# How to Build a Toddler Lexicon: The Importance of Representative Data Sources

by

Jennifer Marie Weber
B.A., University of Wisconsin-Madison, 2016
M.S., University of Colorado Boulder, 2020

A thesis submitted to the
Faculty of the Graduate School of the
University of Colorado in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Psychology and Neuroscience
Institute of Cognitive Science
2024

Committee Members:
Eliana Colunga
Aaron Clauset
David Huber
Pui Fong Kan
James Martin
Lei Yuan

Weber, Jennifer Marie (Ph.D., Psychology and Neuroscience, Cognitive Science)

How to Build a Toddler Lexicon: The Importance of Representative Data Sources

Thesis directed by Prof. Eliana Colunga

The present thesis models young children's vocabulary acquisition using different semantic network structures in hopes of better understanding the underlying cognitive representations children have about the words they know. Throughout each chapter, we predicted the specific word-by-word learning of toddlers using words' centrality in different network representations. First, we found that gathering semantic similarity using distributional methods from a child-directed language corpus better predicts specific learning than using an adult language corpus. We then investigated three different semantic representations, the same distributional method, word associations, and feature norms. Each semantic type had both a child-based and adult-based version. We found modest confirmation that child-based sources of semantic similarity are indeed better than adult-based sources, and further found that association-based semantics best capture the knowledge young children have and use to learn new vocabulary. The current work lays a foundation for understanding the underlying structure of young children's vocabulary knowledge.

# Acknowledgements

First, I would like to thank the first influence in my life, my parents, for instilling in me a love of learning. I would also like to thank those professors that started me on my research journey, including Vanessa Simmering and Timothy Rogers, for instilling in me a love of psychology, development, and language. I could not have navigated research or graduate school without my advisor, Eliana Colunga, as well as other professors from both my department and the rest of the university. Similarly, the friendships I made at CU Boulder helped keep me sane and motivated. And of course, I would like to thank my husband, Josh, for supporting me through every season college, graduate school, and this dissertation has brought. Finally, Charlotte, for showing me why I began this work in the first place, and what life can hold after graduate school.

# Table of Contents

# Tables

# Figures

# Chapter 1

# Introduction

The present thesis models young children's vocabulary acquisition using different semantic network structures in hopes of better understanding the underlying cognitive representation children have about the words they know. By using different network structures to predict word-by-word learning over time, we hope to better understand which representations children have and use to learn new words. As such, this thesis heavily utilizes network science techniques; but what is network science and why is it useful for this endeavor?

Network science, or the use of network representations, has become increasingly relevant for studying psychological phenomena. Networks have been used to represent the structure of numerous systems inherent in our world, including biological (Gosak et al., 2018; Panni et al., 2020), social (McCulloh & Carley, 2011; McPherson et al., 1992), technological (Tu et al., 2018), epidemiological (Pastor-Satorras et al., 2015; Wang et al., 2016), and of course cognitive systems (e.g., Kenett & Hills, 2022). In networks, nodes often represent a discrete entity, whether it be a person or animal, a neuron, or a word; and the edges or connections between these nodes represent some sort of relationship between these entities. This previous work not only uses networks to understand the structure of systems, but to model complex processes such as longitudinal trajectories or change in these systems (Boccaletti et al., 2006; Barabási & Albert, 1999). Much of this work examines how to represent these systems to best understand how they operate (e.g., Baronchelli et al., 2013; Siew, 2020).

## 1.1 Networks and Their Applications

Networks can be harnessed to understand changing systems (e.g., Boccaletti et al., 2006). For example, in social networks, research has shown how changes in the structure of these

networks, such as broken connections or members joining/leaving, impacts team effectiveness and causes cascading effects This work further shows that we may be able to predict significant structural changes based on current network structure (McCulloh & Carley, 2011; McPherson et al., 1992). Epidemiological studies have often used networks to model the spread of diseases, predicting how best to stem an epidemic (Pastor-Satorras et al., 2015). The applications for network science to predict change have been shown in a wide variety of fields, from the analysis of sports strategies and team dynamics to predict future league winners to the analysis of bank transactions to predict accumulated savings (Mattsson & Takes, 2021). Research has found consistencies in network change across fields, showing how general structural features can impact propagation speed, synchronizability, and computational power in changing networks (Watts & Strogatz, 1998). In particular, small world characteristics, such as high clustering of nodes and short paths between nodes, has been shown to drive such changes. in 1999, Barabási and Albert suggested that change and growth in natural networks may be driven by a process termed preferential attachment, whereby new nodes are added preferentially to already well-connected nodes. Further research in diverse fields has shown this principle to hold (e.g., Capocci et al., 2006; Newman, 2001). This capacity to capture the way that representations and processes drive change in a system is critical for the work in this thesis, which is concerned with the growth of early lexicons.

With an influx of new methods and techniques to examine network architecture, many cognitive scientists have looked to network science as a way to understand cognitive processes. Many advances in our understanding of cognition have come about from network modeling techniques. From attempting to visualize the entire mind as a network (Hills & Kenett, 2021), to case studies such as mapping specific brain networks (e.g., Sporns et al., 2004) or understanding

the cognitive costs of sequential processing (Park & Levy, 2009), there are many demonstrations that network science can help elucidate human cognition and its underlying processes (Siew et al., 2019; Hills & Kenett, 2021). Network science has shown that we may map the complex networks of the world in our own minds, indicating how much those network properties influence our thought and behavior (Lynn & Bassett, 2020). Further, just as one can visualize and gain insights from the changing dynamics of biological and social systems using networks, so can we model cognitive phenomena dynamically or over time (Siew et al., 2019; Baronchelli et al., 2013). In fact, a recent special issue in TopiCS was dedicated to reviewing the contributions of networks science to our understanding of human cognition (see Kenett & Hills, 2022).

One area of cognitive science in which network science has a successful decades-long history is in representing and understanding the mental lexicon (e.g., Collins & Quillian, 1969, Collins & Loftus, 1975). That is, understanding how we represent words and their meanings in our minds. In these networks, nodes typically represent individual words, and the edges between them typically represent how similar or related the underlying meanings of those words are (for review, see Beckage and Colunga, 2016). In general, semantic networks created from multiple sources all show the same small-world structure of other natural systems (Steyvers & Tenenbaum, 2005). Previous work further aimed to use network representations to investigate how semantic structure impacts processing in the mind. For example, both spreading activation in semantic networks as well as the path length between words in networks predicts response times in lexical and semantic behavioral tasks (Rotaru et al., 2018; Marko & Riečanský, 2021; Kenett et al., 2017). Network representations of spoken language also predict speech errors during behavioral tasks (Chan & Vitevitch, 2010). In addition, research not just investigates how

language is represented in the mind, but how it changes over time. Evidence suggests that people's idiosyncratic experiences impact their internal lexical representation and consequently their cognitive performance. For example, the placement of nodes and their connections in an individual's network strongly relate to that person's responses on both verbal and memory tasks (Wulff et al., 2022). Further, younger and older participants had significantly different lexical network structures. In general, networks can help us examine how representations impact processing as we experience disease, and particularly relevant for the current thesis, as we age (Wulff et al., 2022; Vitevitch, 2021; Castro, 2021).

### 1.2 What Makes a Good Cognitive Model

Though the following chapters will delve deeper into how we can use networks in the pursuit of understanding vocabulary development in particular, we next want to turn to what we need to include in such models so we can evaluate their ability to capture the cognitive phenomenon we are attempting to model. Here that is a young child's developing semantic space, or the representation they have about the words in their budding vocabulary.

Christiansen & Chater (2001) propose three criteria to evaluate computational models of language processing: 1) data contact, the degree to which the model captures human psycholinguistic data; 2) task veridicality, the match between the task given to the model and the task given to humans; and 3) input representativeness, the match between the information available to the model and the information available to humans. Sims and colleagues (2013) add another criteria particularly relevant for developmental work: 4) temporality, the degree to which the model captures continuous change including the processes that drive that change. The work throughout this thesis takes into account each of these criteria, but with a special emphasis on input representativeness. Specifically, we will return to these criteria at the end of the thesis to

discuss which of these criteria we have captured well, and which still need improvement. We will also discuss future avenues to continue to better capture each criterion.

### 1.3 The Present Thesis

The following chapters of this dissertation build upon each other to explain a body of work examining how the representativeness of the data we use in our models of early vocabulary development impact our results and conclusions. Though we will explain our individual methods in more depth throughout each chapter, here we give just a high-level overview of our research.

Chapter 2 aimed to better represent the semantic space of the typical toddler compared to the often-used metrics of past work. We did this by creating and evaluating a new corpus representing different types of language input toddlers typically hear. To evaluate the new corpus, we created a vocabulary network using distributional semantics over the corpus, and predicted individual words learned by the average American toddler using network centrality measures. The accuracy of the resulting predictions suggested that lexical networks created from child sources better represent young children's language learning compared to networks created from an adult language corpus. We concluded that the toddler corpus and the use of distributional semantics were indeed capturing something about the semantic structure children use in their word learning.

Chapter 3 builds on the findings of Chapter 2 by using the same two toddler and adult distributional networks and the same prediction method, but predicting 84 individual children's word-learning month-by-month over one year, from the ages of about one and a half to two and a half years. The findings of Chapter 3 enhance the previous results, and showed that across many individual children, the toddler network resulted in statistically better predictions than those from the adult network. It also showed that predictions based on a child's specific, in-the-moment,

vocabulary structure performs better than predictions based on general network structure. Both Chapter 2 and 3 show that having representations that more align with a child's language input, as compared to adult's language, better represent their word knowledge and future word learning.

Finally, Chapter 4 presents recent work investigating not only whether a representation created specifically to reflect a child's knowledge is better than one based on adults, but whether certain types of semantic representations are better than others at capturing a child's underlying semantic space. This work builds on the past chapters by introducing other ways of capturing semantic similarity in addition to the distributional one used in the past two chapters. Though it becomes a bit more complex, the work in Chapter 4 suggests that overall, semantic data created with children in mind continues to be more representative of a young child's internal word knowledge than that created solely based on adults, and that semantics based on word associations may best capture that internal knowledge structure compared to other semantic representations.

In sum, in this body of work we strive to understand how different representational choices impact the accuracy of our model of a child's underlying vocabulary knowledge. We believe this body of work exemplifies the importance of having data sources that represent the population of interest, as other sources could result in less optimal representations, leading to the wrong conclusions and inaccurate theories. Creating or finding data representative of the population at hand is important when measuring other psychological phenomena, and likely has a place in other fields beyond psychology, such as linguistics, speech pathology, ecology, business practices, machine translation, and more.

# Chapter 2

## Representing the Toddler Lexicon: Does the Corpus and Semantics Matter?

Chapter adapted from: **Weber, J.**, & Colunga, E. (2022, June). Representing the Toddler Lexicon: Do the Corpus and Semantics Matter?. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 3960-3968).

### 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 GoogleNews. 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

## 2.1 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.

### 2.1.1 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

GoogleNews (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.

### 2.1.2 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 (Fernald & Marchman, 2012; Moyle et al., 2007), 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.

## 2.2 Methods

### 2.2.1 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 2.1.

**Table 2.1 Toddler Corpus Statistics**

| | 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 |

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.

### 2.2.2 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.

### 2.2.3 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.

### 2.2.4 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.

## 2.3 Results

### 2.3.1 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 2.2. 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.

**Table 2.2 Overlap Between Top 50 and Bottom 50 Ranked Words From Each Network**

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

### 2.3.2 Average Child Prediction

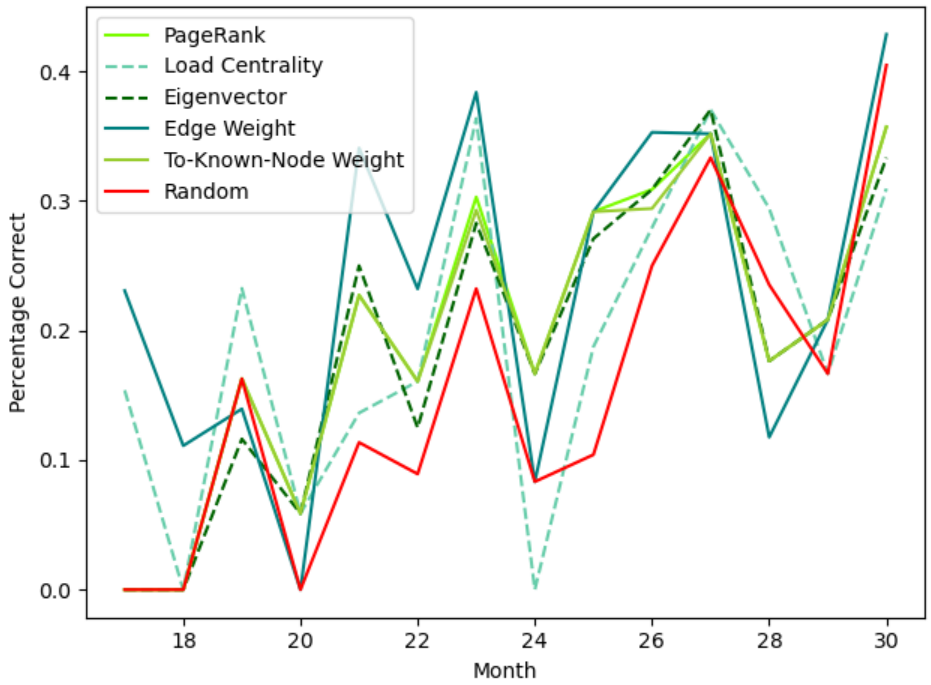
In order to gain a sense of predictive power, we used the average age at which the percentage of children who produced each word reached 50 percent, based on data from WordBank (Frank et al., 2017). This measure provided us with monthly knowledge, and thus word learning, for an "average child" from 16 to 30 months old. 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 2.1.

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 2.1 Toddler Network Compared to Random**



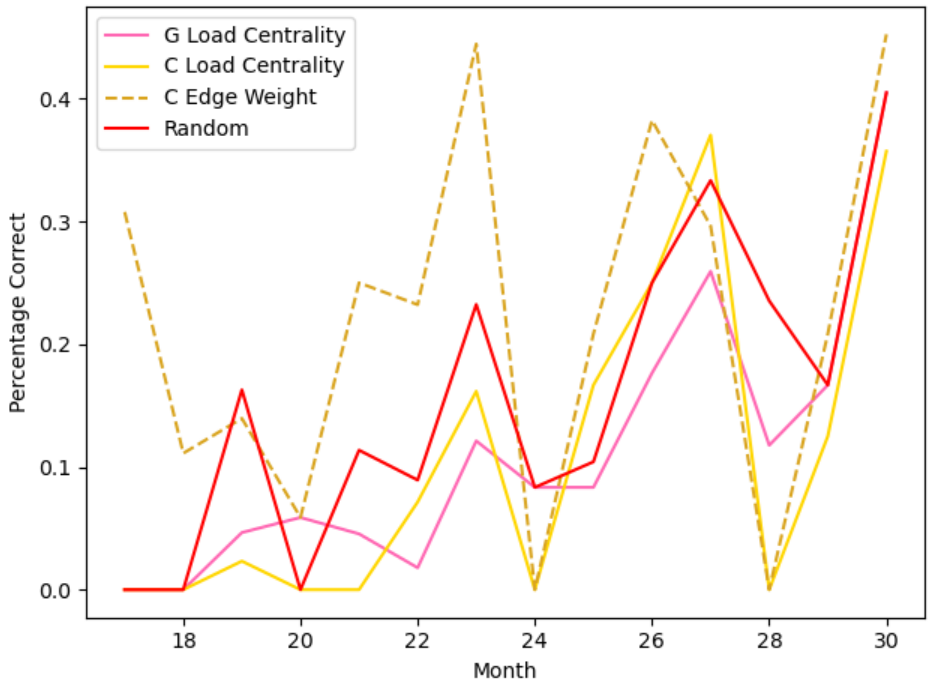

***Note.*** For the Toddler network, Pagerank, Edge Weight, and To-Known-Node Weight all perform better than random prediction.

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 GoogleNews and Combined networks can be seen in Figure 2.2.

**Figure 2.2 GoogleNews and Combined Versus Random**



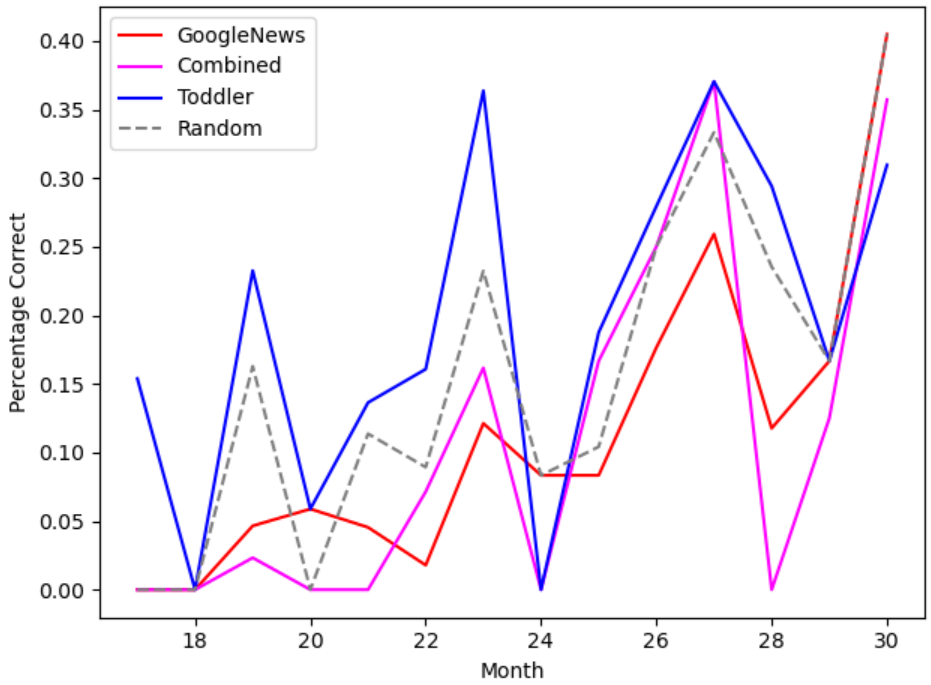

***Note.*** No GoogleNews or Combined network measures perform better than random prediction.

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 2.3.

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 2.3 Load Centrality Network Comparison**

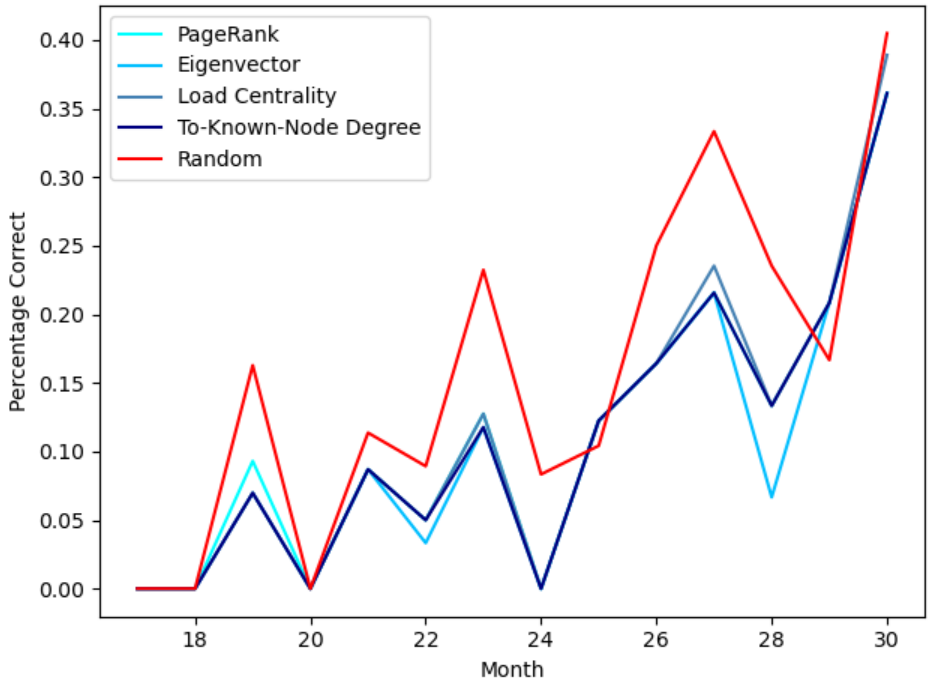



***Note.*** For Load Centrality, the Toddler network performs better than the GoogleNews and Combined networks.

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-word 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 2.4.

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 2.4 Sliding Window (5) Versus Random**




***Note.*** All Sliding Window network measures performed worse than random prediction.

## 2.4 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.

### 2.4.1 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.

### 2.4.2 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.

### 2.4.3 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.

### 2.4.4 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.

# Chapter 3

# Predicting Individual Vocabulary Learning: The Importance of Approximating Toddlers' Linguistic Environment

Chapter adapted from: **Weber, J.**, & Colunga, E. (under revision). Predicting Individual Vocabulary Learning: The Importance of Approximating Toddlers' Linguistic Environment. *Canadian Journal of Experimental Psychology*.

## Abstract

Using network representations of the lexicon has expanded our understanding of vocabulary growth processes and vocabulary structure during early development. Though past investigations of children's vocabulary growth used both adult- and child-based metrics for connecting words in lexical networks, we know that child-directed and adult-directed speech have different properties (Hills, 2013). More recently, Weber & Colunga (2022) demonstrated that predictions of early vocabulary norms can be improved by using network representations based on a corpus incorporating language a young child might typically hear. The present work goes a step further by evaluating the accuracy of network representations based on embeddings gathered from either an adult or child language corpus. We predicted the specific words that individual children add to their vocabulary over time, using a longitudinal dataset of 86 monolingual English-speaking toddler's vocabulary growth from 18 to 30 months of age. The toddler-based network predicted word-learning more accurately than the adult-based network. Further, prediction methods that took into account the individual child's particular network structure rather than overall network

connectivity were more accurate as well. These results highlight the importance of tailoring representational and processing choices to the population of interest.

*Keywords:* language acquisition; semantic networks; word embeddings

## 3.1 Introduction

Network representations have been used for decades to understand the way that our minds represent the words we know and the processes that operate on these representations, the so-called mental lexicon (e.g., Collins & Quillian, 1969, Collins & Loftus, 1975). In these networks, nodes typically represent individual words, and the edges between two words typically represent the relationship between those two words (for review, see Beckage & Colunga, 2016). These network representations are useful to understand both static lexical structure and how we process language. For example, both spreading activation as well as the path length between words in networks predicts response times in lexical and semantic behavioral tasks (Rotaru et al., 2018; Marko & Riečanský, 2021; Kenett et al., 2017) and network representations of spoken language predict speech errors during behavioral tasks (Chan & Vitevitch, 2010). Further, network science techniques have also been used to examine how our lexicons change over time. Evidence suggests that people's idiosyncratic experiences as they age or cope with disease impact their internal lexical representation and, consequently, their cognitive performance (Wulff et al., 2022; Vitevitch, 2022; Castro, 2022). For example, the placement of nodes and their connections in an individual's network strongly relate to that specific person's responses on verbal and memory tasks, and older adults have significantly different lexical network structure than younger adults (Wulff et al., 2022).

### 3.1.1 Lexical Networks and Language Development

Modeling the way lexical networks change over time is particularly interesting In the context of language development. Using network representations of the lexicon has expanded our understanding of vocabulary growth processes and vocabulary structure during the first few years of life (e.g., Hills et al., 2009; Hills et al., 2010; Beckage et al., 2011), and how this may be different for different populations, for example children who are learning more than one language (Bilson et al., 2015) or children who lag in their language development compared to their peers (Jimenez & Hills, 2017, Beckage et al., 2010). In this work we model the way the lexicons of individual children grow over time, and examine the way in which tailoring representational choices to specific populations (young children) and specific individuals impacts the accuracy of these models. Specifically, we compare the prediction accuracies of network representations based on representations derived from two different corpora: 1) an adult-directed corpus comprised of GoogleNews articles and 2) a toddler-directed corpus that incorporates transcripts of child-directed conversations, the text of picture books written for preschoolers, and dialog transcripts from G-rated movies.

### 3.1.2 Approximating Which Words a Child Knows

Building a model of a child's lexicon involves first approximating which words a child knows, and then estimating the relationship between these words. Many researchers have used the MacArthur-Bates Communicative Development Inventory Words and Sentences (CDI) as the basis for accomplishing the first step (Fenson et al., 1994). The CDI is a 680-word parent checklist that asks parents to report, for each word, whether their child says it or not. Vocabulary checklists filled out by parents have been found to be a good method to approximate toddlers' productive vocabularies at this age, when children's vocabularies are still relatively small. The

CDI in particular has been shown to be valid and reliable for toddlers between 16 and 30 months of age, and has been normed for these ages on a cross-sectional sample of children from the United States (Fenson et al., 1994). A large repository of CDI's has been created and utilized for developmental research in multiple languages, further solidifying the utility of the CDI as a tool for examining vocabulary development (Frank et al., 2017; Fourtassi et al., 2020).

Work using the CDI as the basis for children's vocabulary knowledge has led to important insights about language development. These include the relationship between lexical development, syntax, and processing speed (Moyle et al., 2007; Fernald & Marchman, 2012; Borovsky, 2022), and the relationship between lexical variability and language proficiency in toddlers (MacRoy-Higgins et al., 2016; Kim & Yim, 2018). Work using the CDI as an approximation of children's vocabulary has shown how the words in a child's vocabulary shape how children learn new words (Gershkoff-Stowe & Smith, 2004; Perry & Samuelson, 2011; Colunga & Sims, 2017). Indeed, researchers have evaluated the potential of the CDI as a diagnostic tool to predict future language outcomes. For example, using a child's percentile based on the total number of words reported by their parent as an early screener for possible language disorders (Heilmann et al., 2005) or using the relative proportions of different types of words a child knows in a similar way (Fasolo & D'Odorico, 2002; Weber & Colunga, 2021; Perry et al., 2023). Indeed, knowing the specific words – not the number of words or the types of words, but the specific words – a child knows out of the CDI increases a neural network's ability to predict future specific word learning, over and above knowing other child demographic information (Beckage et al., 2020).

At a glance, lexical networks based on the CDI are usually unweighted and sparse, and show evidence of the same small-world and scale-free structure of adult lexicons (Beckage et al.,

2011). As with other natural systems and adult lexical networks (see Barabási and Albert, 1999), these networks likely add new words based on properties of the network such as node degree (Hills et al., 2009; Hills et al., 2010). Further, this structure can be observed across languages: children tend to learn words that are highly connected in these networks earlier than words with smaller neighborhood densities regardless of the specific language they are learning (Fourtassi et al., 2020). In fact, the structure of networks based on the CDI closely align with that of networks built using children's language samples as the source of their vocabulary (Amatuni & Bergelson, 2017). In short, the work using CDI-based vocabularies to model early word learning has been fruitful for understanding the mechanisms of vocabulary growth, suggesting that the CDI is a good way to approximate a child's vocabulary. Next, we review the different ways in which the relationship between these words can be estimated, particularly for young children.

### 3.1.3 Estimating the Relationships between Words

Earlier work modeling children's lexicons often used adult-based similarity measures to estimate the relationship between the words in the network. These similarity measures can come from a variety of sources, such as association norms, feature norms, or distributional semantics. For example, Steyvers and Tenenbaum (2005) showed that networks created from either adult word associations, thesaurus entries, or WordNet all share the same structural properties; those that are also seen in other complex, natural systems. Additionally, specific connectivity properties of words in all three networks, like number of neighbors, align with words' predicted estimated age of acquisitions. In modeling early vocabulary growth, CDI-based networks that connect words based on adult feature norms (Hills et al, 2009a) or based on adult association norms match words' age of acquisition (Hills et al., 2009b; Bilson et al., 2015). Distributional semantics is another way to estimate the relationship between words. In this case, large corpora

are used to infer the meanings of words from the way in which words are used in the corpus. Recent work uses neural network-based distributional semantics, or algorithms such as word2vec and GloVe, to approximate the meanings of words. These methods learn a vector representation (or an embedding) for each word using a neural network that predicts the next word in the source corpus, and the resulting embeddings have been shown to capture aspects of human performance (Mikolov et al., 2013). Investigating children's lexical growth, Amatuni and Bergelson (2017) showed that networks created using GloVe embeddings show the same structural properties of semantic networks created using other similarity metrics for children between 6 and 30 months of age.

More recent work investigating children's lexicons has made a shift to utilizing child-based similarity metrics to create semantic network representations for a few reasons. First, as Christianson and Chater (2001) suggest, language models should strive for good input representativeness; the information available to a model should match the information available to the entity we are trying to model. Indeed, recent work has shown that child-directed and adult-directed speech have different properties (Jones at al, 2023), and that properties like contextual diversity in child-directed speech are correlated with age of acquisition (Hills, 2013). Second, 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.

There is also empirical evidence suggesting that aligning semantic representations to children's experiences can be advantageous in modeling children. For example, Howell et al. (2005) showed that neural networks that include sensorimotor features (size, color, has legs, is edible), or features presumably more in line with the knowledge a child has, better predict word learning than models without such features. Similarly, a recent study created child-based association norms by querying adults to name associates that would help a toddler learn words, and found these child-oriented norms (i.e., given *star*, respond *bright*) to be more predictive of normative vocabulary growth than adult-oriented associations (i.e., given *star*, respond *luminous*; Cox & Haebig, 2022). More recently, Weber & Colunga (2022) demonstrated that predictions of early vocabulary norms can be improved by using network representations based on a corpus that is more representative of the language environment of young children as compared to a corpus of adult-directed text such as GoogleNews.

Most past work attempting to create semantic networks more in line with children's experiences has utilized co-occurrence based distributional metrics over child-oriented corpora. Specifically, researchers have taken advantage of the CHILDES corpus, a repository of caregiver-child conversations contributed to by many researchers (MacWhinney, 2000). Using the co-occurrence of words in the corpus to construct lexical networks, Hills et al. (2010) produced networks with characteristics that closely align with age-of-acquisition norms. Also using co-occurrence of words in the CHILDES corpus to create lexical networks, Beckage et al. (2011) and Jimenez & Hills (2017) found structural differences in the networks of individual children with varying degrees of language proficiency, expanding our understanding of lexical structure from that of a normative child to individuals who may deviate from these norms. Some work has suggested that neural network-based distributional semantics might be particularly well

suited for modeling early vocabularies (Huebner & Willits, 2018; Asr et al., 2016). In fact, Weber & Colunga (2022) compared distributional semantics that were derived either through sliding-window co-occurrences or by using word2vec on the same child-directed corpus, including CHILDES. They found that network-based distributional semantics outperform co-occurrence-based semantics in predicting normed vocabulary growth.

In sum, using child-based metrics is beneficial in representing the lexical knowledge of the average child. The current study uses neural network-based distributional semantics on a toddler corpus that includes CHILDES corpus and other media sources American toddlers may be exposed to, to try and create a semantic space representative of that of a toddler's. Further, we explore this question in the context of modeling the changing lexical networks of individual children over time.

### 3.1.4 Predicting Which Words Children Will Learn

Past work investigating children's lexicons has explored how the lexical networks of an average child develop over time. Hills and his colleagues have worked extensively comparing multiple growth algorithms and exploring a variety of semantic network representations to better understand the processes underlying normative vocabulary growth. For this work, Hills and colleagues create normative vocabularies for each month between the ages of 16 to 30 months by assuming the vocabulary for each month is composed of those words which are known by at least 50% of the children in the normative sample. These monthly vocabularies, put together, serve as an approximation of the developmental trajectory of the average child's vocabulary. The findings using these normative data suggest that different growth algorithms worked best for networks created using different similarity metrics (Hills et al, 2009; Hills et al, 2010). Other work suggests the perceptual features of nouns may be particularly predictive of words' order of

acquisition (Peters & Borovsky, 2019). Weber and Colunga (2022) also suggest using network centrality measures, such as degree or eigenvector centrality, to add new words to the semantic network of a normed child more accurately captures specific word learning than randomly adding words to the lexicon. Moving from normative data to individual children's data, recent work indicates that lexical networks can be harnessed to model individual vocabulary trajectories, finding that different growth algorithms or networks created using different similarity metrics are better for different stages of development (Beckage & Colunga, 2019). Further, neural networks that take into account a child's specific word knowledge best predict that child's future word-by-word learning (Beckage et al., 2020).

When making predictions about the specific word learning of individual children, there are multiple ways to make such predictions. In the current work we will compare two methods that take advantage of individual words' placement or connectivity in a child's network to determine whether each word should be predicted to be learned or not by the model. More details will be provided in the methods. Of importance though is that one method, what we term static growth, involves making predictions about a word based on where it lies in the full CDI network, thus arguably using information about a word's semantics in the external environment. The other method, which we term dynamic growth, bases predictions on a word's connectivity within a child's individual vocabulary network, or using information about a word's semantics in a child's mental lexicon. Past work has suggested that both the structure of the linguistic environment and the current state of the individual's mental lexicon matter, but that information about the external environment may be particularly useful (Hills et al., 2009; Hills et al., 2010). Note that these two methods differ dramatically in terms of their computational complexity. The dynamic growth is more individualized, and arguably more representative of the individual

child's lexicon, but requires computing a word's connectivity in each child's network separately as network structure is different. In contrast, for the static growth algorithm word's connectivity in the overall network is the same across children, thus requiring less individual computation.

### 3.1.5 The Current Study

To summarize, our main research question is: would more accurately representing the language toddlers' actually encounter increase our ability to predict young children's specific word learning? To answer this question, we will predict individual words learned using semantic networks derived from two different sources: a toddler-directed and an adult-directed corpus. If input representativeness increases predictive accuracy, we expect that predictions based on the toddler networks will be more accurate than predictions based on the adult networks, as the former corpus is more representative of the knowledge of our population of interest, young children, than the later corpus. On the other hand, it is possible that the amount of data the two corpora are trained on matters more, as many of the most recent advances in language models are the direct result of training on extremely large corpora (e.g., Kalyan, 2023). Many adult English corpora are quite large, on the scales of billions of words. However, corpora representing minority populations, such as under-resourced languages or, in the present case, child language, are much smaller. For example, the entirety of the English CHILDES was only 3.5 million tokens in 2010 (Bååth, 2010). Here, the adult corpus contains approximately 100 billion tokens, and the toddler corpus only contains 5.75 million tokens. Thus, it is possible that there is a tradeoff; even if the adult corpus is less representative of a child's language environment, the sheer breadth of the information contained therein might be enough to yield more accurate predictions.

A secondary question we ask is whether further personalizing the prediction method to the individual trajectory being modeled impacts accuracy. That is, does dynamically making predictions based on each individual child's vocabulary structure increase our accuracy in predicting the next words learned by that child, over using the overall network structure to result? Again, if representativeness matters, one would expect that the dynamic, individualized, method will outperform the static, one-size-fits-all, method.

## 3.2 Methods

### 3.2.1 Participants

We will utilize a longitudinal, observational dataset to make predictions. To understand individual word learning properly, we need to use longitudinal metrics within child to make and then verify model predictions; collapsing across children and using cross-sectional data obscures important heterogeneity in children's specific word-learning. The dataset consists of 86 monolingual children beginning between 15.38 and 19.33 months of age ($M$ = 17.67, $SD$ = 0.97) and followed for 12 consecutive months. Each visit was spaced as close to one month as possible barring scheduling restrictions. This consists of up to 11 Pre-Post pairings per child we can utilize for model prediction and verification. At each visit, parents filled out the CDI indicating what words their child produced. Some children were not able to participate in every visit, and so months in which there is not a pre and post CDI were excluded, so that all predictions constitute a one-month period of learning. Data was constrained so that a child could not "forget" a word once it had entered their vocabulary.

### 3.2.2 The Corpora

In order to compare networks created using adult and child-oriented semantics, we need to first find or create such corpora. Though other developmental lexical research has used

CHILDES, we know that children receive language input in many other forms besides conversations with caregivers. Young children growing up in the United States experience many forms of media, including movies, TV shows and music. Media 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. Children also receive input during story time. Most parents report reading storybooks with their children, and we know children receive different types of language from shared reading than they would during a typical 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, suggesting storybook language likely influences a toddler's vocabulary (Flack et al., 2018). We used the same corpus of child-directed language created and used in Weber and Colunga (2022). This toddler corpus contains conversational input (Mother and Father transcripts from the CHILDES database), picture books (transcripts created in the lab), and G-rated movies (fan-created transcripts available online). Though there are many other possible sources of data, these covered a range of input young children typically receive. Corpus statistics are shown in Table 3.1. Though CHILDES constitutes the majority of the corpus, this is likely representative of many children's actual input.

**Table 3.1 Toddler Corpus Statistics**

| | CHILDES | Books | Movies |
|---|---|---|---|
| **Number** | 5,751 | 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 |

We contrast semantics taken from this toddler corpus to those from an often-used adult corpus, the GoogleNews corpus. The GoogleNews corpus has been used successfully to model human semantics in past work (e.g., Beckage et al., 2020; Handler, 2014; Church, 2017). We will term this corpus and the networks created using this corpus as the "Adult" network, in contrast to those networks created using the toddler corpus and termed the "Toddler" network. The adult corpus consists of 100 billion tokens, and 3 million types, including both individual words as well as some phrases.

### 3.2.3 Creating the Lexical Networks

We used a neural network-based distributional method, the word2vec algorithm, to calculate the similarities between vocabulary words in order to create lexical networks for each child at each timepoint. The word2vec algorithm takes into account both the co-occurrence of words in the corpus and the context they appear in, and has been used successfully in past developmental work (Beckage et al., 2020; Weber & Colunga, 2022). To use the word2vec algorithm to create lexical networks, we take the pairwise cosine similarity between all the resultant embeddings for words in our vocabulary, based on the CDI. The Adult network was created using the pre-trained GoogleNews word2vec model (available at https://code.google.com/archive/p/word2vec/). To create the Toddler network in a similar manner, we trained a word2vec model using the python gensim package with all the same training parameters as the GoogleNews model but trained on the toddler corpus described in Section 3.2.2.

The cosine similarities between the embeddings were used as the weights between words in our lexical networks. Thus, our networks are weighted and fully connected. In other words, every node (or word) is 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 CDI words and phrases. Excluded words include words that appear on the CDI as both a verb and a noun (which are not differentiated in these distributional models), multi-word phrases (we do not believe the sum or average of the embeddings for 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 networks. We matched the Adult and Toddler networks to have the same words (640) and number of edges (204,480), so any structural and predictive differences are based solely on differences in the way words are used within the original corpora.

### 3.2.4 Prediction Methods

In addition to the two networks we are comparing (the Adult and Toddler networks), we also have two ways in which we can make predictions. The first is based on the full network structure, as if the child knew every word, and the second is based on a child's individual network and word knowledge at that specific timepoint. Arguably, the first method, based on the complete network's connectivity, is similar to using the structure in the learning environment *external* to the child to make predictions, whereas the second uses the structure of a child's *internal* knowledge. We term the first method "Static", as predictions are made using full network connectivity in the same way child-to-child. The second is termed "Dynamic", as a word's prediction score changes based on the child's current vocabulary structure.

In the Static approach, each unknown word receives a "score" based on how strongly that word is connected to all other words in the full network. For example, the unknown word *dog* will have the same "score" regardless of if the child knows *puppy* and *cat* or knows *horse* and

*baby* due to *dog's* placement in the full 640-word network. Thus, this prediction is "static" across children. The Dynamic approach is more complex and computationally heavy. For each unknown word, we add it and all its connections into the child's individual network of known words for the given timepoint. We then calculate the unknown words' "score" before removing it and moving to the next unknown word. Thus, each word's score is based "dynamically" on the child's current vocabulary structure. For example, dog would have a higher score if the child knew *puppy* and *cat* than if the child knew *horse* and *baby,* as dog is more strongly connected or more similar to the words *puppy* and *cat.* All words are scored independently, so each word's prediction score depends only on the child's current network structure, not on any other unknown words' scores, for a given timepoint.

After completing these predictions, every unknown word has a score based on how central it would be in the child's current network. To calculate the final accuracy for that child for that timepoint, we take the top scoring predicted words (the same number as how many words that individual child actually learned based on our follow-up data), and compare those to the words the child did in fact learn, to compute the proportion of words that were accurately predicted. Let's say a child learned 20 words between the initial and follow-up timepoints; we then compare the top 20 highest scoring words to the 20 words the child actually learned, assessing what proportion the model accurately predicted. We do not expect model accuracy to be high, as there are many other environmental factors that we cannot control (e.g., the child visited the grocery store frequently with mom that month and experienced more food words, or visited the zoo that month and experienced more animals), but we assume that the more accurate the resulting predictions from the network are, the more representative that network is of the semantic knowledge that child has and uses to learn new words.

### 3.2.5 Network Measures

There are many different measures we can use to examine network structure and base predictions off of. We are not only interested in visualizing the structural 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. As noted in the previous section, our predictions are based on the idea that children may be more ready to learn words that complement, or are more similar, to words they already know.

All three measures are considered 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. The first measure is each unknown word's, or node's, weighted degree, which is the weighted number of connections to other nodes. A higher weighted degree means that word is more similar to a larger number of other words. The second measure is PageRank, which is the algorithm that Google uses to organize its search results; PageRank uses quantity and quality (edge weight) of links to determine a centrality "score". The final measure is Eigenvector Centrality, defined as the influence of a node; a node is more central if it's more strongly connected to other high centrality nodes. Though other measures exist, preliminary modeling suggested others either resulted in poor predictive accuracies or were too closely aligned with one of the three measures ultimately chosen. We also felt these three measures captured a range of ways to "get at" centrality.

### 3.2.6 Analyses

First, we will analyze structural differences between our two networks by calculating the top and least most central words in each full network using our three centrality measures. We do this to verify that these two networks are indeed structurally different. We then move to our main analysis of interest investigating representativeness. Statistically, we have two main manipulations, corpus (Toddler vs. Adult) and prediction method (Static vs. Dynamic), to create 2x2 linear mixed effects models, conducted for each of the three centrality measures. In all analyses, the dependent variable is the proportion of correctly predicted words. Multiple visits are nested within child, and each analysis also controls for the visit number, the child's CDI percentile at that visit, the number of words learned in that month, and the number of unknown words the child still had. Months in which a child learned less than 10 words or had less than 10 words left to learn were excluded from analyses. Further post hoc analyses were conducted investigating each manipulation (corpus or prediction method) separately. All analyses were conducted using the lmer package in R.

Finally, all accuracies were compared to a baseline control. To create this control, we randomly selected words from those the child has left to learn. Random selection was performed 1000 times, with average accuracy over those 1000 runs used as a baseline measure to compare against the predictions made by the centrality measures. In other words, if the prediction accuracies made using the centrality measures were no better than chance selection of words to-be-learned, these measures are not capturing individual word learning in any meaningful way. Each paired t-test was Bonferroni-corrected for the number of comparisons done.

### 3.3 Results

First, we investigated how many words in the top and bottom 50 most central words overlapped between the two networks, to see how much these networks might have in common. Only three to six words (depending on the centrality measure) were common between the two networks in their 50 most central words, and five to six overlapped in the least central words. That is, these two networks did indeed have different structures. A paired t-test further verified that the cosine similarities (or edge weights) in the two networks were significantly different from each other, $t(408{,}959) = 43.57$, $p < .001$. The words that appeared in the top 50 most central words in both networks include *butt*, *spoon*, *sock*, *monkey*, *pillow*, and *popsicle*. Interestingly, many common and early learned words such as *ball*, *kitty*, *car* and many foods were not in the top 50 for these networks.

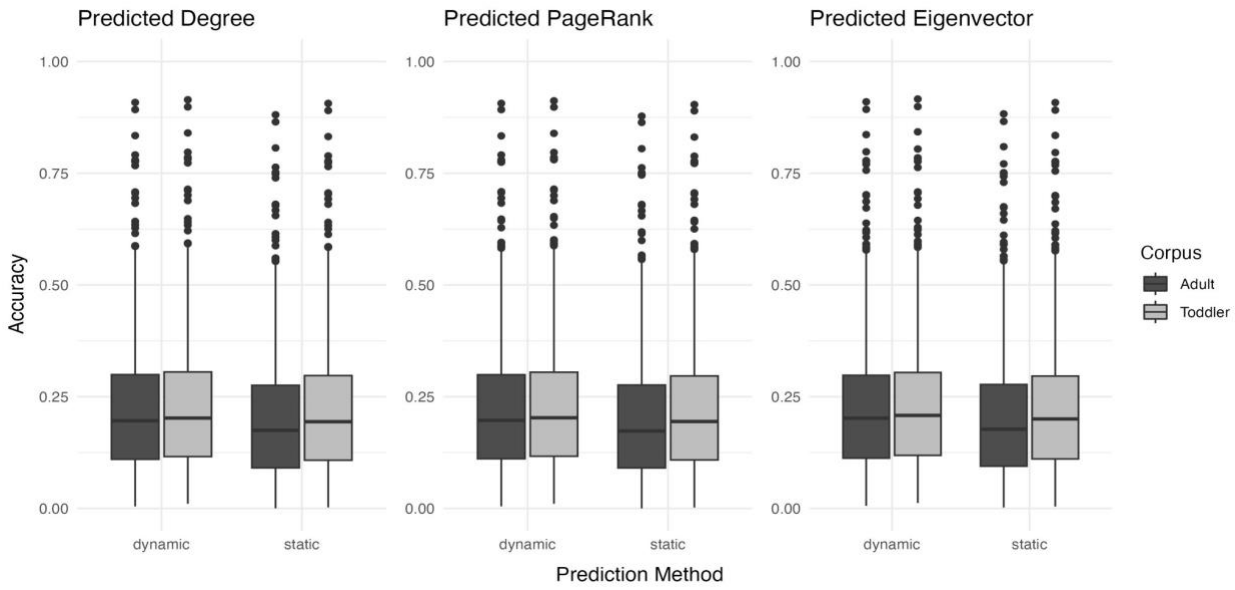
Next, we statistically compared the accuracy of model predictions for both networks and both prediction methods. For all three measures, both of the main effects and the interaction between them were significant in the full 2x2 model. In each, the Toddler network outperformed the Adult network (Degree: $t(2088) = 2.11$, $p < .05$; PageRank: $t(2088) = 1.97$, $p < .05$; Eigenvector: $t(2088) = 2.25$, $p < .05$). Further, the Dynamic prediction method outperformed the Static method (Degree: $t(2088) = 9.69$, $p < .001$; PageRank: $t(2088) = 9.98$, $p < .001$; Eigenvector: $t(2088) = 9.70$, $p < .001$). The interaction between the two manipulations was also significant, in that there was a larger difference in accuracies between the two corpora (the Toddler network more so outperformed the Adult network) when using the Static prediction method than when using the Dynamic prediction method (Degree: $t(2088) = 4.83$, $p < .001$; PageRank: $t(2088) = 5.00$, $p < .001$; Eigenvector: $t(2088) = 4.83$, $p < .001$). Each interaction is represented in Figure 3.1.

Delving deeper into the interaction, post hoc analyses investigated each manipulation separately. Within the Toddler networks only, Dynamic predictions outperformed Static predictions (Degree: $t(696) = 6.79$, $p < .001$; PageRank: $t(696) = 6.90$, $p < .001$; Eigenvector: $t(696) = 6.52$, $p < .001$). The same held true when running the same analysis within the Adult networks only (Degree: $t(696) = 14.71$, $p < .001$; PageRank: $t(696) = 15.40$, $p < .001$; Eigenvector: $t(696) = 13.93$, $p < .001$). Within Static predictions, the Toddler networks continued to significantly outperform the Adult networks (Degree: $t(696) = 7.66$, $p < .001$; PageRank: $t(696) = 7.72$, $p < .001$; Eigenvector: $t(696) = 7.78$, $p < .001$). However, within Dynamic predictions, we find that though the Toddler network significantly outperforms the Adult network when using Eigenvector centrality ($t(696 = 1.98$, $p < .05$), the Toddler networks only marginally outperform the Adult networks for Degree ($t(696) = 1.87$, $p = .063$) and PageRank ($t(696 = 1.74$, $p = .083$). This finding further illuminates the findings from the interaction in the larger 2x2 analysis.

Finally, we compared each type of network and prediction method to a random baseline, for a total of 12 comparisons with Bonferroni-corrected p-values. All the network predictions using the dynamic method were significantly better than our random control. Both the Adult Dynamic networks (Degree: $t(696) = 11.78$, $p < .001$; PageRank: $t(696) = 11.85$, $p < .001$; Eigenvector: $t(696) = 11.72$, $p < .001$) and the Toddler Dynamic networks (Degree: $t(696) = 12.53$, $p < .001$; PageRank: $t(696) = 12.31$, $p < .001$; Eigenvector: $t(696) = 12.95$, $p < .001$) were better than random prediction. The same was not true for the static prediction method. After correction, the Adult Static networks were no different from random prediction (Degree: $t(696) = 2.00$, $p = .551$; PageRank: $t(696) = 1.44$, $p = 1$; Eigenvector: $t(696) = 2.46$, $p = .168$). However, the Toddler Static networks were significantly better than random (Degree: $t(696) = 10.36$, $p <$

.001; PageRank: t(696) = 10.05, p < .001; Eigenvector: t(696) = 10.66, p < .001), indicating that overall the Toddler networks, no matter the prediction method, were more accurate than a baseline control of randomly selecting words to learn.

**Figure 3.1 Corpus by Prediction Method Results**




***Note.*** The main effect for Prediction Method and Age are significant for each metric. Each interaction is significant; the difference between adult and child is larger for static than dynamic networks.

## 3.4 Discussion

Our main question was whether, in building the semantics of children's lexicons, there would be a predictive advantage of using a corpus that is more representative of children's language over using an adult-directed corpus. Our findings suggest that, indeed, building lexical networks from toddler-directed corpora improves word-learning prediction over using larger but less representative adult corpora. This is remarkable given the fact that the adult-directed training corpus was multiple orders of magnitude larger than the child-directed training corpus. This highlights the importance of using data sources that are representative of the population of

interest. Conversely, this suggests that one may want to proceed with caution when expanding results based on assumptions that are true for a majority of children to children who are part of a linguistic minority.

A secondary question of this work was whether using a predictive method that takes into account a child's current vocabulary structure would provide more accurate word-learning predictions than using only full network connectivity. Here, again, we find that taking the more personalized approach is advantageous: the dynamic growth method was overall more accurate than the static growth. However, this advantage of the dynamic prediction method needs to be balanced against the greater computational demands that this method requires over the static one. In practical terms, making predictions for one child at one time point using the static method takes approximately one second, whereas making the same predictions with the dynamic method takes up to half an hour, or a difference of three orders of magnitude. Depending on the situation, this computational cost could make some applications prohibitive. Our results suggest a possible compromise; although making predictions dynamically based on the individual child's vocabulary structure was overall advantageous, the difference between the dynamic and the static methods was minimal for the networks with semantics based on the more representative child-directed corpus. This suggests that taking care to tailor the network representations to the population at hand may be a way to decrease the computational burden while achieving similarly accurate results.

### 3.4.1 Other Considerations

Another finding that highlights the importance of having semantic representations that match those of the population being modeled is in the comparisons to a random control. The two networks crossed with the two prediction methods were compared individually to our random

baseline control, in which the predicted words were simply selected at random from the set of words that child did not know at that time point. Here again we find that networks built based on the toddler corpus, regardless of predictive method, outperform the random baseline, but for the networks based on the adult corpus, only the dynamic predictions that take into account individual children's vocabulary structure outperform the control.

From these findings, one might conclude that, generally speaking, focusing on individualizing the data representations leads to more accurate conclusions about that individual. And that individualizing in one arena might make up for adapting a more generic stance in a different arena. In other words, more individualization, whether this is individualizing the data source or the prediction method, seems to lead to higher accuracy. These broad generalizations need to be tempered with a few caveats. First, the predictive ability of each network is relatively low, a little over 20% on average. In spite of the relatively low performance, most of the models outperform the random baseline control, with only accuracies from the adult-based, static prediction method failing to clear this bar. The relatively low performance of these networks is not entirely unexpected. After all, we are asking them to predict future word-by-word learning for individual children based solely on the child's current vocabulary, which is very little information. One way to interpret these predictions, then, is that the predicted words are words that the child is likely, or ready, to learn. However, though a child might have an easy time learning any specific predicted word, say the name of another animal, that child may not actually have the opportunity during that month to encounter and learn that particular word. In other words, we do not expect these networks to achieve very high accuracy on observational data; we simply want to know if these models are finding important structure in the networks and how the choice of representation and prediction method impacts their accuracy. In the future, these

models can be tested experimentally by ensuring that children encounter opportunities to learn these words, for example by having children actively taught the chosen words to compare against control words.

A second consideration is that all the control variables: age, vocabulary size, percentile, played significant roles in network accuracy. Prediction accuracy was higher for older children, for children with larger vocabularies, for children who fell in higher percentiles of the CDI, and for children who learned more words in a given month, though of course all these variables are intertwined. Future work should look at any systematic differences in model accuracy for specific subpopulations. For example, making accurate predictions for children of different ages might require a more fine-grained age adjustment to their semantic space beyond our coarse child-adult dichotomy. Similarly, having child-directed based semantics may be particularly critical for late talkers, or children who have relatively small vocabularies compared to their same age peers. Conversely, accurately predicting word learning for late talkers might require a completely different approximation of their semantic space.

### 3.4.2 Future Work

Future work, then, could examine alternative semantic representations beyond distributional semantics. We are currently working on testing semantic networks that are based on association norms geared toward adults or children (e.g, Nelson et al., 2004; Cox & Haebig, 2022) and feature norms geared for adults or children (Mcrae et al., 2005; Howell et al., 2005; Borovsky et al., 2023). We do not know as of yet what form young toddlers are storing semantic information in, whether it be similar to a hierarchy based on what features go with what concepts (i.e., feature norms); what words are most associated with others regardless of word type or type of relation (i.e., word association norms); similar to an embedding with multiple dimensions for

each word, including both semantic and grammar information (i.e., distributional methods); or some conglomeration of all three (e.g., Stella et al., 2017). Another dimension that could be explored in future work is the algorithm that drives how words get added to the network. This work focused on network centrality measures to predict specific word learning, but there are other suggested learning algorithms in the field that could be tested as well in the current setup. For example, past research has explored network growth algorithms such as preferential attachment, preferential acquisition, and lure of the associates (e.g., Steyvers & Tenenbaum, 2004; Hills et al., 2009; Hills et al., 2010; Beckage & Colunga, 2019). These algorithms use different criteria to add words into the lexicon and previous work suggests that the choice of growth algorithm can interact with the similarity representation chosen for the network (Hills et al, 2010; Beckage & Colunga, 2019).

### 3.4.3 Conclusions

In short, this work highlights the power of tailoring representational and processing choices to the population at hand. When modeling children's language development trajectories, using child-based semantics is better. Further, this promising work has the potential to be refined to accommodate children with different language profiles. For example, we could extend this work to predict vocabulary for children who are at risk of language disorders, or children who must learn more than one language. 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). Although, the underlying mechanisms by which children who lag in vocabulary growth learn new words may be different than that of typically developing children and models that capture the learning of children in these populations will have to be developed and refined, work like this

might inform targeted interventions that are tailored specifically to the individual, to support such children.

# Chapter 4

# How to Build a Toddler Lexicon: The Importance of Representative Data Sources


## Abstract

The current study builds upon previous work investigating where to source lexical representations of toddlers' vocabularies from. First, Weber & Colunga (2022) demonstrated that using network representations based on distributional semantics from a toddler language corpus can improve predictions of early vocabulary norms. Further, Weber & Colunga (under revision) extended that work to a large dataset of individual children's longitudinal vocabulary growth from 18 to 30 months of age, also showing that distributional semantics from a toddler corpus better predict specific word learning than semantics from a common adult corpus. The present work continues to extend this past work by not only continuing to compare adult and child sources of data, but by doing so over three different types of semantics: the same distributional method, as well as associative and feature-based semantic representations. The current analyses also utilize two different longitudinal datasets to make word-learning predictions with, and use a new method for statistically comparing across different network structures. The results suggest that lexical structures based on child-like word associations may be most similar to the underlying representations toddlers have and use to learn language.


*Keywords:* vocabulary acquisition; semantic networks; word associations, feature norms, word embeddings

## 4.1 Introduction

The past two chapters have provided evidence that child semantic structure is best represented with child-directed language input, rather than adult-directed. In the current and final study, we examine a different aspect of representativeness, other methods for capturing the semantics of the lexicon in addition to the previously used distributional semantics. This will help answer the question, how do children capture and represent the word internally? What methods do they use to represent the similarity between the concepts that they know? The next section 4.1.1 will review the classic ways of representing semantic similarity that we will use to tackle these questions. Further, we aim to extend our previous findings that networks created using child sources perform better than adult networks for these new types of semantic representation.

The current study will continue to build on Christiansen & Chater's (2001) criteria of input representativeness for evaluating computational models of language processing. That is, the match between the information available to the model and the information available to humans. We also believe another criterion proposed by Sims and colleagues (2013) is important to consider in light of our specific application to word learning, which is temporality, or the degree to which the model captures continuous change including the processes that drive that change. The following two sections review prior work on using network science to model the lexicon: first the fully-formed lexicon of adults and then the developing lexicons of young children.

### 4.1.1 Capturing the Semantics that Determine Lexical Structure

In order to model adult lexicons, researchers have used a variety of data sources to infer the way that words are connected to each other. There are three main methods for gathering this

information that are of interest in the current dissertation: association norms, feature norms, and distributional semantics. The first two rely on collecting data from adult participant judgments; the last one is based on the distributions of words in text or speech corpora.

The first method is association norms, where participants list the first word that comes to mind when given a cue word (e.g., given fish: water). Association norms have a long record in psychological memory research, dating back decades. In general, association norms are thought to capture everyday experience, for example when searching the internet, finding the right synonym for our writing, or answering broad questions (Nelson et al., 2000). Furthermore, associations can be dynamically impacted by experience, such that both gradual knowledge accumulation and temporary priming will impact the associates that first come to mind (Nelson et al., 2004). Arguably, associations capture not just associative knowledge but the meaning of words as well, and in fact, attempts to disentangle association from meaning has proved difficult if not impossible (e.g., Lucas, 2000). Though multiple studies have collected association norms over the years, the most recent set was produced by Nelson and his colleagues in 2004, and contains over 72,000 word pairs. In a network based on association norms, words will be connected to each other to the extent that people are likely to produce one when cued with the other.

The second method for capturing semantic knowledge is feature norms. Feature norms can be obtained a few different ways but in general depend on asking participants about the features of the concept represented by a word. For example, one can ask participants to list the determining features of a cue word (e.g., given fish: has scales, swims) or one can ask participants to rate the extent to which a cue word has a specific feature (e.g., on a scale of 1-9, how alive are fish?). As with association norms, feature norms have found their place in decades

of past memory research. Both similar and very different from association norms, feature norms are thought to directly index the meanings of words through the hierarchical structure of knowledge (Collins & Loftus, 1975). Features directly access the defining features, or perhaps meanings of words, whereas associations may capture many other aspects of world knowledge. Recent feature norm sets have been created by McCrae and his colleagues (1997, 2005), as well as Vinson and Vigliocco (2002). In a feature-based network, words will be connected to each other to the extent that they share similar features.

A third method for representing semantic knowledge uses large corpora of language to infer the meaning or usage of words based on a word's distribution in a corpus. A simple version of this is sliding window co-occurrence, in which words that appear near each other in the text are considered to be related to each other. In Latent Semantic Analysis, for example, words are considered to be similar to each other to the extent that they appear in similar contexts (Landauer & Dumais, 1997). Based on this idea of extracting semantics from the distribution of words in large corpora, neural network-based methods have been developed to compute distributional metrics such as Word2Vec and GloVe that learn vector representations (or an embedding) for each word. Rather than collecting responses from human participants, these methods are able to represent the similarities among words using a vector representation from large corpora. These corpora-derived representations have been shown to capture aspects of human performance. For example, Mikolov et al. (2013) showed that Word2Vec representations approximate human performance on a word similarity tasks.

These three methods of capturing the relationship between words have been successfully used in research on the relationship between the structure of lexicons and human performance in psycholinguistic tasks. For example, both the strength of word association norms and feature

norms predict adult semantic processing speed, in that words posited by these norms to be more related to each other are in fact more quickly produced or recognized by participants in tasks such as fill-in-the-blank or word-by-word reading (Nelson et al., 2009; McRae et al., 2005; Hutchison, 2003; Pexman et al., 2003). Similarly, researchers have also shown a relationship of semantic processing speed to co-occurrence metrics derived from various corpora (Lund & Burgess, 1996; Vankrunkelsven et al., 2018). Previous work has shown both that undirected Word2Vec and GloVe networks show degree distributions similar to that of human semantic networks (Kajić & Eliasmith, 2018) and that Word2Vec cosine similarities outperform co-occurrence statistics in predicting adult-rated word associations (Hofmann et al., 2018). On the other hand, some work suggests such distributional metrics are capturing something different than, and perhaps not on par with, adult word association data (Nematzadeh et al., 2017).

### 4.1.2 Modeling a Toddler's Lexicon

Building a model of a child's lexicon involves first determining the words a child knows and approximating the relationship between these words. Many researchers have used the MacArthur-Bates Communicative Development Inventory Words and Sentences (CDI) as the basis for modeling early lexicons. The CDI is a 680-word parent checklist used to assess whether a child produces a word at a given age or timepoint. Vocabulary checklists filled by parents have been found to be a good method to approximate toddlers vocabularies, when their vocabularies are still relatively small. The CDI in particular has been shown to be valid and reliable for toddlers between 16 and 30 months of age. The CDI has been normed for children between the ages of 16 and 30 months on a cross-sectional sample of 745 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. These include the relationship between lexical development, syntax, and processing speed (Moyle et al., 2007; Fernald & Marchman, 2012), lexical variability in toddlers with different levels of language proficiency (MacRoy-Higgins et al., 2016; Kim & Yim, 2018), and even how the words in a child's vocabulary shape how children learn new words (Gershkoff-Stowe & Smith, 2004; Perry & Samuelson, 2011; Colunga & Sims, 2017). Networks based on the CDI are usually unweighted and sparse, and show evidence of small-world structure (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., 2008; Hills et al., 2010). In short, the work using a child-based vocabulary to model early word learning has been fruitful for understanding the mechanisms of vocabulary growth.

A lot of work modeling early lexical networks has used a variety of adult-based metrics to model semantic knowledge. Steyvers and Tenenbaum (2005) showed that networks created from adult word associations, thesaurus entries, and WordNet all share structural properties common to other natural complex systems, such as sparse connectivity, strong local clustering, and a degree distribution that follows power law. These same characteristics have been shown to emerge from the preferential attachment growth process proposed by Barabási and Alberts (1999). As mentioned previously, building up lexical networks using preferential attachment and similar growth processes aligns with both adult-predicted and CDI-based age of acquisition (AoA) for nouns (Steyvers & Tenenbaum, 2005; Hills et al., 2008). Further, networks based on adult association norms also indicate the structure of language in the environment drives vocabulary growth (Fourtassi et al., 2020; Hills et al., 2008). Though there is some mixed

evidence for the validity of adult-determined perceptual feature norms (McRae et al., 2005) when using preferential attachment or similar processes to grow the lexical network of the average child (Hills et al., 2008), other work indicates feature norms perform significantly better than chance when predicting *individual* children (Beckage et al., 2020; Beckage & Colunga, 2019). Beckage et al. (2020) further found that predicting individual word learning using simple (single hidden layer) neural networks, in which the input was Word2Vec embeddings from a common adult GoogleNews dataset, performed similarly or better than other input representations, such as feature norms and phonology. Though this work did not directly build lexical networks, it did show that different representations of children's word knowledge is not just useful for evaluating growth in terms of general AoA, but can predict *individual* children's *word-by-word* learning.

Modeling early lexical development using adult-based norms has proved fruitful, but an adult's semantic space might look very different from a toddlers'. As Christianson and Chater (2001) eloquently pointed out in their description of input representativeness, the information available to a model should match the information available to the entity we are trying to model. 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. In our previous work using the readily available pre-trained GoogleNews Word2Vec vectors to create semantic network representations, words with multiple meanings have proven problematic. For example, the GoogleNews Word2Vec-based networks have *chair* and *head* strongly connected, with *chair*

only weakly connected to other instances of furniture such as *table* and *couch*. Presumably, this is because the two senses of chair as head of a committee vs. chair to sit on; only one of these senses is likely to be known by a 2-year-old.

In the current work, we contrast adult-oriented semantics with child-oriented semantics for each of the most common methods used to establish semantic links between words. This is important because each of these methods relies on different sources of information. For example, association norms come from the sense we have of how things go together in our daily, multimodal experience. This experience includes all sorts of information, the actions we perform on and with objects, what our senses perceive about these objects and actions, and what we say or hear about them. For example, the associations we have with *dog* can include ideas like *pet* and *cat*, but also the smell of a wet dog, and thus *stinky* as well as clichéd phrases, like *good dog*. All that information is somehow processed into people's responses to a cued word in an association task. Previous work investigating the difference between adult and child semantic representations suggests child-oriented word association norms are more predictive of normative vocabulary growth than adult-oriented associations (Cox & Haebig, 2022).

On the other hand, distributional semantic methods only have access to the information present in text. Although there is a lot of implicit information about the perceptual world in large corpora, the representation is limited by the things we choose to talk about with language and by the way our language expresses information. My previous work has shown that networks created from distributional metrics over child-directed corpora predict normative word learning more accurately than those over adult-directed corpora (Weber & Colunga, 2022).

Feature-based metrics, in contrast, take a more analytical approach. The specific features that are judged or produced by participants are explicitly stated. This allows us to decide, a

priori, which types of features might be available to adults and which might be available to toddlers. For example, both toddlers and adults may be able to perceive that *apples* and *firetrucks* are *red*, but features like *expensive* or *is a mammal* are not accessible to very young children. Previous work shows that neural networks that include sensorimotor features, or features more in line with the knowledge a child has, better predict word learning than models without such features (Howell et al., 2005).

Previous work attempting to approximate the linguistic environment of a young child without using adult-created metrics typically uses CHILDES (MacWhinney, 2000), a repository of caregiver-child conversations contributed to by many researchers. Using co-occurrence statistics on the CHILDES corpus, scholars have created networks to compare the lexical structure of children with high or low levels of vocabulary knowledge, revealing structural differences between the lexicons of the two populations (Beckage et al., 2011; Jimenez & Hills, 2017; Nematzadeh et al., 2014). Looking more specifically at lexical development and how it changes over time, 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. This corpus of work suggests using child-based sources of data is both fruitful and warranted. However, despite of the documented benefits of using distributed representations in modeling human performance (Rogers & McClelland, 2008; Hoffman et al, 2018; Beckage et al, 2020), only one pre-print has investigated using distributional metrics such as Word2Vec, rather than co-occurrence, on a child-directed corpus. Amatuni and Bergleson (2017) found that lexical networks created using GloVe vectors share the same structural properties of other semantic networks, in that they high clustering, low average path lengths, and a degree distribution following power law. This suggests that using Word2Vec embeddings,

another distributional metric, on child-based corpora could also be fruitful in addition to other common metrics such as association and feature norms.

### 4.1.3 Preliminary Work

Prior work presented in Chapter 2 has shown that input representativeness as discussed above plays a role in the ability of models to accurately capture early lexical development (Weber & Colunga, 2022). As a reminder, we created and compared an Adult and Toddler Network using Word2Vec embeddings trained on adult and toddler corpora. Not only did we find that these two networks were structurally different, but they had different predictive ability when predicting word-learning for the average, normed child (Frank et al., 2017). The Toddler network was significantly or marginally more accurate than randomly choosing words to learn, whereas the Adult network did not perform significantly differently than random chance. Further, though sliding-window co-occurrence metrics have been used in past child modeling work (e.g., Hills et al., 2010), our analyses showed our Word2Vec Toddler Network outperformed a Co-occurrence Toddler Network, created in the same way as Hills et al. (2010).

In work presented in Chapter 3, we compared the two network representations but now by predicting the specific words learned by individual children aged 18-30-months-old in an observational, year-long longitudinal dataset collected at monthly intervals. Predictions were compared against each child's actual word learning during each month-long period to create an accuracy score. Linear mixed effects models revealed that the Toddler Network significantly or marginally outperformed the Adult Network after controlling for multiple child characteristics (e.g., number of words learned, age).

This work shows that input representativeness matters – representing semantics using child-based sources is best for modeling child development. However, these findings also

suggest the need for a more thorough exploration of adult versus child-sourced metrics. The differences between these predictions were small and half of the used prediction metrics were only marginally significant. Further, both the Adult and Toddler Networks performed well above random word prediction in the longitudinal data sample, indicating the adult network is also capturing semantic information relevant to a child.

Additionally, other past explorations into the representation of specific populations have shown mixed results. Using AoA's based on averaging together 1,800 children's learning from the CDI results in more accurate predictions than using AoA's based on almost 2,000 adult ratings of when they think they learned a particular word. Even though the adult ratings were more fine-grained (differences in acquisitions of a few days) compared to the CDI norms (difference of one month), these findings were in line with the network results, suggesting child-sourced data results in more representative models. However, using our network prediction method to investigate 3-5-year-old Cantonese-English bilingual vocabulary growth showed that networks derived from an English corpus better predicted both English *and* Cantonese word learning as compared to a network derived from a Cantonese corpus. As with the Toddler-Adult comparison, the English corpus contained more language samples than the Cantonese corpus, suggesting sometimes there is more at play than representation alone. In other words, this question of child- versus adult-sourced semantic representation need to be further explored in other network representations.

### 4.1.4 Current Study

In the current work we ask two questions. Are some methods of capturing semantic information more representative of young language learners than others? And are sources of information gathered with children's knowledge in mind more representative than those from

adults? To answer the first question, we investigate the different sources outlined in Section 4.1.2., termed **semantic type**. Previous work has investigated aspects of these semantic types in relation to young children's vocabulary learning, but none have directly compared these representations' ability to accurately predict *individual* children's learning of *specific* words as we do here, to understand which method may best represent a young child's lexical knowledge. Past developmental work does not strongly suggest one semantic type will result in better word predictions than others. However, some have found better success with association norms compared to others (e.g., Hills et al., 2008). Association norms have also been suggested as the best representation for adult semantic structure (De Deyne & Storms, 2008; Rapp, 2014). Because of this, we predict association norms to be the best performing network. In previous work, both distributional semantics (e.g., Beckage et al., 2020) and feature-based semantics (e.g., Beckage & Colunga, 2019) have successfully been used, though perhaps less so for feature-based (e.g., Hills et al., 2008; Beckage et al., 2020). Because of this, we might expect distributional semantics to outperform feature-based semantics, even though the distributional methods are newer to the literature.

The second factor of interest is what we term **age type**, or the comparison of adult- versus child-based metrics. Though more detail will be given in the methods, each of the above representations can be sourced using adult data (e.g., querying adults, using adult corpora), and using child (approximated) data (e.g., querying adults as if they were children, using child corpora). Based on our own past work presented in Chapters 2 and 3, as well as general intuitions about the need for representative data sources, we expect that child-based semantic representations will outperform adult-based representations.

To evaluate these different lexical representations, we will use two longitudinal datasets of young toddlers' vocabulary growth gathered in our lab – an observational year-long study and a two-month intervention study. We will use multiple network centrality measures to predict next words learned for each individual child and calculate accuracy by comparing predictions to the words each child in fact learned during that given time period. We will then compare the accuracy of the different network representations, *with the assumption that sources with higher input representativeness result in more accurate word-learning predictions.*

## 4.2 Methods

### 4.2.1 Participants

We utilized two different longitudinal datasets of young children's vocabulary acquisition previously collected within our lab. The two will be analyzed separately. The first is larger and observational, and the second will be analyzed to corroborate findings from the first set of analyses. See Table 4.1 for some summary statistics.

The first dataset is the same observational dataset used for the analyses in Chapter 3. This dataset consists of 86 monolingual children beginning between 15.38 and 19.33 months of age ($M$ = 17.67, $SD$ = 0.97) and followed for 12 consecutive months. Each visit was spaced as close to one month as possible barring scheduling restrictions. This consists of up to 11 Pre-Post pairings per child we can utilize for model prediction and verification. At each visit, parents filled out the CDI indicating what words their child produced. Some children were not able to participate in every visit, and so months in which there is not a pre and post CDI were excluded, so that all predictions constitute a one-month period of learning. For example, if a child did not come in for the fifth month's visit, that child would only have nine datapoints (1-2, 2-3, 3-4, 6-7,

7-8, etc.). In total, we analyzed 831 pre-post pairings. Data was constrained so that a child could not "forget" a word once it had entered their vocabulary.

The second dataset consists of 55 monolingual children who participated in a three-visit intervention study. Fifty-four of those children completed all three visits and one completed the first and third visits. Children began the study between 18.86 and 22.06 months of age ($M$ = 20.51, $SD$ = 0.84). At each visit, parents filled out the CDI. Between visits one and two, lasting two weeks, children learned 16 words individualized to them based on their initial vocabularies using a book read to them by their parents. These words were chosen in one of two ways. Children either received words that were predicted to have a high probability of being learned by them based on their vocabulary structure, or words predicted to have a low probability of being learned. Between visits two to three, lasting six more weeks, families went about their routine as usual, with no intervention. To gather pre-post pairings needed to make and verify model predictions, we used children's vocabularies at their first visit (prior to any intervention) to predict their third visit vocabulary, about eight weeks or two months apart. As with the other dataset, data was constrained so that a child could not "forget" a word once it had entered their vocabulary.

**Table 4.1 Demographics for each Dataset**

| Source | N | Initial Age | Length (mo.) | Initial no. of Words | Final no. of Words | Initial CDI Percentile | Final CDI Percentile |
|---|---|---|---|---|---|---|---|
| **Obs.** | 86 | 17.68(0.97) | 12 (1) | 74.14(83.73) | 560.28(130.41) | 36.40(29.66) | 67.84(30.11) |
| **Exp.** | 55 | 20.51(0.84) | 2 (2) | 122.36(141.31) | 189.33(189.03) | 30.85(30.45) | 36.54(32.70) |

### 4.2.2 Sources of Semantic Information

The CDI contains 680 words and phrases developed to assess typical language development for children aged 16 to 30 months (Fenson et al., 1994). The full CDI is listed in Table A.1 in Appendix A. The CDI makes up the basis for the vocabulary, or nodes, that we use

for each of the semantic networks. However, four phrases and three proper nouns were excluded from every network due to their idiosyncratic nature. The phrases are: *give me five!*, *gonna get you!*, and *this little piggy*. The proper nouns are: *babysitter's name*, *child's own name*, and *pet's name*. Table 4.3 provides an overview of the structures of each network.

#### *4.2.2.1 Word Associations: Adult*

The most often-used adult association norms are the University of Florida Norms, gathered by Nelson et al. (2004). Nelson and his colleagues queried participants on thousands of cue words, gathering over 72,000 Cue-Target associations. Each participant provided associates for 25-30 cue words, listing their one top associate for each. Nelson et al.'s norms mainly consists of nouns (76% of their normed words), verbs (7%), or adjectives (13%). However, because participants could give any word in their free response, there were "duplicate words" as responses that are also found on the CDI. Duplicates are words with multiple meanings (e.g., *slide* appears as both a noun and verb, *orange* appears as both a noun and adjective). In each case, we used the first instance of the word as it appeared on the CDI (e.g., *slide* as a noun appears before *slide* as a verb, so all instances in the dataset were treated as the noun, and only the noun appears in our network). More information about how this was done during cleaning can be found in Section 4.2.3.1. Five hundred and ninety one of the six hundred eighty words are represented in the Nelson norms, covering just over 86% of the CDI. Further, their use in past developmental studies makes their use all the more warranted (Hills et al., 2008, 2009; Beckage & Colunga, 2019). See Table A.4 for those words from the CDI missing in the Nelson norms.

#### *4.2.2.2 Word Associations: Toddler*

There are no known word associations gathered by querying child participants of any age. However, Cox and Haebig (2022) attempted to gather word associations that better approximate

a child's knowledge by prompting participants to create child-oriented responses, specifically created from the CDI. They asked participants, "Imagine you are playing a game with a toddler (a 2- to 3-year-old child). In this game you draw a card with a word on it and say the first 3 words that come to mind. Over the course of the game, you will draw many cards and expose the toddler to many related words. Each of the following screens is like a card draw. You should type the first three related words you would say to the toddler." They presented instructions with an image of a toddler to help orient participants to the target age. All participants came from Amazon Turk or Prolific. To assess if their manipulation worked, the authors compared adult-oriented associations to those created with the child manipulation on a variety of metrics. Cox and Haebig found that the child-oriented responses were typically higher frequency words, were shorter, and had younger age-of-acquisitions. These responses also had higher contextual diversity and correlations among the cue-target responses suggest the manipulation did produce a different semantic structure. Though not a perfect example of the word associations children have, these associations have proven to be the best approximation we can use to represent children associate knowledge.

The Cox and Haebig (2022) set of association norms use 675 of the 680 words and phrases. The only phrases that do appear in the Cox norms are those not included in any dataset, making this the dataset with the most CDI coverage. To note, the idiosyncratic CDI term *pet's name* became *pet* in the cues presented to participants, and some statistics reported in Cox and Haebig (2022) are not reflected in their raw data (e.g., excluding the term "so big!", having only 672 cue words). Queried with similar instructions to Nelson et al. (2004), participants on Amazon Turk listed their top three associates to any given cue word (rather than only one as in the Nelson norms). Not all participants provided targets for every cue.

#### *4.2.2.1 Distributional: Adult*

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 GoogleNews (3 billion words). Past work has found success using the widely known pre-trained GoogleNews Word2Vec model to model human semantics (Beckage et al., 2020; Handler, 2014; Church, 2017). As implied, the GoogleNews corpus contains thousands of news articles in English. This dataset contains both more language than other known corpora and is pre-trained and ready to query, making it the optimal adult-based Word2Vec representation. The GoogleNews Word2Vec is a skip-gram neural network trained using a window size of 5 (considers the five words before and after the target word) and a minimum word count of five (the word must appear five times in the corpus to receive an embedding). Further, the trained embedding is a 300-length vector, which is believed to be the optimal length for semantic specificity and computational power, performing better than other length vectors on a variety of tasks (e.g., Hartmann et al., 2017). Networks created from the GoogleNews Word2Vec will henceforth be referred to as the adult distributional network.

From the pre-trained vectors, 647 of the 680 words on the CDI can be found. Aside from the phrases removed from every network, some other multi-word phrases and words were removed as they were not found in the corpus and no semantic matches to query in their place could be determined. Duplicate words were also removed and the first instance used as in the Adult Association norms. Because these distributional models do not differentiate words during training based on part of speech, these duplicate words would have the exact same embeddings and thus the exact same place (connections to other words) in the network (and would therefore

be predicted at the exact same rate by our predictive method). Table A.2 in Appendix A indicates words that were removed or changed.

#### *4.2.2.4 Distributional: Toddler*

In addition to naturalistic conversation, young children growing up in the United States experience many forms of media, including movies, TV shows and music. Children also receive input during story time. Though CHILDES alone exists and has been used to compute co-occurrence-based networks in previous work (e.g., Beckage et al., 2011; Hills et al., 2010), we created our own corpus that contained both conversations from CHILDES as well as other sources of language children might typically encounter, including movies and storybooks. Details about the creation of this corpus can be found in Section 2.2.1 of Chapter 2. Table 4.2 provides summary statistics of the sources of information in our toddler corpus.

**Table 4.2 Toddler Corpus Statistics**

| | CHILDES | Books | Movies |
|---|---|---|---|
| **Number** | 5,751 | 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 |

Using the newly created Toddler Corpus, we trained a Word2Vec model using the python gensim package that we then used to gather embeddings for our vocabulary of interest, the CDI. The training parameters were the same as those used in the pre-trained GoogleNews Word2Vec model. Similar to the adult Word2Vec, multi-word phrases were removed from the vocabulary as they are not found in the resultant embeddings, and some words were changed to find semantic matches within the corpus. Again, duplicates were removed due to a lack of differentiation in the corpus. The toddler distributional network has 646 of 680 words from the CDI. Despite the similar size, these words differ slightly from those contained in the adult network. See Table A.3

in Appendix A for the relevant words that were removed or changed to gather words for the final full toddler distributional network.

#### *4.2.2.5 Features: Adult*

The most widely used feature norms were created by McRae et al. (2005). McRae and colleagues collected features for 541 concepts, where 30 participants listed features for each concept. Each participant could list up to 10 features for any given concept, and were told to list different types of features, including functional, perceptual, and categorical or encyclopedic features. McRae and his colleagues organized these results into 2,526 total features ascribed to the different cue words. However, only 147 nouns from the CDI are found in the McRae feature norms, a very small subsection of the 680 total words on the CDI. The adult feature network is thus be the smallest network analzed. Some of the words were found by querying synonyms or the singular/plural version of the original CDI words. See Table A.6 for a list of the 147 CDI words found in the McRae norms and the synonym substitutes we used to query the norms.

#### *4.2.2.6 Features: Child*

Features known to a child likely look vastly different to the ones listed by adult participants, especially given that participants in McRae et al.'s study were told they could use encyclopedic-type features. This was the motivations for Howell et al. (2005) to create a set of sensorimotor features likely readily perceptually available to a toddler just beginning to learn about word meanings. These are the only known feature norms that are geared toward representing the knowledge young children have about certain words.

To create the child norms, Howell and his colleagues first narrowed down the original features created by McRae et al. (1997) to 97 feature dimensions which were reviewed by an independent developmental psychologist for appropriateness. Three-hundred-and-forty-five

nouns from the CDI were then rated for each feature dimension by 200 participants. See Table A.7 for the list of nouns included in the Howell norms. To note, this is large deviation from the way McRae queried participants, and may have a significant impact on how features were gathered. Rather than free response, participants instead rated a noun on all 97 features on a scale. Even though a participant may never freely respond that a snake *climbs*, they may still rate *climbs* slightly higher than *runs* given that *runs* is even less a feature of a snake. Despite these differences, both datasets are unique to the literature in the semantics that they represent.

**Table 4.3 Network Properties**

| Method | Age | Nodes | Edges |
|---|---|---|---|
| Word Association | Adult (Nelson) | 591 (86.91%) | 2,834 (1.63%) |
| | Toddler (Cox & Haebig) | 675 (99.26%) | 14,760 (6.49%) |
| Feature | Adult (McRae) | 147 (21.62%) | 3,322 (30.96%) |
| | Toddler (Howell) - Original<br>Toddler (Howell) - Used | 345 (50.74%)<br>345 (50.74%) | 59,340 (100%)<br>18,372 (30.96%) |
| Word2Vec | Adult (GoogleNews) | 647 (95.15%) | 208,981 (100%) |
| | Toddler (self-created) | 646 (95.00%) | 208,335 (100%) |

### 4.2.3 Network Structures

Above we described the individual data sources and how the authors gathered that data, as well as what words from our CDI vocabulary we were able to gather from said sources. Here we describe the manner in which we use that data to create the different networks. In other words, we will next describe how we connect the words in each network together, or how we find the edge weights between nodes. As a structural note, all networks are undirected and weighted, with no self loops (e.g., dog cannot be connected with itself).

A preliminary investigation of accuracy in these networks among different density levels and between weighted and unweighted versions of the networks showed no significant differences based on these structural changes, and in fact different networks showed differing trends in what densities/weightedness performed better. Due to these preliminary results, as well

as past work indicating arbitrary cutoff's for density can drastically alter network structure and subsequent conclusions (e.g., Connor et al., 2017), the decision was made to keep networks in their "full" state, minus one network that was changed after initial evaluation and reanalyzed due to poor performance. We altered the child feature network to contain a density cutoff matching its adult peer, of 30.96%, after it was found the fully connected network performed worse than controls. We will expand on this reasoning in the discussion, Section 4.4. Network similarity matrices depicting the edge weights as a heatmap are shown in Appendix B. Due to differing network structures, all network accuracies will be compared to a corresponding control model, rather than directly to each other.

#### *4.2.3.1 Word Association Networks*

For each of the word association networks, we first pulled the relevant associations from the full list of cleaned (either by the original authors or by us), cue-target data for all the possible words in our vocabulary of interest. In other words, all associations in which both the Cue and Target were in our CDI vocabulary were gathered from original data. The weight between words is the proportion of participants who indicated that response (out of the total responses for that cue word), resulting in a weight from 0.001 to 1 (a weight of 0 cannot exist – if no one indicted a response it does not appear in the associations). So, if there are 300 total target responses for the cue word *dog*, and 50 were *cat*, the weight between *dog-cat* would be 0.167. For Cue-Target pairs that appeared in both directions (e.g., *cat-dog* and *dog-cat*), the two proportions are averaged to create one weight, so this network can be undirected.

Nelson et al. (2004) cleaned and provided these proportions in their cleaned data. However, because we wished to also include synonyms we recalculated proportions ourselves after further cleaning. Replaced words include some plural/singular forms of words (e.g., *bubble*

for *bubbles*) and shortened versions of words (e.g., *fridge* for *refrigerator*) in addition to the synonyms. Table A.5 in the appendix lists these changes. After recalculation and averaging bidirectional associations, the resultant adult association network has 2,834 edges, or 1.63% of the total possible 174,345 edges if this network was fully connected. The Nelson network is the sparsest network of those analyzed.

Cox and Haebig (2022) only provided their raw data with every individual response listed for each of the 675 cue words from the CDI (~300-350 responses per cue). Multiple cleaning steps were taken to create our networks. First an initial spelling pass was taken using the built-in functions of Excel. Second, synonyms were replaced in the targets according to Table A.5 in Appendix A. Third, multiple forms of verbs were replaced with the form found on the CDI. In a next step, duplicate words that appear on the CDI as two different word types were cleaned. Though the cues differentiated based on word type, the targets given by participants did not most of the time. In this step, any target given for a cue word that was a duplicate was assumed to be the other word type (if *slide* was given for *slide (noun)* it was assumed the participant meant *slide (verb)*). Further, targets given for cues where it was not immediately clear which word type was meant were evenly split between the two word types. Hand cleaning was also done to find any versions of the word that may not appear in a typical search (e.g., some indicating *slide the thing* for *slide (noun)*). Finally, all targets that were the same as the cue word (minus those for multiple word types) were removed. After this, the proportions were calculated and bidirectional associations were averaged. The resultant child association network consisted of 14,760 edges, or 6.49% of the possible edges in the 675-word network.

#### *4.2.3.2 Distributional Networks*

Both networks created using semantic embeddings calculated using the Word2Vec algorithm were created in the same way. To connect words in our networks, we take the cosine similarity between each word in our vocabulary. These cosine similarities become the weight of the edge between the two words. Because cosine similarities are symmetrical (the calculated similarity between *cat-dog* is the same as *dog-cat*), the resultant networks are undirected, with only one weight representing the relationship between two words (e.g., there is not a separate edge weight connecting *dog* à *cat* than *cat* à *dog*).

Cosine similarities range from -1 to 1, with a value closer to 1 meaning two vector embeddings are highly similar and -1 meaning the two embeddings were opposites. A similarity of 0 means there was no relationship between the two embeddings. In general, a higher magnitude cosine similarity (whether toward -1 or 1) indicates the two vectors were more similar in some way to each other. Because these would be the only networks among the six that would have a negative weights, and based on work using social balance theory (see Kaur & Singh, 2016), we used the absolute value of all the cosine similarities to remove negative weights.

Because every word's embeddings contained all non-zero numbers, all cosine similarities and thus weights between words are non-zero. So, both of the distributional networks are fully connected, with every word connected to every other with some weight. For the adult network of 647 words, this results in a total 208,981 edges, and for the toddler network of 646 words, this results in 208,335 edges.

#### *4.2.3.3 Feature Networks*

Both sets of feature norms are used to create edge weights for the networks in the same way. Each vocabulary word has a corresponding feature vector to describe it, with a value for

each feature the authors used. For example, in the McRae norms, each word has a 2,526-length feature-vector. As with the vector embeddings from the distributional networks, we take the cosine similarity between the vectors for every word, pairwise, to gather weights between words. In the data McRae et al. (2005) offer, these cosine similarities were already calculated and are provided in matrix format. All the cosine similarities, and thus weights between words, are between 0 and 1, since the feature vectors all contained positive values. This technique resulted in 3,322 edges in the adult feature network (30.96% of the possible 10,731 edges if this network was fully connected). Unlike the distributional embeddings, many of the values in each feature vector are zero, as that feature was not used to describe that word, resulting in many cosine similarities whose value was zero. A zero weight is the same as having no edge between two words.

In much the same way as McRae did, we can take the cosine similarity between the 97-lenth feature vectors in the Howell norms to find the weight between words. However, every value in the vector is a non-zero, positive number, because participants rated each feature on a scale indicating how well that feature described the word. Because of this, every cosine similarity between vectors is non-zero and positive as well, resulting in a fully connected network of 59,340 edges.

### 4.2.4 Predicting Words Learned

Here we describe the actual methods for making specific word predictions for each child and calculating accuracies by which to compare the network representations. This method is the same as the dynamic method described in Chapter 3 Section 3.2.4. For each network above (see Table 4.3 for a summary), we have both the full network, containing all possible words and connections, representing the full extent of what a child *could know*, and for each individual

child at each timepoint we also have that child's individual network, a subset of the words and connections the child *currently knows.* We also have the follow-up network, or the words and connections the child knows at the next timepoint, so we can see what words the child *learned*. Recall that one of the considerations driving this work is that a developmental model must capture the mechanism of change. Specifically, we assume that **unknown words that are more similar to words already known will be more likely learned by the child.** To calculate which words best fit this criterion, we use network centrality measures.

Our prediction method is as follows: for each unknown word, we add it and all its connections into the child's individual network of known words for any given timepoint for which we have follow-up data for. We then calculate the unknown words influence through each measure before removing it and moving to the next unknown word. Each words' score is based on the child's current vocabulary structure. For example, the centrality score for *bus* would be much higher when adding it into a network of vehicles than if adding it to a network of mostly animal words, and vice versa for the word *kitty*. All words are scored independently, so each word's prediction score depends only on the child's current network structure, not on any other unknown words' scores, for a given timepoint.

After completing these predictions, every unknown word has a score based on how central it would be in the child's initial network. To calculate the final accuracy for that child for that timepoint based on these predictions, we take the top __ scoring words, the same number as how many words that individual child actually learned in the following timeframe (based on our follow-up data). Let's say a child learned 20 words between the initial and follow-up timepoints; we then compare the top 20 highest scoring words to the 20 words the child actually learned, assessing what proportion (or percent) the model accurately predicted (what proportion

overlapped with the actual words learned). We do not expect model accuracy to be high, as there are many other environmental factors that we cannot control (e.g., the child visited the grocery store frequently with mom that month and experienced more food words, or visited the zoo that month and experienced more animals), but **we assume that a network that more accurately predicts words an individual child actually learned is more representative of the semantic knowledge that child has and used to learn new words.** A more *accurate* network is a more *representative* network.

#### *4.2.4.1 Network Measures*

To calculate each unknown word's score if it were to be added into the child's network, we used multiple different network measures. To note, all networks were built using the networkx package in python, and network measures were calculated using the corresponding, built-in weighted version of these networkx algorithms.

The first measure was each unknown word's **Degree**. However, this was 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 (as all the measures are). The degree is the weighted number of connections to other nodes; in a semantic network nodes have a higher degree if they are more similar to a larger number of other nodes. For our purposes, a higher weighted degree meant that word was more likely to be learned by the child, as it was more highly connected (and more similar to) those words already in the child's network subset. Similarly the next centrality measure we used was **PageRank**, which is the algorithm that Google uses to organize its search results. PageRank uses quantity and quality (the edge weight) of links to determine centrality. The last centrality measure we used was **Eigenvector Centrality**, which measures the influence of a node. In this algorithm, a node is more central if

it's more strongly connected to other high centrality nodes. These three centrality measures, though similar, also capture separate aspects of centrality. Finally, **Clustering Coefficient** was also used to make predictions, though it is not technically a centrality measure. The clustering coefficient of a word describes the degree to which a node's close connections are closely connected among themselves, or how tightly knit that node is in the network.

### 4.2.5 Analyses

#### *4.2.5.1 How to Statistically Compare Networks*

The first issue to tackle in a statistical analysis is the differing scales of our networks, ranging from 147 to 675 nodes and 1.6% to 100% density. Most past work comparing different network representations do not equate their networks (e.g., Steyvers & Tenenbaum, 2005; Hills et al., 2008; Beckage & Colunga, 2019). However, preliminary work in Chapter Three where we controlled for the number of words in a child's vocabulary when statistically comparing networks suggests the number of possible words to choose from during prediction has a large impact on accuracy. Simply controlling for this factor in statistical analyses does not fully ameliorate this confound. If differences were to be found in a direct comparison, a confound occurs as to whether that difference was due to the differing density/size or due to our manipulation of interest, the semantic representation, operationalized through differing edges and their weights.

To avoid confounds of network structure, we compared each network accuracy for each child to a corresponding control network's accuracy for that same child. Control networks had the same nodes and contained the same number of edges with the same weights, but these edges were randomized between nodes. Each child at each month (each datapoint) had its own randomized network for word prediction, so every control was different for every datapoint.

Thus, every experimental accuracy was paired with its corresponding control accuracy for statistical comparison. Statistically, a comparison of these comparisons is analyzed (is the difference between the experimental and control model for the association adult network significantly larger than the difference between the experimental and control model for the association child network). Using this control comparison method, we assessed two main questions relating to the type of semantic representation (association, distributional, feature) and the age (adult, child). This resulted in an overall three (semantic type: association, distributional, feature) by two (age type: adult, child) by two (experimental type: experimental, control) ANCOVA, controlling for child-level variables; age, CDI percentile, and words known and learned. To avoid floor and ceiling effects (the centrality measure getting an accuracy of zero because the child didn't learn any words or didn't have any words left to learn), we removed any datapoints where a child learned less than 1% or more than 99%  of the words still left to learn in their network.

#### *4.2.5.2 Post Hoc Analyses*

We also conducted ANCOVAs to investigate accuracies relative to other variables of interest. For each of our post hoc variables, we created categories to compare. Specifically, in our observational dataset we were interested in how a child's age and language proficiency played a role in different networks' accuracies. For example, past work has shown that some types of networks perform better at younger ages and others at slightly older ages (e.g., Beckage & Colunga, 2019). Although we have a relatively small age range, it is important to note that this is a period of both rapid lexical growth and rapid cognitive change in development. It could be that children at younger ages or with smaller relative vocabularies are better represented by say

association norms, that is, relatively more driven by world experience, but older toddlers are better represented by distributional embeddings, relatively more driven by language use.

In our analyses for language proficiency, we will divide children up based on their CDI percentiles. For every CDI the parents fill out, a child also receives a percentile score that ranks their vocabulary size based on their age and gender. There is evidence this percentile score can be predictive of language delay and later academic difficulties (e.g., Heilmann et al., 2005; Rescorla, 2009). Further, children who fall in the lower percentiles on the CD, termed late talkers in the literature, may both have different vocabulary structures compared to their typically developing peers (e.g., Perry et al., 2023; Weber & Colunga, 2021), and may learn language through different mechanisms (e.g., Jimenez & Hills, 2017; Colunga & Sims, 2017; Beckage et al., 2011). For example, late talkers may have less connected vocabularies compared to their peers, suggesting they may have a preference to learn more "oddball" (less semantically related) words (Beckage et al., 2011). If late talkers acquire words through different mechanisms, our centrality prediction measure which focuses on predicting *more* related words may not work as well for this group. Because of these interesting differences found in past work, we endeavored to understand its role in our current network predictions.

Specifically in the intervention dataset, we will again look at language proficiency to corroborate the results from the observational dataset, as well as investigate the intervention condition children were assigned to. To note, in both studies we purposely over-recruited toddlers with lower language proficiencies so that the resulting data includes a large range of vocabulary sizes at each time point.

These post hoc analyses will be conducted with similar ANCOVAs, each with the additional post hoc factor of interest. For each analysis, we will first look for significance in the

overall ANCOVA (e.g., 3x2x2x2 ANCOVA for language proficiency), and then continue to delve deeper into how the trends for our semantic and age types differ within our factor groups.

### 4.3 Results

For all the following analyses, the measure of interest is the performance of the experimental networks compared to their respective control networks. For the rest of the paper, when referring to a particular network's "performance", we are referring to the difference between the experimental and control versions of that network. Thus, when describing one network (e.g., associative child) as performing better than another (e.g., distributional child), we are referring to a significant interaction between the experimental and control networks for those two representations.

We have two main research questions. First, are some semantic types better at capturing early vocabulary growth? Second, do child-based networks perform better than adult-based, supporting previous work but with different semantic types and using this new evaluation method (comparison to control networks)? Of further interest, we also ask, are there different differences in semantic types for each age type? To answer these questions, we ran a 3 (semantic type) x 2 (age type) x 2 (experimental type) ANCOVA for each dataset, controlling for child age, CDI percentile, and the number of words known and learned. We first investigate the larger observational dataset, and then attempt to replicate the results within the intervention dataset.

Though we used multiple different network measures to predict specific word learning for our datasets, we report results for only PageRank. After collecting all our predictions, PageRank was found to perform the best. The other centrality measures made very similar predictions and resulted in pretty similar accuracies overall. However, clustering coefficient, the

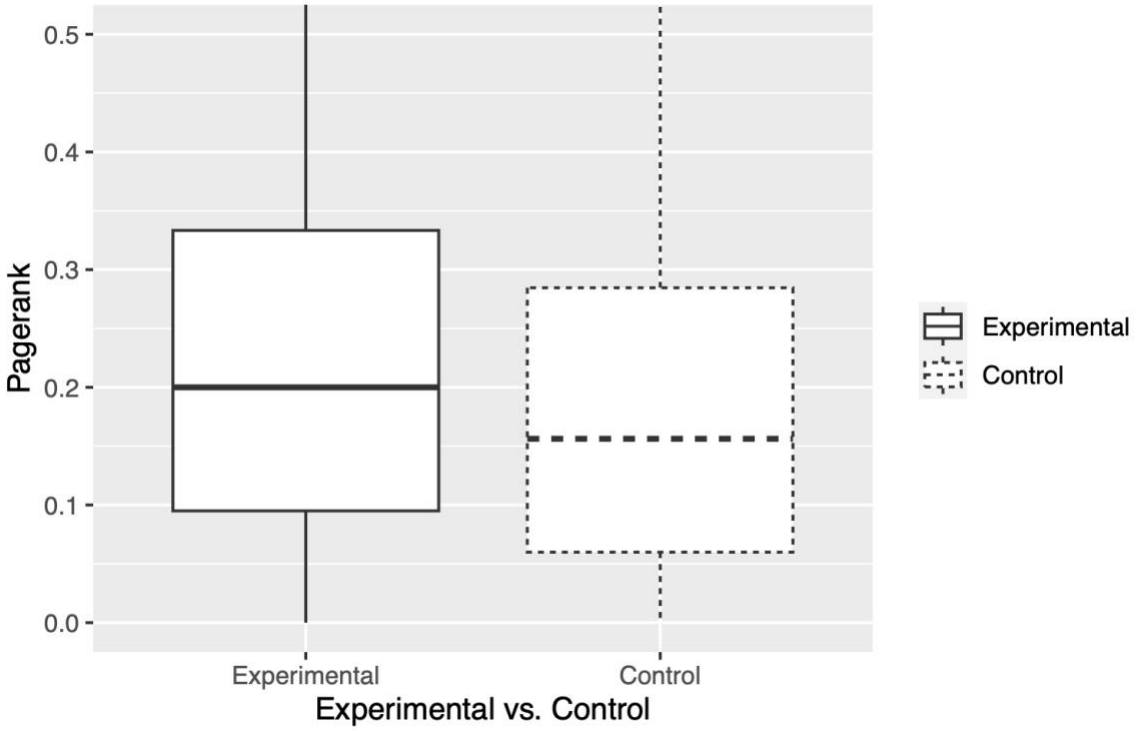
network we used that was not a centrality measure, performed worse, and word-by-word predictions did not closely align with the other measures as they did with each other.

As a final note, we will be reporting marginally significant results ($.05 < p > .10$). Though we do not take such results as significant, when we find marginal interactions of interest we will continue to unpack them through further analysis, particularly when part of a pre-planned analysis.

### 4.3.1 Observational Dataset

We begin by investigating the 3 (semantic type) x 2 (age type) x 2 (experimental type) ANCOVA in the observational dataset. The only main effect of interest is that of experimental type (experimental vs. control networks), which is significant, $F(1,8946) = 187.151$, $p < .001$. The main effect confirms that across age and semantic types, experimental networks outperform their respective control networks, as shown in Figure 4.1.

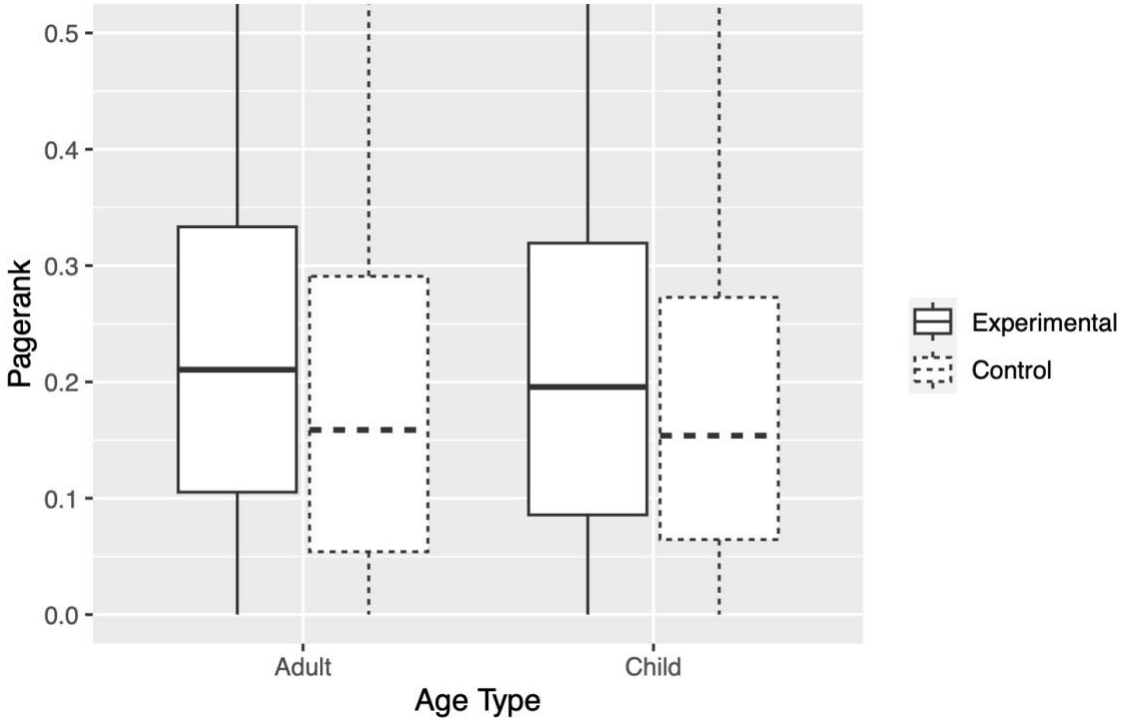
**Figure 4.1 Observational: Main Effect of Experimental Type**




***Note.*** Across all network types, experimental networks perform better than their controls.

The interactions of interest are those that involve experimental type, indicating a difference in performance. The interaction between age type and experimental type was not significant, $F(1,8946) = 1.792$, $p = .181$. Across semantic types, there is no difference in performance between the child and adult networks, illustrated in Figure 4.2. However, as we will note later, this is likely driven by fact that the feature networks' performance deviate from the other two.

**Figure 4.2 Observational: Comparison of Age Types**



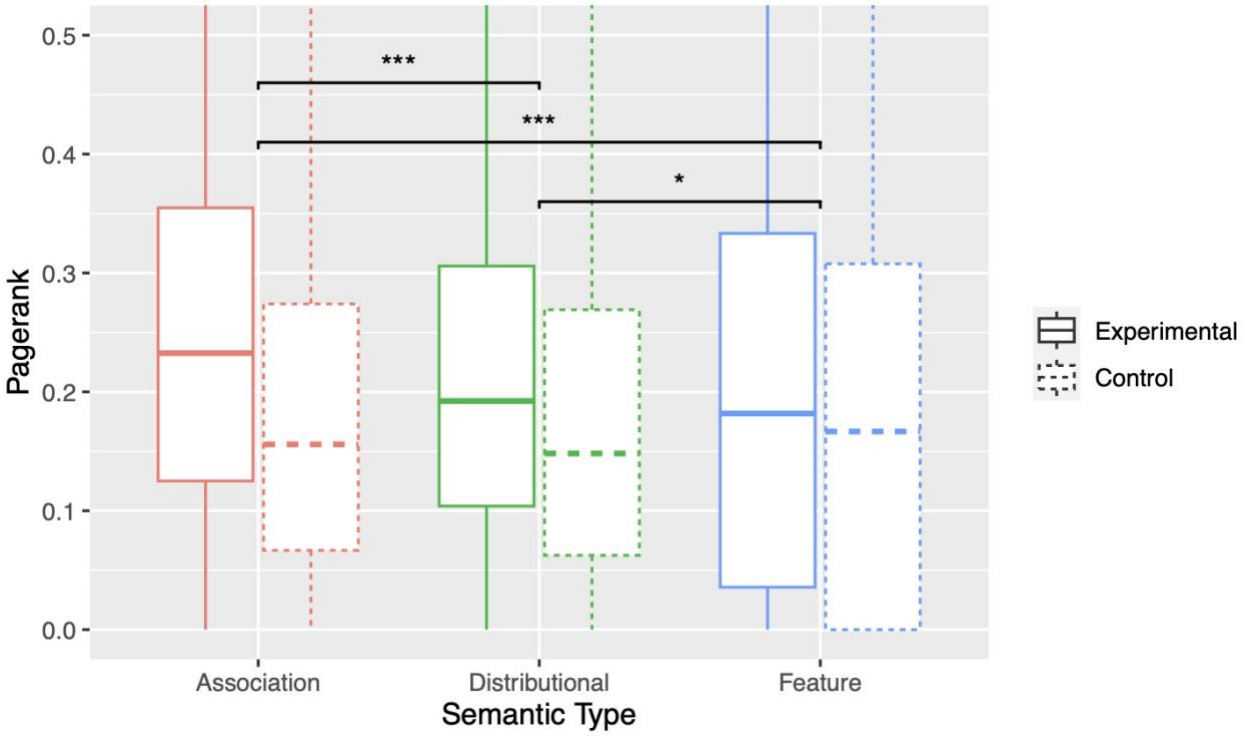

***Note.*** the age type by experimental type interaction is not significant in the overall model.

The interaction between semantic type and experimental type was significant ($F(2,8946) = 29.670$, $p < .001$), indicating a difference in the performance of the three semantic types. This effect is shown in Figure 4.3. To compare each pair of semantic types, we conduct three individual 2 (semantic type) x 2 (experimental type) ANCOVA's, averaged across both age types. This keeps results in terms of the experimental-control difference. These show that the association networks perform better than both the distributional ($F(1,6020 = 36.388$, $p < .001$) and the feature networks ($F(1,5894) = 50.797$, $p < .001$). The distributional networks also perform better than the feature networks, $F(1,5895) = 5.998$, $p < .05$.

All main effects and interactions are subsumed by a three-way interaction between experimental type, semantic type, and age type, ($F(2,8946) = 10.896$, $p < .001$. To unpack this three-way interaction, we next conduct separate 3 (semantic type) x 2 (experimental type)

ANCOVAs for the child and adult type networks separately. This interaction, along with differences found in the next sections, are found in Figure 4.4.

**Figure 4.3 Observational: Comparison of Semantic Types**



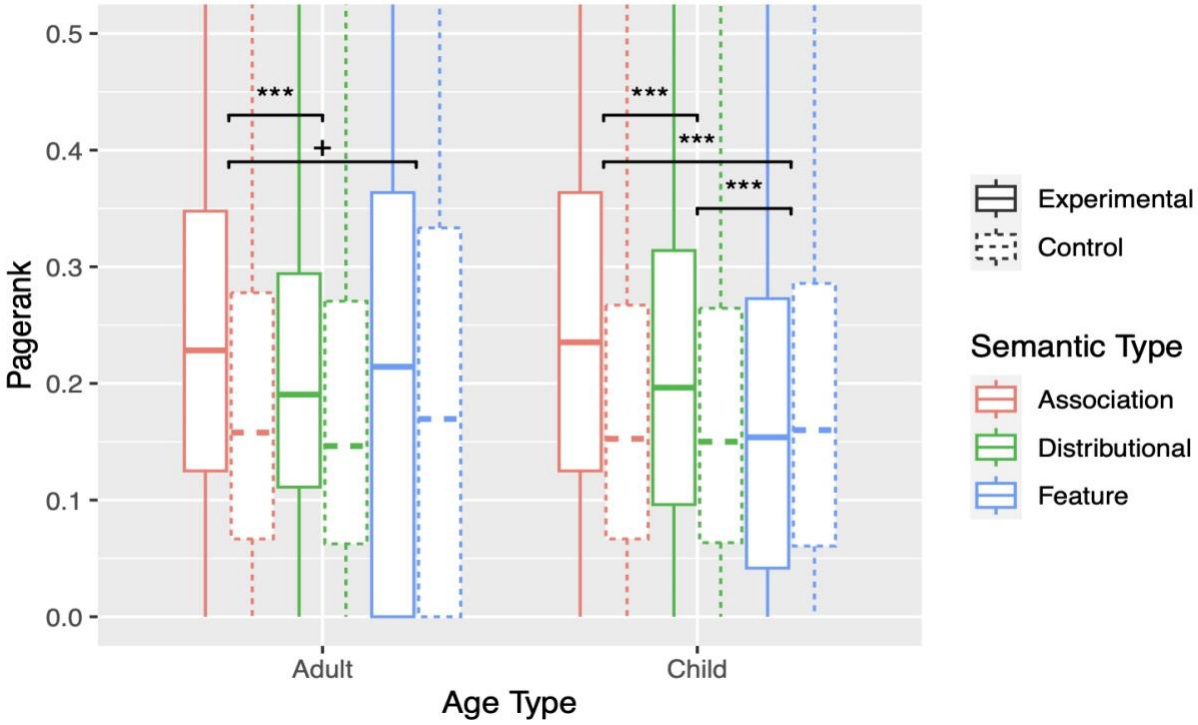

***Note.*** The semantic type by experimental type interaction is significant when averaged across the child and adult type networks. The association type outperforms the distributional and feature types, and the distributional also outperform the feature type.

#### *4.3.1.1 Adult*

Within Adult networks, the 3 (semantic type) x 2 (experimental type) ANCOVA controlling for the same participant factors reveals a significant interaction, $F(2, 4395) = 3.783$, $p < .05$. Next, we perform pairwise comparisons between the three semantic types by running separate 2 (semantic type) x 2 (experimental type) ANCOVAs to compare each pair of semantic types. The first of these analyses revealed a significant interaction indicating the association

networks outperform the distributional networks, $F(1,2959) = 11.938$, $p < .001$. Association networks were found to only marginally outperform feature networks, $F(1,2867) = 3.436$, $p = .0639$. However, performance for distributional and feature networks did not significantly differ, $F(1,2870) = 0.281$, $p = .5959$. For means and a depiction of these results, see Table 4.4 and Figure 4.4. These findings suggest that within the adult type networks, association-based networks best predict children's word-learning. Further, there's no difference in performance between the distributional and feature-based networks.

**Figure 4.4 Observational: Comparison of Semantic Types within Age Types**




***Note.*** For the adult networks, the association networks are better than the distributional networks. For the child networks, the association networks perform better than both the distributional and feature networks, and the distributional also perform better than the feature networks.

#### *4.3.1.2 Child*

Within the Child networks, the 3 (semantic type) x 2 (experimental type) ANCOVA revealed a significant semantic type by experimental type interaction, $F(2,4458) = 45.879$, $p < .001$. We then analyze the 2x2 ANCOVAs comparing each semantic type pairwise. As in the adult networks, association networks outperform distributional networks, $F(1,2967) = 26.825$, $p < .001$. Association networks also outperformed feature networks, $F(1,2929) = 88.136$, $p < .001$. In fact, the child feature networks are the only ones in which the experimental networks do not do better than control networks[1]. Finally, unlike in adult networks, child distributional representations outperformed child feature networks, $F(1,2926) = 22.474$, $p < .001$. In summary, within the child network type, association networks best predict children's word-learning, followed by distributional networks and in contrast with feature networks, which do not perform better than their controls.

#### *4.3.1.3 Association*

To further break down the overall 2x3x2 ANCOVA, we also investigated the difference in age types within each semantic type in order to answer our second research question, despite the lack of a age type by experimental type interaction in the overall ANCOVA. We conducted 2x2 ANCOVAs controlling for the same child-level factors as before. For the association networks, we find a main effect of experimental type, indicating that across ages, the experimental networks perform better than their controls as expected. We also find a significant interaction, showing that the child networks perform better than the adults, $F(1,2964) = 4.722$, $p < .05$. Differences can be seen in Figure 4.5.

---

[1] Traditional Bonferroni-corrected post hoc comparisons find that for the six network types analyzed, the experimental networks were significantly better than their controls at a p-value of less than .001, except for the child feature networks, where there was no difference between the experimental and control networks, $p = .3083$.

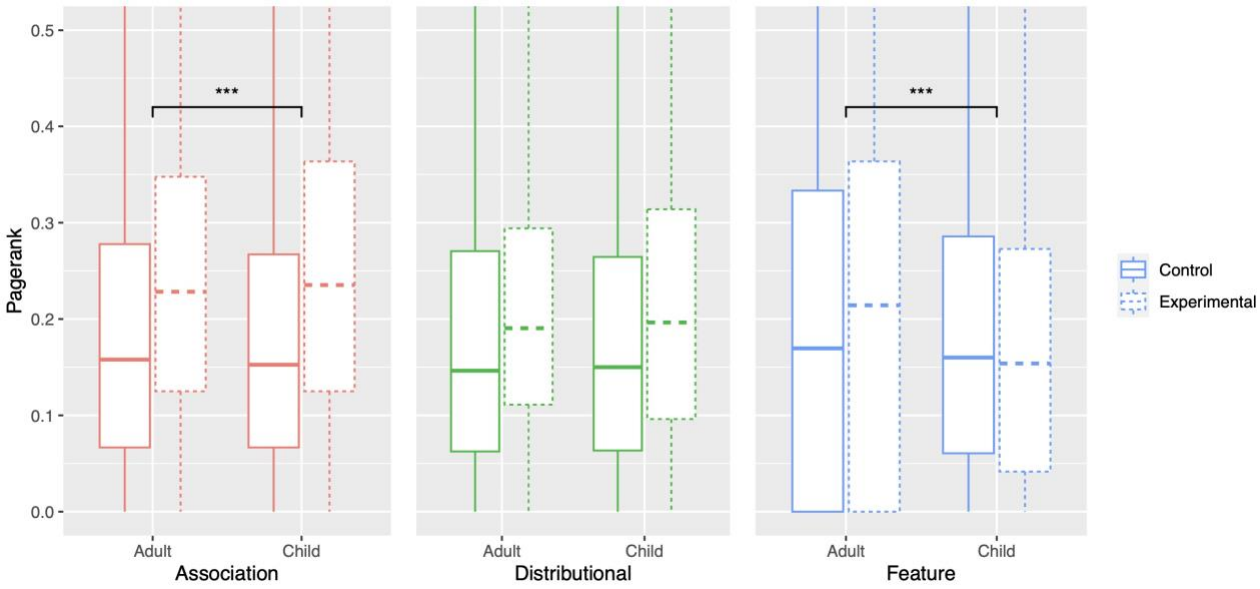
#### *4.3.1.4 Distributional*

Within the distributional networks only, we find that across age types, experimental networks perform better than their controls, $F(1,2962) = 65.791$, $p < .001$. However, we do not find a significant interaction, suggesting that the child and adult networks perform the same, $F(1,2962) = 0.299$, $p = .5843$. Results are shown in Figure 4.5.

#### *4.3.1.5 Feature*

Within the feature networks, experimental networks perform significantly better than controls, $F(1,2830) = 6.470$, $p < .05$. We also find a significant interaction, indicating the adult feature networks perform better than the child ones, $F(1,2830) = 13.971$, $p < .001$. Figure 4.5 depicts these differences.

**Figure 4.5 Observational: Comparison of Age Types within Semantic Types**




***Note.*** Within the association networks, child type networks perform better than adult type. Within the feature networks, adult type networks perform better than child type.

**Table 4.4 Observational: Overall Means Comparing Semantic and Age Types**

| | Adult | | Child | |
|---|---|---|---|---|
| | **Experimental** | **Control** | **Experimental** | **Control** |
| **Association** | 0.2514 | 0.1969 | 0.2621 | 0.1918 |
| **Distributional** | 0.2199 | 0.1913 | 0.2243 | 0.1916 |
| **Feature** | 0.2499 | 0.2155 | 0.1867 | 0.1932 |

***Note.*** Differences in red indicate networks in which the experimental networks do not outperform their control counterparts.

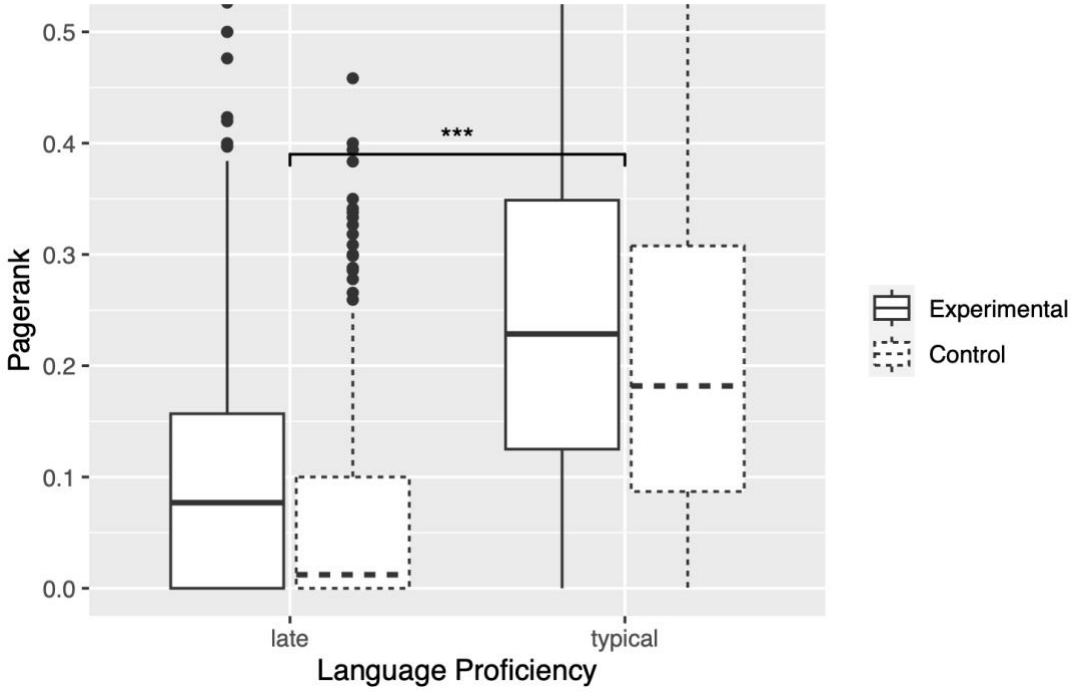
### 4.3.2 Post Hoc – Language Proficiency

After our initial findings, we next were interested in evaluating whether a child's language proficiency plays a role in network accuracy. To do so, we divided children into two groups based on their profiles (at the same visit their vocabulary structures are taken from). There are children who we consider late talkers, or those that fall more than one standard deviation below the normed mean on their CDI percentiles (or those in the 0 to 16th percentile), and their typically developing peers (or those in the 16th to 100th percentiles).

First, to see if children's proficiency plays a role, we investigate the effect by adding it into the statistical model, creating a 3 (semantic type) x 2 (age type) x 2 (experimental type) x 2 (language proficiency) ANCOVA. Here, we are interested in any main effects and interactions containing our factor of interest, language proficiency, and the needed experimental type difference. We still find a main effect of experimental type, in which the experimental networks outperform their controls, $F(1,8934) = 98.623$, $p < .001$. We also find a main effect for proficiency, indicating that both typical talker experimental and control networks have significantly higher accuracies than late talker networks, $F(1,7134) = 94.801$, $p < .001$. This is shown in Figure 4.6. However, no interactions with both factors of interest are significant,

though the semantic type by experimental type by proficiency interaction is marginal, $F(2,8934) = 2.579$, $p = .0759$. Despite a lack of significant interactions of interest, we still break down our model into the 3 (semantic type) x 2 (age type) x 2 (experimental type) interaction within our two proficiency designations, per our proposed analyses.

**Figure 4.6 Observational: Comparison of Language Proficiencies**




***Note.*** Typically developing children overall have higher accuracies than late talkers, but there is no language proficiency by experimental type interaction.

#### *4.3.2.1 Typically Developing*

First, we look at the typically developing children, who make up the majority of the sample. Likely because of this, we replicate the findings from the overall statistical model. First, within the typically developing children the 3 (semantic type) x 2 (age type) x 2 (experimental type) interaction was significant ($F(2, 3717) = 4.622$, $p < .001$). We then investigate this interaction within each age type separately. See Table 4.5 for means.

Within the adult type networks, the 3 (semantic type) x 2 (experimental type) interaction was significant, warranting pairwise analysis, $F(2,3717) = 4.622$, $p < .01$. The association networks outperform the distributional networks, $F(1,2530) = 15.424$, $p < .001$. Association networks were found to only marginally outperform feature networks, $F(1,2431) = 3.691$, $p = .0548$. However, performance for distributional and feature networks did not significantly differ, $F(1,2430) = 0.595$, $p = .4406$.

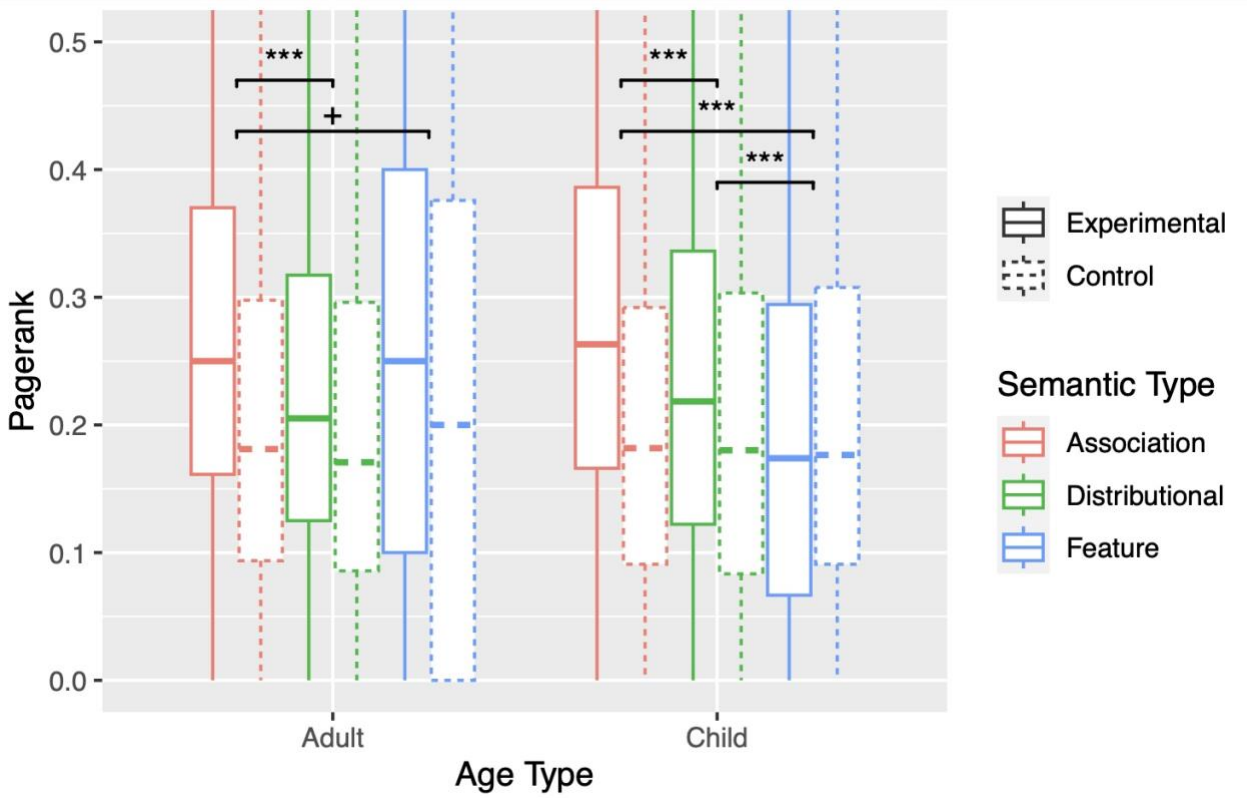
Within the child type networks we also find a significant 3 (semantic type) by 2 (experimental type) interaction, $F(2,3813) = 39.928$, $p < .001$. The association networks perform significantly better than both distributional ($F(1,2505 = 76.404$, $p < .001$) and feature networks ($F(1,2538) = 24.740$, $p < .001$), and the distributional networks also perform better than the feature networks ($F(1,2504) = 17.474$, $p < .001$). For the child feature networks, the control networks do not outperform their control counterparts. These results can be seen in Figure 4.7, and the means in Table 4.5.

**Table 4.5 Observational: Semantic and Age Type for Typical and Late Talking Children**

| | **Typically Developing** | | | | **Late Talkers** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Adult** | | **Child** | | **Adult** | | **Child** | |
| | **Exp.** | **Con.** | **Exp.** | **Con.** | **Exp.** | **Con.** | **Exp.** | **Con.** |
| **Association** | 0.2779 | 0.2212 | 0.2890 | 0.2151 | 0.0942 | 0.0531 | 0.1032 | 0.0537 |
| **Distributional** | 0.2381 | 0.2146 | 0.2467 | 0.2147 | 0.1113 | 0.0525 | 0.0913 | 0.0546 |
| **Feature** | 0.2747 | 0.2416 | 0.2073 | 0.2141 | 0.1155 | 0.0740 | 0.0610 | 0.0656 |

***Note.*** Differences in red indicate networks in which the experimental networks do not outperform their control counterparts.

**Figure 4.7 Observational Typical: Comparison of Semantic Types within Age Type**

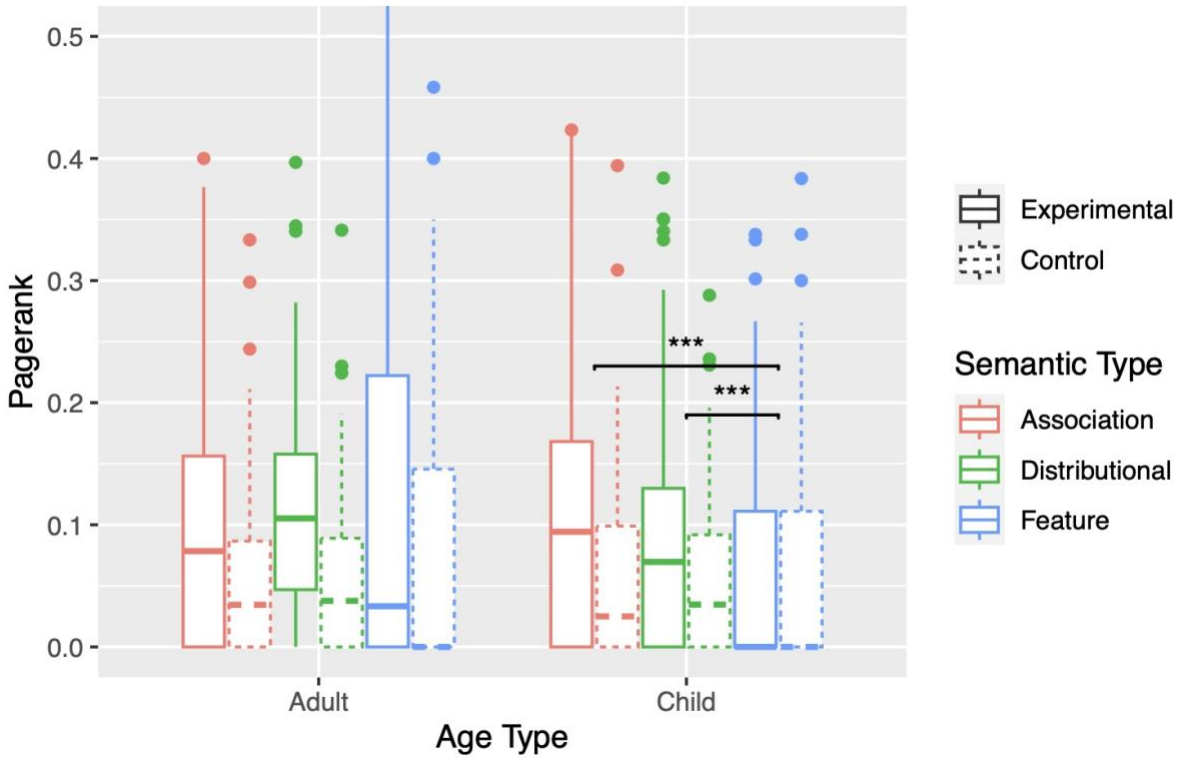



***Note.*** For the adult networks, the association networks are better than the distributional networks. For the child networks, the association networks perform better than both the distributional and feature networks, and the distributional also perform better than the feature networks.

#### *4.3.2.2 Late Talkers*

Next, we investigate our findings within children considered late talkers. We find a significant 3 (semantic type) x 2 (age type) x 2 (experimental type) interaction, $F(2,1269) = 4.770$, $p < .01$. See Table 4.5 and Figure 4.8 for the means and significant differences within our late talking sample.

**Figure 4.8 Observational Late Talkers: Comparison of Semantic Types within Age Type**




***Note.*** For the child networks, both the association networks and the distributional networks perform better than the feature networks and feature networks, but the association and distributional networks perform the same.

We next investigate this interaction in late talkers by breaking it down by age type next. Within the adult networks, there is no significant 3 (semantic type) x 2 (experimental type) interaction, suggesting within late talkers, all the adult networks perform equally, $F(2,626) = 0.975$, $p = .378$. Within the child networks, however, we do see a significant semantic by experimental type interaction ($F(2,624) = 15.520$, $p < .001$), indicating we should investigate the pairwise ANCOVAs. Both the association networks ($F(1,408) = 27.371$, $p < .001$) and the distributional networks ($F(1,402) = 16.532$, $p < .001$) outperform the feature networks, but the association networks and the distributional networks perform equally ($F(1,434) = 1.808$, $p =$

.1795). Again, for the child feature networks, the control networks do not outperform their control counterparts.

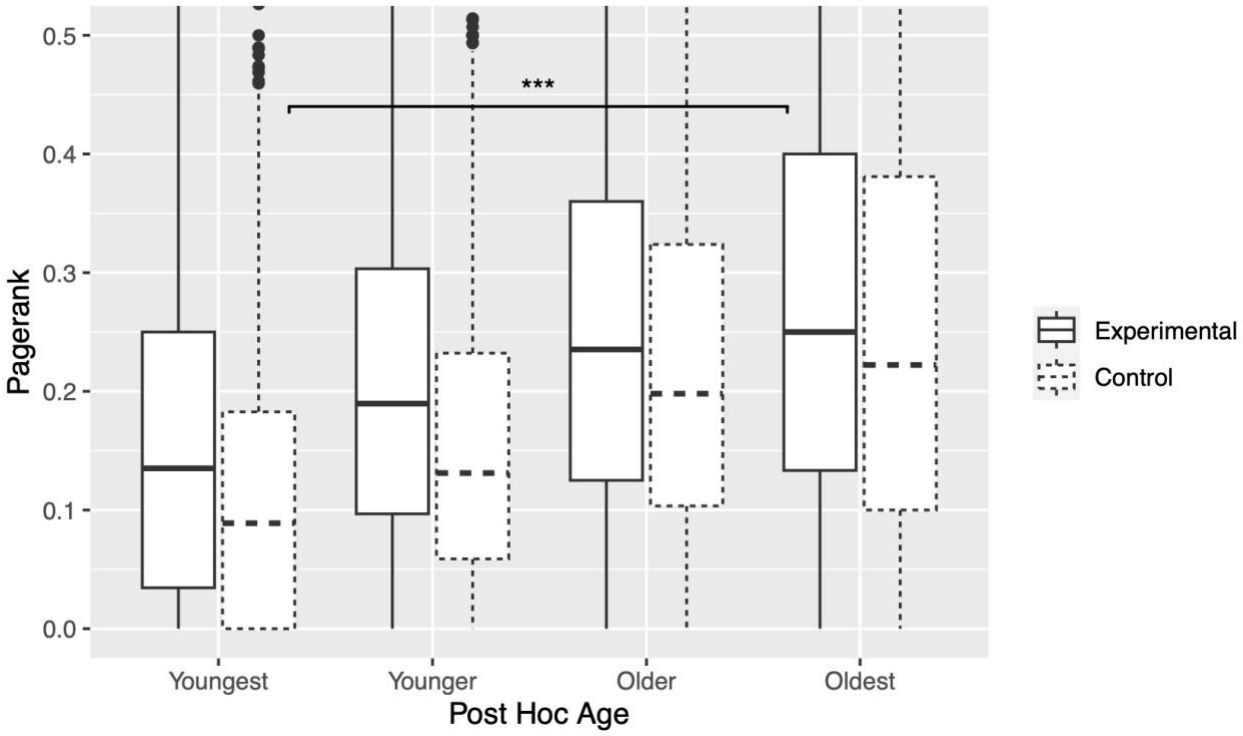
### 4.3.3 Post Hoc – Age

We were also interested in how the age of the child plays a factor in network prediction accuracies. First, we know that in our full models, age as a continuous factor is a significant predictor of accuracy; as age increases, so does accuracy, $F(1,5042) = 151.610$, $p < .001$. However, to delve deeper into this finding, we broke children up into four categories to investigate the interactions of interest within each: youngest ($M$ = 18.41 months, $SD$ = 0.94), younger ($M$ = 21.08 months, $SD$ = 0.77), older ($M$ = 23.93 months, $SD$ = 0.87), and oldest ($M$ = 27.33 months, $SD$ = 1.30). To create our categories, we divided children so that each category had an equal number of children. Because our total datapoints was not divisible by four, the youngest and oldest groups contain one more datapoint than the younger and older groups.

As with language proficiency, we first add our age factor into the overall model to see if differences might exist, creating a 3 (semantic type) x 2 (age type) x 2 (experimental type) x 4 (age group) ANCOVA. We find a main effect age group, suggesting there are overall differences in between our four age groups across both the experimental and control networks and all other factors, $F(3,8087) = 20.762$, $p < .001$. The trends suggest that as children age, overall accuracy increase. We further find an experimental type by age group interaction, suggesting performance differs among our groups, $F(3,8911) = 5.123$, $p < .01$. Figure 4.9 illustrates this finding. Though there are differences in performance, we do not see a trend of performance increases or decrease as age increases. Finally, we find a 3 (semantic type) x 2 (experimental type) x 4 (age group) interaction, further suggesting the differences among semantic type performance differs among our age groups, $F(6,8911) = 3.013$, $p < .01$. However, we do not find the omnibus four-way

interaction, $F(6,8911) = 0.226$, $p = .9649$. These findings warrant a deeper investigation, next looking within each age group separately. For each age group, accuracy means are provided in Table 4.6.

**Figure 4.9 Observational: Post Hoc Comparison of Ages**



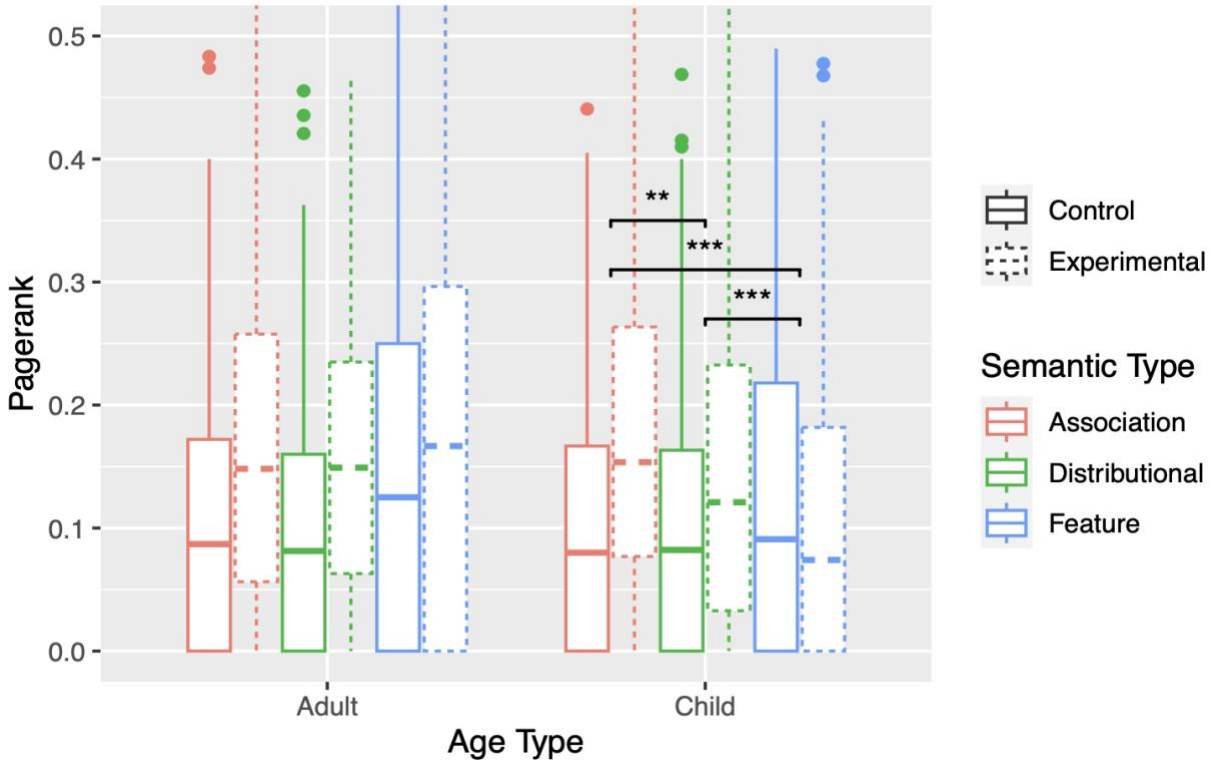

***Note.*** There is a significant difference among the ages in their performance. There are also differences between semantic types among the ages.

#### *4.3.3.1 Youngest*

Within the youngest group of children, we find a significant semantic type by age type by experimental type interaction, warranting further analysis, $F(2,1970) = 7.043$, $p < .001$. Within the adult networks though, we do not see a significant semantic type by experimental type interaction, suggesting that for the youngest children, there is no difference between the three adult semantic types when predicting specific word learning. There is a significant semantic type

by experimental type interaction within the child type networks, however, which is shown in Figure 4.10, $F(2,953) = 40.572$, $p < .001$. Pairwise ANCOVAs reveal that the child association networks perform better than both the distributional ($F(1,632) = 7.846$, $p < .01$) and the feature networks ($F(1,601) = 71.042$, $p < .001$). The distributional networks are also better than the feature networks, $F(1,5966) = 38.974$, $p < .001$.

**Figure 4.10 Observational Youngest Ages: Comparison of Semantic Types within Age Type**



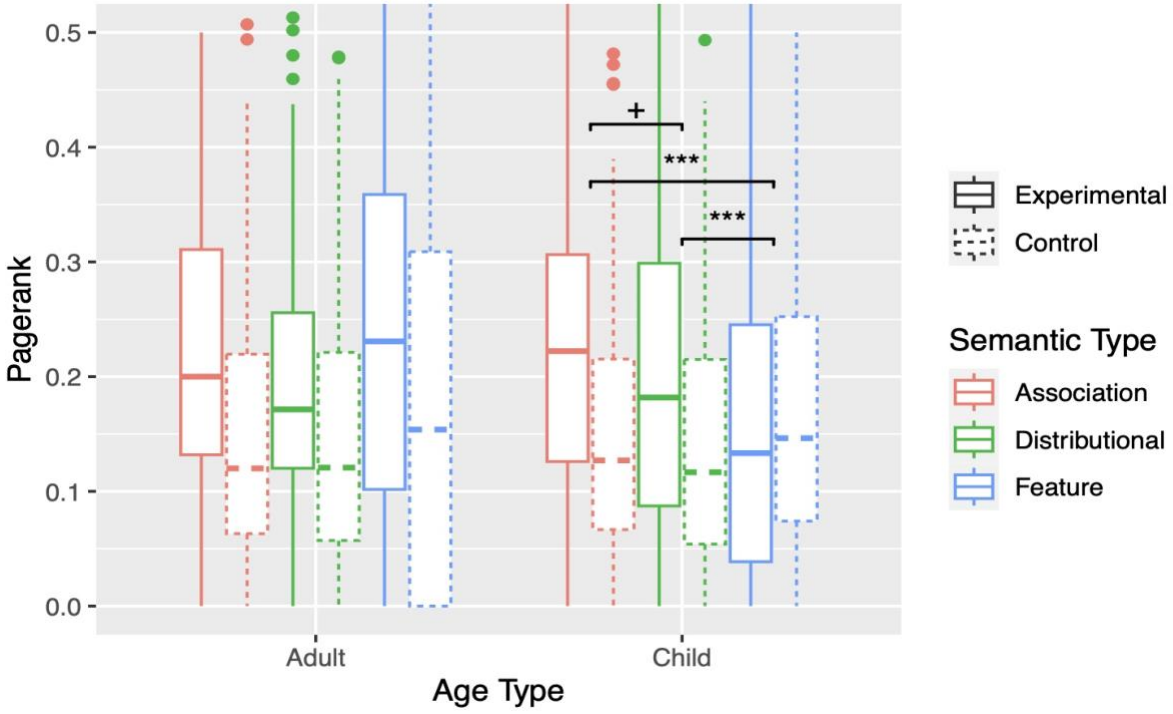

***Note.*** For youngest age group and the child type networks, the association networks perform better than both the distributional and feature networks, and the distributional also perform better than the feature networks.

#### *4.3.3.2 Younger*

For children in the next youngest group, the younger children, we again see a significant semantic type by age type by experimental type interaction, $F(2,2211) = 8.786$, $p < .001$. We

next investigate the semantic by experimental type interaction within the adult and child type networks separately. There is no difference among the semantic types within the adult type networks, $F(2,1027) = 0.987$, $p = .3724$. Within the child type networks, however, we do see a semantic by experimental type interaction, warranting further pairwise analyses, $F(2,1016) = 28.479$, $p < .001$. Further analysis indicates that both the association networks ($F(1,635) = 44.376$, $p < .001$) and the distributional networks ($F(1,631) = 29.711$, $p < .001$) perform better than the feature networks. There is no difference between the association and distributional networks, $F(1,675) = 2.982$, $p = .0847$. These results are depicted in Figure 4.11.

**Figure 4.11 Observational Younger Ages: Comparison of Semantic Types within Age Type**

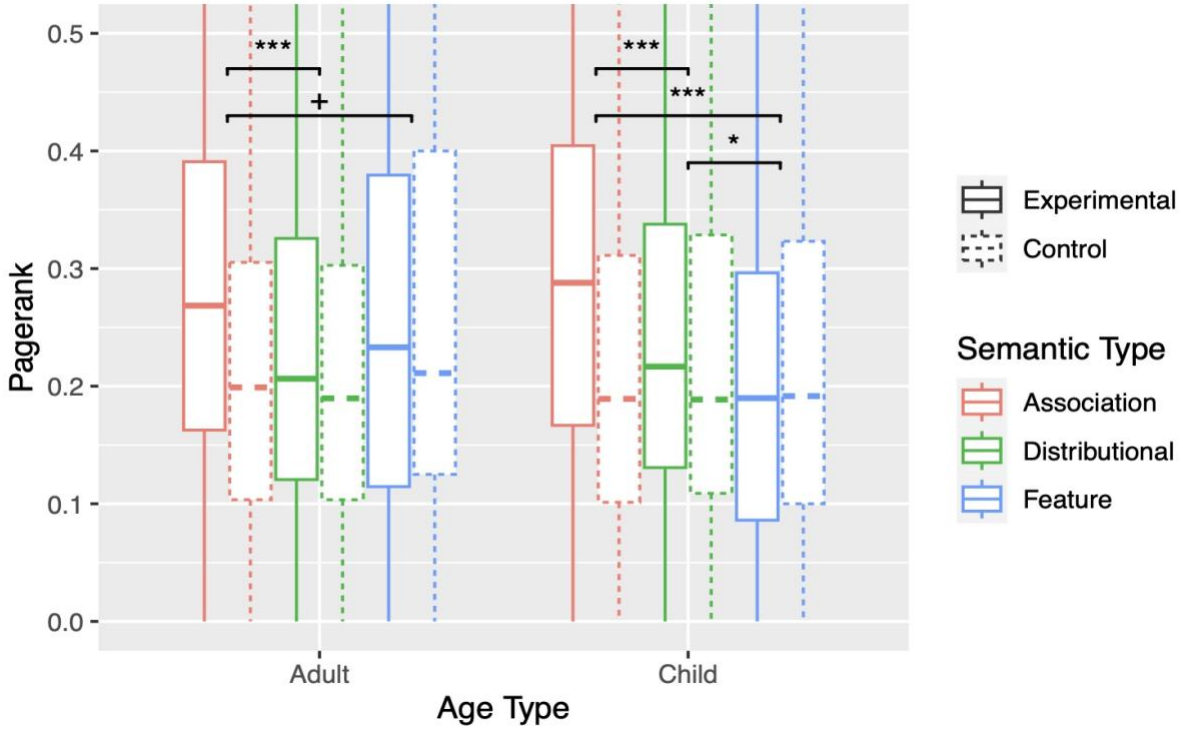



***Note.*** For the child networks, both the association networks and the distributional networks perform better than the feature networks and feature networks, but the association and distributional networks perform the same.

#### *4.3.3.3 Older*

For children in the third oldest group, we find a marginally significant semantic type by age type by experimental type interaction, which we believe warrants further analysis, $F(1,2231) = 2.845$, $p = .0584$. Investigating the semantic by experimental type interaction within the adult type networks, we find it to be significant, $F(2,1041) = 3.548$, $p < .05$. Delving deeper into the pairwise differences, we find that the adult association networks are significantly better than the adult distributional ($F(1,703) = 13.596$, $p < .001$), but only marginally better than the adult feature ($F(1,676) = 3.854$, $p = .0500$). There is no difference between the adult distributional and feature networks, ($F(1,663) = 0.043$, $p = .8352$). Within the child type networks, we also find a significant semantic type by experimental type interaction, $F(2,1071) = 20.566$, $p < .001$. All pairwise differences are significant, showing that the child association networks are better than both the child distributional networks ($F(1,707)=18.603$, p <.001) and the child feature networks ($F(1,689) = 38.812$, $p < .001$), and the child distributional networks are also better than the child feature networks ($F(1,693) = 6.269$, $p < .05$). The results for both adult and child type networks are found in Figure 4.12.

**Figure 4.12 Observational Older Ages: Comparison of Semantic Types within Age Type**



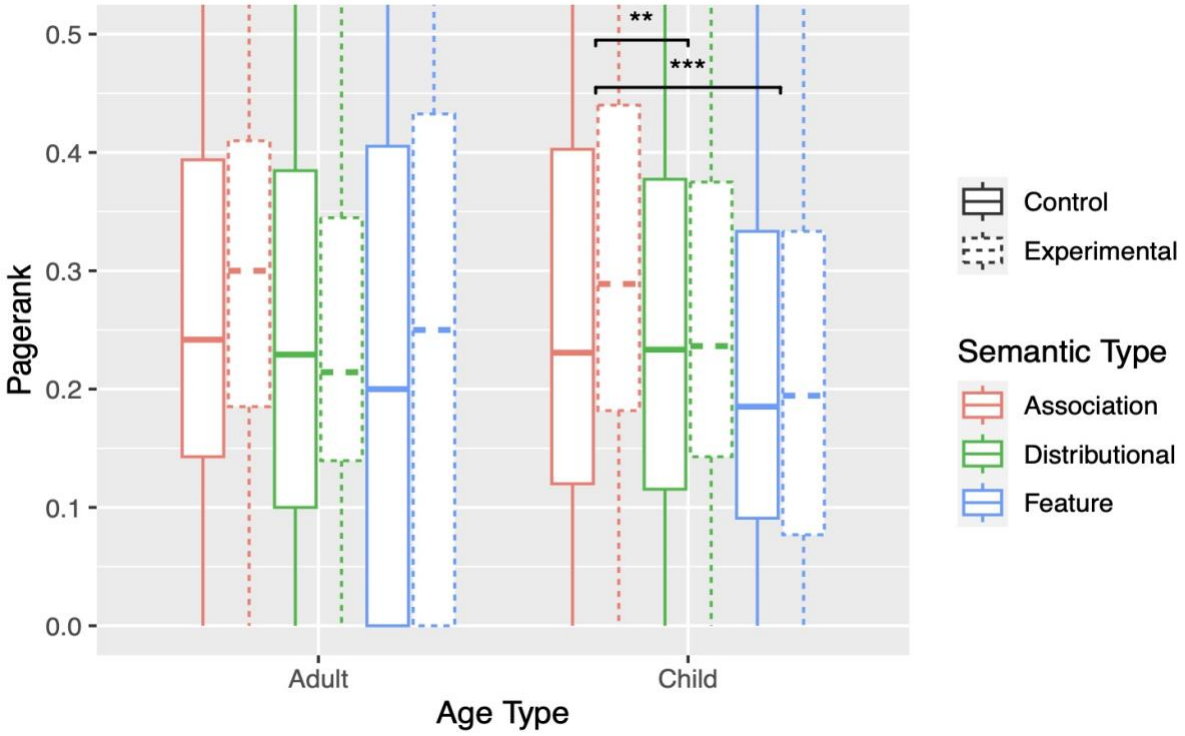

***Note.*** For the adult networks, the association networks are better than the distributional networks. For the child networks, the association networks perform better than both the distributional and feature networks, and the distributional also perform better than the feature networks.

#### *4.3.3.4 Oldest*

Our final age grouping contains the oldest children in the observational study. In this group, there is no significant semantic type by age type by experimental type interaction $F(2,2203) = 1.377$, $p = .2526$. However, we are still interested in understanding what might be happening within this age group by examining the adult and child type networks separately, as we did with the other age groups. The adult networks show no differences among the semantic types, $F(2,1033) = 1.789$, $p = .1677$. The child type networks do show significant differences among the three semantic types, warranting pairwise comparison, $F(2,1081) = 5.990$, $p < .01$. The child association networks show better performance than both the child distributional

($F(1,709) = 10.924$, $p < .001$) and child feature networks ($F(1,688) = 7.1669$, $p < .01$), but there is no difference between the child distributional and feature networks ($F(1,688) = 0.164$, $p = .685$). Results are shown in Figure 4.13. Curiously, for the oldest children we see the expected pattern in which the experimental child feature networks show higher accuracies than controls, though we do not know if this is significant.

**Figure 4.13 Observational Oldest Ages: Comparison of Semantic Types within Age Type**




***Note.*** For the child networks, the association networks perform better than both the distributional and feature networks, but the distributional and feature networks perform the same.

**Table 4.6 Observational: Comparing Semantic and Age Types within the Four Age Groups**

| | Youngest | | | | Younger | | | |
|---|---|---|---|---|---|---|---|---|
| | Adult | | Child | | Adult | | Child | |
| | Exp. | Con. | Exp. | Con. | Exp. | Con. | Exp. | Con. |
| **Association** | 0.1721 | 0.1157 | 0.1799 | 0.1079 | 0.2260 | 0.1563 | 0.2338 | 0.1610 |
| **Distributional** | 0.1624 | 0.1064 | 0.1514 | 0.1032 | 0.2063 | 0.1549 | 0.2105 | 0.1551 |
| **Feature** | 0.1865 | 0.1535 | 0.1175 | 0.1308 | 0.2509 | 0.1981 | 0.1717 | 0.1826 |
| | Older | | | | Oldest | | | |
| | Adult | | Child | | Adult | | Child | |
| | Exp. | Con. | Exp. | Con. | Exp. | Con. | Exp. | Con. |
| **Association** | 0.2852 | 0.2274 | 0.2985 | 0.2170 | 0.3131 | 0.2790 | 0.3269 | 0.2716 |
| **Distributional** | 0.2410 | 0.2282 | 0.2581 | 0.2311 | 0.2627 | 0.2651 | 0.2684 | 0.2660 |
| **Feature** | 0.2816 | 0.2643 | 0.2212 | 0.2330 | 0.2787 | 0.2445 | 0.2288 | 0.2195 |

***Note.*** Differences in red indicate networks in which the experimental networks do not outperform their control counterparts.

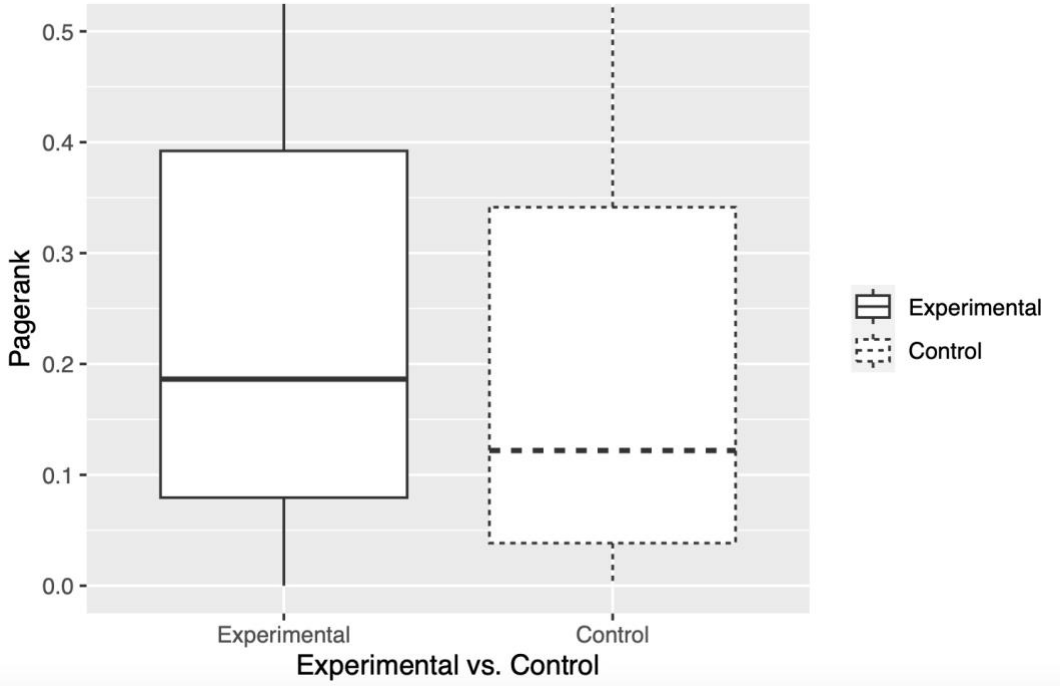
### 4.3.4 Intervention Dataset

Our intervention dataset is a unique opportunity to try and corroborate our previous findings within a new set of children that have pre and post CDI data in the same age group from the same Colorado population. We also can investigate a new factor, whether a short word-learning intervention where children received personalized vocabulary words impacts how our models perform. Of particular interest is the fact that this personalized vocabulary came from semantic structural information similar to that found in our child feature model, with some children receiving highly predicted words (related to the way we are predicting specific words here), and some receiving the least-likely predicted words. Before investigating that possibility, we will first look to replicate our previous findings with the observational dataset.

Examining the overall 3 (semantic type) x 2(age type) x 2 (experimental type) ANCOVA, we are only interested in effects with experimental type as a factor. First, we find a significant

experimental type main effect, indicating that overall, the experimental networks continue to outperform their controls, $F(1,493) = 75.843$, $p < .001$[2]. See Figure 4.14.

**Figure 4.14 Intervention: Main Effect of Experimental Type**



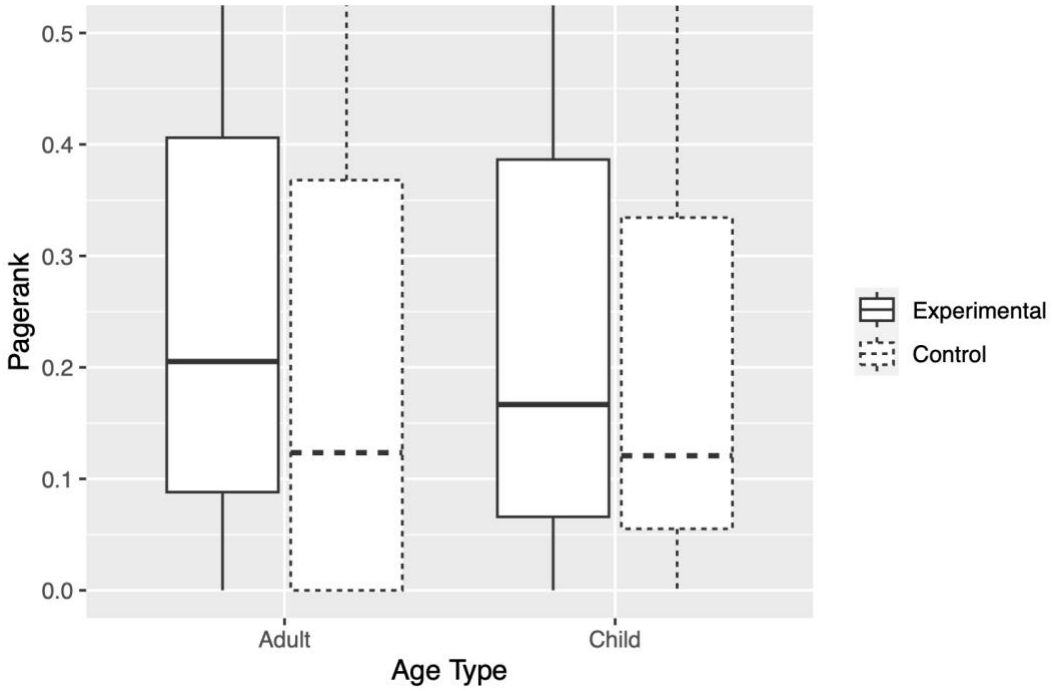

***Note.*** Across all network types, experimental networks perform better than their controls.

Though we do not find an age type by experimental type interaction ($F(1,493) = 1.764$, $p = .1848$), suggesting no differences between the adult and child type networks when collapsing across semantic type, we do find a semantic type by experimental type interaction ($F(2,493) = 13.076$, $p < .001$), suggesting a performance difference in semantic types across the two age types. These findings replicate those found in the overall ANCOVA for the observational dataset. See Figures 4.15 and 4.16. However, in this dataset we do not find a significant three-

[2] Traditional Bonferroni-corrected post hoc comparisons find that for the six network types analyzed, the experimental networks were significantly better than their controls at $p < .05$, except for the child feature networks, where there was no difference between the experimental and control networks, $p = .3668$.

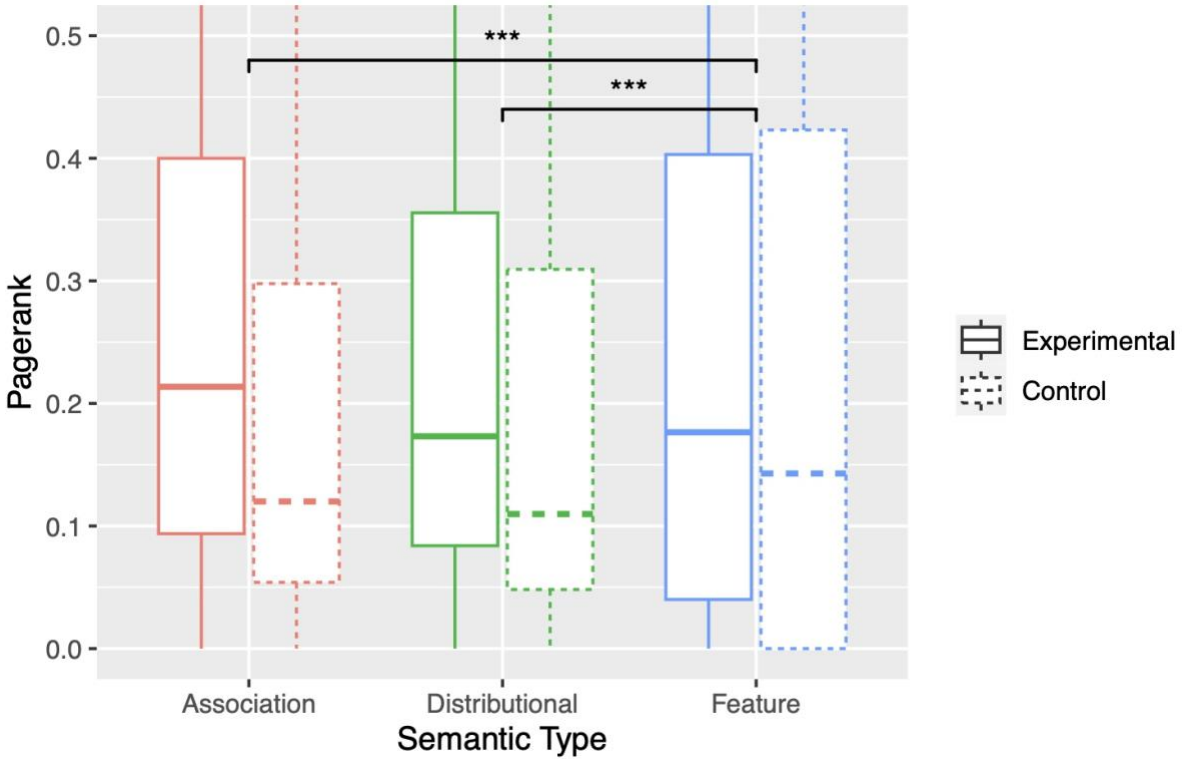
way interaction between semantic type, age type, and experimental type, $F(2,493) = 1.933$, $p = .1459$. The means for each semantic type and age type can be found in Table 4.7.

**Figure 4.15 Intervention: Comparison of Age Types**




***Note.*** There is no difference in performance between the adult and child networks.

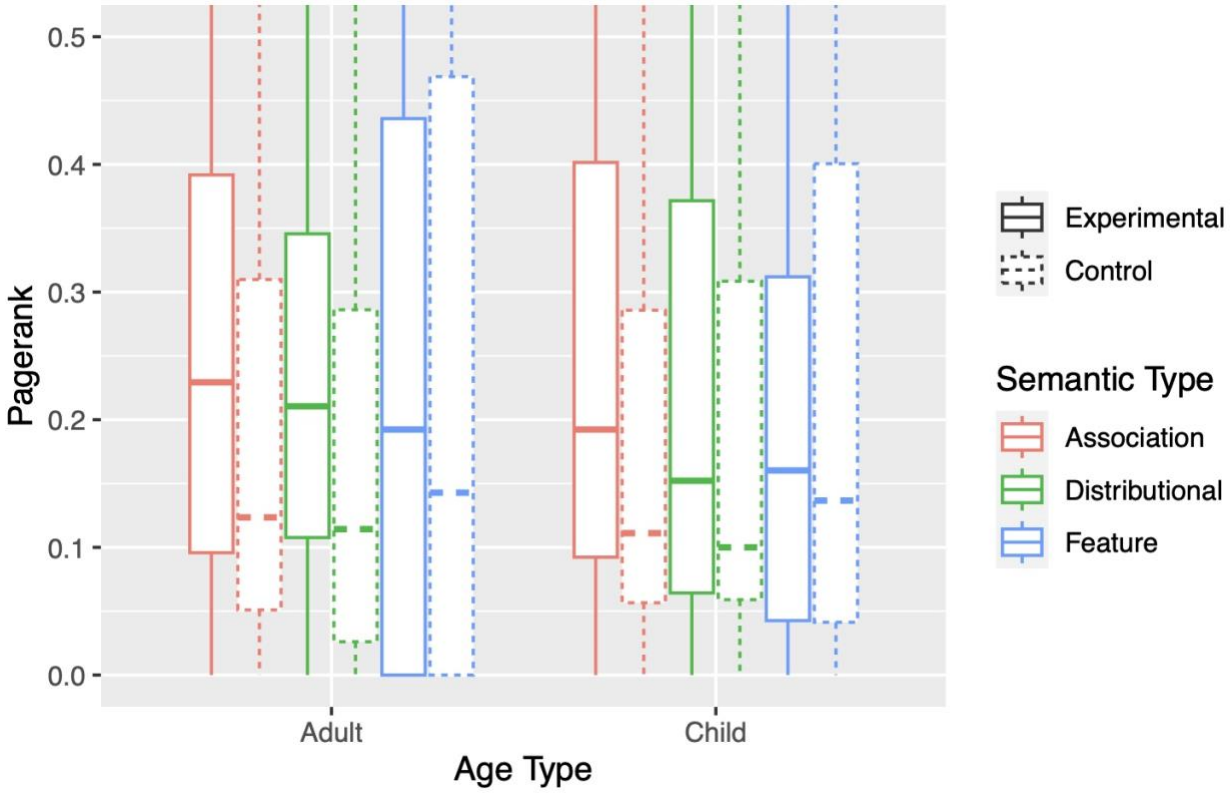
**Figure 4.16 Intervention: Comparison of Semantic Types**




***Note.*** The association and the distributional networks outperform the feature networks.

To investigate the significant semantic type by experimental type interaction further, we next compare each semantic type pairwise using 2 (semantic type) x 2 (experimental type) ANCOVAs, averaging across the age types (because there was no significant semantic type by age type by experimental type interaction). Both the association ($F(1,316) = 17.734$, $p < .001$) and the distributional type networks ($F(1,319) = 13.609$, $p < .001$) show better performance (a larger difference between their experimental and control networks) than the feature networks. There is no difference between the association and the distributional networks, however, $F(1,318) = 1.337$, $p = .2485$. Though we did not statistically compare the semantic types within age type due to the lack of a significant three-way interaction, the observed trends follow those of the observational dataset; see Figure 4.17.

**Figure 4.17 Intervention: Comparison of Semantic Types within Age Types**




***Note.*** We did not statistically compare semantic types within age types due to a lack of a significant semantic type by age type by experimental type interaction.

**Table 4.7 Intervention: Overall Means Comparing Semantic and Age Types**

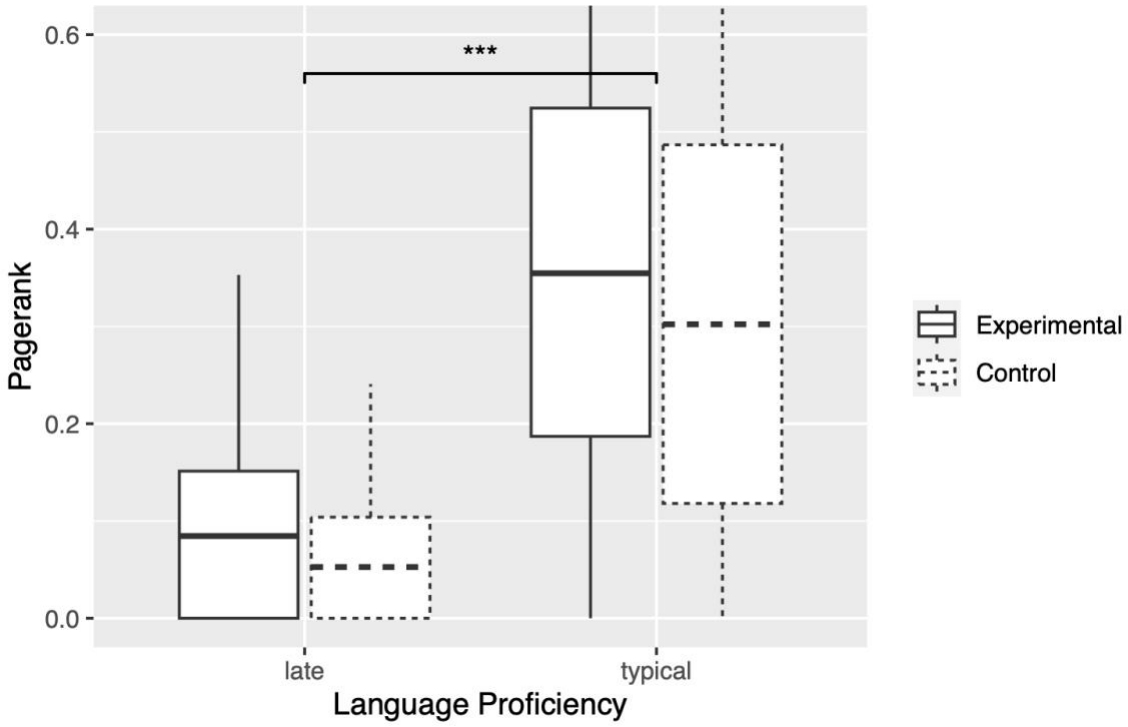
| | **Adult** | | **Child** | |
|---|---|---|---|---|
| | **Experimental** | **Control** | **Experimental** | **Control** |
| **Association** | 0.2697 | 0.2037 | 0.2632 | 0.1876 |
| **Distributional** | 0.2474 | 0.1841 | 0.2448 | 0.1923 |
| **Feature** | 0.2879 | 0.2500 | 0.2208 | 0.2326 |

***Note.*** Differences in red indicate networks in which the experimental networks do not outperform their control counterparts.

### 4.3.5 Post Hoc – Language Proficiency

We next examine language proficiency within the intervention dataset in an attempt to replicate findings from the observational. We again divided children into late talkers (greater than one standard deviation below the mean for CDI percentile) and their typically developing peers. We first add the language proficiency categorical variable into our model, creating a 3 (semantic type) x 2 (age type) x 2 (experimental type) x 2 (language proficiency) ANCOVA. We find a significant main effect of language proficiency ($F(1,46) = 28.035$, $p < .001$), showing that, across experimental and control networks, the models perform better for typically developing children than for late talkers. See Figure 4.18. There is also a marginal experimental type by language proficiency interaction ($F(1,485) = 3.732$, $p = .0540$), and a significant semantic type by experimental type by language proficiency three-way interaction ($F(2,485) = 3.480$. $p < .05$). These findings are subsumed by the four-way interaction between semantic type, age type, experimental type, and language proficiency, $F(2,485) = 3.015$, $p < .05$. These findings are unlike those found in the observational dataset. To begin to unpack these interactions, we first investigate typically developing children only.

**Figure 4.18 Intervention: Comparison of Language Proficiencies**



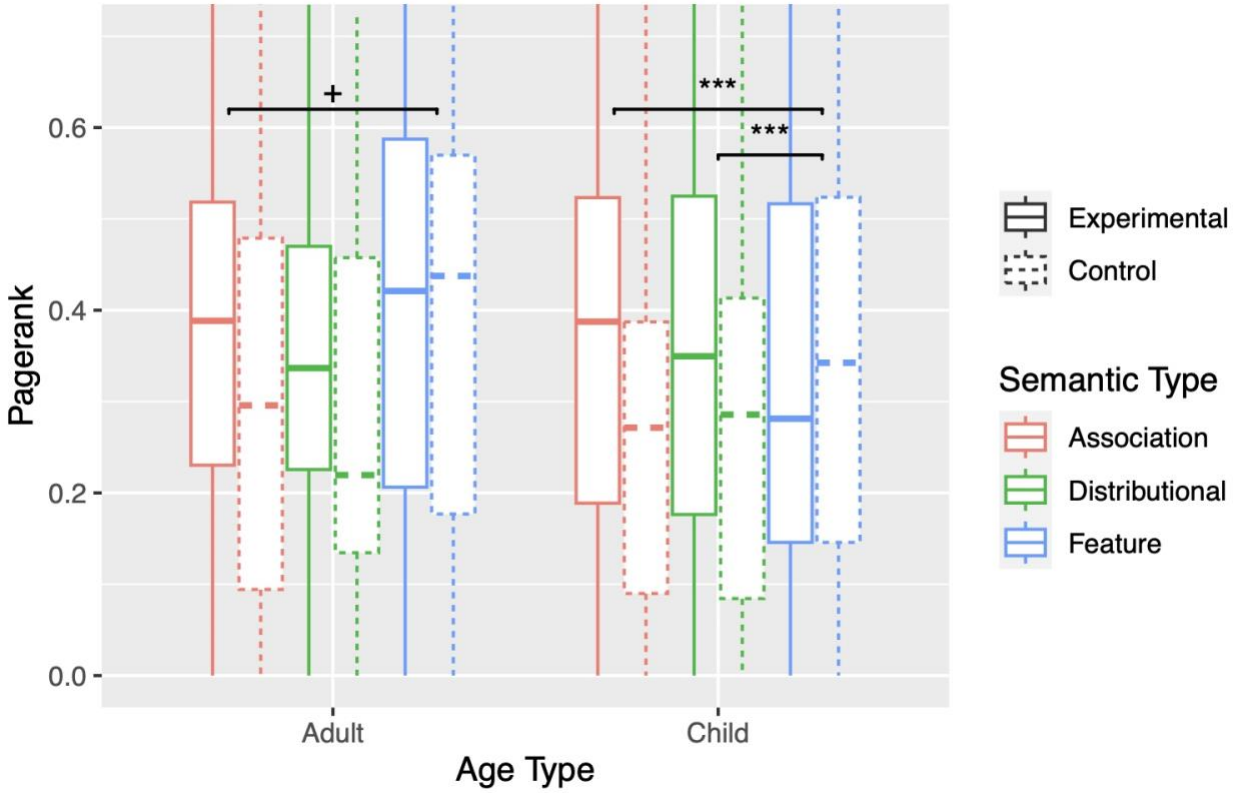

***Note.*** Typically developing children overall have higher accuracies than late talkers.

#### *4.3.5.1 Typically Developing*

Within typically developing children, we find a marginally significant semantic type by age type by experimental type interaction, $F(2,283) = 2.629$, $p = .0739$. We believe this warrants further investigation. Means can be found in Table 4.8. First, within the adult type networks, we see only a marginally significant semantic type by experimental type interaction, $F(2,126) = 2.435$, $p = .0917$. However, we will still unpack this further by conducting pairwise ANCOVAs. We only find that the adult association networks marginally outperform the adult feature, $F(1,71) = 3.844$, $p = .0538$. There are no differences in performance between the adult distributional and adult association ($F(1,77) = 1.324$, $p = .2535$) or adult feature ($F(1,72) = 1.608$, $p = .2088$). Trends can be seen in comparison to the child networks in Figure 4.19. In the child type networks, there is a significant semantic by experimental interaction that is unpacked further

with pairwise comparisons, $F(2,119) = 17.089$, $p < .001$. The child association ($F(1,59) = 19.074$, $p < .001$) and distributional networks ($F(1,59) = 25.915$, $p < .001$) both outperform the feature networks. There is no difference between the association and distributional networks, $F(1,78) = .001$, $p = .9791$. Trends follow that of the typically developing children in the observational dataset, though not all the differences that were statistically significant in the observational dataset were significant for the smaller intervention dataset.

**Figure 4.19 Intervention Typical: Comparison of Semantic Types within Age Type**



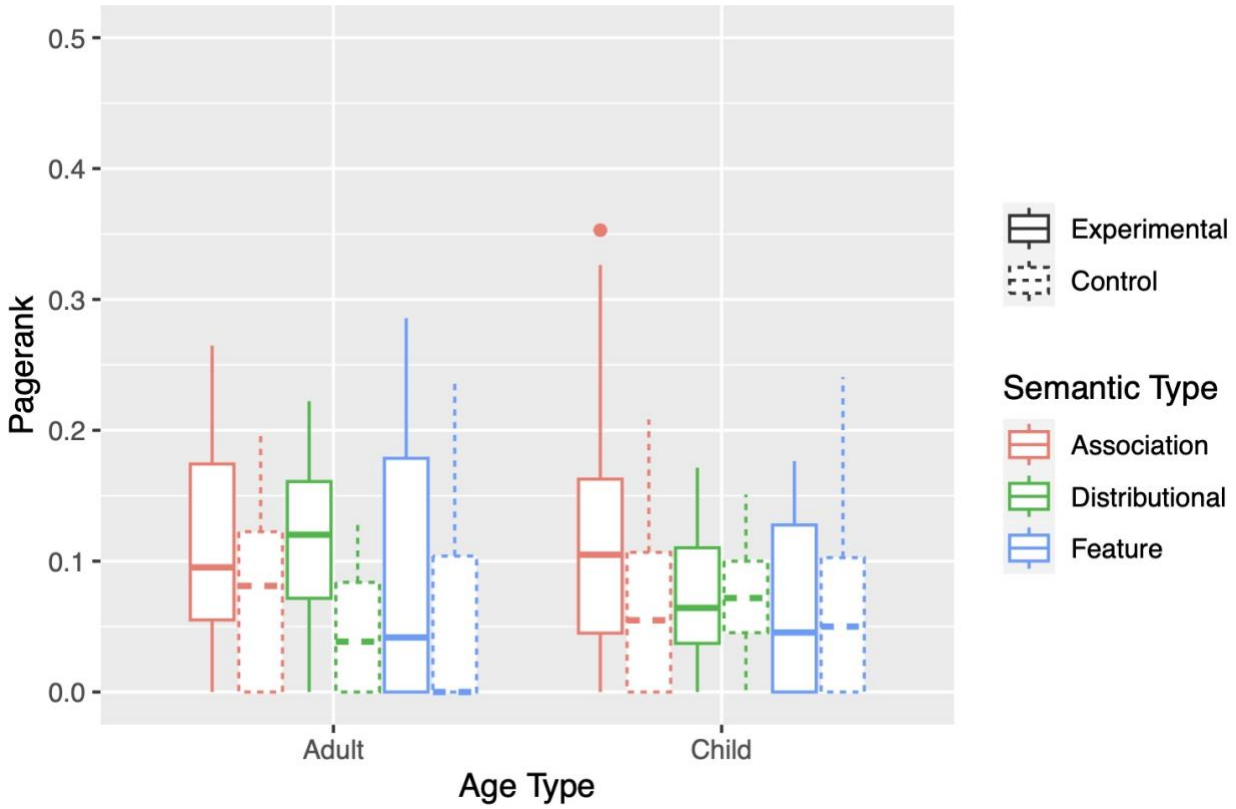

***Note.*** Within the child type networks, both the child association and distributional networks outperform the feature networks.

#### *4.3.5.2 Late Talkers*

Next, we investigate late talkers. Within late talkers, the 3 (semantic type) x 2 (age type) x 2 (experimental type) interaction is not significant, $F(2,205) = 2.287$, $p = .1042$. Though we do

not statistically unpack this interaction, we do visualize trends within the semantic types and age types in Figure 4.20. Means for each network type can be found in Table 4.8.

**Figure 4.20 Intervention Late Talkers: Comparison of Semantic Types within Age Type**




***Note.*** There was no significant semantic type by age type by experimental type interaction, so no other comparisons were explored.

**Table 4.8 Intervention: Semantic and Age Type for Typical and Late Talking Children**

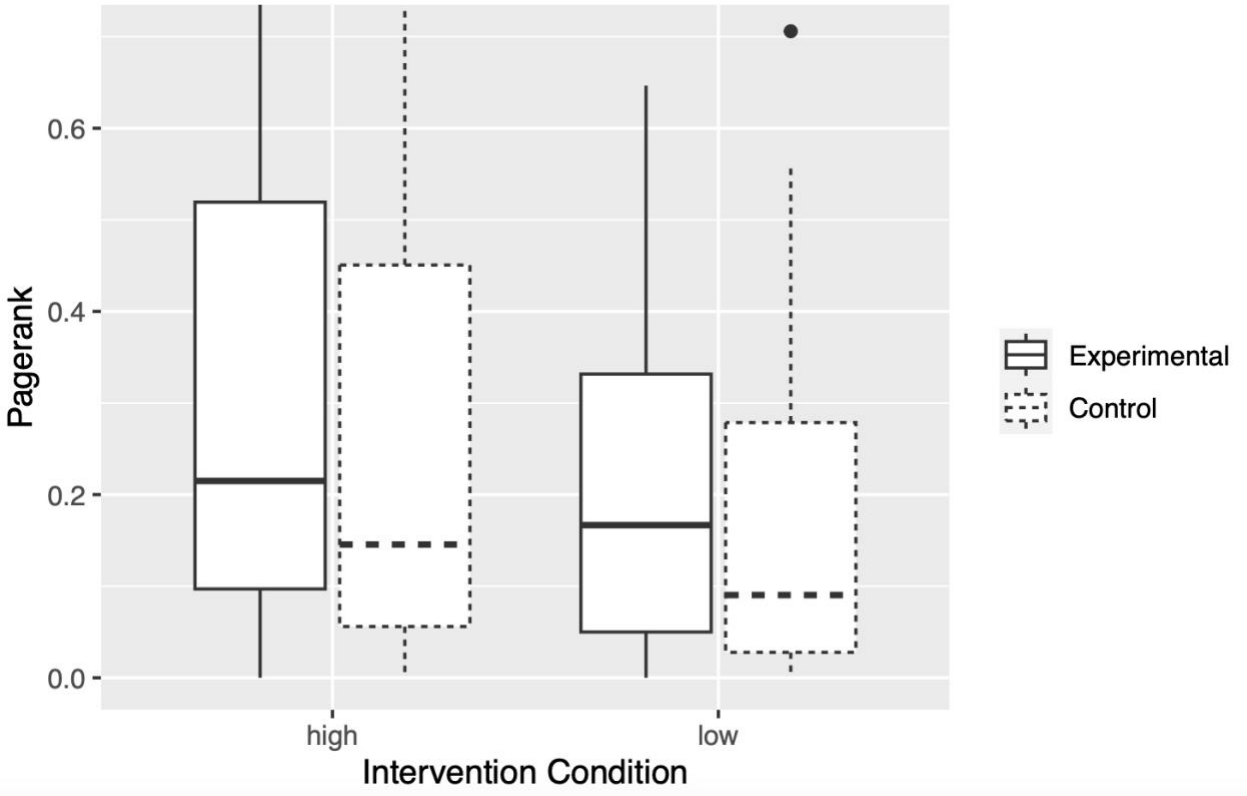
| | Typically Developing | | | | Late Talkers | | | |
|---|---|---|---|---|---|---|---|---|
| | Adult | | Child | | Adult | | Child | |
| | Exp. | Con. | Exp. | Con. | Exp. | Con. | Exp. | Con. |
| **Association** | 0.3811 | 0.2971 | 0.3709 | 0.2815 | 0.1114 | 0.0709 | 0.1179 | 0.0607 |
| **Distributional** | 0.3468 | 0.2858 | 0.3723 | 0.2823 | 0.1132 | 0.0467 | 0.0727 | 0.0708 |
| **Feature** | 0.4361 | 0.4114 | 0.3296 | 0.3528 | 0.0859 | 0.0518 | 0.0663 | 0.0617 |

***Note.*** Differences in red indicate networks in which the experimental networks do not outperform their control counterparts.

### 4.3.6 Post Hoc – Intervention Condition

Our final post hoc analysis evaluates the effect of which intervention condition a child was assigned to. As described earlier, some children received a short intervention in which they were exposed to words that were predicted to be more likely (high probability) or less likely (low probability) to learn based off their vocabulary structure. Though these predictions were different in nature to the method of centrality we use currently, the underlying structure of the vocabulary was very similar to our child feature structure. We believe intervention condition is an interesting aspect to consider because the intervention used vocabulary structure, and thus may have altered vocabulary structure to post test, affecting our network predictions, However, a 3 (semantic type) x 2 (age type) x 2 (experimental type) x 2 (intervention condition) ANCOVA revealed no main effect of intervention condition ($F(1,44) = 2.283$, $p = .1380$) or any significant interactions containing intervention condition and experimental type. Condition trends cans be seen in Figure 4.21. Means for each semantic and age type within each intervention condition can still be found in Table 4.9, though we did not explore these differences statistically.

**Figure 4.21 Intervention: Intervention Condition by Experimental Type**




***Note.*** The intervention condition by experimental type interaction is not significant.

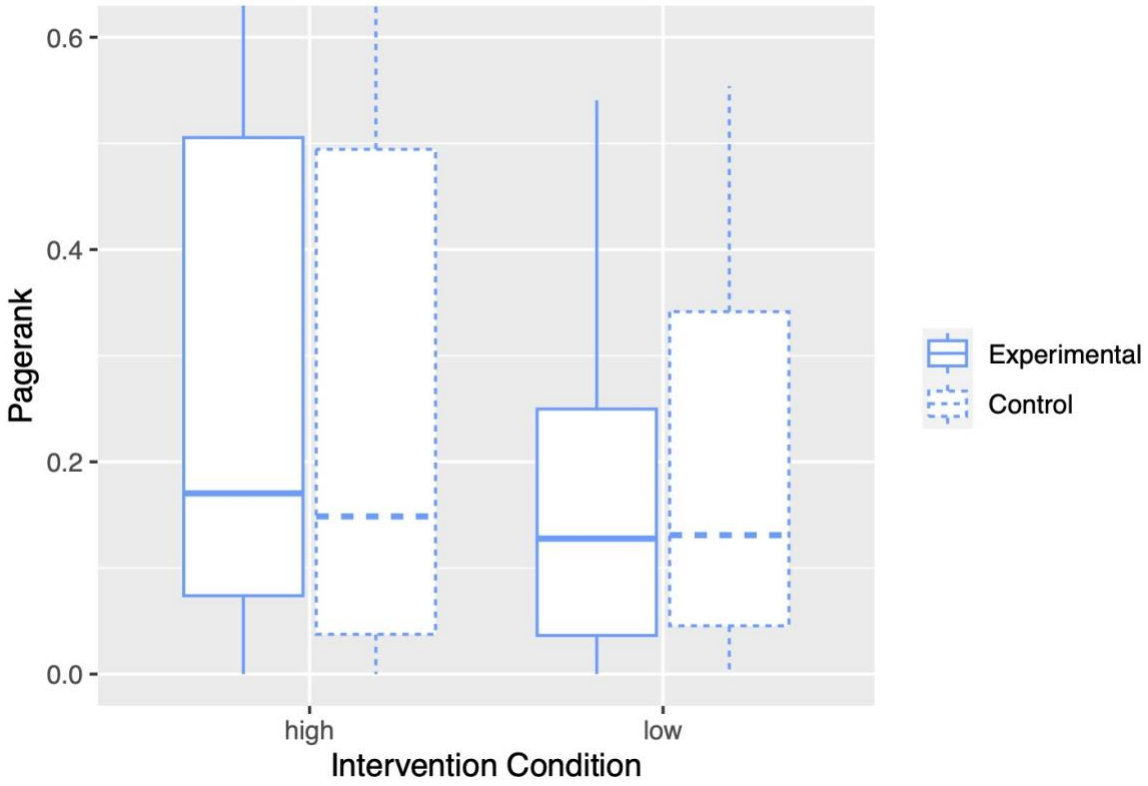
**Table 4.9 Intervention: Differences in Condition Between the Semantic and Age Types**

| | **High** | | | | **Low** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Adult** | | **Child** | | **Adult** | | **Child** | |
| | **Exp.** | **Con.** | **Exp.** | **Con.** | **Exp.** | **Con.** | **Exp.** | **Con.** |
| **Association** | 0.3181 | 0.2548 | 0.3131 | 0.2283 | 0.2212 | 0.1525 | 0.2112 | 0.1450 |
| **Distributional** | 0.2865 | 0.2210 | 0.2972 | 0.2409 | 0.2066 | 0.1455 | 0.1900 | 0.1417 |
| **Feature** | 0.3263 | 0.3042 | 0.2784 | 0.2729 | 0.2294 | 0.1935 | 0.1580 | 0.1887 |

We might specifically expect the child feature networks to perform well in the high condition and poorly in the low condition, given a similar structure was used to provide target words for children, most of which did learn said target words during the intervention. In other words, we might expect our child feature network to predict many of the same words as in the

high condition of the original study, helping produce higher accuracies since children learned many of those words. However, within our child feature networks, we conducted a 2 (experimental type) x 2 (intervention condition) ANOVA and found no main effects or interaction. Across the high and low conditions, there was no difference between the experimental and control networks ($F(1,44) = 1.307$, $p = .2592$), and the differences between experimental types did not differ based on intervention condition ($F(1,44) = 2.722$, $p = .1061$). This lack of difference in performance between the intervention conditions for the child feature networks is illustrated in Figure 4.22. Means can be found in Table 4.9.

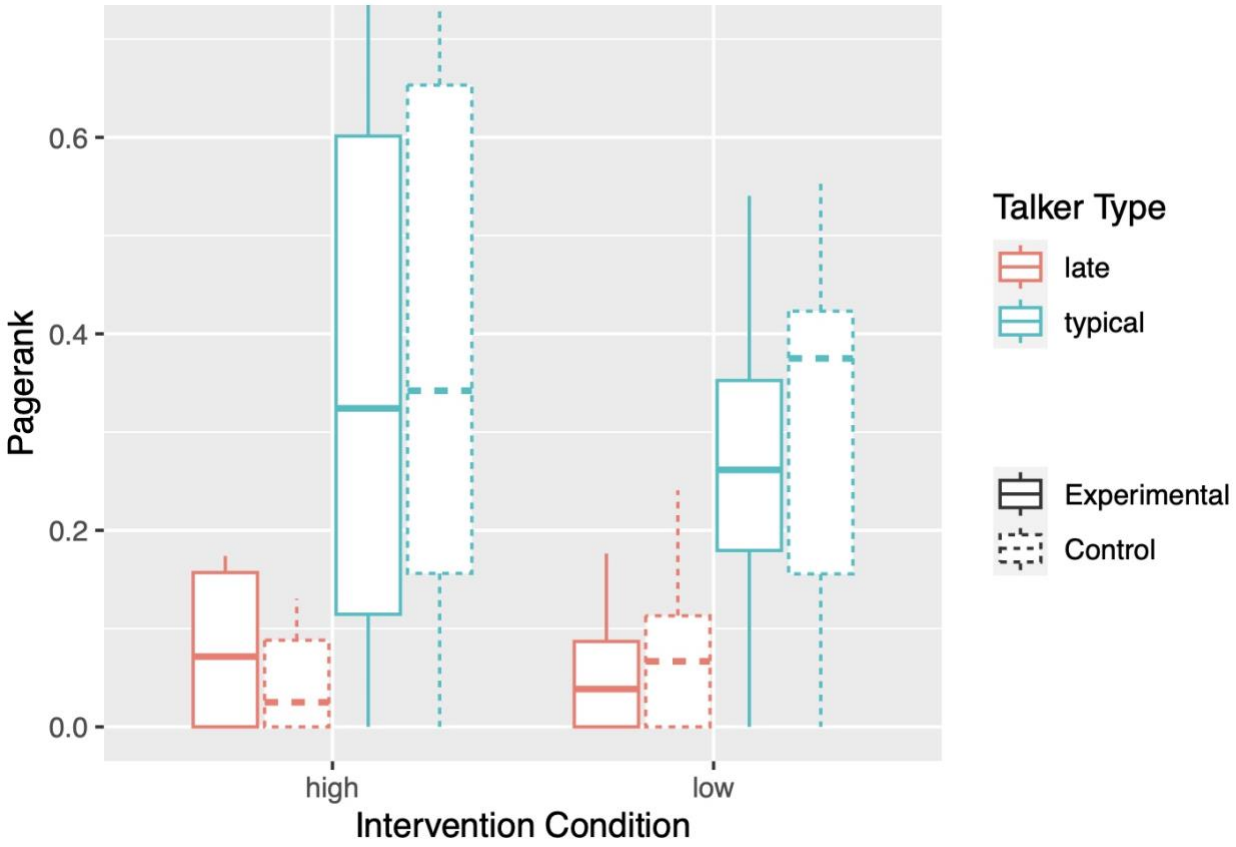
**Figure 4.22 Intervention: Condition by Experimental Type Within Child Feature Networks**




***Note.*** The intervention condition by experimental type interaction within the child feature networks is not significant.

Within the original study, there was a finding that the effects of the conditions differed based on language proficiency, so we lastly investigated this possibility within our predictions. We solely investigated this possibility within the child feature networks. We conducted a 2 (intervention condition) x 2 (language proficiency) x 2 (experimental type) ANCOVA. However, we do not find support that there is an interplay between intervention condition and language proficiency in performance, $F(1,42) = 0.048$, $p = .8273$. The pattern of results is depicted in Figure 4.23, with means in Table 4.10.

**Figure 4.23 Intervention Child Feature: Intervention Condition and Language Proficiency**




***Note.*** The intervention condition by language proficiency interaction is not significant within the child feature networks.

**Table 4.10 Intervention Child Feature: Intervention Condition and Language Proficiency**

| | High | | Low | |
|---|---|---|---|---|
| | **Experimental** | **Control** | **Experimental** | **Control** |
| **Typically Developing** | 0.3778 | 0.3856 | 0.2595 | 0.3052 |
| **Late Talkers** | 0.0797 | 0.0474 | 0.0566 | 0.0721 |

***Note.*** Differences in red indicate networks in which the experimental networks do not outperform their control counterparts.

## 4.4 Discussion

We wished to investigate how the sources of data used to create semantic network representations of young children's vocabulary knowledge impact the networks' abilities to predict future specific word learning on an individual basis. We investigated two research questions related to the sources of data, 1) do different semantic representations better or worse represent young children's knowledge?, and 2) do representations sourced specifically to capture child knowledge indeed better represent children's knowledge than those sourced from adults? As we will soon discuss, we found interesting predictive differences among our network representations of interest, as well as finding general support for using our network method to predict individual vocabulary learning. We believe these questions and their answers help us both better understand the underlying vocabulary representations of young children just beginning to speak, and better approximate this knowledge for future research.

### 4.4.1 Experimental versus Control

First and foremost, we found that overall, our experimental networks had significantly better word-by-word predictive accuracy than their control counterparts. This means that overall, the networks are capturing the semantic structure needed to predict future vocabulary learning better than a network of the same size and density but with randomly distributed edges (e.g., the

same network but lacking semantic structure). There is one caveat to this finding to keep in mind when interpreting many of the following results; the child feature networks, created using the Howell feature norms, did not show any difference in accuracies between the experimental and control networks, across both datasets. We will discuss the possible reasons for this in the limitations section. Aside from the child feature networks, this finding suggests the utility for using this network prediction method, both in future modeling work as well as in behavioral studies with children.

### 4.4.2 Semantic Type

Next, we delve into our first research question by discussing the differences we found among of three semantic representations, word association-based, distributional, and feature-based. Summarizing across the many analyses conducted across different datasets and subgroups of children, we find modest support for our hypothesis that association-based metrics best represent and predict very young children's vocabulary acquisition. Across the two age types, the association networks performed better than both other representations in the observational dataset, and better than the feature networks in the smaller intervention dataset. Though we find less support that the adult association network is best compared to the other adult-based networks, we do see that for the majority of analyses the child association network was the better than the other child semantic types. This finding has implications for our understanding of the underlying representations children at this age have and use to learn new vocabulary. If we remember, the target association word given for the cue word can be anything that first comes to mind. This can be synonyms, antonyms, features, noun-verb relationships, or other relational associations. These associative pairs can cross word types (e.g., noun-noun vs. noun-adjective), and thus have an enormous range of possible connections. Further, association norms are

typically thought of as reflecting everyday experience, and capture not just associative knowledge but meaning as well (Nelson et al., 2000; Lucas, 2000). Our findings reflect decades of using association norms as the gold standard for representing adult semantic knowledge (e.g., Nelson et al., 2009; Rapp, 2014), likely extending to children's budding internal knowledge structures as well.

The finding that association norms best represent children's vocabulary knowledge makes sense on an intuitive level as well. From the very beginning, children are making associations. Learning to segment speech, the first step in language learning and vocabulary acquisition and an act of statistical learning, is in its nature partly an associative process as well (Barakat et al., 2012). Infants are capturing these statistical patterns in the input, likely from day one (Hay et al., 2011; Saffran et al., 1996). Further, as children age, associative processes have been shown to help explain word learning (e.g., Golinkoff & Hirsh-Pasek, 2006; Colunga & Smith, 2005), social learning (e.g., Lind et al., 2019), causal reasoning (e.g., Sobel & Kirkham, 2006), and more (e.g., Crivello et al., 2019; Colunga & Smith, 2003), and they can explain such learning better than other models, for example Bayesian models of learning (Benton et al., 2023). Because associative learning is involved in so many aspects of young children's learning, it makes sense that associations underlie their developing mental lexicon as well.

What of the other two semantic types? In the overall analyses collapsing across age type as well as in the multiple analyses within the child type networks, distributional was on par or worse than association, but better than feature. However, we know that the child feature networks are likely problematic, possibly for a more computational reason than a representational one, and is diminishing the predictive accuracy of the adult feature networks when averaged together. Across our different analyses within the adult type networks, the feature

networks are never statistically different than the other two, and its performance is in between that of the association and distributional types. Can we make anything of this then? Could we assume a different way of representing child feature networks would also perform in between the other two child semantic types? Future research will have to help disentangle these two semantic representations further. For example, though some past research suggests distributional metrics such as the one we used (word2vec) is on par with adult word associations (e.g., Hofmann et al., 2018), other suggest they might be less representative of internal lexical structures (Nematzadeh et al., 2017). Both of these findings can be corroborated by our current results. Though past developmental work with networks have still found support for feature norms (e.g., Beckage et al., 2020; Beckage et al., 2019), other work suggests they may still be worse for representing young children than either association (e.g., Hills et al., 2008) or distributional metrics (e.g., Beckage et al., 2020). Further, this work does not speak to the likelihood that each of these types of semantics still play at least some role in children's internal vocabulary representations. We took the first step in using pure representations of each semantic type and comparing them, but future work must continue to investigate the likely interplay between these and other representations, such as phonological similarity, grammaticality, and so on.

The differences between our semantic types also held within our post hoc analyses. Though there were differences among the four age groups in these patterns, there was no clear trend as children aged. For example, in all but the third oldest age group (older children), there were no differences among the three adult versions of each semantic type. Similarly, though in the first and third age groups (youngest and older children), association outperformed distributional, in the second and fourth groups (younger and oldest), they performed on par with each other, again showing no clear trends. In our language proficiency analyses, we also found

similar trends between the semantic types as with the overall analyses. Of interest perhaps is that we did find less significant differences between our semantic types in the late talking group, suggesting that late talkers may not have as clear of a vocabulary structure that they can utilize in learning new words. This does corroborate past findings investigating late talkers using network representations. Late talkers have different vocabulary structures in general (e.g., Perry et al., 2023; Jiménez et al., 2020; MacRoy-Higgins et al., 2016), and these differences lead to less well-connected lexical networks with different properties (Beckage et al., 2011; Kim & Yim, 2018). For example, late talkers do not show the same extent of small-world structure as their typically developing peers (Beckage et al., 2011). This lack of a clear structure, coupled with the heterogeneity that is known to be found among late talkers (e.g., Perry & Kucker, 2019; Desmarais et al., 2008), likely contributes to the lack of differences we found among the semantic types in predicting individual word learning. Perhaps this group would benefit from a hybrid structure with contributions from multiple semantic types, or perhaps an even deeper analysis would find different profiles within the late talking group reflecting different semantic structures. Here, we do not have enough data to begin to analyze such a possibility.

### 4.4.3 Age Type

Our second research question asked if networks created from metrics specifically sourced to represent children outperform typical adult metrics. We also wanted to corroborate, with our newer experimental-control comparison analysis method, our previous findings that child distributional type networks outperform adult distributional networks. We find some support for the first part of our question, but we do not corroborate our past findings. Specifically, we found that in both datasets, across semantic types, there was no difference in the two age types in their

performance. However, we again believe this overall result might be occluded by the poor performance of our child feature networks.

In the larger observational dataset, we found that the child association networks did outperform the adult association ones. There was no difference between the two age types within distributional networks though. This does not corroborate our past findings in the previous two chapters. We could argue our newer experimental-control comparison method is more rigorous than our past method of constraining our two networks to have the same nodes and number of edges. This also introduces a new factor into our statistical models, making them more complex. but whether one agrees with that or not, this new method does add another variable to our statistical model and requires more power. Finally, the adult feature networks performed better than the child feature networks. Is this last finding actually because feature norms sourced from children capture word-by-word learning to a lesser degree than adult norms, or is our current child feature metric problematic in some way? Excluding the feature networks from consideration, we then have modest support that child-based networks better represent the vocabulary structure of children than adult-based ones.

There is another reason we may not find strong support for our second research question. Though we claim our child-based metrics are just that, child-based, none of them are actually sourced from children themselves. Arguably, our distributional networks are the closest to being truly child-based, as they only contain naturalistic conversation from mothers and fathers directed toward children under five years, as well as g-rated movies and children's storybooks. However, this is actually a representation of a child's input, not of the children themselves. We cannot collect distributional metrics over the children's speech themselves. At these ages many

of the children's speech is telegraphic or fragmented enough that there is simply not enough data or full sentences to gather the necessary statistics from.

Similarly, our child feature networks were created based on having *adults* rate sensorimotor features. Arguably, these sensorimotor features are the main features children have access to and use during language learning (Howell et al., 2005), over other features that may be present in the adult norms (encyclopedic, taxonomic). Further, these norms have been successfully used in prior developmental modeling work (Beckage & Colunga, 2019; Stella et al., 2017; Beckage et al., 2015). However, these norms were still created by using adult ratings, and it may also be the case that their norms are excluding other types of features aside from sensorimotor ones that children do actually have and use for new word learning.

Finally, though we did find that the child association networks outperformed the adult ones, the child norms we used are still problematic in the same way as the other metrics. In this case, the associations were still created by *adults*, but they did so with child language learning in mind. However, these child association norms were the only metric used that was created specifically with the CDI and child vocabulary modeling in mind. This could be the ultimate reason it performed the best at predicting children's word learning as measured thorough the CDI. Though we opted to use the adult association norms most used in previous literature (i.e., Nelson et al., 2004), the same study that collected the child association norms collected their own set of adult norms with the same words as the child ones (Cox & Haebig, 2022). The authors found that the child version of their norms were more predictive of early lexical growth using other methods for capturing said growth, but this would be a valuable comparison to model in the future using our own methods.

Despite a lack of strong evidence from the current study indicating that network representations sourced from child-based semantics are better than adult-based ones, we still believe researchers should continue to strive to create and use metrics as representative of their population as possible. In our line of research, that means possibly querying children themselves, despite the difficulties that come with such an endeavor.

### 4.4.4 Other Interesting Findings

There were a few other interesting findings not related to our hypotheses. First, though we did not report on our control variables, they were normally significantly predictive of accuracy within our statistical models. Our first control variable, CDI percentile, reflect our language proficiency findings, in that as children's CDI percentile increases, so do all network accuracies, regardless of if they are experimental or control. Similarly, as age increases so do accuracies. This finding is reflected in our categorical age analyses. Also related, as children's vocabularies increase, so do accuracies. Finally, as children learn more words in a month, prediction accuracies also increase. These four control variables are all partially confounded, but all also get at separate aspects of word learning. For example, children of the same age can have wildly varying vocabulary sizes, proficiencies (CDI percentiles), and numbers of words they learned. And children of the same percentile will still have varying ages, vocabulary sizes, and numbers of learned words. Children of the same vocabulary size may all learn a different number of words in that month. However, it is also true that children who are older tend to have larger vocabularies, and children with larger vocabularies tend to have higher CDI percentiles. Further, children of higher percentiles tend to learn more words. Interestingly, children with larger vocabularies do not necessarily learn more words in a month. This was after we removed children who learned less than 1% of the words they had left to learn or more than 99% of those

words left to learn (to remove floor and ceiling effects). This means that though percentile score can be indicative of word learning ability, raw vocabulary size is not. This corroborates work suggesting that clinicians and researchers must look at more than just the number of words in a child's vocabulary, but rather vocabulary structure, to get a sense of a child's abilities (Perry et al., 2023; Weber & Colunga, 2021). This is especially important to consider for diagnosis of language impairment (Czaplewska, 2016).

Though it is true that overall these factors were typically significant, there were interesting instances where these factors were not predictive. Within our late talking samples, these variables, particularly age and vocabulary size, were less likely to play a role in network accuracies. Though not the same as percentile, the lack of significance for vocabulary size is likely due to late talkers having a more homogeneous number of words in their vocabularies as compared to typically developing children. This lack of relationship between vocabulary size and accuracy was also found in the youngest age group's model. The results of the youngest age group resemble the late talker results in multiple ways, and both groups do have smaller vocabularies in general. However, the finding that age did not predict accuracies in late talkers is likely not explained by homogeneity. Late talkers had just as much variability in their ages as the overall sample, so something else must be driving this lack of relationship.

Finally, the lack of condition differences in the intervention group is unexpected, even when we analyzed accuracies in much the same way learning was analyzed in the original study; that is, looking for an intervention condition by language proficiency interaction. Even when we looked within the child feature networks, which were similar to the ones used to make predictions in the original study, we found no effects. However, there could be a few reasons for this finding. First, this dataset is much smaller than the observational one, and we may have

simply been underpowered to find such a difference. Second, as we found earlier, simply having an intervention at all and boosting learning could overshadow any smaller condition effects. Finally, we know there might be something problematic about the child feature networks. Investigating other findings within predictions coming from a problematic source is itself problematic, and could be the reason nothing was found in those analyses.

### 4.4.5 Limitations

There are a few limitations to keep in mind when analyzing our results. First and foremost are any comparisons to our child feature network, which did not show experimental performance above its controls. Though it is possible a sensorimotor feature representation is simply very bad at predicting word-by-word learning, it is more likely that something else about the design of this representation is driving poor accuracy. As we discussed in Section 4.2.2.6, rather than free responses like in the adult feature norms, participants were forced to rate a noun on every possible feature using a scale of one to ten. This could very much alter the behavior of participants and their resultant feature ratings. For example, even though a participant may never freely respond that a snake *climbs*, they may still rate *climbs* slightly higher than *runs* given that *runs* is even less a feature of a snake. Thus, a cat and a snake could end up being more related than if using free responses, because both climb. Besides possibly inflating some features and deflating others, this also creates a fully-connected network. Though that in itself is not problematic, it is highly juxtaposed to the adult feature network. We did attempt to use cutoffs to create a sparser network with the same density as the adult feature network, which did slightly improve results (the experimental networks went from being significantly worse than controls to not significantly different). However, by doing this not only did we alter the child feature network from its "natural" state, which we wanted to avoid, but when we thresholded it the

smallest edge weight became 0.72. With edge weights being on a scale from zero to one, this basically turned our network into a binary one. The structure of all our networks and their connections are visualized in Appendix B; the two child feature networks are in Figure B.6 In other words, we worry that how these norms were collected is the cause for its poor results, and we do not want to form conclusions about child-based, feature-based semantics from these norms alone.

Another limitation is one that plagues all research, its generalizability. We did extend our findings from one developmental dataset to another, with overall trends staying the same between the two. However, there are not many available longitudinal samples we can further compare to. Further, other aspects of our approach may not be as generalizable. We used only one source of data for each of our types of networks. We did not investigate multiple adult association or child distributional networks to check if initial differences were robust. There are some other datasets available that could also be tested to see if the same trends hold with a different network but of the same semantic and age type (Cox & Haebig, 2022; Davis, 2010; Vinson & Vigliocco, 2002). However, there simply aren't many other datasets available, particularly for our child type designation. This greatly limits our ability to generalize our findings. Finally, we used a very specific method for testing representativeness, and built our models on the assumption that a network that better predict future specific word learning is more representative of a child's internal vocabulary structure. Because we cannot access our children's exact mental lexicons, we have to use an assumption such as this. As seen in previous literature though, there are other prediction methods we could use and compare to in order to get a sense of the robustness of our results. Further, these comparisons could help further disentangle underlying language learning mechanisms. For example, past literature has used growth

algorithms such as preferential attachment, where more well-connected words already in the lexicon preferentially acquire new words (Barabási & Albert, 1999; Steyvers & Tenenbaum, 2005), preferential acquisition, where words with more contextual diversity in the environment are acquired preferentially (Hills et al., 2010; Siew & Vitevitch, 2020; Hills et al., 2009), and lure of associates, where words that are more related to already known words are acquired preferentially (Hills et al., 2010; Siew & Vitevitch, 2020). Though our own prediction method is based on the same principles (particularly that of lure of associates), in that the connectedness of words should predict their entry into the network (learning by children), all these prediction methods are executed in different ways, and may thus lead to different conclusions.

### 4.4.6 Future Directions

As noted in the limitations, further modeling work could continue to investigate children's knowledge representations by extending the current work to new longitudinal datasets, new sources of knowledge representation, and new prediction methods. Though we believe our study meets and expands upon Christensen and Chater's (2001) criterion of input representativeness, as well as Sims and colleagues (2013) added criterion of temporality, testing new sources could help continue the search for the best representation, and extending to other longitudinal datasets and other growth algorithms could further solidify that this type of work meets the temporality criterion.

In the midst of this modeling work is the eventual need to test these representations in behavioral studies. Similar to the motivations behind our predictions of already collected data, if we predict target words to be learned using two different representations, and children learn more of the targeted words predicted by one representation than the other, this could lend evidence to that representation indeed being the underlying lexical representation children have

and are using to learn new words. Though we believe these studies begin to get at another one of Christiansen and Chater's (2001) criteria, that of data contact, in that we use actual vocabulary data from individual children, studying this phenomena in a learning study rather than in previously collected data would better meet this criterion. Similarly, the differences we see in our models, though significant, are often quite small, and do not necessarily translate to meaningful differences. Behavioral studies are needed to better understand how these different representations operate during ongoing development.

### 4.4.7 Conclusions

In short, the present chapter extends the previous work found in chapters two and three to new semantic representations. It also further emphasizes the need to tailor the underlying representations used in psychological work by using data sources designed for the population of interest. We found modest support that children's vocabularies best fit an association-based structure. We also still believe that finding the correct child-based representations is of more benefit than using adult-based ones. Though we find that these network models are overall worse at accurately predicting word-by-word learning of children with low language proficiency and younger children, they are still capturing these groups' semantic structure better than our control. This suggests we can continue to refine these representations to better fit these and other new populations. Both by continuing to refine the current models and by testing competing representations in situ, we can better understand children's underlying knowledge representations and learning mechanisms.

# Chapter 5

# Conclusions

The current body of work used network science methods to evaluate the underlying representations of young children's lexicons. We did so by predicting the specific word-by-word learning of toddlers using words' centrality in different network representations, to understand which representations best capture the knowledge children have and use to learn new vocabulary. Chapter 2 first showed the utility of using distributional methods over a child-directed language corpus to represent semantic similarity for the average young child. Chapter 3 extended those results using a larger longitudinal dataset of individual children's vocabulary learning, showing these distributional semantics based on child-directed language better predict future specific word learning than adult-based semantics. Chapter 3 further showed the advantages of individualizing predictions based on each child's particular vocabulary structure. Finally, Chapter 4 extended the previous results by investigating the representativeness of different methods for capturing the semantic similarity between words aside from the previously used distributional one, namely semantics based on association norms and feature norms. The final study made two contributions. First, the results suggest that young children represent their knowledge of concepts through a series of associations, rather than as sets of shared features or in a distributed, vectorized manner. Second, modest evidence supports the findings of Chapters 2 and 3, and further implies that sourcing computational representations directly from children would be beneficial, and an ideal future avenue in order to build the most representative models of language development.

## 5.1 Christiansen and Chater's Criteria

We turn to an overarching discussion as to whether the current models meet criteria proposed for evaluating such models. In particular, we will evaluate our models in light of Christiansen & Chater's (2001) three criteria, as well as Sims and colleagues (2013) additional one.

The first criterion is data contact, or the degree to which the model captures human psycholinguistic data. We use a normed parent checklist of common vocabulary words that approximates overall vocabulary. Though we admit this is not perfect data contact, in that 1) parents are indicating the words their children produce, and 2) it does not capture the entirety of their possible vocabulary, the CDI checklist has been shown to be reliable, valid, and related to other child language measures (Fenson et al., 1994). Additionally, from Chapter 3 on we are using CDI data from individual children, rather than data norms which arguably have less contact with true child data. Furthermore, capturing the entirety of every individual child's in situ word-learning is close to impossible, or at least time consuming, labor intensive, and invasive; the very nature of trying to collect such data could alter children's word-learning trajectories (e.g., Cook, 1967; Bağçeli Kahraman & Başal, 2015). Thus, we still think the CDI is one of the best measures we can use due to both its ease of collection and its history in the developmental literature (e.g., MacRoy-Higgins et al., 2016; Moyle et al., 2007; Gershkoff-Stowe & Smith, 2004), and we believe we are successfully meeting the criterion of data contact. However, future work could do more in attempts to increase data contact. For example, having parents fill out daily word diaries (in which they write down the new words they have heard their child produce) would ameliorate our second point above, in that parents could disclose their child's entire vocabulary rather than a sampling (Place & Hoff, 2016; Rescorla, 1980).

Next is the criterion of task veridicality, which is the match between the task given to the model and the task given to humans. We attempt to approximate the act of a child learning new words by having our models score and rank unknown words using network centrality. We then take the top scoring words (based on the number the corresponding child learned) and consider those as "learned" by the model. However, our models are not actually learning or building knowledge, and every word is scored independently, based only on where it alone would be placed in the network. Though we could argue the dynamic method of prediction we use better captures the criterion as compared to the static method presented in Chapter 3 we know word learning is a much more dynamic task than what we give it credit for with our methods here. Though at its very core we could argue both children and our model are "tasked" with adding new words to their vocabulary, we know our method is still a far cry from meeting the criterion of task veridicality. Future work can continue to build models in specific light of this criterion, focusing on capturing the *way* in which children learn words (Siew & Vitevitch, 2020; Beckage & Colunga, 2019; Colunga & Sims, 2017; Hills et al., 2009), just as we focused on the underlying representations and the criterion of input representativeness here.

This brings us to the third criterion of input representativeness, which was a main motivation throughout our analyses. The original authors define input representativeness as the degree to which the model captures human psycholinguistic data. If we consider the actual learning of children to be our psycholinguistic data, then the accuracy our models reach is our proxy for how well our models represent the typical toddler. Overall, our models may seem as though they perform poorly, only accurately predicting specific word learning on average 26 percent of the time, though they range in accuracy from zero to one hundred percent. However, we must remember that the endeavor to predict the next words a child will learn is no easy feat.

Word learning is influenced by many individual factors. The largest and most heterogeneous of these factors is a child's moment-by-moment experiences. Even if our models included more environmental factors such as socioeconomic status, cognitive environment (e.g., Mahli et al., 2018), and even word-learning biases (e.g., Colunga & Smith, 2008; Landau et al., 1988), we cannot account for idiosyncratic experiences such as one child going to the zoo one month or another accompanying their mother to the grocery store more often. Therefore, these models are not expected to reach high levels of accuracy on already-collected learning data, but rather these are the preliminary steps needed to produce the best models for use in behavioral intervention studies, where vocabularies are still malleable.

In that case, we are not interested in comparing our models to an unrealistic standard of 100 percent accuracy, but rather in comparing different representations to each other as we did throughout this thesis. Though we believe the question of finding the best representation of children's world knowledge is not fully answered, we have come many steps closer to determining what children's underlying representations consist of, and what information and methods we need to represent that structure computationally. However, in the process of better meeting the criteria of input representativeness, one contribution of this work has been to unpack not only where best to source information from (i.e., young children), but what methods of capturing semantic similarity are more accurate (i.e., associations).

Finally, a fourth criterion was introduced subsequently by Sims et al. (2013), named temporality. This is the degree to which the model captures continuous change including the processes that drive that change. We believe our studies achieve this to some degree. We use longitudinal datasets of individual children's growth, and further conduct post hoc analyses looking particularly at how model accuracy changes with age. However, though in our statistical

models we nest multiple visits within each child, our predictions themselves treat each pair of contiguous timepoints individually, and do not consider if the input vocabulary structure is from the same child. Therefore, though our models are capturing continuous change better than other past research, we are still not fully accounting for the criterion of temporality.

### 5.2 Limitations and Future Work

There are a few limitations of this overarching work we wish to touch on here. First, though not a limitation in itself, as a design choice we only investigated semantic representations. There are, of course, other representations that also partially account for language learning, such as phonological or grammatical representations (e.g., Storkel & Lee, 2011; Monaghan & Mattock, 2012). Additionally, though we tested each representation individually, the underlying knowledge structures of children may contain an amalgamation of multiple different representations, both semantic in nature and otherwise. Even the individual representations we tested presently may contain aspects of these other types; for example, distributional networks capture syntactic and morphological information as well as semantic (Limisiewicz & Mareček, 2020; Cotterell & Schütze, 2019). Future work can test different combinations, perhaps using multiplex or stacked approaches (Stella et al., 2017; Cao et al., 2017; Yang et al., 2016). As we alluded to at the beginning of this chapter, new sources of semantic similarity could be created. Specifically, our sources of child semantics did not come from children themselves. Though a painstaking endeavor, perhaps the field should attempt to gather similarity metrics through experimental paradigms with children. This may need to be the next steps if we want to continue building more and more accurate representations of children's knowledge.

Second, we use a specific prediction method for evaluating our representations of vocabulary structure. Again, we fully acknowledge that there are, 1) other ways to evaluate network representations, and 2) other prediction methods or algorithms than the one we use in the current thesis. For example, there are other methods for analyzing different representations besides having networks predict new vocabulary words, such as correlational methods with general word acquisition norms, or deeper investigations into specific differences between network structures. On this second note, we argue that the current statistical evaluation we used comparing experimental to control networks helps ameliorate the confounds of comparing different network structures that some past work has run into. There are also other growth algorithms that we have briefly described in previous sections (see Section 4.4.5). We believe it would be fruitful to use both the current representations and current experimental-control comparison technique but with other growth algorithms to better understand the underlying processes driving vocabulary acquisition in young children.

## 5.3 Conclusion

The present body of work focused on creating the most accurate representations of young toddlers' lexicons. We evaluated multiple different representations, investigating whether semantic similarity was obtained from child or adult sources as well as investigating different methods for representing the similarity between words. Our work helped illuminate the underlying structure of young children's vocabulary knowledge. Though more work is needed to continue understanding the development of the lexicon, we have laid the foundation here.

# References

Amatuni, A., & Bergelson, E. (2017). Semantic networks generated from early linguistic input. *bioRxiv*, 157701.

Asr, F. T., Willits, J., & Jones, M. N. (2016, August). Comparing Predictive and Co-occurrence Based Models of Lexical Semantics Trained on Child-directed Speech. In *CogSci*.

Bağçeli Kahraman, P. & Başal, Handan. (2015). The effects of preschool education program with enriched parental involvement dimension on school readiness of 5-6 year old children. *Akademik Sosyal Araştırmalar Dergisi*. *3*. 1-10.

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

Barakat, B., Seitz, A., & Shams, L. (2012). There is more to statistical learning than associative learning: Predictable items are enhanced even when not predicted. *Journal of Vision*, *12*(9), 694-694.

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,* 1-17.

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. (2010). Semantic network connectivity is related to vocabulary growth rate in children. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 32, No. 32).

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

Benton, D. T., Kamper, D., Beaton, R. M., & Sobel, D. M. (2023). Don't throw the associative baby out with the Bayesian bathwater: Children are more associative when reasoning retrospectively under information processing demands. *Developmental Science*, e13464.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., & Hwang, D. U. (2006). Complex networks: Structure and dynamics. *Physics reports*, *424*(4-5), 175-308.

Borovsky, A. (2022). Lexico-semantic structure in vocabulary and its links to lexical processing in toddlerhood and language outcomes at age three. *Developmental psychology*, *58*(4), 607.

Borovsky, A., Peters, R. E., Cox, J. I., & McRae, K. (2023). Feats: A database of semantic features for early produced noun concepts. *Behavior Research Methods*, 1-21.

Bilson, S., Yoshida, H., Tran, C. D., Woods, E. A., & Hills, T. T. (2015). Semantic facilitation in bilingual first language acquisition. *Cognition*, *140*, 122-134.

Cao, W., Song, A., & Hu, J. (2017). Stacked residual recurrent neural network with word weight for text classification. *IAENG International Journal of Computer Science, 44*(3), 277-284.

Capocci, A., Servedio, V. D., Colaiori, F., Buriol, L. S., Donato, D., Leonardi, S., & Caldarelli, G. (2006). Preferential attachment in the growth of social networks: The internet encyclopedia Wikipedia. *Physical review E, 74*(3), 036116.

Castro, N. (2022). Methodological considerations for incorporating clinical data into a network model of retrieval failures. *Topics in Cognitive Science*, *14*(1), 111-126.

Chan, K. Y., & Vitevitch, M. S. (2010). Network structure influences speech production. *Cognitive science*, *34*(4), 685-697.

Christiansen, M. H., & Chater, N. (2001). Connectionist psycholinguistics: Capturing the empirical data. *Trends in Cognitive Sciences*, *5*(2), 82-88.

Church, K. W. (2017). Word2Vec. *Natural Language Engineering*, *23*(1), 155-162.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of verbal learning and verbal behavior*, *8*(2), 240-247.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological review*, *82*(6), 407.

Colunga, E. (2019, March). A Model-based Individualized Vocabulary Intervention in Typically-developing children and Late Talkers. Talk presented in the Symposium on *The Organization of Knowledge: Development and Modeling of Early Semantic Networks*, at the *2019 Biennial Meeting of the Society for Research on Child Development.* Baltimore, MD.

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.

Colunga, E., & Smith, L. B. (2003). The emergence of abstract ideas: Evidence from networks and babies. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 358*(1435), 1205-1214.

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

Colunga, E., & Smith, L. B. (2008). Knowledge embedded in process: The self-organization of skilled noun learning. *Developmental Science, 11*(2), 195-203.

Connor, N., Barberán, A., & Clauset, A. (2017). Using null models to infer microbial co-occurrence networks. *PloS one, 12*(5), e0176751.

Cook, D. L. (1967). The Impact of the Hawthorne Effect in Experimental Designs in Educational Research. Final Report.

Cotterell, R., & Schütze, H. (2019). Morphological word embeddings. *arXiv preprint arXiv:1907.02423*.

Cox, C. R., & Haebig, E. (2023). Child-oriented word associations improve models of early word learning. *Behavior research methods, 55*(1), 16-37.

Crivello, C., Grossman, S., & Poulin-Dubois, D. (2021). Specifying links between infants' theory of mind, associative learning, and selective trust. *Infancy, 26*(5), 664-685.

Czaplewska, E. (2016). Children with Language Disorders or Late Bloomers–the problem of differential diagnosis. *Polish Psychological Bulletin*, (3).

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and linguistic computing*, *25*(4), 447-464.

De Deyne, S., & Storms, G. (2008). Word associations: Norms for 1,424 Dutch words in a continuous task. *Behavior research methods*, *40*(1), 198-205.

Desmarais, C., Sylvestre, A., Meyer, F., Bairati, I., & Rouleau, N. (2008). Systematic review of the literature on characteristics of late-talking toddlers. *International journal of language & communication disorders*, *43*(4), 361-389.

Fasolo, M.., & D'Odorico, L. (2002). Vocabulary development of Late-Talking children: a longitudinal research from eighteen to thirty months of age. *Rivista di Psicolinguistica Applicata*, *3*, 13-21.

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*, *54*(7), 1334.

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.

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.

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

Golinkoff, R. M., & Hirsh-Pasek, K. (2006). Baby wordsmith: From associationist to social sophisticate. *Current directions in psychological science*, *15*(1), 30-33.

Handler, A. (2014). An empirical study of semantic similarity in WordNet and Word2Vec.

Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., & Aluisio, S. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *arXiv preprint arXiv:1708.06025*.

Hay, J. F., Pelucchi, B., Estes, K. G., & Saffran, J. R. (2011). Linking sounds to meanings: Infant statistical learning in a natural language. *Cognitive psychology*, *63*(2), 93-106.

Heilmann, J., Weismer, S. E., Evans, J., & Hollar, C. (2005). Utility of the MacArthur—Bates Communicative Development Inventory in identifying language abilities of late-talking and typically developing toddlers.

Hills, T. (2013). The company that words keep: comparing the statistical structure of child-versus adult-directed language. *Journal of child language*, *40*(3), 586-604.

Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. B. (2008). Is There Preferential Attachment in the Growth of Early Semantic Noun Networks?. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 30, No. 30).

Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009). Categorical structure among shared features in networks of early-learned nouns. *Cognition*, *112*(3), 381-396.

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.

Hofmann, M. J., Biemann, C., Westbury, C., Murusidze, M., Conrad, M., & Jacobs, A. M. (2018). Simple co-occurrence statistics reproducibly predict association ratings. *Cognitive Science*, *42*(7), 2287-2312.

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.

Huebner, P. A., & Willits, J. A. (2018). Structured semantic knowledge can emerge automatically from predicting word sequences in child-directed speech. *Frontiers in Psychology*, *9*, 133.

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

Jiménez, E., Haebig, E., & Hills, T. T. (2021). Identifying areas of overlap and distinction in early lexical profiles of children with autism spectrum disorder, late talkers, and typical talkers. *Journal of Autism and Developmental Disorders*, *51*(9), 3109-3125.

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

Jones, G., Cabiddu, F., Barrett, D. J., Castro, A., & Lee, B. (2023). How the characteristics of words in child-directed speech differ from adult-directed speech to influence children's productive vocabularies. *First Language*, *43*(3), 253-282.

Kajic, I., & Eliasmith, C. (2018). *Evaluating the psychological plausibility of word2vec and GloVe distributional semantic models*. Technical Report CTN-TR-20180824-012. Centre for Theoretical Neuroscience, University of Waterloo. Waterloo, ON, Canada. doi: 10.13140/RG. 2.2. 25289.60004.

Kalyan, K. S. (2023). A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Natural Language Processing Journal*, 100048.

Kaur, M., & Singh, S. (2016). Analyzing negative ties in social networks: A survey. *Egyptian Informatics Journal*, *17*(1), 21-43.

Kenett, Y. N., Levi, E., Anaki, D., & Faust, M. (2017). The semantic distance task: Quantifying semantic distance with semantic network path length. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(9), 1470–1489. https://doi.org/10.1037/xlm0000391

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

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive development*, *3*(3), 299-321.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211.

Limisiewicz, T., & Mareček, D. (2020). Syntax Representation in Word Embeddings and Neural Networks--A Survey. *arXiv preprint arXiv:2010.01063*.

Lind, J., Ghirlanda, S., & Enquist, M. (2019). Social learning through associative processes: a computational theory. *Royal Society open science*, *6*(3), 181777.

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

Lucas, M. (2000). Semantic priming without association: A meta- analytic review. *Psychonomic Bulletin & Review*, **7**, 618-630.

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.

Malhi, P., Menon, J., Bharti, B., & Sidhu, M. (2018). Cognitive development of toddlers: Does parental stimulation matter?. *The Indian Journal of Pediatrics, 85*, 498-503.

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. *Applied psycholinguistics*, *23*(1), 1-21.

Marko, M., & Riečanský, I. (2021). The structure of semantic representation shapes controlled semantic retrieval. *Memory*, *29*(4), 538-546.

McCulloh, I., & Carley, K. M. (2011). Detecting change in longitudinal social networks. *Journal of social structure*, *12*(3), 1-37.

McPherson, J. M., Popielarz, P. A., & Drobnic, S. (1992). Social networks and organizational dynamics. *American sociological review*, 153-170.

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.

McRae, K., De Sa, V. R., & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, *126*(2), 99.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Monaghan, P., & Mattock, K. (2012). Integrating constraints for learning word–referent mappings. *Cognition*, *123*(1), 133-143.

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., & Dennis, S. (2000). What is free association and what does it measure?. *Memory & cognition*, *28*(6), 887-899.

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.

Nematzadeh, A., Meylan, S. C., & Griffiths, T. L. (2017, July). Evaluating Vector-Space Models of Word Representation, or, The Unreasonable Effectiveness of Counting Words Near Other Words. In *CogSci*.

Newman, M. E. (2001). Clustering and preferential attachment in growing networks. *Physical review E*, *64*(2), 025102.

Perry, L. K., & Kucker, S. C. (2019). The heterogeneity of word learning biases in late-talking children. *Journal of Speech, Language, and Hearing Research*, *62*(3), 554-563.

Perry, L. K., Kucker, S. C., Horst, J. S., & Samuelson, L. K. (2023). Late bloomer or language disorder? Differences in toddler vocabulary composition associated with long-term language outcomes. *Developmental science*, *26*(4), e13342.

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

Peters, R., & Borovsky, A. (2019). Modeling early lexico-semantic network development: Perceptual features matter most. *Journal of Experimental Psychology: General*, *148*(4), 763.

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

Place, S., & Hoff, E. (2016). Effects and noneffects of input in bilingual environments on dual language skills in 2 ½-year-olds. *Bilingualism: Language and Cognition*, *19*(5), 1023-1041.

Rescorla, L. A. (1980). Overextension in early language development. *Journal of child language*, *7*(2), 321-335.

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

Rogers, T. T., & McClelland, J. L. (2008). Précis of semantic cognition: A parallel distributed processing approach. *Behavioral and Brain Sciences*, *31*(6), 689-714.

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 906-914.

Rotaru, A. S., Vigliocco, G., & Frank, S. L. (2018). Modeling the structure and dynamics of semantic processing. *Cognitive science*, *42*(8), 2890-2917.

Pastor-Satorras,R., Castellano, C., Van Mieghem, P., & Vespignani, A. (2015). Rev. Mod. Phys. 87, 925

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

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.

Siew, C. S. (2020). Applications of network science to education research: Quantifying knowledge and the development of expertise through network analysis. *Education Sciences*, *10*(4), 101.

Siew, C. S., & Vitevitch, M. S. (2020). An investigation of network growth principles in the phonological language network. *Journal of Experimental Psychology: General*, *149*(12), 2376.

Sims, C. E., Schilling, S. M., & Colunga, E. (2013). Beyond modeling abstractions: learning nouns over developmental time in atypical populations and individuals. *Frontiers in psychology*, *4*, 871.

Sobel, D. M., & Kirkham, N. Z. (2006). Blickets and babies: the development of causal reasoning in toddlers and infants. *Developmental psychology*, *42*(6), 1103.

Stella, M., Beckage, N. M., & Brede, M. (2017). Multiplex lexical networks reveal patterns in early word acquisition in children. *Scientific reports*, *7*(1), 46730.

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., & Lee, S. Y. (2011). The independent effects of phonotactic probability and neighbourhood density on lexical acquisition by preschool children. *Language and cognitive processes*, *26*(2), 191-211.

Tu, H., Xia, Y., Iu, H. H. C., & Chen, X. (2018). Optimal robustness in power grids from a network science perspective. *IEEE Transactions on Circuits and Systems II: Express Briefs*, *66*(1), 126-130.

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).

Vinson, D. P., & Vigliocco, G. (2008). Semantic feature production norms for a large set of objects and events. *Behavior Research Methods*, *40*(1), 183-190.

Vitevitch, M. S. (2022). What can network science tell us about phonology and language processing?. *Topics in Cognitive Science*, *14*(1), 127-142.

Wang, Z., Bauch, C. T., Bhattacharyya, S., d'Onofrio, A., Manfredi, P., Perc, M., ... & Zhao, D. (2016). Statistical physics of vaccination. *Physics Reports*, *664*, 1-113.

Weber, J., & Colunga, E. (2021). A longitudinal analysis of vocabulary struc- ture in persistent late-talkers, late-bloomers, and typically-developing toddlers. *Society for research in child development, virtual*.

Weber, J., & Colunga, E. (2022, June). Representing the Toddler Lexicon: Do the Corpus and Semantics Matter?. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 3960-3968).

Wulff, D. U., De Deyne, S., Aeschbach, S., & Mata, R. (2022). Using network science to understand the aging lexicon: Linking individuals' experience, semantic networks, and cognitive performance. *Topics in Cognitive Science*, *14*(1), 93-110.

Yang, Z., He, X., Gao, J., Deng, L., & Smola, A. (2016). Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 21-29).

# Appendix A

## Words in Each Network

Appendix A first contains the full 680-word CDI. The rest of the tables in Appendix A indicate which words were removed or changed during queries of the original datasets for use in each of the networks. In the case of the feature norm sets, tables indicate which words were included rather than excluded.

**Table A.1 The Full CDI (used as the basis for our lexicons)**

| | | | | |
|---|---|---|---|---|
| baa baa | potato | dryer | catch | old |
| choo choo | potato chip | garage | chase | orange(adjective) |
| cockadoodledoo | pretzel | high chair | clap | poor |
| grr | pudding | kitchen | clean(verb) | pretty |
| meow | pumpkin | living room | climb | quiet |
| moo | raisin | oven | close | red |
| ouch | salt | play pen | cook | sad |
| quack quack | sandwich | porch | cover | scared |
| uh oh | sauce | potty | cry | sick |
| vroom | soda/pop | refrigerator | cut | sleepy |
| woof woof | soup | rocking chair | dance | slow |
| yum yum | spaghetti | room | draw | soft |
| alligator | strawberry | shower | drink(verb) | sticky |
| animal | toast | sink | drive | stuck |
| ant | tuna | sofa | drop | thirsty |
| bear | vanilla | stairs | dry(verb) | tiny |
| bee | vitamins | stove | dump | tired |
| bird | water(food) | table | eat | wet |
| bug | yogurt | tv | fall | white |
| bunny | beads | washing machine | feed | windy |
| butterfly | belt | window | find | yellow |
| cat | bib | backyard | finish | yucky |
| chicken(animal) | boots | cloud | fit | after |
| cow | button | flag | fix | before |
| deer | coat | flower | get | day |
| dog | diaper | garden | give | later |
| donkey | dress | grass | go | morning |
| duck | gloves | hose | hate | night |
| elephant | hat | ladder | have | now |
| fish(animal) | jacket | lawn mower | hear | time |
| frog | jeans | moon | help | today |
| giraffe | mittens | pool | hide | tomorrow |
| goose | necklace | rain | hit | tonight |
| hen | pajamas | rock | hold | yesterday |
| horse | pants | roof | hug | he |
| kitty | scarf | sandbox | hurry | her |
| lamb | shirt | shovel | jump | hers |
| lion | shoe | sidewalk | kick | him |

| monkey | shorts | sky | kiss | his |
|---|---|---|---|---|
| moose | slipper | slide(noun) | knock | i |
| mouse | sneaker | snow | lick | it |
| owl | snowsuit | snowman | like | me |
| penguin | sock | sprinkler | listen | mine |
| pig | sweater | star | look | my |
| pony | tights | stick | love | myself |
| puppy | underpants | stone | make | our |
| rooster | zipper | street | open | she |
| sheep | ankle | sun | paint | that |
| squirrel | arm | swing(noun) | pick | their |
| teddybear | belly button | tree | play | them |
| tiger | buttocks/bottom | water(outside) | pour | these |
| turkey | cheek | wind | pretend | they |
| turtle | chin | beach | pull | this |
| wolf | ear | camping | push | those |
| zebra | eye | church | put | us |
| airplane | face | circus | read | we |
| bicycle | feet | country | ride | you |
| boat | finger | downtown | rip | your |
| bus | hair | farm | run | yourself |
| car | hand | gas station | say | how |
| fire truck | head | home | see | what |
| helicopter | knee | house | shake | when |
| motorcycle | leg | movie | share | where |
| sled | lips | outside | show | which |
| stroller | mouth | park | sing | who |
| tractor | nose | party | sit | why |
| train | owie/boo boo | picnic | skate | about |
| tricycle | penis | playground | sleep | above |
| truck | shoulder | school | slide(verb) | around |
| ball | tooth | store | smile | at |
| balloon | toe | woods | spill | away |
| bat | tongue | work(place) | splash | back |
| block | tummy | yard | stand | behind |
| book | vagina | zoo | stay | beside |
| bubbles | basket | aunt | stop | by |
| chalk | blanket | baby | sweep | down |
| crayon | bottle | babysitter | swim | for |
| doll | box | babysitter's name | swing(verb) | here |
| game | bowl | boy | take | inside/in |
| glue | broom | brother | talk | into |
| pen | brush | child | taste | next to |
| pencil | bucket | clown | tear | of |
| play dough | camera | cowboy | think | off |
| present | can(noun) | daddy | throw | on |
| puzzle | clock | doctor | tickle | on top of |
| story | comb | fireman | touch | out |
| toy | cup | friend | wait | over |
| apple | dish | girl | wake | there |
| applesauce | fork | grandma | walk | to |
| banana | garbage | grandpa | wash | under |
| beans | glass | lady | watch(verb) | up |
| bread | glasses | mailman | wipe | with |
| butter | hammer | man | wish | a |
| cake | jar | mommy | work(verb) | all |

| | | | | |
|---|---|---|---|---|
| candy | keys | nurse | write | a lot |
| carrots | knife | child's own name | allgone | an |
| cereal | lamp | people | asleep | another |
| cheerios | light | person | awake | any |
| cheese | medicine | pet's name | bad | each |
| chicken(food) | money | police | better | every |
| chocolate | mop | sister | big | more |
| coffee | nail | teacher | black | much |
| coke | napkin | uncle | blue | not |
| cookie | paper | bath | broken | none |
| corn | penny | breakfast | brown | other |
| cracker | picture | bye | careful | same |
| donut | pillow | call (on phone) | clean(adjective) | some |
| drink(noun) | plant | dinner | cold | the |
| egg | plate | give me five! | cute | too |
| fish(food) | purse | gonna get you! | dark | am |
| food | radio | go potty | dirty | are |
| french fries | scissors | hi | dry(adjective) | be |
| grapes | soap | hello | empty | can(verb) |
| green beans | spoon | lunch | fast | could |
| gum | tape | nap | fine | did/did ya |
| hamburger | telephone | night night | first | do |
| ice | tissue/kleenex | no | full | does |
| ice cream | toothbrush | patty cake | gentle | don't |
| jello | towel | peekaboo | good | gonna/going to |
| jelly | trash | please | green | gotta/got to |
| juice | tray | shh/shush/hush | happy | hafta/have to |
| lollipop | vacuum | shopping | hard | is |
| meat | walker | snack | heavy | lemme/ let me |
| melon | watch(noun) | so big! | high | need/need to |
| milk | basement | thank you | hot | try/try to |
| muffin | bathroom | this little piggy | hungry | wanna/want to |
| noodles | bathtub | turn around | hurt | was |
| nuts | bed | yes | last | were |
| orange(noun) | bedroom | bite | little | will |
| pancake | bench | blow | long | would |
| peanut butter | chair | break | loud | and |
| peas | closet | bring | mad | because |
| pickle | couch | build | naughty | but |
| pizza | crib | bump | new | if |
| popcorn | door | buy | nice | so |
| popsicle | drawer | carry | noisy | then |

**Table A.2 Missing or changed words in the Adult Distributional Network**

| Missing (removed) | Altered (original – changed) | | Duplicates (removed) |
|---|---|---|---|
| Cockadoodledoo | Potato chip | chips | Chicken (food) |
| Uh oh | French fries | fries | Fish (food) |
| Green beans | Thank you | thanks | Drink (verb) |
| Peanut butter | Turn around | turn | Orange (adjective) |
| Rocking chair | Belly button | belly | Water (outside) |
| Gas station | Buttocks | butt | Can (verb) |
| Go potty | Washing | washer | Watch (verb) |
| All gone | Lawn mower | mower | Slide (verb) |
| Next to | Night night | goodnight | Swing (verb) |
| Of | | | Work (verb) |
| On top of | | | Clean (adjective) |
| To | | | Dry (adjective) |
| A | | | |
| And | | | |

**Table A.3 Missing or changed words in the Toddler Distributional Network**

| Missing (removed) | Altered (original – changed) | | Duplicates (removed) |
|---|---|---|---|
| Cockadoodledoo | Potato chip | chips | Chicken (food) |
| Uh oh | French fries | fries | Fish (food) |
| Green beans | Thank you | thanks | Drink (verb) |
| Peanut butter | Turn around | turn | Orange (adjective) |
| Vagina | Belly button | belly | Water (outside) |
| Living Room | Buttocks | butt | Can (verb) |
| Play Pen | Washing | washer | Watch (verb) |
| Rocking chair | Lawn mower | mower | Slide (verb) |
| Gas station | Night night | goodnight | Swing (verb) |
| Go potty | | | Work (verb) |
| All gone | | | Clean (adjective) |
| Next to | | | Dry (adjective) |
| On top of | | | |
| A lot | | | |
| Do not | | | |

**Table A.4 Words missing from the Nelson Word Association Norms**

| | | | | | |
|---|---|---|---|---|---|
| baa baa | melon | backyard | so big! | those | could |
| choo choo | popsicle | snowman | clap | us | does |
| cockadoodledoo | pretzel | sprinkler | allgone | we | don’t |
| grr | bib | church | naughty | your | gonna |
| uh oh | snowsuit | gas station | windy | which | gotta |
| vroom vroom | tights | babysitter | hers | at | hafta |
| woof woof | chin | bye | my | by | lemme |
| kitty | owie | go potty | our | into | was |
| fire truck | penis | night night | them | next to | were |
| sled | vagina | patty cake | these | on top of | would |
| applesauce | high chair | peekaboo | they | a lot | if |
| green beans | play pen | hush | this | each | |

***Note.*** Duplicates were removed as they were in the distributional networks. Words that were changed include the same words changed in the distributional queries, as well as yum yum to yum, stroller to buggy, tissue to Kleenex, rocking chair to rocker, yucky to yuck, need to to need, try to to try, and wanna to want.

**Table A.5 Words queried and combined with original target in both association norms**

| Original Word | Combined Word | Original Word | Combined Word |
|---|---|---|---|
| Airplane | Plane | Lawn Mower | Mower |
| Alligator | Gator | Mittens | Mitten |
| Baa Baa | Bahh, Baabaa | Mommy | Mom, Mother |
| Beads | Bead | Pancake | Pancakes |
| Beans | Bean | Potato chip | Chips |
| Bicycle | Bike | Potty | Toilet |
| Boo Boo | Booboo, Owie | Refrigerator | Fridge |
| Boots | Boot | Rocking chair | Rocker |
| Bubbles | Bubble | Slipper | Slippers |
| Buttocks | Butt | Teddybear | Teddy, Teddy Bear |
| Bye | Bye Bye, Goodbye | Telephone | Phone |
| Carrots | Carrot | Thank You | Thanks |
| Choo Choo | Choo, Chu Choo | Tricycle | Trike |
| Daddy | Dad, Father | Tummy | Stomach |
| Donut | Doughnut | Vitamins | Vitamin |
| Eye | Eyes | Vroom Vroom | Vroom |
| Foot | Feet | Washing Machine | Washer |
| French Fries | Fries | Yum | Yummy,Yummy Yummy |
| Grapes | Grape | | |

**Table A.6 CDI Words included in the McRae norms (with substitutions in parentheses)**

| | | | | |
|---|---|---|---|---|
| airplane | bunny (rabbit) | fish (cod) | orange | sled |
| alligator | bus | fork | oven | socks |
| ant | butterfly | frog | owl | sofa |
| apple | cake | garage | pajamas | spoon |
| ball | car | garbage (bin) | pants | squirrel |
| balloon | carrot | giraffe | peas | stick |
| banana | cat | gloves | pen | stone |
| basement | chair | goose | pencil | stove |
| basket | cheese | hammer | penguin | strawberry |
| bat | chicken | hat (cap) | pickle | stroller (buggy) |
| bathtub | church | helicopter | pig | sweater |
| beans | clock | horse | pillow | table |
| bear | closet | hose | plate | tape |
| bed | coat | house | pony | telephone |
| bedroom | comb | jacket | potato | tiger |
| bee (wasp) | corn | jar | potty (toilet) | tights (leggings) |
| belt | couch | jeans | pumpkin | toy |
| bench | cow | key | radio | tractor |
| bicycle (bike) | crayon | knife | raisin | train |
| bird (robin) | cup | lamb | refrigerator (fridge) | tray |
| boat | deer | lamp | rock | tree (oak) |
| book | dish | lion | rocking chair (rocker) | tricycle |
| boots | dog | melon (cantaloupe) | rooster | truck |
| bottle | doll | mittens | scarf | tuna |
| bowl | donkey | moose | scissors | turkey |
| box | door | motorcycle | sheep | turtle |
| bread | dress | mouse | shirt | zebra |
| broom | duck | napkin | shoes | |
| brush | elephant | necklace | shovel | |
| bucket | fish (goldfish) | nut (walnut) | sink | |

**Table A.7 CDI Words included in the Howell Norms**

| | | | | | |
|---|---|---|---|---|---|
| alligator | cheek | friend | medicine | potato | story |
| animal | cheerios | frog | melon | potato chip | stove |
| ankle | cheese | game | milk | potty | strawberry |
| ant | chicken | garage | mittens | present | street |
| apple | child | garbage | mommy | pretzel | stroller |
| applesauce | chin | garden | money | pudding | sun |
| arm | chocolate | gas station | monkey | pumpkin | sweater |
| aunt | church | giraffe | moon | puppy | swing |
| baby | circus | girl | moose | purse | table |
| babysitter | clock | glass | mop | puzzle | tape |
| backyard | closet | glasses | motorcycle | radio | teacher |
| banana | cloud | gloves | mouse | rain | teddybear |
| basement | clown | glue | mouth | raisin | telephone |
| basket | coat | goose | movie | refrigerator | tiger |
| bat | coffee | grandma | muffin | rock | tights |
| bathroom | coke | grandpa | nail | rocking chair | tissue |
| bathtub | comb | grapes | napkin | roof | toast |
| beach | cookie | grass | necklace | room | toe |
| beads | corn | green beans | noodles | rooster | tongue |
| beans | couch | gum | nose | salt | tooth |
| bear | country | hair | nurse | sandbox | toothbrush |
| bed | cow | hamburger | nuts | sandwich | towel |
| bedroom | cowboy | hammer | orange | sauce | toy |
| bee | cracker | hand | outside | scarf | tractor |
| bellybutton | crib | hat | oven | school | train |
| belt | cup | head | owie | scissors | trash |
| bench | daddy | helicopter | owl | sheep | tray |
| bib | deer | hen | pajamas | shirt | tree |
| bicycle | diaper | highchair | pancake | shoe | tricycle |
| bird | dish | home | pants | shorts | truck |
| blanket | doctor | hose | paper | shoulder | tummy |
| block | dog | house | park | shovel | tuna |
| book | doll | ice | party | shower | turkey |
| boots | donkey | ice cream | peanut butter | sidewalk | turtle |
| bottle | donut | jacket | peas | sink | tv |
| bowl | door | jar | pencil | sister | uncle |
| box | downtown | jeans | penguin | sky | underpants |
| boy | drawer | jello | penis | sled | vacuum |
| bread | dress | jelly | penny | slide | vagina |
| broom | drink | juice | people | slipper | vanilla |
| brother | dryer | keys | person | sneaker | vitamins |

| | | | | | |
|---|---|---|---|---|---|
| brush | duck | kitchen | pickle | snowman | walker |
| bucket | ear | kitty | picnic | snowsuit | washing machine |
| bug | egg | knee | picture | soap | watch |
| bunny | elephant | knife | pig | sock | water |
| bus | eye | ladder | pillow | soda | wind |
| butterfly | face | lady | pizza | sofa | window |
| buttocks | farm | lamp | plant | soup | wolf |
| button | feet | lawnmower | plate | spaghetti | woods |
| cake | finger | leg | playdough | spoon | work |
| camera | fireman | light | playground | sprinkler | yard |
| camping | firetruck | lion | playpen | squirrel | yogurt |
| candy | fish | lips | police | stairs | your babysitter |
| car | flag | living room | pool | star | your pet |
| carrots | flower | lollipop | popcorn | stick | zebra |
| cat | food | mailman | popsicle | stone | zipper |
| cereal | fork | man | porch | store | zoo |
| chair | french fries | meat | | | |

# Appendix B

## Similarity Matrices for each Network

**Figure B.1 Adult Association Heatmap**

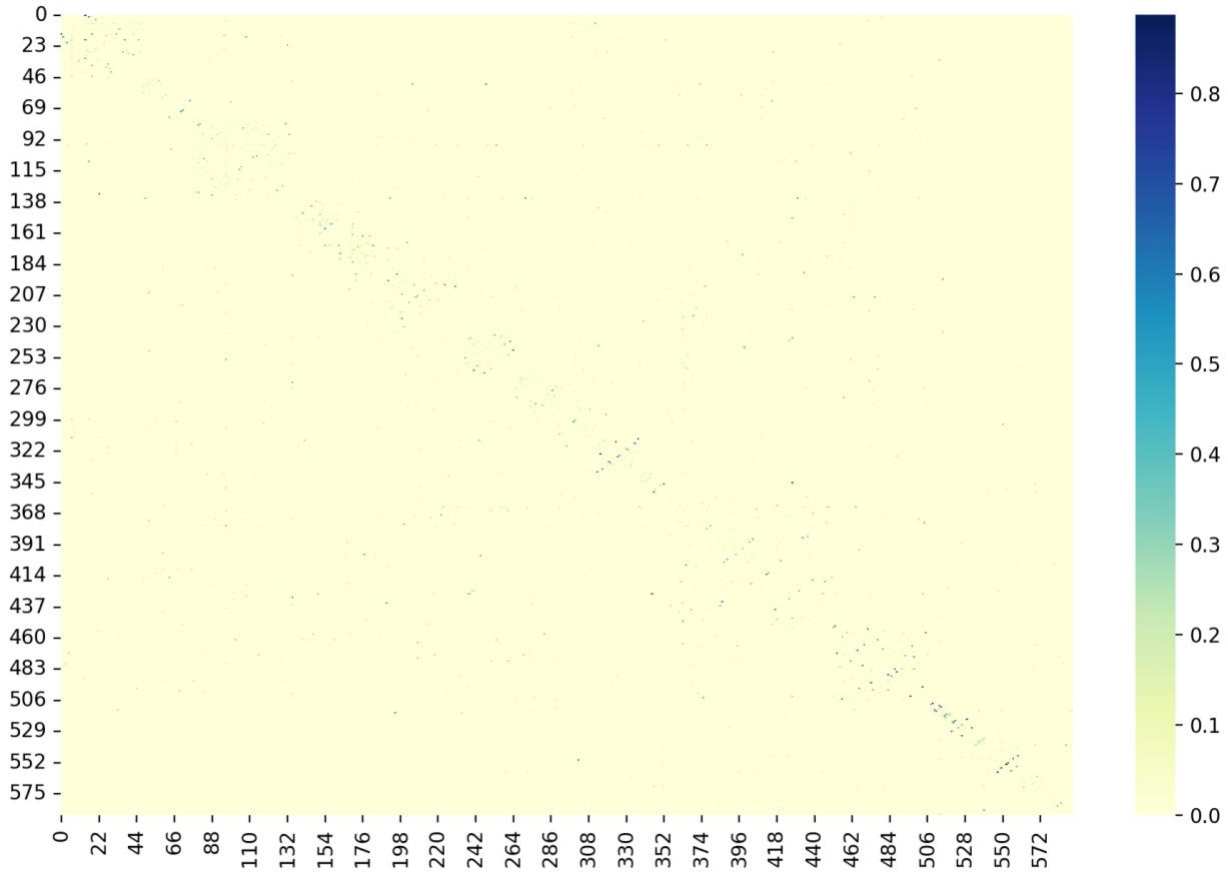


**Figure B.2 Child Association Heatmap**

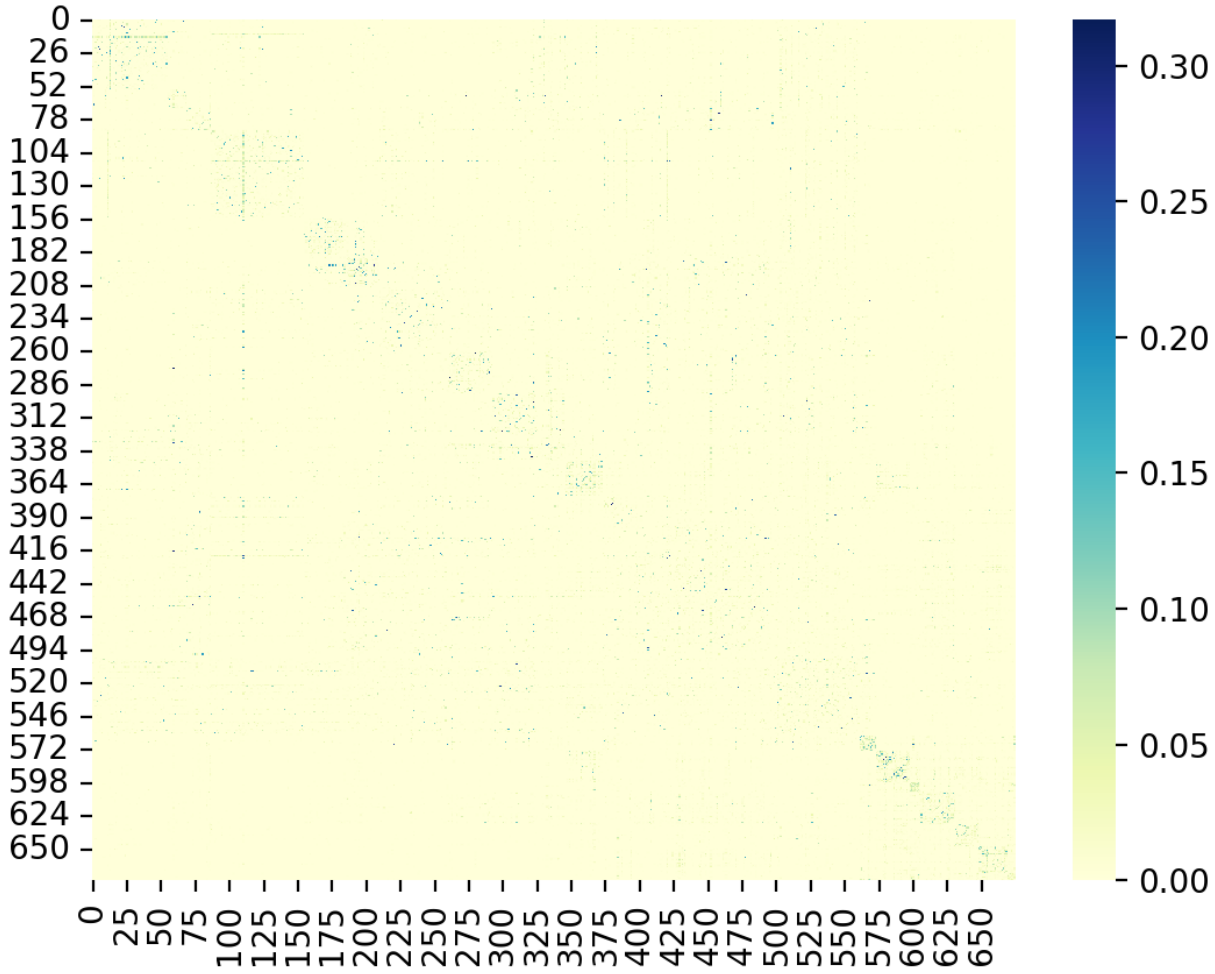


**Figure B.3 Adult Distributional Heatmap**

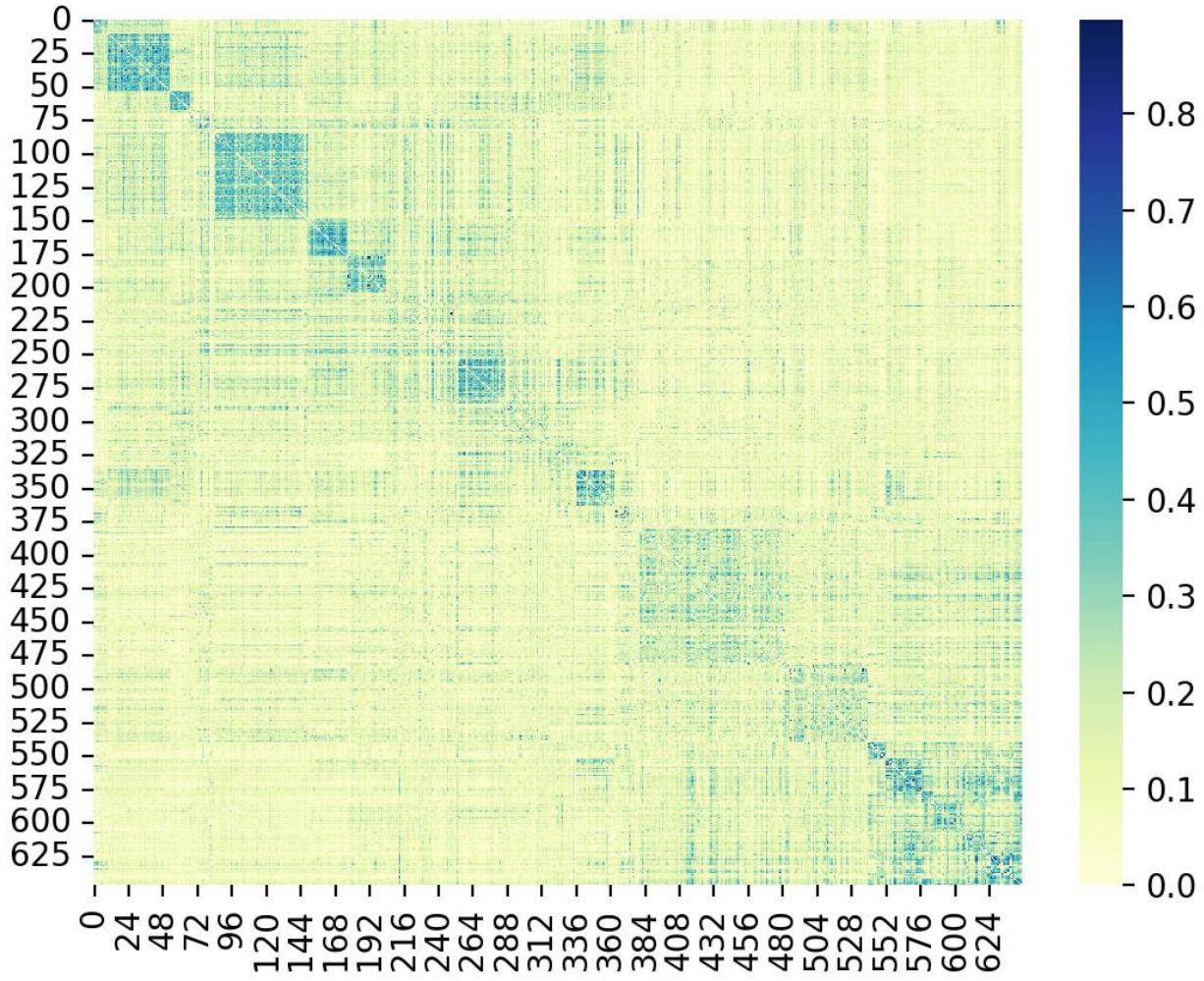


**Figure B.4 Child Distributional Heatmap**

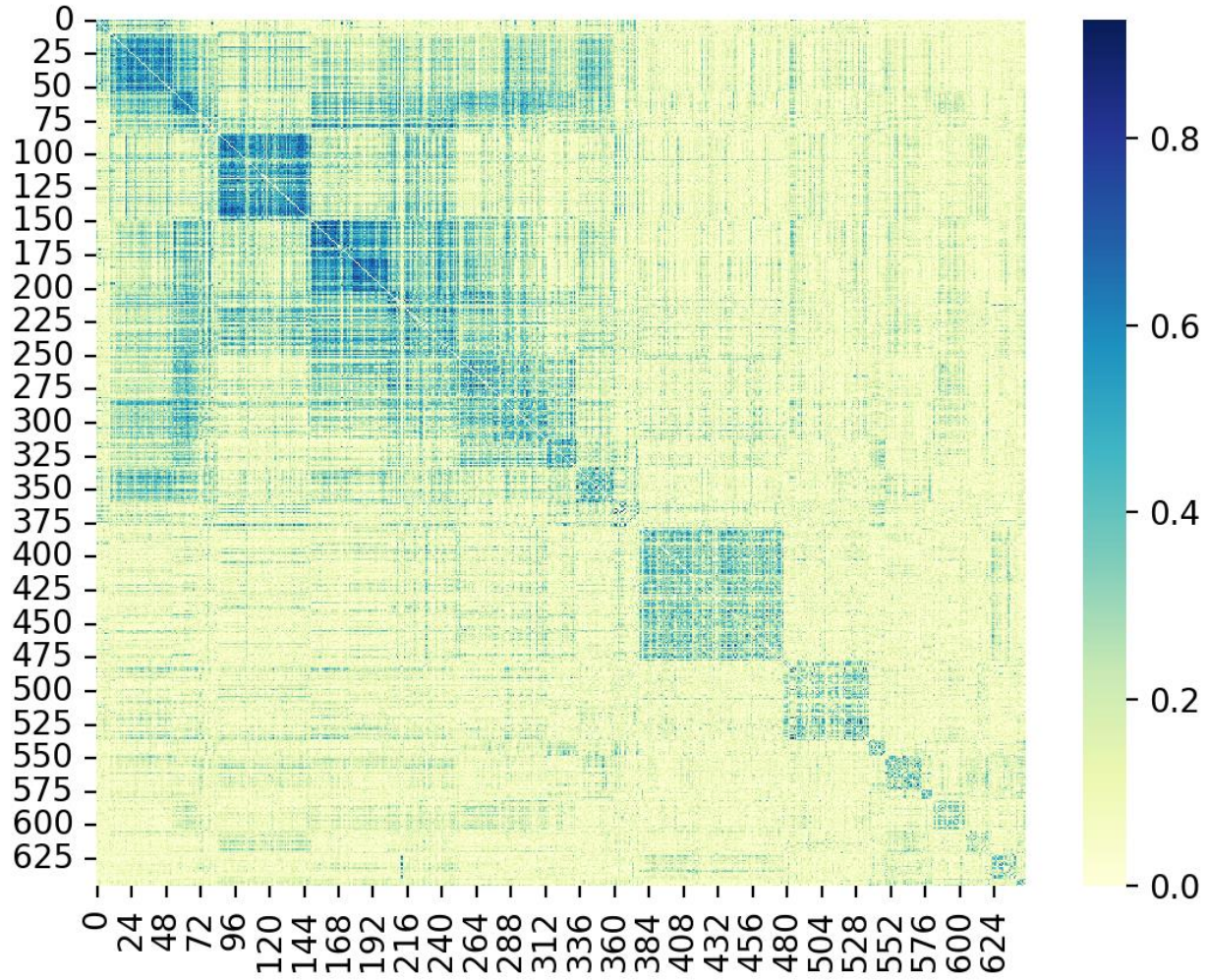


**Figure B.5 Adult Feature Heatmap**

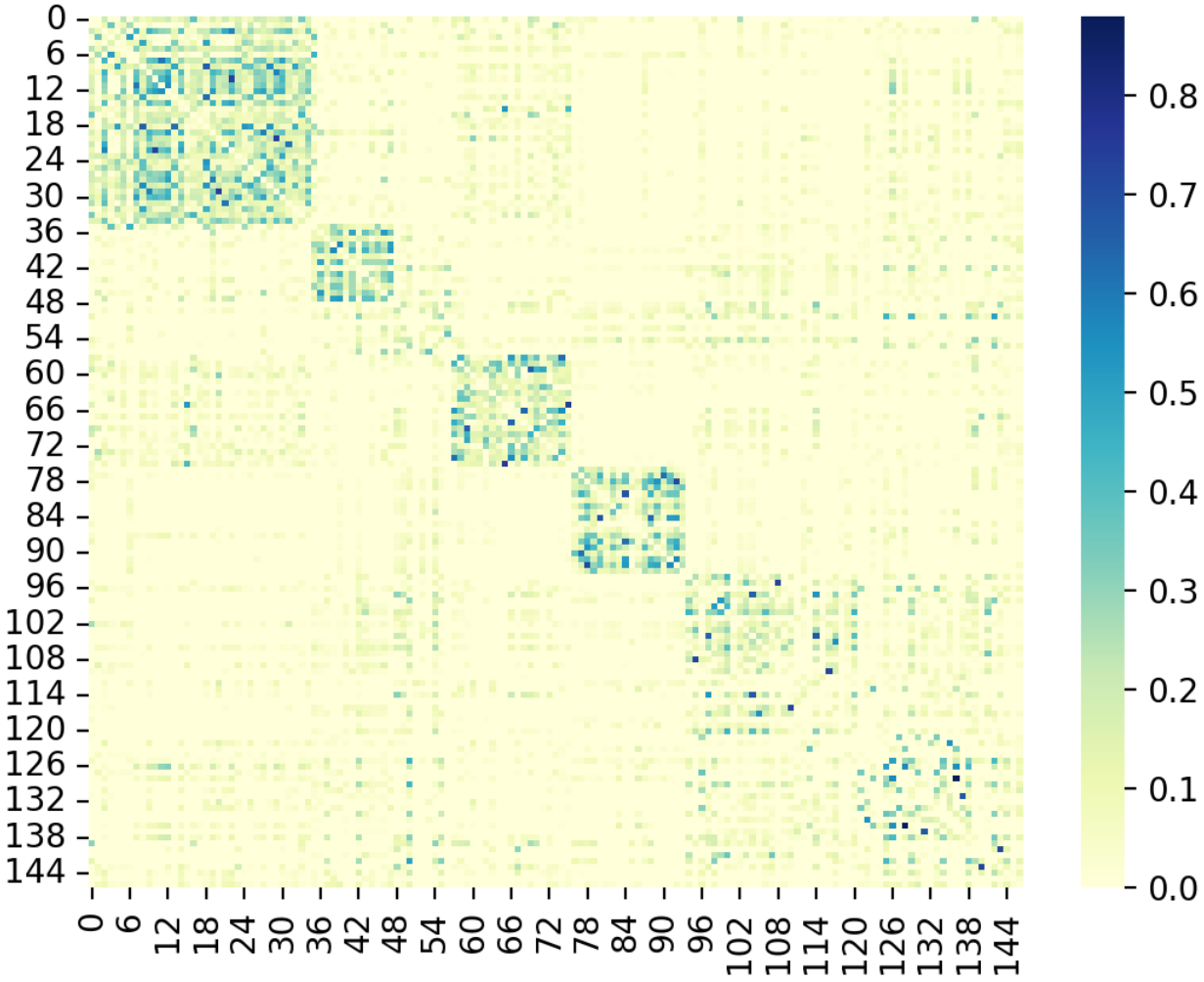


**Figure B.6 Child Feature Heatmap, Original and Used**