Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3960–3968
Marseille, 20-25 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

Representing the Toddler Lexicon: Do the Corpus and Semantics Matter?

Jennifer M. Weber, Eliana Colunga
Department of Psychology and Neuroscience
345 UCB Boulder, CO 80309 USA
{jennifer.m.ellis, eliana.colunga}@colorado.edu

Abstract

Understanding child language development requires accurately representing children's lexicons. However, much of the past work modeling children's vocabulary development has utilized adult-based measures. The present investigation asks whether using corpora that captures the language input of young children more accurately represents children's vocabulary knowledge. We present a newly-created toddler corpus that incorporates transcripts of child-directed conversations, the text of picture books written for preschoolers, and dialog from G-rated movies to approximate the language input a North American preschooler might hear. We evaluate the utility of the new corpus for modeling children's vocabulary development by building and analyzing different semantic network models and comparing them to norms based on vocabulary norms for toddlers in this age range. More specifically, the relations between words in our semantic networks were derived from skip-gram neural networks (Word2Vec) trained on our toddler corpus or on Google news. Results revealed that the models built from the toddler corpus were more accurate at predicting toddler vocabulary growth than the adult-based corpus. These results speak to the importance of selecting a corpus that matches the population of interest.

Keywords: Word2Vec, corpora, semantic networks, word-learning

1. Introduction

In the field of first language learning, there has been an increasing interest in modeling lexical development as a way to both understand and predict language development. There is enormous variability in the vocabularies of toddlers in their second year of life, which makes it difficult to tease out possible patterns in vocabulary acquisition and resulting lexical knowledge. To model vocabulary growth, and to predict future growth, we must accurately represent children's current word knowledge and how the words children know relate to each other. Most previous research attempting to model young children's lexicons has relied on adult metrics to create semantic networks. But the knowledge adults have about words may not be an accurate proxy for child knowledge. Further, advances in natural language processing techniques and network science provide the opportunity to create richer models than have previously been investigated. Thus, the main research question of this paper is whether the language samples from which we derive lexical representations matter for accurately modeling toddler vocabulary growth. Of additional interest is whether different distributional models of semantics provide richer representations that lead to more accurately capturing child development. To accomplish our goals, we will build and analyze networks representing the relationships between the words a typical English-speaking, American child knows at different ages between 16 and 30 months, investigating whether networks created using a toddler corpus more accurately represent and predict language growth compared to those created from adult corpora.

1.1 Representing a Lexicon

Networks are often used to model vocabularies. In these models, nodes typically represent the individual words in the lexicon, and edges, or the connections between nodes, represent how "related" those two words are to each other. Relatedness can be determined from co-occurrence, similarity in sound (e.g., cat-bat), or similarity in meaning (e.g. cat-dog), among other things (see Beckage & Colunga, 2016 for a review). In the network models considered in this work, we will limit ourselves to semantic networks. Thus, edges will represent semantic similarity. Edges can be weighted, here meaning the edge connecting two words that are more similar will have a higher value (weight) than the edge between two words that are less similar to each other. However, most networks used in vocabulary acquisition research treat edges as unweighted or binary, such that words are either connected (1) or not connected (0), implying that all directly connected words are equally similar.

Considerable work investigating adults has established the utility of multiple metrics for representing adult lexical structure. For example, the strength of word association norms (first word that comes to mind when presented with the word "bat") and feature norms (e.g., fish: has scales, swims, has gills) predict adult semantic processing speed (Nelson et al., 2009; McRae et al., 2005; Hutchison, 2003; Pexman et al., 2003). Researchers have also calculated co-occurrence metrics from varying corpora to establish word relatedness, again showing such metrics relate to adult semantic processing (Lund & Burgess, 1996; Vankrunkelsven et al., 2018). However, creating semantic networks that accurately represent the vocabulary of a child may require relatedness metrics derived specifically for children. Previous work tends to use adult-derived metrics to model child lexicons, one reason being pure practicality. In the lab, unlike young children, adults are much more compliant during tedious long laboratory tasks. So, it is possible to have adults list features for hundreds of words (McRae et al., 2005) or make judgments on whether dozens of features apply to each of those words (Howell et al., 2005). Further, there already exist very large adult-produced and adult-directed corpora with which to compute distributional metrics, such as COCA (1 billion words), Wikipedia (1.9 billion words), and Google News (100 billion words). However, children's and adult's experiences are very different both quantitatively and qualitatively, and thus one would expect their semantic knowledge about the world to be different as well. The associations an adult makes may not be readily obvious to a child, or even available at all based on the experience the child has had with words and their referents. For example a child might only understand the literal but not the figurative meaning of a word. Children may have absent, partial, or completely different semantic representations compared to those of adults.

1.2 Child Language and Networks

Building a model of a child's lexicon involves determining the words a child knows and approximating the relationship between these words. Through parent vocabulary checklists, we can see which words any one child knows at that given time, rather than relying on adult age-of-acquisition norms to guess the words children might produce. As a proxy of early vocabulary, many researchers have used the MacArthur-Bates Communicative Development Inventory (CDI) as the basis for early lexicons. The CDI is a 680-word parent checklist and has been found to be reliable, related to child performance language measures, and has been normed across children from the United States (Fenson et al., 1994).

Work using the CDI as a basis for modeling children's lexicons has led to important insights about language development. This includes insights into the relationship between lexical development, syntax, and processing speed (Moyle et al., 2007; Fernald & Marchman, 2012), lexical variability based on language mastery (MacRoy-Higgins et al., 2016; Kim & Yim, 2018), and even how the words in a child's vocabulary shape how they learn new words (Gershkoff-Stowe & Smith, 2004; Perry & Samuelson, 2011; Colunga & Sims, 2017). CDI-based networks are usually unweighted and sparse, and show evidence of small-world structure, seen in high local clustering and short average path lengths (Beckage et al., 2011). Research using these networks shows that across multiple languages, children tend to learn words that are highly connected in these networks earlier than words with smaller neighborhood densities (Fourtassi et al., 2020; Hills et al., 2009; Hills et al., 2010). Further, different growth algorithms have been compared using CDI-based networks, including preferential attachment, preferential acquisition, and lure of associates (Steyvers & Tenenbaum, 2005; Hills et al., 2010; Beckage et al., 2011; Beckage et al., 2015; for review see Beckage & Colunga, 2019), with their findings suggesting that different populations might be more accurately modeled using different growth algorithms. In short, the work using a child-based vocabulary to model early word learning has been fruitful for understanding the mechanisms of vocabulary growth. However, these different studies have utilized different similarity information to model the semantic similarity between each word and finding an accurate metric to model child semantic networks is crucial.

Though using the CDI allows us to approximate which words a child knows, previous work often still models what the child knows about each word using adult semantic measures. Steyvers and Tenenbaum (2005) showed that networks created from adult word associations, thesaurus entries, and WordNet all share the same structural properties. Further, networks with edges determined using adult association norms show this connectivity drives vocabulary growth, particularly preferential acquisition (Fourtassi et al., 2020; Hills et al., 2009). Networks created using adult-determined perceptual feature norms fared less well when predicting child language growth, though they still perform significantly better than chance (Beckage et al., 2020; Beckage et al., 2015; Beckage & Colunga, 2019; Hills et al., 2009; McRae et al., 2005). Jaccard similarity has also been used to relate children's comprehension and production networks (Kim & Yim, 2018).

A noted exception to all this work using adult-derived metrics is the use of CHILDES (MacWhinney, 2000), a repository of caregiver-child conversations contributed by many researchers, as a way to approximate the linguistic environment of a young child. Using co-occurrence statistics on the CHILDES corpus, scholars have created networks to compare the lexical structure of children with different levels of vocabulary knowledge (i.e. typically developing vs. late talkers), revealing structural differences between the two populations (Beckage et al., 2011; Jimenez & Hills, 2017). Hills et al. (2010) produced networks with characteristics that closely align with age-of-acquisition data by using a window size of five to calculate co-occurrences in the CHILDES database.

Different metrics have been shown to have different predictive power when it comes to capture early vocabulary development. Beckage et al. (2020) used neural network models that had access to different types of information to predict future word learning, resulting in a comparison of six models representing word knowledge in different ways. From the simplest model utilizing child demographic information such as age, sex, CDI percentile and vocabulary size, to models with information like perceptual features, phonology, semantic category, to a final model including semantic feature vectors derived from Word2Vec (a skip-gram neural network). Beckage and her colleagues found that predicting actual toddler vocabulary growth with a neural network using Word2Vec embeddings outperforms the other representations, indicating semantic feature vectors trained using skip-gram models may be particularly useful, especially for children with small vocabulary sizes.

To go one step further than Beckage et al. (2020), taking the cosine similarities between the Word2Vec embeddings offers a richer range of similarities compared to the binary similarity associations used from sliding window co-occurrence statistics and other previous methods. Word2Vec models arguably offer a richer semantic representation by not only considering words similar if they appear together, but also if they appear in similar contexts, something co-occurrence models miss. However, many researchers, if using Word2Vec or similar neural network embeddings (GloVe, FastText), use the readily-available vectors that were trained on adult corpora (e.g., Beckage et al., 2020). In previous work using the pre-trained GoogleNews Word2Vec vectors to create semantic network representations of children's language, we have noticed that some words are misclassified due to multiple senses of words, with one or more of those meanings not being one that a child would likely have in their own representation. For example, the GoogleNews word2Vec-based networks have *chair* and *head* strongly connected, with *chair* not connected to other instances of furniture such as *table* and *couch*. Presumably, this is because of the two senses of *chair* as head of a committee or a department vs. *chair* to sit on. Only one of these senses is likely to be known by a 2-year-old. Further, Word2Vec networks used in the past also threshold the similarities, thus treating similarity as a binary connection; either two words are related or they are not. To our knowledge, the only other instance of research to utilize Word2Vec embeddings derived from a child corpus investigated the semantic network's ability to form categories, but did not directly investigate true toddler vocabulary growth, as we do here (Asr et al., 2016).

To summarize, our main research question is: Can we understand early language development better by better approximating the language a young child might actually encounter? A secondary question is whether we can use a predictive neural network model to derive more accurate network representations than sliding window co-occurrence models. Once we accurately model toddler vocabulary growth, we can then apply this knowledge to model possible future growth trajectories.

2. The Corpus

Our first goal was to use a broader range of toddler language input than has been seen in other language modeling studies. Thus, in the current paper we strived to create a richer, more representative, corpus to derive semantic networks from. Specifically, we focused on North American English-speaking children, to avoid confounding cultural and language differences within our work. No matter how exact the language processing technique is that derives word similarities, if the resource used initially is unrepresentative, the final product will be too.

Though some past work has utilized a corpus of child language input (CHILDES) previously, young children receive language input in many other forms besides conversations with caregivers. Nowadays, young children experience many forms of media, including movies, TV shows, music, and more. TV or movie input, though perhaps not as rich as in-person conversation, still provides children with rich, grammatical language and contextual scenes to capture the child's attention. In fact, toddlers exposed to child-directed programming had better outcomes at age four than those that viewed adult-directed TV (Barr et al., 2010). Further, video viewing can be related to both better receptive and expressive vocabularies (Linebarger & Walker, 2005). Children also receive input during story time. Most parents report reading storybooks with their children, many daily (Sénéchal & LeFevre, 2002). We also know shared storybook reading is a significant predictor of later language skills, and children receive different types of language than they would during normal conversation (Sénéchal & LeFevre, 2002; Fletcher & Finch, 2015). Parents also treat story time as an opportunity to model new vocabulary and ask children vocabulary-related questions (Flack et al., 2018). In other words, the language in story books geared towards young children most likely has a large influence on toddlers' current vocabulary. In short, children receive language from multiple sources involving multiple media, and their input from these varied sources has been shown to have an impact on a child's vocabulary development.

Using publicly available resources, we created a new corpus using caregiver input. Specifically, transcripts of the language said by MOTHER or FATHER (but not, for example, siblings, experimenters, or the child themselves), were taken sentence by sentence from the CHILDES database, limiting to conversations with children less than five years old (MacWhinney, 2000). The reasoning was two-fold. First, we wanted to retain transcripts that supplied the most naturalistic type of language representative of typical input, and many experimenter transcripts were related to scripted surveys and tasks. Second, many of the transcripts supplied by siblings and the children themselves are telegraphic in nature, lacking the necessary grammaticality to pull good quality semantics from using distributional models. Further, these transcripts also contain a large number of unknown words, further complicating the type of semantic content that would be found using our models.

Transcripts of popular young children's picture books were transcribed in the lab, separated by sentence, and G-rated movie transcripts created by fans for public use were also used for the corpus. All the books included featured full sentences, rather than, for example, word books, which lack the necessary sentence structure to pull semantic vectors from. The average sentence length was between nine and ten words. Books were chosen if they were geared toward young children and toddlers (roughly ages one to five) and were considered widely available. These books were transcribed for in-lab use. As more children's books are published, more transcripts may be added, especially to include those geared toward minority communities.

Movie transcripts were scraped from parent-geared online sites and were created by fans. Sites include foreverdreaming.org and fandom.com. Because they are not original scripts, small errors may be present. Only spoken language within each movie was used, including songs. We excluded any visual cues ("[character] walked down the street") and speaker designations, as this is not spoken within the movie itself and so not language a child would hear while watching the movie. All movies were rated G and in English, with most being created in the United States. Though most were animated, some were live action. Future expansions will likely include TV shows as well. As with the children's books, we hope to expand the representation within our dataset to include minority-geared media for future work.

Though there are other possible sources of linguistic input to a child, the sources compiled here covered a range of input that North American children growing up in English-speaking families typically receive. Corpus statistics are shown in *Table 1*.

	CHILDES	**Books**	**Movies**
Number	many	1,039	81
Sentences	1,105,870	54,213	92,919
Tokens	4,716,063	510,312	507,625
Types	27,337	5,895	5,822

Table 1: Toddler Corpus Statistics

Though there are large variations in the amount of screen media, book, and conversational exposure toddlers receive, we attempted to create a corpus representing that of the average North-American two-year-old. Here, we wanted this to represent the demographics of the children whose data we were modeling, which were for the most part, middle to upper middle class monolingual English-speaking children. As our results will show, the match between the population being modeled and the assumed linguistic environment is important. In the future, it will be important to expand this toddler corpus to other populations, as well as to diversify the demographics of our laboratory samples. In addition, future iterations of this work may also investigate different relative contributions to represent, for example, children with more movie and less book exposure, or less conversational exposure.

3. The Similarity Metric

Not only are the contents of the corpus important, the technique or algorithm used to calculate similarity statistics from that corpus could be as well. Here, we posit advanced language processing models (specifically neural networks) can provide richer similarity metrics with which to model semantic network structure, as compared to metrics used in previous work.

Though some past research investigating child semantic networks area has used CHILDES over adult-created measures, these studies only pulled sliding window co-occurrence statistics as the similarity measure (see Hills et al., 2010; Beckage et al., 2011). Hills et al. (2010), in particular, built networks by connecting any pair of words that had co-occurred (within a 5-word window) at least once in the corpus. Predictive neural networks designed to learn embeddings based on both context and co-occurrence may better capture semantic similarity. We used the cosine similarities between embeddings from skip-gram neural networks to quantify the degree of similarity between any two words. Using the created toddler corpus, we trained a Word2Vec model using the python gensim package, to compare against the pre-trained embeddings from the popular GoogleNews Word2Vec model, using all the same training parameters. The GoogleNews model has been used to find similarities in other child language studies utilizing skip-gram models (see Beckage et al., 2020). GoogleNews will be considered adult content versus the child content in our toddler corpus. Finally, the embeddings from the GoogleNews model were used to create a new Word2Vec that we then continued training using the toddler corpus. In other words, we wanted to fine-tune the adult model using toddler data. Because the GoogleNews Corpus is 100 billion words (as compared to ~5 million), this combined model may combine the breadth gained from such a large corpus with the depth, or context, of the toddler corpus.

4. The Lexical Networks

Now that we have accurate representations of the words children might know, we need to create networks representing the connections between these words, or the child's lexical structure. There are many ways to use the posited cosine similarity metric to create these networks. As noted previously, most studies investigating lexicons this way threshold, or only connect two words if their similarity score is above a certain value. Then each connection in the network is represented as binary, or unweighted. However, to utilize the richness Word2Vec embeddings, we were interested in creating fully connected, weighted networks. In other words, every node (or word) will be connected to every other node with a specific value representing how similar the two words are. Though we cannot feasibly know every word a child knows, we can get a subset of common words that a child knows using the CDI. From this checklist, we were able to pull word embeddings and calculate cosine similarities for 640 of the 680 words. Excluded words include words that appear on the CDI as both a verb and a noun, which are not differentiated in skip-gram models, multi-word phrases (we do not believe the sum or average of these multi-word phrases were equivalent to their whole meaning) or were stop words the GoogleNews corpus removed. The networkx package in python was used to create the semantic network. This fully connected, weighted network consists of 204,480 edges. There are three semantic networks, one using similarities derived from the toddler corpus (T), one from the adult or GoogleNews corpus (G), and one using the combined embeddings (C).

Finally, to investigate if skip-gram (Word2Vec) models characterize semantic similarity more richly than the other prevailing language processing technique in child network science, sliding window co-occurrence algorithms, we gathered co-occurrence statistics using a window size of five for the toddler corpus only (for reasoning, see Beckage et al., 2011). The network created from the resulting co-occurrence matrix was neither fully connected or weighted, but rather two words were connected within the network if they both appeared within the same five-word window at least once throughout the corpus or were not connected otherwise. This follows the work of previous researchers (Hills et al., 2010; Beckage et al., 2011). This network consists of the same 640 words and contains 84,035 edges. All network measures will be calculated similarly to the other networks, but without considering edge weight in the calculations (e.g., weighted degree vs. degree). The five-gram co-occurrence network (5) will only be compared to the toddler (T) network, as the other two networks cannot be directly compared to this one.

4.1 Network Measures

There are many different measures we can use to examine network structure, though not all can be performed on a fully connected network. We are not only interested in visualizing differences between the networks, but also if we can use network measures to predict words a specific child is ready to learn next. Further, we want to compare this predictive ability between the networks, assuming their structures are in fact different. At any point in time, a child's lexicon contains a subset of the network; there is that subset as well as the rest of the full network of words and connections the child does not know, based on their CDI. Based on the idea that children may be more ready to learn words that complement, or are more similar, to words they already know, some of the measures chosen took advantage of the connections between "known" and "unknown" words.

The first measure is each unknown word's, or node's, weighted degree. However, this is not the word's overall weighted degree in the full network, but rather that unknown word's weighted degree if it was connected in the child's subset of the network. A higher weighted degree means that word is more likely to be learned by the child, as it is more similar to those words already in the child's network subset. Hence, this measure is named "To-Known-Node Weight". Similarly, the five-word window network used simple degree (and could be considered "To-Known-Node Degree"). The highest weighted nodes are then taken in accordance with the number the child actually learned by the next timepoint, to calculate accuracy.

The second measure also utilizes edge weight, by sorting every edge between a known node and an unknown node by its weight. To use this as a predictive measure, the algorithm goes through the sorted list, from highest weight to lowest, choosing that edge and in turn the unknown node associated it with it *if* that node hadn't been chosen previously. This ensures the algorithm doesn't pick the same node multiple times, and predicts the same number the child in question actually learned, so predictive accuracy can be calculated. This measure is simply named "Edge Weight". This measure cannot be calculated for the unweighted five-word model.

The remaining network measures examined were centrality measures, the reasoning being that if a word is more central to the network, based upon being more related to many other words in the network, it has a higher probability to be learned than those with smaller-weighted connections. Here, for every unknown word, that node is placed into the child's known subset of the network, and every node's centrality is predicted. This gives a sense of how central that word is not overall, but in specific relation to the child's lexicon. Once every unknown node's centrality was calculated, the algorithm chose the top most-central ones, choosing the same amount the child learned by the next timepoint, to calculate accuracy. The centrality measures used were "PageRank", "Eigenvector Centrality", and "Load Centrality". Though other centrality measures exist, they were excluded because they too-closely aligned with the predictions of one of the considered measures (e.g., Eigenvector Centrality). All accuracies were compared to random selection of nodes. Random selection was performed multiple times, with average accuracy at each timepoint used for comparison.

5. Results

5.1 Network Overlap

A within-subjects ANOVA was conducted to compare the three Word2Vec networks' edges for each combination of two nodes. Pairwise t-tests revealed all comparisons to be significant (TvG: t(408,59) = 43.57, p<.001; GvC: t(408,959) = 1719.6, p<.001; TvC: t(408,959) = 1497.2, p<.001). We could not directly compare against the five-word network in this way. Additionally, over 200,000 weights are too many to sort through and inspect by hand to find trends or patterns of differences. Instead, we calculated multiple centrality measures, which essentially assign each word a score representing how important, or central, that word is to the overall network. We calculated a subset of the different possible centrality algorithms to get a range of measures. Though the scores themselves were all significantly different in ANOVA's, we were more interested in the overlap in the actual words ranked as highly important using these measures. From the scores, we gathered both the top and bottom 50 scoring words from each network, for each measure. Overlap between the networks can be found in *Table 2*. This overlap measures if any unique word appears in one of the other two corpora (or both), but the word does not have to appear in all three to be counted in this overlap measure. Expected examples of words that appeared in multiple top 50 lists include: monkey, balloon, cookie, blanket, puppy, and spoon. Some words were more unexpected, such as tractor, pumpkin, and rooster. These are surprising given more common animals (kitty), vehicles (car), and food (cheese, peas), were not common among the three networks.

Top 50					**Bottom 50**				
Centrality Measure	**TvG**	**TvC**	**GvC**	**Tv5**	**Centrality Measure**	**TvG**	**TvC**	**GvC**	**Tv5**
PageRank	3	9	7	7	*PageRank*	5	5	0	5
(Weighted) Degree	6	9	7	0	*(Weighted) Degree*	5	3	0	0
Clustering Coefficient	5	11	9	0	*Clustering Coefficient*	6	2	0	1
Load Centrality	3	7	9	1	*Load Centrality*	4	9	2	10
Eigenvector Centrality	5	11	9	0	*Eigenvector Centrality*	6	2	0	1

Table 2: Overlap Between the Top 50 and Bottom 50 Ranked Words From Each Network

5.2 Average Child Prediction

In order to gain a sense of predictive power, we used the average of CDI's collected in our lab to calculate accuracy on an "average child" from 16 to 30 months old. These We have vocabulary data for each month. We investigated not just how accurate each network measure was, but how that accuracy compared between the created networks. First, we investigated how each network measure compared to random choice, within network.

We first compared the Toddler network to random. Individual paired t-tests were Bonferroni-corrected for multiple comparisons. Over the 14-month period the lab has CDI data for, three prediction measures performed significantly better than random at accurately predicting the next words an average child would learn between months, and the other two performed marginally better. The accuracy rates per month can be seen in *Figure 1*.

On average across months, random only accurately predicted 15.54% of learned words correctly. In comparison, PageRank was correct 19.81% (t(13)=2.47, p<.05), Edge Weight 23.38% (t(13)=2.85, p<.05), and To-Known-Node Weight performed at 19.63% (t(13)=2.39, p<.05). Of the two marginal results, Eigenvector performed accurately 19.06% of the time (t(13)=1.93, p=.076), Load Centrality 19.38% (t(13)=2.07, p=.059).

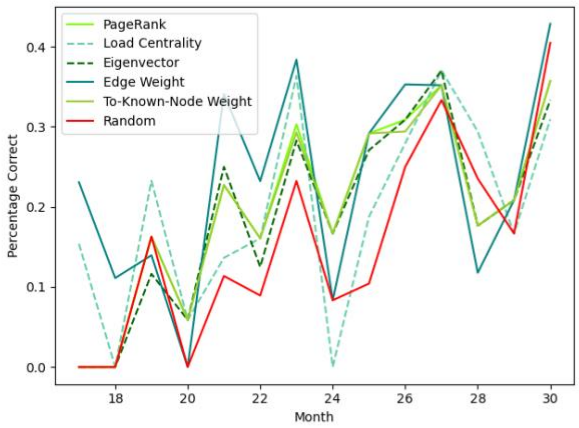

Figure 1: Toddler Network Compared to Random

Comparing random to the predictions associated with the GoogleNews network, we find one significant comparison, but in the wrong direction. Load Centrality (11.30%) actually performed worse than random at predicting learned words (t(13)=-2.90, p<.05). All other measures performed at an average of 15.24% to 18.66%.

The investigation of the combined network revealed similar results to that of the GoogleNews network. Again, Load Centrality (10.90%) performed significantly worse than random (t(13)=-2.23, p<.05). All other measures showed no significant difference from random, ranging from 17.41% accuracy to 17.65%, though Edge Weight (22.08%) was marginally better (t(13)=1.84, p=.089). Significant and marginal differences for the Google and Combined networks can be seen in *Figure 2*.

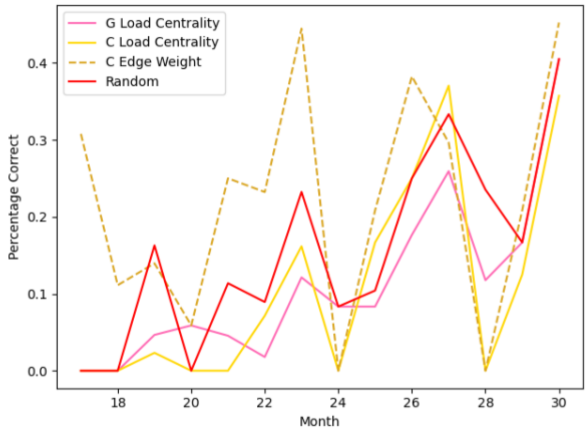

Figure 2: GoogleNews and Combined Versus Random

Next, we compared network accuracies to each other, to find if one network and measure outperforms other child lexicon representations. As before, all p-values were Bonferroni corrected for multiple pairwise comparisons. Only significant differences will be discussed. Unsurprisingly, the toddler network's Load Centrality accuracy was significantly better than that of both the GoogleNews and Combined networks' (t(13)=2.97, p<.05; t(13)=3.18, p<.01, respectively). See *Figure 3*.

None of the other measures' predictions were significantly different between networks, though there were a few marginal differences found. The toddler network marginally outperformed the GoogleNews network on Eigenvector Centrality (t(13)=1.89, p=.081) and Edge Weight (t(13)=1.94, p=.075).

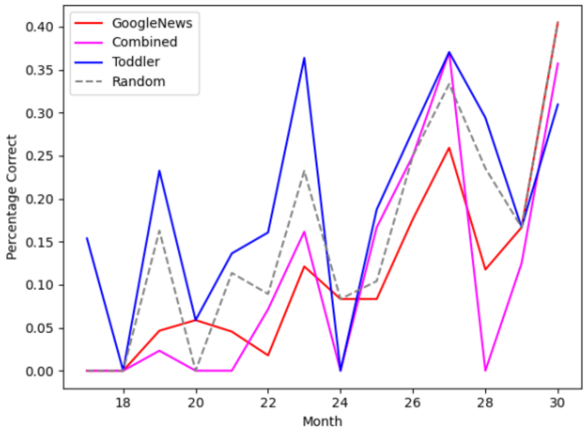

Figure 3: Load Centrality Network Comparison

Finally, we were interested in not comparing different training corpora but rather different similarity algorithms, specifically investigating an often used sliding window co-occurrence approach. The predictions based on the five-words window approach performed significantly worse than random on every measure investigated. PageRank was correct only 11.16% (t(13)=-3.21, p<.01), Eigenvector performed accurately 10.33% of the time (t(13)=-3.20, p<.01), Load Centrality 11.33% (t(13)=-3.09, p<.01), and To-Known-Node Weight (Degree) performed at 10.92% (t(13)=-3.25, p<.01). These results can be seen in *Figure 4*.

In accordance with the previous results, the toddler window approach's network also performed significantly worse than the toddler Word2Vec network on every measure (PageRank: t(13)=4/61, p<.001; Eigenvector: t(13)=4.49, p<.001; Load Centrality: t(13)=3.38, p<.01; To-Known-Node Weight: t(13)=4.71, p<.001).

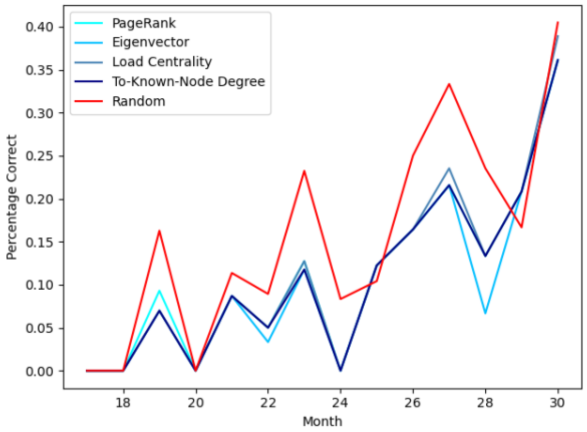

Figure 4: Edge Weight Network Comparison

6. Discussion

Overall, the present work provides evidence that using toddler input corpora, word embeddings and similarities drawn from neural network models such as Word2Vec, and fully-connected, weighted networks can provide a level of accurate word-learning prediction better than random chance, embeddings trained on adult-language corpora, and toddler sliding-window co-occurrence similarities. Though further work is needed to generalize this work for use with individual children, it leads us one step closer to understanding the mechanisms underpinning vocabulary and semantic growth. In addition, the present research adds rich information that could be used to create individualized learning materials and inform interventions for at-risk children. More importantly for the language research community as a whole, this work highlights the need to pay particular attention to the language source used depending on the population of interest.

6.1 Available Resources

Moving to the creation of the toddler-input corpus, we expect that improvements to the corpus itself will be made iteratively as new data becomes available. The corpus is small by today's standards. Possible future additions include TV show or YouTube video transcripts, music lyrics, PG-rated Movies, and possibly even learning materials or language used during game-play. Further, only readily-available resources were used in the current iteration of the toddler corpus. Future iterations could hopefully include data and resources collected from other child-development labs.

As previously stated, our present work investigated an average North American, English-speaking, monolingual toddler. However, there is enormous variability in the amount and type of input even just American children receive. Future work can investigate the relative contributions of different types of input (e.g., movies versus storybooks). Though out of reach right now, individualized corpora based on reported child input could be used to model individual development. Further, future research should expand on these findings by including other cultures and other languages. Based on our conclusion that the corpora of language used in modeling studies does matter, materials representative of minorities should be included to model more inclusive samples, and similarly large toddler corpora in other languages need to be created in order to expand this research across languages.

In a similar vein, we can only include words from the CDI, which we know doesn't fully capture a child's lexicon. This is important because a child's semantic knowledge or internal network may be quite different than the ones we are creating and using to model their language development. Though other routes may include using individual child data from CHILDES, or parent word diaries, these routes may still only provide a snapshot of a child's total vocabulary, and can be time-intensive.

6.2 Predictive Accuracy

Readers may have noticed that even though above random, the predictive ability of the network measures for an average child only range around 15% to 20% accuracy, compared to what the child actually learned. However, we want to model what words a child is most *ready* to learn, based on how that word complements their current vocabulary. The CDI data used is purely observational, and cannot account for the environment that child is actually in. Though a child might have an easy time learning the word for another animal, there may be no opportunity during that month to see and actually learn a new animal name. Instead, that child may encounter and learn many clothing items. In other words, we do not expect these models to achieve very high accuracy on observational data; we simply want to know if these models are understanding something about language learning that is helpful. In the future, these models can be tested in the field, where we can initiate more controlled scenarios in which the children are actively taught the chosen words to compare against randomly selected words.

Further, we have predicted only the trajectory for an average child as of now, based on aggregate data. Though aggregate data provided a simple preliminary analysis, it is possible the systematicity one might expect to see over time in a single child may not be present in aggregate data. Our ultimate goal is to predict what words any single child is ready to learn next.

6.3 Future Work

Currently, we are pursuing logistic regression as another predictive model, using longitudinal CDI's collected monthly in the lab for typically-developing children between 18-30 months. Using the network measures discussed previously, as well as other child and network characteristics, we are in the process of gathering features and selecting ideal training parameters. Here, the goal is to predict whether any individual word should be learned by the child or not, based on collected features. Because we know a specific subset of words learned each month, this dataset is ideal for a predictive language model. Preliminary results suggest logistic regression is a promising avenue for node-by-node prediction, regardless of the child the data comes from.

Once more fine-tuned, this new predictive model will be compared to the other prevailing growth model used in the field of language modeling, preferential attachment (Steyvers & Tenenbaum, 2004; Hills et al., 2010; Beckage et al., 2011). Instead of choosing each node individually, this algorithm successively chooses unknown nodes to-be-learned with probability proportional to their degree. Multiple iterations are averaged, which start by randomly choosing one of the known nodes in the network. Then an unknown node connected to this node is chosen, and an unknown node connected to that node is chosen, and so on. The theory behind this algorithm hinges on the notion that all successive nodes' ability to be learned depend on its previous counterpart being learned. New growth models with different approaches may perform better.

The ultimate goal is to create a generalizable, predictive model. It is highly possible that children at different ages, even month to month, require different model features and parameters to predict growth. Further, we are interested in predicting vocabulary for children who are at risk of language disorders, or bilingual children who must learn two lexicons. We know that a significant proportion of children who have relatively small early vocabularies will continue to have persistent language delays and even poorer language skills through adolescence (Manhardt & Rescorla, 2002; Rescorla, 2009). We think this work can inform targeted interventions tailored specifically to the individual, to support such children. However, the underlying mechanisms by which children who lag in vocabulary growth learn new words may be quite different than that of monolingual, typically developing children and models that capture the learning of children in these populations will have to be developed and refined.

7. Conclusions

The present work shows multiple improvements over past work. We not only showed that similarity measures based on toddler corpora perform better than random while adult or even combined models do not, we also showed that using Word2Vec skip-gram similarities over sliding-window co-occurrence similarities provide richer and more accurate predictions (compare to Beckage et al., 2011 or Hills et al., 2010). This improvement can also be attributed to the use of a fully connected, weighted network. The sliding-window, and most other semantic models, threshold and binarize their networks, whereas here we utilize the range of similarities provided by Word2Vec. We do not know of any other research that utilized skip-gram embeddings to create weighted networks using cosine similarity for modeling vocabulary growth, though some work has suggested utilizing Word2Vec embeddings in predictive neural networks or calculating cosine similarities from LSA (Beckage et al., 2020; Steyvers & Tenenbaum, 2005).

The present work is a promising new direction in the pursuit of understanding language development. Despite the forward progress still required to implement the proposed model, valuable insights cam be gained for future modeling attempts. Our analyses not only suggest the need to be mindful when choosing similarity metrics or semantic network structure, but highlight the importance of the degree of representativeness that corpora from different populations achieve, based on the question of interest.

7. Bibliographical References

Asr, F. T., Willits, J., & Jones, M. N. (2016, August). Comparing predictive and co-occurrence based models of lexical Ssmantics trained on child-directed speech. In *CogSci*.

Barr, R., Lauricella, A., Zack, E., & Calvert, S. L. (2010). Infant and early childhood exposure to adult-directed and child-directed television programming: Relations with cognitive skills at age four. *Merrill-Palmer Quarterly (1982-)*, 21-48.

Beckage, N., Aguilar, A., & Colunga, E. (2015, July). Modeling Lexical Acquisition Through Networks. In *CogSci*.

Beckage, N. & Colunga, E. (2016). Language Networks as Models of Cognition: Understanding Cognition through Language. In Mehler, A., Blanchard, P., Job, B., & Banisch, S. (Eds) *Toward a Theoretical Framework for Analyzing Complex Linguistic Networks.* (pp. 3-28) Springer Series on Understanding Complex Systems. Springer. 348.

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

Beckage, N. M., Mozer, M. C., & Colunga, E. (2020). Quantifying the role of vocabulary knowledge in predicting future word learning. IEEE Transactions on Cog. and Dev. Sys., 12(2), 148-159.

Beckage, N., Smith, L., & Hills, T. (2011). Small worlds and semantic network growth in typical and late talkers. PloS one, 6(5), e19348.

Colunga, E., & Sims, C. E. (2017). Not only size matters: Early-talker and late-talker vocabularies support different word-learning biases in babies and networks. *Cognitive science*, *41*, 73-95.

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, i-185.

Fernald, A., & Marchman, V. A. (2012). Individual differences in lexical processing at 18 months predict vocabulary growth in typically developing and late-talking toddlers. *Child development*, *83*(1), 203-222.

Flack, Z. M., Field, A. P., & Horst, J. S. (2018). The effects of shared storybook reading on word learning: A meta-analysis. Developmental psychology.

Fletcher, K. L., & Finch, W. H. (2015). The role of book familiarity and book type on mothers' reading strategies and toddlers' responsiveness. Journal of Early Childhood Literacy, 15(1), 73-96.

Fourtassi, A., Bian, Y., & Frank, M. C. (2020). The growth of children's semantic and phonological networks: Insight from 10 languages. *Cognitive Science*, *44*(7), e12847.

Gershkoff-Stowe, L., & Smith, L. B. (2004). Shape and the first hundred nouns. Child Dev. 75(4), 1098-1114.

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.

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.

Howell, S. R., Jankowicz, D., & Becker, S. (2005). A model of grounded language acquisition: Sensorimotor features improve lexical and grammatical learning. *Journal of Memory and Language*, *53*(2), 258-276.

Hutchison, K. A. (2003). Is semantic priming due to association strength or feature overlap? A microanalytic review. *Psychonomic bulletin & review*, *10*(4), 785-813.

Jimenez, E., & Hills, T. T. (2017). Network Analysis of a Large Sample of Typical and Late Talkers. In CogSci.

Colunga, E., & Smith, L. B. (2005). From the lexicon to expectations about kinds: a role for associative learning. Psyc. Review, 112(2), 347.

Kim, Y., Yim, D., Kim, Y., & Yim, D. (2018). Differences of early semantic relatedness between late talkers and typically developing children. Com Sci & Dis., 23(4), 845-857.

Linebarger, D. L., & Walker, D. (2005). Infants' and toddlers' television viewing and language outcomes. *Am Behav Sci*. *48*(5), 624-645.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior research methods, instruments, & computers*, *28*(2), 203-208.

MacRoy-Higgins, M., Shafer, V. L., Fahey, K. J., & Kaden, E. R. (2016). Vocabulary of toddlers who are late talkers. *Journal of Early Intervention*, *38*(2), 118-129.

MacWhinney, B. (2000). The CHILDES Project: Tools for analyzing talk. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Manhardt, J., & Rescorla, L. (2002). Oral narrative skills of late talkers at ages 8 and 9. Appl. Psycholinguistic, 23(1), 1-21.

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.

Moyle, M. J., Weismer, S. E., Evans, J. L., & Lindstrom, M. J. (2007). Longitudinal relationships between lexical and grammatical development in typical and late-talking children.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402-407.

Perry, L. K., & Samuelson, L. K. (2011). The shape of the vocabulary predicts the shape of the bias. *Frontiers in Psychology*, *2*, 345.

Pexman, P. M., Holyk, G. G., & Monfils, M. H. (2003). Number-of-features effects and semantic processing. *Memory & Cognition*, *31*(6), 842-855.

Rescorla, L. (2009). Age 17 language and reading outcomes in late-talking toddlers: Support for a dimensional perspective on language delay. J. Speech Lang. Hear.

Sénéchal, M., & LeFevre, J. A. (2002). Parental involvement in the development of children's reading skill: A five-year longitudinal study. Child development, 73(2), 445-460.

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cog. Sci.*, *29*(1), 41-74.

Vankrunkelsven, H., Verheyen, S., Storms, G., & De Deyne, S. (2018). Predicting lexical norms: A comparison between a word association model and text-based word co-occurrence models. *Journal of cognition*, *1*(1).