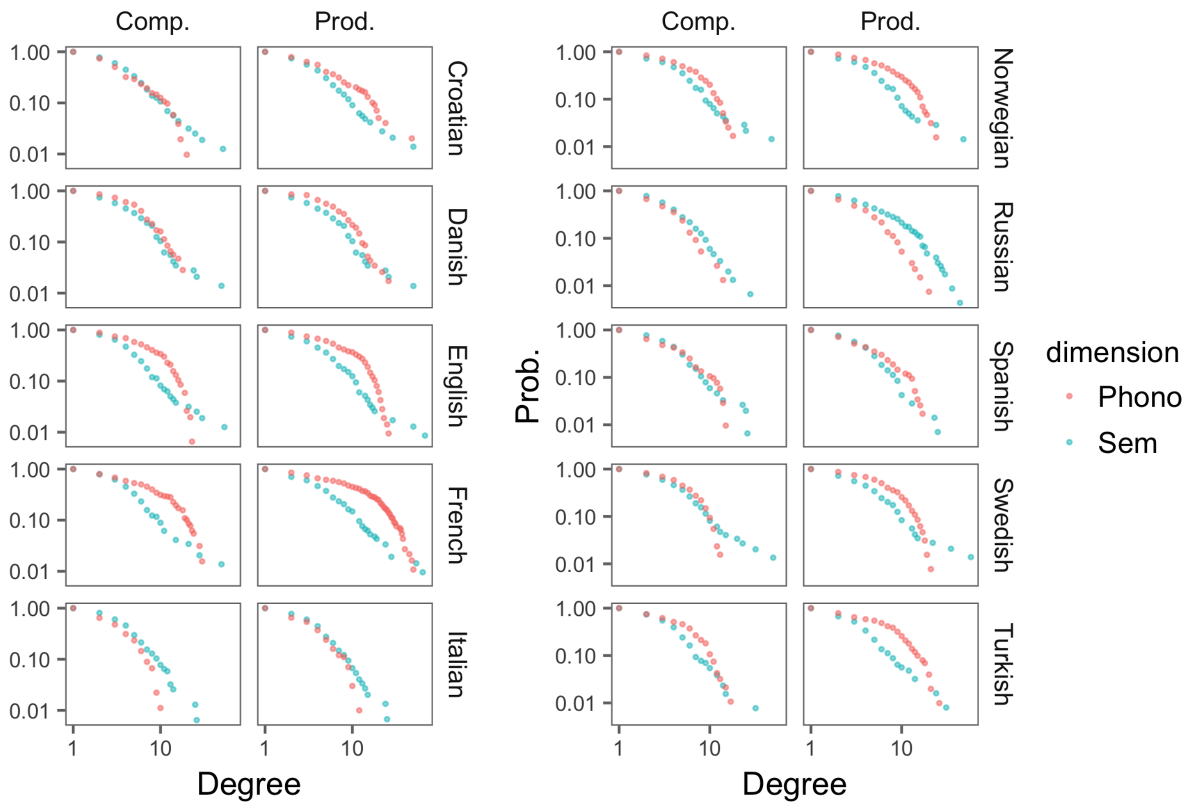

Figure 4. Log-log plot of the cumulative degree distribution function for the global phonological and semantic networks across languages. The figure shows the results for both production and comprehension data. A perfect power-law distribution should appear as a straight line in this graph.

Power-law degree distribution. We also analyzed the global network's degree distribution. The shape of this distribution is particularly relevant to INT 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 end-state network displays this property, this fact would suggest, but not prove, an INT-like generative process. If, however, the degree distribution is very different from a power law, this would significantly weaken the case for INT. The log-log plots are shown in Figure 4. We fit a power law to each empirical degree distribution following the procedure outlined in Clauset, Shalizi, and Newman (2009) and using a related R package (poweRlaw, Gillespie, 2015).

Table 2

Results of fitting a power law model to the degree (i.e., connectivity) distribution in each model for production data. Numbers indicate the cut-off degree, the scaling parameter alpha, and the p-value which quantifies the plausibility of the power law hypothesis. If the p-value is close to 1, a power law cannot be rejected as a plausible fit for the data.

	Phono.			Sem.		
Language	cut-off	alpha	p-value	cut-off	alpha	p-value
Croatian	4	2.18	0.123	4	2.55	0.881
Danish	11	4.55	0.858	4	2.38	0.001
English	20	9.14	0.511	5	2.66	0.132
French	20	3.75	0.112	8	2.81	0.133
Italian	9	9.45	0.780	4	2.93	0.608
Norwegian	15	6.28	0.744	5	2.88	0.201
Russian	8	4.20	0.541	24	5.61	0.723
Spanish	13	8.75	0.736	4	2.98	0.460
Swedish	11	4.68	0.103	4	2.49	0.171
Turkish	8	3.26	0.375	4	2.87	0.925

In brief, the analysis consisted in two steps. First, we derived the optimal cut-off, k_{min}, above which the distribution is more likely to follow a power law,[5] and we estimate the corresponding scaling parameter α. Second we calculated the goodness-to-fit, which resulted in a p-value quantifying the plausibility of the model. The results are shown in Table 2 for production data, and in Table 3 for comprehension data.

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

Table 3

Results of fitting a power law model to the degree distribution in each model for comprehension data. Numbers indicate the cut-off degree, the scaling parameter alpha, and the p-value which quantifies the plausibility of the power law hypothesis. If the p-value is close to 1, a power law cannot be rejected as a plausible fit for the data.

	Phono.			Sem.		
Language	cut-off	alpha	p-value	cut-off	alpha	p-value
Croatian	2	2.06	0.020	5	2.67	0.895
Danish	5	2.98	0.136	4	2.39	0.005
English	13	5.16	0.235	4	2.64	0.765
French	18	5.58	0.336	4	2.63	0.330
Italian	8	10.27	0.909	4	2.88	0.688
Norwegian	13	7.65	0.440	5	2.87	0.433
Russian	5	3.97	0.854	8	3.91	0.952
Spanish	5	3.01	0.085	5	3.11	0.552
Swedish	9	6.75	0.102	5	2.81	0.713
Turkish	9	5.73	0.958	4	3.13	0.887

Overall, we could not reject the null hypothesis of a power-law distribution: The p-value was generally above 0.1 in almost all languages for both production and comprehension. That said, phonological networks had relatively larger cut-offs than semantic networks. As was suggested by Arbesman et al. (2010), these “truncated” power-laws in phonological networks may reflect the various constraints that exist on word formation in the phonological domain such as the number of phonemes in the language, the phonotactics (i.e., the way sound sequences are arranged in words), and the length of words. Such constraints may limit the number of words that are phonologically similar, thus leading to distributions which decay faster than a non-truncated power law.

In sum, the static properties of the end-state network are *a priori* compatible with both INT and EXT. In order to decide between these two developmental scenarios, we need to fit explicit growth models to the data.

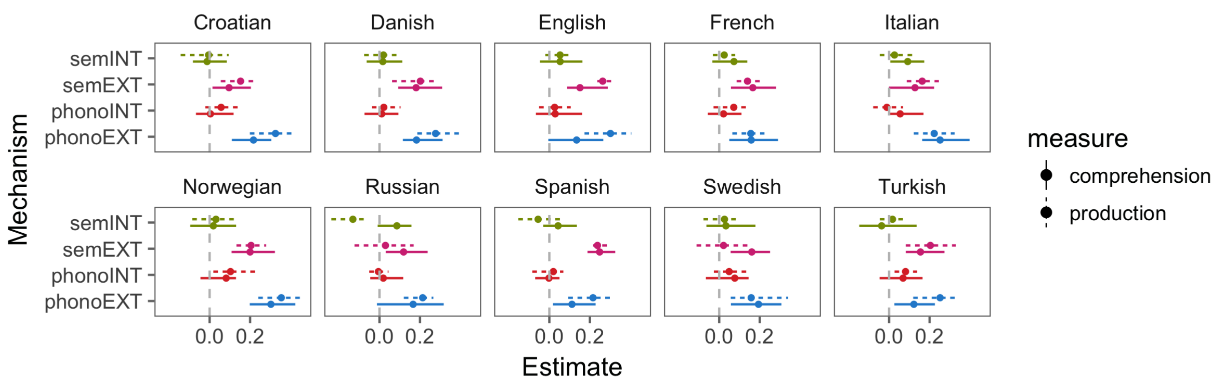

Figure 5. Evaluation of growth scenarios (EXT: externally-driven, INT: internally-driven) for both semantic and phonological networks. Each point represents the mean of the posterior distribution of the growth parameter, with ranges representing 95% credible intervals. Positive values mean that learning proceeds according to the predictions of the growth scenario, whereas negative values mean that learning proceeds in opposition to the predictions of the growth scenario.

Network growth models

To test the network growth scenarios, we fit two growth models to the data. We calculated the probability that a word w_i, with a utility value u_i would enter the lexicon at a given month, using a softmax function:

$$p(w_i) = \frac{e^{\beta u_i}}{\sum_j e^{\beta u_j}} \tag{1}$$

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 utility values u_i are acquired first, and a negative value means that words with lower utility values are acquired first (see Figure 1 for an illustration of how utilities values u_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 & Stuhlmuller, 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 5.

First, the results replicate Hills et al.'s original finding regarding the semantic network in English and the production-based vocabulary, which is that this network grows by EXT, not by INT. Second, our results show that, generally speaking, this finding generalizes to comprehension-based vocabulary, and holds across languages. This generalization was obtained in both the semantic[6] and phonological domains.

Comparison to other predictors of age of acquisition

Above we showed that the way semantic and phonological information is structured in the learning environment contributes to noun learning (via EXT) 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

[6] One could imagine that the fact of using English free association norms cross-linguistically would decrease the effect of non-English semantic networks because of possible cultural differences. However, our findings do not support this assumption, rather it supports our initial approximation about the universality of the semantic similarity measure. That said, this apprimation is not perfect; for example there is evidence that a small part of the variance in free assication data can be explained by phonological similarity (Kachergis, Cox, & Jones, 2011; Matusevych & Stevenson, 2018), thus leading to possibly minor cross-linguistic differences.

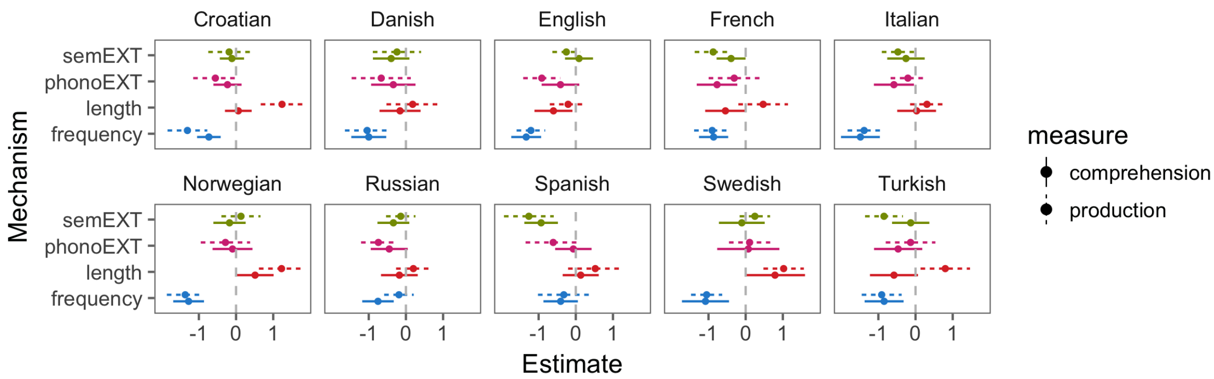

Figure 6. Estimates of the relative contribution of each predictor of AoA in the regression model in each language. Results are shown for both production and comprehension data. Ranges indicate 95% confidence intervals. Positive values indicate a positive relationship (e.g. longer words tend to have a higher AoA), while negative values indicate a negative relationship (e.g. words with higher frequency tend to have a lower AoA).

in a reliable fashion (e.g. J. C. Goodman et al., 2008). The second factor is word length, which was shown to correlate with phonological connectivity: Shorter words are more likely to have higher connectivity (D. B. Pisoni, Nusbaum, Luce, & Slowiaczek, 1985; Vitevitch & Rodríguez, 2005).

Since we found INT to be uninformative, we dropped it from this analysis, keeping only EXT. This simplified the model because we no longer needed to fit growth month-by-month. The latter was a requirement only for INT where the words' utilities varied from month to month, depending on how connectivity changed in the growing internal network. A more direct way to assess and compare the contribution of EXT 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.

For word length, we counted the number of phonemes in our generated IPA transcription. For word frequency, we used the frequency estimates from Braginsky et al. (2016) where unigram counts were derived based on CHILDES corpora in each language

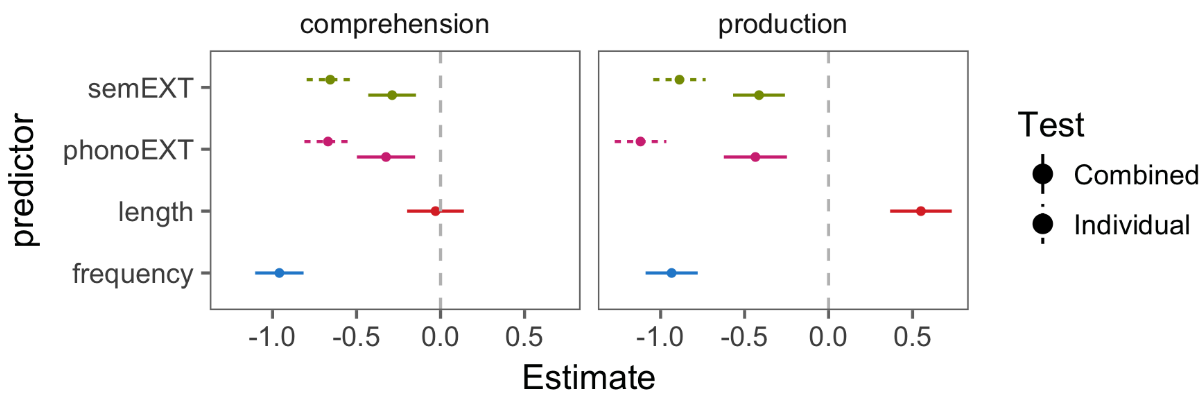

Figure 7. Estimates of the relative contribution of each predictor of AoA in the combined mixed-effects model with language as a random effect. Results are shown for both production and comprehension data. Ranges indicate 95% confidence intervals. Dotted ranges indicate the estimates for the predictor in a separate model that includes only this predictor as a fixed effect.

(MacWhinney, 2014). 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"). Although these frequency counts use transcripts from independent sets of children, they are based on large samples, and this allows us to average out possible differences between children and the specificities of their input (see J. C. Goodman et al., 2008 for a similar research strategy).

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 6 shows the coefficient estimate for each predictor in each language for production and comprehension data. Figure 7 shows the coefficient estimates for all languages combined (all predictors were centered and scaled).

The findings for the new predictors were as follows. Overall, frequency is the largest and most consistent predictor of age of acquisition in both comprehension and production data and across languages, endorsing 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 is more predictive for production than comprehension (and this difference is very clear in the global model), replicating previous work (Braginsky et al., under review). Thus, word length seems to reflect the effects of production's constraints rather than than comprehension's constraints, i.e., longer words are harder to articulate but they may not be significantly more difficult to store and access.

As for the factors of interest, i.e., semantic and phonological connectivity, we found cross-linguistic differences. Connectivity contributes to learning in some languages but not in other. In particular, 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, this might be due to our using a slightly different model (which included word length as a covariate) and a larger dataset. That said, and despite these apparent cross-linguistic differences, both phonological and semantic connectivity are significant predictors in the combined model (Figure 7).

Discussion

This study provided an analysis of network growth during development. We compared two network growth scenarios described in the pioneering work of Steyvers and Tenenbaum (2005) and Hills et al. (2009). The first scenario, INT (originally called Preferential Attachment), described a rich-get-richer network growth model in which the current structure of the learner's internal network determines future growth; the other, EXT (originally called Preferential Acquisition) described a model in which the external, global environmental network structure determines learners' growth patterns. These two mechanisms represent two fundamentally different accounts of lexical growth: One suggests that future word knowledge is primarily shaped by the children's past knowledge and its organization, whereas the other suggests that learning is shaped, rather, by salient properties in the input regardless of how past knowledge is organized. The present study tested the generality of previous findings (Hills et al., 2010, 2009) by 1) investigating phonological networks together with semantic networks, 2) testing both comprehension- and production-based vocabularies, and 3) comparing the results across 10 languages.

We found that the original findings reported in Hills et al. (2009) generalize well across all these dimensions. First, just like semantic networks, phonological networks grow via the externally-driven scenario (EXT), not by the internally-driven mechanism (INT). Second, comprehension-based vocabularies grow in a way similar to production-based vocabularies. Finally, the findings were, overall, similar across the 10 languages we tested. Although we find some cross-linguistic variation when semantic and phonological networks were pitted against frequency and length, this variability is to be taken with a grain of salt as it might be exaggerated in our study by the limited and partially-overlapping sample of nouns for each language. In fact, both phonological and semantic connectivity are significant predictors above and beyond frequency and length when data are pooled across languages.

These findings corroborate the hypothesis that children start by learning words that have high similarity to a variety of other words in the learning environment, not in the child's available lexicon. This hypothesis implies that children are sensitive to highly connected words although they do not initially have access to the full network, thus raising some important questions: What mechanism allows children to distinguish highly connected words from other words? Besides, why would highly connected words be easier to learn?

One possibility is that these patterns emerge from children's use of statistical learning abilities (Aslin & Newport, 2012; Saffran, Aslin, & Newport, 1996; Smith & Yu, 2008). The term "statistical learning" has been used in the developmental literature to describes the process by which one acquires information about their environment through tracking the frequency distribution of some elements (e.g., words) in different contexts. An important property of this kind of learning is that it occurs without explicit instructions and through mere exposure to the input. Previous work in this line of research has documented specific mechanisms which can explain the patterns found in the current study.

For example, in the semantic domain, growth according to EXT can be explained by a mechanism similar to cross-situational learning (Pinker, 2013; Smith & Yu, 2008; Yurovsky & Frank, 2015). According to this mechanism, children track the co-occurrence of concrete nouns with their possible semantic referents. The referent of a word heard in only one naming situation can be ambiguous (e.g., when the word "ball" is heard for the first time in the presence of both a ball and a chair), but hearing the same word in a diversity of semantic contexts allows the learner to narrow down the set of possible word-object mappings. In our case, free association (used to determine semantic network connectivity) is related to contextual co-occurrence (Fourtassi & Dupoux, 2013; Griffiths, Steyvers, & Tenenbaum, 2007), meaning that highly connected words will tend to occur in a variety of speech and referential contexts. This fact makes such words easier to learn because they have more referential disambiguating cues across learning contexts. Crucially, children can learn these words without necessarily knowing the meaning of all other words with witch they co-occur (hence the similarity with EXT). This possibility is supported by the finding that words' diversity of occurrence in child directed speech predicts their age of learning (Hills et al., 2010).

In the phonological case, network growth according to EXT is also compatible with a scenario whereby children are tracking lwo level statistical patterns, e.g., high probability sound sequences. Indeed, connectivity in the phonological network is inherently correlated with phonotactic probability (Vitevitch, Luce, Pisoni, & Auer, 1999). That is, highly connected words tend to be made of frequent sound sequences. Children are sensitive to local phonotactic regularities (Jusczyk, Luce, & Charles-Luce, 1994) and this sensitivity might lead them to learn higher-probability words more easily (Storkel, 2001). This explanation is supported by computational simulations that show how learning general phonotactics patterns create "well-worn paths" which allow the models to represent several distinct but phonologically neighboring words (Dell, Juliano, & Govindjee, 1993; Takac, Knott, & Stokes, 2017).

Besides using their own statistical learning skills, children could also benefit from the way their caregivers speak. Perhaps the caregivers put more emphasis on the words that are highly connected in *their* mature lexical network. This emphasis would guide children to learn first these highly connected words even though children do not have access to the distribution of words' connectivity in the final network. Investigating this possibility would require further research on caregiver-child interaction (MacWhinney, 2014; B. C. Roy et al., 2015), examining what words are introduced over development and the extent to which children's uptake is influenced by this input (Clark, 2007; Hoff & Naigles, 2002).

This work shares a number of limitations with previous studies using similar research strategy and datasets. Chief among these limitations is the fact that the age of word acquisition is computed using different children at different ages (due to the fact that available CDI data is mainly cross-sectional). Although this measure has proven highly consistent (Fenson et al., 1994), it led us to focus on studying the learning mechanism of the "average" child. Individual trajectories, however, could lead to different netwtok strcutures and show different learning patterns. For example, using longitudinal data Beckage, Smith, and Hills (2011) found differences between typical and late talkers in terms of the semantic network structure. Besides, although our study endorses the externally-driven account of network growth, this does not mean individual children never use some variant of INT or some combination of both INT and EXT (Beckage and Colunga, under review). For example, some children develop "islands of expertise", that is, well organized knowledge about a certain topic (e.g., birds or dinosaurs). This prior knowledge enables these children to learn new related words more easily (e.g., Chi & Koeske, 1983).

To conclude, our work validates and generalizes previous results in early network development. It suggests that the advantage of highly connected words may result, at least in the early stages of word learning, from the operation of simpler mechanisms in both the semantic and phonological domains. One question for future experimental work is whether such correlational 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.

Disclosure statement

None of the authors have any financial interest or a conflict of interest regarding this work and this submission.

References

Altvater-Mackensen, N., & Mani, N. (2013). Word-form familiarity bootstraps infant speech segmentation. *Developmental Science*, *16*(6).

Arbesman, S., Strogatz, S. H., & Vitevitch, M. S. (2010). The structure of phonological networks across multiple languages. *International Journal of Bifurcation and Chaos*, *20*(03), 679–685.

Aslin, R. N., & Newport, E. L. (2012). Statistical learning: From acquiring specific items to forming general rules. *Current Directions in Psychological Science*, *21*(3).

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

Bates, E., & MacWhinney, B. (1987). Competition, variation, and language learning. *Mechanisms of Language Acquisition*, 157–193.

Beckage, N. M., & Colunga, E. (2016). Language networks as models of cognition: Understanding cognition through language. In *Towards a theoretical framework for analyzing complex linguistic networks* (pp. 3–28). Springer.

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.

Benedict, H. (1979). Early lexical development: Comprehension and production. *Journal of Child Language*, *6*(2), 183–200.

Borovsky, A., Ellis, E. M., Evans, J. L., & Elman, J. L. (2016). Lexical leverage: Category knowledge boosts real-time novel word recognition in 2-year-olds. *Developmental Science*, *19*(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*.

Carlson, M. T., Sonderegger, M., & Bane, M. (2014). How children explore the phonological network in child-directed speech: A survival analysis of children?s first word productions. *Journal of Memory and Language*, *75*, 159–180.

Chi, M. T., & Koeske, R. D. (1983). Network representation of a child's dinosaur knowledge. *Developmental Psychology*, *19*(1).

Clark, E. V. (2007). Young children's uptake of new words in conversation. *Language in Society*, *36*(2).

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

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6).

Cristia, A., Dupoux, E., Gurven, M., & Stieglitz, J. (2017). Child-directed speech is infrequent in a forager-farmer population: A time allocation study. *Child Development*.

Dell, G. S., Juliano, C., & Govindjee, A. (1993). Structure and content in language production: A theory of frame constraints in phonological speech errors. *Cognitive Science*, *17*(2), 149–195.

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).

Fourtassi, A., & Dupoux, E. (2013). A corpus-based evaluation method for distributional semantic models. In *51st annual meeting of the association for computational linguistics proceedings of the student research workshop* (pp. 165–171).

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., & Stuhlmuller, 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.

Hoff, E., & Naigles, L. (2002). How children use input to acquire a lexicon. *Child Development*, *73*(2).

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology*, *27*(2).

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.

Kachergis, G., Cox, G. E., & Jones, M. N. (2011). OrBEAGLE: Integrating orthography into a holographic model of the lexicon. In *International conference on artificial neural networks* (pp. 307–314). Springer.

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., … Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science*, *277*(5326), 684–686.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, *19*(1).

MacWhinney, B. (2014). *The childes project: Tools for analyzing talk, volume ii: The database*. Psychology Press.

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

Matusevych, Y., & Stevenson, S. (2018). Analyzing and modeling free word associations. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.

McMurray, B., Horst, J. S., & Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review*, *119*.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*(4), 547–559.

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/

Pinker, S. (2013). *Learnability and cognition: The acquisition of argument structure*. MIT press.

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.

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.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Slobin, D. I. (2014). *The crosslinguistic study of language acquisition* (Vol. 4). Psychology Press.

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

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.

Swingley, D., & Humphrey, C. (2018). Quantitative linguistic predictors of infants? Learning of specific english words. *Child Development*, *89*(4), 1247–1267.

Takac, M., Knott, A., & Stokes, S. (2017). What can neighbourhood density effects tell us about word learning? Insights from a connectionist model of vocabulary development. *Journal of Child Language*, *44*(2).

Vitevitch, M. S. (2008). What can graph theory tell us about word learning and lexical retrieval? *Journal of Speech, Language, and Hearing Research*, *51*(2), 408–422.

Vitevitch, M. S., & Rodríguez, E. (2005). Neighborhood density effects in spoken word recognition in spanish. *Journal of Multilingual Communication Disorders*, *3*(1).

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.

Youn, H., Sutton, L., Smith, E., Moore, C., Wilkins, J. F., Maddieson, I., … Bhattacharya, T. (2016). On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences*, *113*(7), 1766–1771.

Yurovsky, D., & Frank, M. C. (2015). An integrative account of constraints on cross-situational learning. *Cognition*, *145*.