Journal of Memory and Language 132 (2023) 104446

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Using known words to learn more words: A distributional model of child vocabulary acquisition

Andrew Z. Flores *, Jessica L. Montag, Jon A. Willits

Department of Psychology, University of Illinois at Urbana Champaign, USA

ARTICLE INFO

Keywords:
Vocabulary
Age of acquisition
Distributional learning
Prior knowledge
Bootstrapping

ABSTRACT

Why do children learn some words before others? A large body of behavioral research has identified properties of the language environment that facilitate word learning, emphasizing the importance of particularly informative language contexts that build on children's prior knowledge. However, these findings have not informed research that uses distributional properties of words to predict vocabulary composition. In the current work, we introduce a predictor of word learning that emphasizes the role of prior knowledge. We investigate item-based variability in vocabulary development using lexical properties of distributional statistics derived from a large corpus of child-directed speech. Unlike previous analyses, we predicted word trajectories cross-sectionally across child age, shedding light on trends in vocabulary development that may not have been evident at a single time point. We also show that regardless of a word's grammatical class, the best distributional predictor of whether a child knows a word is the number of other known words with which that word tends to co-occur.

Introduction

Learning new words is a complex process, and many studies have examined how basic learning mechanisms and inductive biases promote vocabulary growth. Learning mechanisms such as fast mapping (Carey & Bartlett, 1978), analogical learning (Gentner, 1989), cross-situational statistical learning (Yu & Smith, 2007), distributional learning (Gleitman, 1990; Harris, 1957; Lany & Saffran, 2010), and hypothesis testing (Trueswell, Medina, Hafri & Gleitman, 2013) all provide means by which a learner can develop reasonable knowledge about word-referent mappings, as well as the extensive aspects of word meaning that go beyond word-referent mapping (Wojcik, Zettersten, & Benitez, 2022). In addition, inductive biases may simplify the learning problem by reducing the number of hypotheses a learning mechanism needs to actively consider. Several inductive biases used by children have been identified, including the mutual exclusivity principle (Markman and Wachtel, 1988), and the shape bias (Smith et al., 2002), and attentional biases towards particular social cues like eye-gaze and pointing (Akhtar, Carpenter & Tomasello, 1996; Tomasello & Todd, 1983, Tomasello, 1988; Yu & Ballard, 2007). Understanding the learning mechanisms and inductive biases that allow children to learn language has been a major goal of the field of language development.

Learning mechanisms and inductive biases have typically been studied using two often divergent yet related approaches. The first approach is experimental studies of word learning in controlled laboratory settings. This approach has been used to test many hypotheses about learning mechanisms and inductive biases that may scaffold the learning process. The second approach is the statistical analysis of large naturalistic datasets. This approach has been used by many researchers to identify properties of children's linguistic environments and using those properties to support or criticize different theories of language acquisition (e.g., Cameron-Faulkner, Lieven, & Tomasello, 2003; Huebner & Willits, 2021; Huttenlocher et al., 2007; Lidz, Waxman, & Freedman, 2003; Schwab & Lew-Williams, 2016). One particular use of this corpus-based approach has been to investigate which properties of children's environments predict broad-based measures of children's vocabulary development (Goodman, Dale & Li, 2008; Frank, Braginsky, Yurovsky, Marchman, 2021). In these kinds of studies, the outcome measure is typically based on large datasets of parent-report surveys, such as the proportion of children who say specific words at specific ages. Research has focused on attempting to find correlational predictors of these outcome variables. Both experimental and statistical approaches have contributed to our understanding of how the language learning process unfolds, with each method providing new information, as well as raising new questions. But in many ways these two approaches have proceeded with little crosstalk. In particular, many critical insights

* Corresponding author.
E-mail address: azf2@illinois.edu (A.Z. Flores).

https://doi.org/10.1016/j.jml.2023.104446
Received 23 September 2021; Received in revised form 14 June 2023; Accepted 17 June 2023
Available online 28 June 2023

from the experimental research have not been incorporated into the statistical modeling research.

In the current work, we highlight three insights from behavioral word learning experiments that we believe can inform statistical approaches to studying language development. Incorporating key findings from behavioral work may increase the predictive power and ecological validity of the statistical models that aim to describe early word learning. The first of these insights is the distinction between quantity and quality in linguistic experience, and specifically the question of what constitutes a "high quality" learning episode. The second insight is the important role of a child's prior word knowledge when predicting subsequent word learning, and the way in which language acquisition is an interactive process with many top-down effects. The third insight involves the relationship between grammatical class and vocabulary development, and the necessity – or lack of necessity - of grammatical class-specific learning mechanisms or representations.

We then use these insights to propose a new predictor of word learning that can be used in statistical models of vocabulary development. The innovation of our approach is that this predictor uses children's prior knowledge as a means of quantifying one way in which a learning episode can constitute a "high quality" learning episode. Unlike other statistical approaches to word learning (such as word frequency and contextual diversity), this prior knowledge-based predictor eliminates the need to posit that words from different grammatical classes require different learning mechanisms - or need to be explicitly represented as members of that category - in order to account for differences in early- and later-learned words. We argue that our new predictor can better account for patterns of word learning, and does so in a way that incorporates insights from the behavioral literature into statistical models of word learning.

We first review existing word learning literature and contrast the extent to which insights from behavioral experiments have (or have not) informed statistical models across three dimensions: (1) how the dichotomy of quantity and quality informs our understanding of the learning environments conducive to learning, (2) how prior knowledge bootstraps subsequent word learning, and (3) the role that grammatical class may play in shaping the learning process. We then introduce a statistical predictor that emphasizes the role of prior knowledge, and test its ability to predict child productive vocabulary development.

The quantity vs. quality distinction

Behavioral evidence

Within the language acquisition literature, there is often a distinction made between the quantity and quality of speech that children hear, with different proposals about the role that both quantity and quality of experiences play in language development. There is substantial evidence that both language quantity and quality are associated with language outcomes. Higher quantities of speech to children are associated with positive language outcomes (Dickinson & Tabors, 1991; Shneidman, Arroyo, Levine, Goldin-Meadow, 2013; Hart & Risley, 1995; Hoff, 2003; Huttenlocher et al., 2010; Song et al., 2012; Weisleder & Fernald, 2014).

All other things being equal, it makes sense that hearing a word more times gives more opportunities to learn the word. But in recent years, several researchers have argued that (at least part of) the mechanism by which frequency matters is by increasing the number of times children have the opportunity to hear a word in "high quality" contexts (Tomasello, 1988; Hirsh-Pasek et al., 2015; Hoff, 2006; Yu & Smith, 2007). For example, social contexts that reduce referential ambiguity are thought to contribute to high quality contexts. One factor shown to be predictive of a high-quality learning episodes is reduction of referential ambiguity using socio-visual cues (Cartmill, Armstrong, Gleitman, Goldin-Meadow, Medina & Trueswell, 2013). Another is whether caregiver and child engage in joint attention (Tomasello & Todd, 1983, Tomasello, 1988; Akhtar et al., 1996) or whether the child's attention is sustained on the target item (Yu, Suanda & Smith, 2019).

In addition to identifying instances where referential ambiguity is reduced, a great deal of research has focused on discovering properties of the language input itself that contribute to high-quality learning episodes. Caregiver speech has been shown to possess various prosodic, lexical and syntactic qualities that aid language development. These include lexical diversity (Hoff & Naigles, 2002; Huttenlocher et al., 2010; Pan et al., 2005; Rowe, 2012), syntactic complexity (Cameron-Faulkner, Lieven, & Tomasello, 2003; Huttenlocher et al., 2002; Rowe, Leech & Cabrera, 2017), or speech that is particularly sensitive or responsive to the child's behavior (Harris, Jones & Grant, 1983; Hirsh-Pasek et al., 2015; Tamis-Lemonda, Kuchiro & Song, 2014). Likewise, speech that is child-directed rather than overheard by the child is particularly associated with positive outcomes (Shneidman et al., 2013; Weisleder & Fernald, 2014), as are word contexts that are particularly informative of word meanings, or relationships between multiple concepts (Beals, 1997; Rowe, 2012). Finally, variability in the contexts in which words appear aids generalization of word labels to new exemplars (Goldenberg & Sandhofer, 2013; Vlach & Sandhofer, 2011).

Certain high-quality contexts also facilitate children's ability to segment words from fluent speech. Reliable cues or anchors that occur in highly familiar or frequent contexts help young language learners reduce potential candidates for new words in fluent speech. These include high frequency lexical items like a child's name (Bortfeld, Morgan, Golinkoff & Rathbun, 2005), highly frequent functional morphemes that reliably precede nouns (Shi & Lepage, 2008) and frequent contextual frames that tend to co-occur with nouns and verbs (Willits, Seidenberg & Saffran, 2014).

To summarize, considerable behavioral research has focused both on the quantity of input that a child receives, and on the quality of those experiences. Though effects of frequency clearly appear across multiple dimensions of language learning (Ambridge, Kidd, Rowland & Theakston, 2015), a theme that emerges in the behavioral literature is that it is the quality of experiences (very broadly defined), and not necessarily the raw quantity alone, that is more important for predicting language outcomes (Anderson et al., 2021; Hirsh-Pasek et al., 2015).

Statistical models

Statistical models of word learning typically have been used to look for predictors of differences in language learning outcomes across individuals or across words. Some of these studies have focused on children as the random variable, looking for predictors of vocabulary size. These studies - which include both correlational studies and statistical models or regression models - have found that many demographic factors, such as gender (Huttenlocher et al., 1991), maternal education (Pan, Rowe, Singer & Snow, 2005), birth order (Havron et al., 2019), amount of language input (Hoff, 2003; Huttenlocher et al., 1991; 2010; Weisleder & Fernald, 2014), and lexical processing speed (Hurtado, Marchman & Fernald, 2008), are all predictors of vocabulary size. Other studies have focused on the words as the random variable, looking for predictors of the age at which individual words are likely to be understood or produced. These studies have found that many distributional and semantic properties of words are predictive of an earlier mean age of acquisition, including word frequency (Blackwell, 2005; Frank et al., 2021; Goodman et al., 2008; Naigles & Hoff-Ginsberg, 1998), contextual diversity (Blackwell, 2005; Hills et al., 2010; Hsu, Hadley & Rispoli, 2017; Naigles & Hoff-Ginsberg, 1998), concreteness (Frank et al., 2021; Swingley & Humphreys, 2018), positive valence (Braginsky, Yurovsky, Marchman and Frank, 2019; Moors et al., 2013), and child "relevance" of the word meaning (Perry, Perlman, Winter, Massaro & Lupyan, 2018).

Statistical models have also attempted to address the importance of linguistic quantity and various measures of linguistic quality in predicting vocabulary development outcomes. Linguistic quantity is a relatively straightforward question to investigate, as word frequency (the number of times a child hears a word) is a good proxy for quantity. Linguistic quantity can then be contrasted with other distributional

predictors that are used to operationalize qualitative aspects of learning episodes. For example, researchers have investigated whether earlier learned words are special in terms of their lexical contextual diversity (the number of other words with which a word co-occurs, Blackwell, 2005; Hills et al., 2010; Naigles & Hoff-Ginsberg, 1998), and episode diversity (the number of different episodes in which a word occurs, Harris, Barrett, Jones & Brookes, 1988; Roy, Frank, DeCamp, Miller & Roy, 2015). These studies have tended to find that higher lexical diversity and lower episodic diversity are associated with earlier learned words. Similarly, researchers have found that words that occur more frequently in isolation tend to be learned earlier (Brent & Siskind, 2001), as do words that more frequently occur in shorter utterances (Swingley & Humphrey, 2018) and words that occur more frequently at the beginning and ending of utterances (Braginsky, Yurovsky, Marchman & Frank, 2016).

From one perspective, the statistical literature seems to parallel the behavioral literature quite closely, with both quantity and many measures of quality each predicting language learning outcomes. But in fact, predictors that are more associated with linguistic quality often have very small effect sizes, or even go away, when simple word frequency is controlled. For example, in Braginsky et al.'s (2019) study investigating many distributional statistics' ability to predict MCDI scores in English, the strength of the relationship between a word's frequency in child-directed speech and a child producing the word was approximately $r = 0.45$. This contrasts with the effects of a word appearing in short utterances, appearing alone, and appearing at the end of an utterance, which were approximately $r = 0.30$, 0.15, and 0.03, respectively.

The relative contribution of linguistic quantity versus quality shows a discrepancy between the behavioral and statistical research. The behavioral research emphasizes the importance of higher-quality learning episodes, while the statistical research routinely shows quantity (word frequency) to be the best predictor of easy-to-learn and hard-to-learn words (c.f., Roy et al., 2015). This mismatch between the behavioral and statistical literatures again suggests one of two conclusions. One possibility is that the implications of the statistical work are being undervalued, and along with it the importance of pure quantity as an important factor in vocabulary acquisition. Alternatively, the mismatch could be pointing to the failure of the statistical work to correctly identify, measure, and use adequate proxies for high quality learning episodes. Resolving this inconsistency between the behavioral and statistical research would shed considerable light on mechanisms of vocabulary acquisition.

Prior knowledge

Behavioral evidence

One extremely important contribution of experimental word learning research has been the demonstration of a wide range of ways that word learners use preexisting knowledge of other words to bootstrap the learning of new words. Each word learning episode does not exist in isolation, and both general learning mechanisms and inductive biases take advantage of prior knowledge. The role that prior knowledge plays in driving subsequent learning is a central theme in the word learning literature.

There are many examples of this phenomenon outside of learning about word meanings. For example, infants' sensitivity to the distributional structure of the sounds in their language affects their phonemic discrimination (Maye, Werker, & Gerken, 2002). Infants also have an easier time recognizing, processing, and learning new syntactic structures that match those with which they have previous experience. For example, nonadjacent dependency learning is bootstrapped by prior learning of an adjacent dependency (Lany, Gomez, & Gerken 2007), when the dependencies share phonological overlap (Onnis, Monaghan, Richmond & Charter, 2005), semantic overlap (Willits, Safran, & Lany, 2017), or are cued by known nonadjacent dependencies (Zettersten, Potter, & Saffran, 2020). Likewise, children's ability to produce and understand complex syntactic structures like relative clauses seems to emerge from children's ability to use and understand simpler sentence structures (Brandt, Diessel & Tomasello, 2008).

Within the realm of learning about words and their meanings, there is a tremendous amount of evidence that children and adults lean heavily on pre-existing knowledge while segmenting, recognizing, and learning the meaning of new words. Specifically, by using prior knowledge, children can narrow down potential referential candidates of new words. For example, children can use known object labels to reduce referential candidates through the principle of mutual exclusivity (Merriman, Bowman & MacWhinney,1989; Markman and Wachtel, 1988), and through comparing prior experiences to new experiences in order to discover common abstractions though analogical learning (Gentner, 1989).

There is a great deal of evidence that children keep track of and accumulate knowledge of statistical and structural regularities in the language environment, and can use this information to aid word learning. Studies which examine children's capacity to learn from statistical and structural regularities have shown they can make inferences about a word's semantic category as a result of their patterns of distributional co-occurrence (Lany & Saffran, 2010). Children can also use sentences' syntactic structures to infer meanings of novel verbs (Landau & Gleitman, 1985; Naigles, 1996; Yuan & Fisher, 2009) and novel nouns (Ferguson, Graf & Waxman, 2014) in those sentences. Research with ERPs shows that new words are learned more easily when they occur in semantically supportive contexts (Borovsky, Kutas, & Elman, 2010). Children are also able to apply previous encounters with distributional regularities, such as when children are tasked with rapidly evaluating statistical evidence across individually ambiguous words, to resolve word-referent ambiguities in cross-situational learning tasks (Yu & Smith, 2007).

Children also use prior knowledge of the sounds and phonotactics of their language to aid word recognition and learning. Children more easily recognize novel words that follow their native language's phonotactic (Jusczyk & Aslin, 1995; Nazzi et al., 2005) and stress (Echols, Crowhurst, & Childers, 1997; Houston, Santelmann, & Jusczyk, 2004; Jusczyk, Houston, & Newsome, 1999; Morgan & Saffran, 1995; Nazzi et al., 2005) patterns. Prior experience with phonological forms also assists individuals with mapping novel word forms to references (Estes, Evans, Alibali, & Saffran, 2007; Fennell & Werker, 2003; Ferry, Hespos & Waxman, 2010; Hay, Pelucchi, Graf Estes, & Saffran, 2011). Similarly, infants can recognize and attend to the visual referent of a word at much earlier ages if it is spoken by a familiar voice, such as their mother's (Bergelson & Swingley, 2012).

There is also considerable evidence that on an individual difference level, children with higher vocabularies have very different word learning abilities. For example, vocabulary size predicts children's memory for object names and features (Perry, Axelsson, & Horst, 2016). Children with larger vocabularies also show more associative facilitation in activating lexical concepts (Borovsky & Peters, 2019). Children's vocabulary size changes the nature and strength of the inductive biases children bring to bear on word learning and word recognition (Colunga & Sims, 2017; Perry & Saffran, 2017; Perry & Samuelson, 2011). Differential vocabulary levels in monolingual and bilingual children predicts differences in those children's disambiguation of novel words (Byers-Heinlein and Werker, 2013).

Many forms of prior knowledge that children bring to word learning tasks, including acoustic, lexical and syntactic knowledge, aid in the language learning process. Despite a rich literature citing the importance of prior knowledge for subsequent learning, prior knowledge has not often been incorporated into statistical models of word learning, or has been incorporated in narrow ways.

Statistical models

Despite the widespread acceptance and considerable work showing that prior knowledge is important for understanding word learning in

behavioral studies, prior knowledge is rarely incorporated into statistical models of word learning. A notable exception to this is work are a few studies using growth model analyses to simulate children's developing lexical networks. For example, Siew and Vitevitch (2020) found that children are more likely to learn new words that have less dense phonological neighborhoods. Cox and Haebig (2022) found that growth models employing child-derived word association strength add predictive power to models of vocabulary development.

Of most relevance to our current work is research by Hills et al. (2010). They created 15 separate graphs of children's lexical networks, one each for children from age 16 to 30 months. In these models, nodes were added to the graph for each word produced by at least 50 % of children at that age (according to MCDI parental surveys), and connections were added if the words ever co-occurred within a fixed window size in child-directed speech (in the CHILDES corpus), effectively a measure of the words' lexical diversity (the number of different words with which a word co-occurred).

Hills et al. then tested three hypotheses about how network connectivity predicted the acquisition of new words (i.e., when words crossed the 50 % threshold). The first was the *preferential attachment* hypothesis, that words most likely to be added next were the words that co-occurred with words that co-occurred with many other words. Put another way, some words are like "hubs" in the network, and easily learned words are those that have connections to those hub words with high lexical diversity. The second was the *lure of the associates* hypothesis, that words most likely to learned next were the words connected to the most words that were already known. Put another way, an easily learned word is one connected to the most words you already know, regardless of whether those words' own connectivity structure. The third was the preferential acquisition hypothesis, that the words most likely to be learned next were those with the most connections overall, both amongst known and unknown words.

Under each of these hypotheses, the children's prior knowledge (i.e., the set of words that are known, defined as being in the network) makes different predictions for which words should be acquired next. *Preferential attachment* predicts that words connected to already known contextually diverse words are easily learned. *Lure of the associates* predicts that the contextual diversity of the newly learned words that matters, but calculated only over already known words. *Preferential acquisition* predicts that the contextual diversity of the newly learned words that matters, and that whether the words are known or not does not matter – it is just the diversity in the language statistics alone that matters.

Hills et al.'s analyses found mixed support for several of the hypotheses. They found that the "lure of the associates" hypothesis best predicted overall word acquisition, and also best predicted noun acquisition. But they found that verbs and function words were best predicted by the "preferential acquisition" hypothesis, and that none of the hypotheses involving child-directed language predicted the acquisition of adjectives.

Hills et al.'s analyses (as well as the other network growth analyses by Siew and Vitevich, as well as by Cox and Haebig), are interesting and notable because they are some of the few studies that attempt to take prior knowledge into account when predicting vocabulary development. But Hills et al.'s study also raises many questions. What mechanisms could support saying that contextual diversity matters, regardless of whether children know the word (as it does in the *preferential acquisition* model)? Additionally, how much do the conclusions of Hills et al. depend on the binary way in which contextual diversity was calculated? A third question is, are there ways to incorporate the prior knowledge being used in the *lure of the associates* model into more standard regression approaches that do not make use of graphical growth models? A final question is, why might contextual diversity matter for some grammatical classes and not others? *Preferential acquisition*, a predictor that doesn't account for prior knowledge, worked for verbs and function words, and lure of the associates, which does factor in current knowledge, worked best for nouns? Are qualitatively different learning mechanisms, or distinct representations, being used for different grammatical classes?

To summarize, there exists a strong disconnect between overwhelming experimental evidence that prior knowledge is a very important factor in predicting the acquisition of new words, and statistical modeling work that has had difficulty demonstrating the importance of that factor. One possible explanation is that, as with the quantity vs. quality distinction, the statistical modeling work is suggesting that this factor is not as important as the experimental work has led us to believe. Alternatively, the mismatch could be pointing to the failure of the statistical work to correctly identify, measure, and use adequate proxies for children's prior knowledge. Resolving this inconsistency between the behavioral and statistical research would shed considerable light on mechanisms of vocabulary acquisition.

The role of grammatical class in word learning.

Behavioral evidence

Many behavioral studies that investigate word learning mechanisms have focused on nouns. This includes studies spanning a range of methods and theoretical approaches, including fast-mapping (Carey & Bartlett, 1978), cross-situational word learning (Yu & Smith, 2007) and hypothesis testing (Trueswell, Medina, Hafri & Gleitman, 2013), as well as inductive biases such as mutual exclusivity (Markman & Wachtel, 1998), shape bias (Smith et al., 2002), and social cues (Akhtar, Carpenter & Tomasello, 1996; Tomasello & Todd, 1983, Tomasello, 1988; Yu & Ballard, 2007). However, despite largely being investigated in the context of nouns, most of these learning mechanisms and inductive biases are proposed to be more generally applicable to words of any grammatical class. For example, the distributional statistics of a word's prosodic information, word co-occurrence information, and syntactic information have each been shown to be useful for inferring aspects of meaning of words from multiple grammatical classes (Arias-Trejo & Alva, 2013; Christophe et al., 2008; Fisher, Gertner, Scott, Yuan, 2010; Hills, Maouene, Riordan & Smith 2010; Lany & Saffran, 2010, Lany & Saffran, 2013; Naigles, 1990; Wojcik and Saffran, 2015). Indeed, computational models using language data to learn distributional semantics tend to not make *a priori* distinctions between grammatical classes, and perform well at learning thematic and taxonomic relations across many grammatical categories (Lund & Burgess, 1996; Elman, 1990; Huebner & Willits, 2018; Jones & Mewhort, 2007). Likewise, most proposals involving analogical learning, Bayesian inference, and hypothesis testing that are formally applicable to the word learning process have been shown to apply to learning about the aspects of meaning of words from multiple grammatical classes (Booth & Waxman, 2009; Gentner, 1989; Gentner & Namy, 2006; Sadeghi, Scheutz, Krause, 2017).

In short, while most learning mechanisms and inductive biases have been demonstrated in the context of noun learning (Akhtar et al., 1996; Carey & Bartlett, 1978; Markman & Wachtel, 1998; Smith et al., 2002; Tomasello & Todd, 1983, Tomasello, 1988; Trueswell et al., 2013; Yu & Smith, 2007), they are hypothesized to be at least partially independent of the grammatical class of the word that is being learned, despite nouns being the demonstrated test case.

Statistical models

In contrast to the learning mechanisms proposed and tested in behavioral experiments, the distributional and semantic predictors of a word's age of acquisition are very much *not* independent of grammatical class. Many investigations focus on a single grammatical class of word (e.g., adjectives: Blackwell, 2005; verbs: Hsu, Hadley & Rispoli, 2017; Naigles & Hoff-Ginsberg, 1998). Further, a word's grammatical class is itself a strong predictor of its age of acquisition, with nouns acquired before verbs, and verbs acquired before adjectives, and adjectives acquired before function words (Fenson et al., 1994; Swingley & Humphrey, 2018; Gentner, 1982).

Even more striking is that all the distributional predictors studies so far are themselves dependent on grammatical class when multiple classes are investigated at the same time. For example, the strength of the correlation between word frequency and how likely children are to say a word has been shown to depend on grammatical class (Frank et al., 2021; Goodman et al., 2008). The relationship between frequency and children's productive vocabulary reflects a "Simpson's Paradox" (Simpson, 1951, Goodman et al., 2008). The correlation is non-significant (or even negative) when examined across all words. But word frequency is significantly positively correlated within words of specific grammatical classes. In other words, 24-month-old children are more likely to say *mommy* than *tree*, and more likely to say *the* than *therefore*. But the children are not more likely to say *the* than *mommy*, even though the frequency of *the* is orders of magnitude higher. The strength of the effect of frequency within each class varies as well, with the effect of word frequency being quite strong for nouns, and smaller (though still significant) for verbs, adjectives, and function words (Goodman et al., 2008). Likewise, predictors such as contextual diversity, which measure the count of unique words or contexts with which a given word co-occurs, show a similar sensitivity to word class. As we have already described, Hills et al. (2010) vocabulary growth models made distinctly different predictions about what distributional predictors best predicted nouns versus what best predicted verbs and function words. Findings like these have been interpreted to suggest that different grammatical categories may be learned via different learning mechanisms, or that word learners representing the grammatical class of words and tracking statistics differently for different words.

A particularly clear example of distributional statistical approaches not being independent of grammatical class is work by Chang and Deák (2020). Chang and Deak created word co-occurrence matrices, one of co-occurrences between content words within sentences, and one of the content words with immediately adjacent syntactic frames. They then computed the principal components of those matrices, and used the words' loadings on these principal components as predictors in a regression model of the age at which a threshold number of children comprehend or produce a word. They found that many of these principal component loadings predict MCDI extremely well. Notably, many of the most highly predictive principal components were grammatical in nature. For example, the strongest effect came from a principal component that effectively indexed whether the word was a noun versus a function word.

Across many studies, statistical models of word learning imply, either implicitly (by only investigating a single grammatical class) or explicitly (by finding different statistical predictors of word learning across grammatical classes) that statistical predictors of word learning are best interpreted in conjunction with information about grammatical class. The fact that statistical models of vocabulary development based on distributional information require (or at least benefit from) information about grammatical class, stands in stark contrast to the proposals put forth based on behavioral studies, which have made the case for grammatical class-independent learning mechanisms. This mismatch once again suggests one of two conclusions: either the statistical work is being undervalued, and grammatical class really is special in some way, or the statistical work is missing something important allowing for the discovery of predictors that account for differences in grammatical classes. As with the other two issues, resolving this inconsistency between the behavioral and statistical research would shed considerable light on mechanisms of vocabulary acquisition.

The present study

Reviewing the behavioral and the statistical modeling literature, we have noted three features that emerge distinguishing the two approaches regarding factors important for word learning: (1) the relative importance of quantity vs. quality, (2) the importance of prior knowledge, and (3) the role of grammatical class. We believe that by focusing on these three discrepancies between the two approaches, we can develop a statistical predictor of word learning that both better predicts children's word knowledge than existing statistical measures, and which incorporates key findings from behavioral research into statistical models.

We introduce a predictor variable designed to capture how children's existing word knowledge interacts with distributional properties of words. Put simply, our predictor is a measure of the proportion of a word's occurrences that are with other words that a child already knows. Our measure Pro-KWo (the Proportion of Known Words, the operational definition of which is described in greater detail in the Methods section), instead compares the proportion of times a word co-occurs with already known words, compared to the number of times it co-occurs with unknown words. The intention of the measure is to capture the intuitive sense that words should be easier to learn if they tend to occur in high quality contexts, with this instance of "high quality" defined as "words occurring in contexts where children are able to leverage their prior knowledge".

To give an intuitive example of Pro-KWo, consider the words "where" and "why." One contributing factor to why children may produce "*where*" before "*why*" is that "*where*" tends to co-occur with words children already know and whose location is getting asked about. In contrast, "why" is often part of questions that involve less frequent, and more abstract, and therefore later-learned referents. A language learner should take longer to acquire the word "*why*" because the meanings of the words that co-occur with "*why*" themselves are less likely to be known.

Pro-KWo bears similarities and differences to earlier proposed distributional predictors of vocabulary. It bears a relationship to lexical contextual diversity. A word has high lexical contextual diversity if it occurs with many different word types. But lexical contextual diversity does not consider whether those co-occurrences are with known or unknown words. A word will have a high Pro-KWo score if a high proportion of the word types with which a word co-occurs are already known to the child, and as such a word's contextual diversity score and Pro-KWo score could differ dramatically.

Pro-KWo is conceptually more like the "lure of the associates" measure proposed by Hills et al. (2010), though operationally there are several important differences. A word's "lure of the associates" score depends on its "indegree" within the graphical lexical network; a word scores high if it is connected to more words that are already known. In our measure, it is the *proportion* of overall co-occurrences that matters. Thus, in "lure of the associates", a word that co-occurs with five known words is more likely to be acquired than a word that co-occurs with three known words. For Pro-KWo, the latter word could be predicted to be the earlier-learned word if those three known words represent a greater proportion of the total number of co-occurring words, relative to the proportion of total words that the five known words comprises. Thus, words with a high Pro-KWo score may or may not co-occur with many different words; they may not even co-occur with many different known words. The key feature is that a high proportion of a word's occurrences are with already known words. Thus, Pro-KWo is designed to be a distributional analogue of the behavioral research demonstrating that prior knowledge often aids word learning through mechanisms that rely on children to already know some of the other words in the sentence (whether those mechanisms be bootstrapping mechanisms like syntactic bootstrapping, or constraint-based mechanisms like mutual exclusivity).

Building upon Hills et al. (2010) we explore a novel way in which prior knowledge can be incorporated into statistical models of word learning. We believe this method may be a way to both include key findings from behavioral experiments into statistical models of word learning, as well as to improve the accuracy of statistical models that predict word learning. The open question is whether the Pro-KWo measure, like "lure of the associates", also shows strong interactions with grammatical class, or is independent of it.

Method

In order to predict the words that children know from distributional statistics of child-directed speech, we must first operationalize and compute both measures of child vocabulary and the four key distributional statistics describing patterns in child-available speech, including our new Pro-KWo measure. All data and code used in the analyses described below are available at https://github.com/AzFlores/Pro-KWo.

Dependent Measures: Child vocabulary data and MCDIp

First, we operationalize word knowledge as children's word production, tracked as part of the American English MacArthur-Bates Communicative Development Inventory (Fenson, 2007). To predict word production data, we use multiple predictors computed from the distributional statistics of child-available speech in the American English CHILDES corpus (MacWhinney, 2000), including our new measure, Pro-KWo.

The words used in our analyses are the 680 items from the American English MacArthur-Bates Communicative Inventory of child language production (Fenson, 2007). We obtained the results of MCDI (Words & Sentences) surveys for 7601 parents, available at the Wordbank website (https://wordbank.stanford.edu, Frank et al., 2017), which reports whether a child produces a word at a given age. The data was originally downloaded on June 10, 2023 directly from the website (https://wordbank.stanford.edu/). In our analysis, we excluded duplicate homonyms (i.e., "can"), because the present analyses group homonyms into a single word form so it is impossible to calculate separate statistics for each meaning. We also excluded compound words (e.g., "french fries"), because compound words are not consistently transcribed as such in the corpora we use, so it would be impossible to calculate accurate statistics for the compound. We also excluded word endings (e.g "eat-ing"), again because these word endings are not consistently parsed from their roots in the corpora we use. And finally, a small set of words were excluded for item specific reasons, such as words for private parts. All decisions about which words to exclude were made before any correlational analyses were conducted. The final dataset included 500 words.

For our outcome variables, we used two measures of children's vocabulary knowledge. Previous research trying to predict children's vocabulary development using the MCDI has used the age at which a certain percentage of children say a word as the dependent measure (e.g., 16 months is the age at which at least 50 % of children say "mommy", and 23 months is the age at which at least 50 % of children say "towel"). We elected not to use this measure for two reasons. First, it is somewhat arbitrary what percentage cutoff to use, and the shape of the distribution changes dramatically depending on what cutoff is used. Second, we were interested in looking at how different distributional predictors change across ages, and the cutoff approach doesn't give an easy way to do that.

Instead, we used two other operationalized definitions of child vocabulary development. The first, hereafter MCDIp (MCDI proportion), refers to the proportion of children who produced a particular word at each age. To calculate a word's MCDIp score, we first summed the number of times a word is reported as produced in the MCDI, then we divided that sum by the total number of administrations. This procedure yielded 500 individual MCDIp scores (one for each word) for each of 15 ages (16–30 months). MCDIp can be calculated at each age, and so cross-sectional differences in the distributional predictors can be analyzed. Our second dependent measure was the binary production outcome (child produced or did not produce a word) for all MCDI surveys from the age subsets described above. In this analysis, we effectively performed a large logistic regression, attempting to predict produced/not produced for each word as a function of our predictor variables.

Distributional predictors

All lexical distributional statistics used as predictor variables in our analyses were derived from the CHILDES database, a corpus of speech addressed to and in the presence of children (MacWhinney, 2000). Our dataset includes 49 corpora of American English spoken to 522 children up to 30 months of age. The data was obtained from the Childes-db website (https://childes-db.stanford.edu, Sanchez et al., 2018) on June 10, 2023, using the R package *childesr* (Braginsky, Sanchez & Yurovsky, 2018). Using this dataset, we obtained the distributional statistics for the 500 MCDI words as described below.

Cumulative log frequency (Frequency)

Each MCDI word's log frequency was computed by counting the number of times it occurred in the CHILDES corpus for children up to a given age, and then performing a $\log_{10}$ transformation. This resulted in 15 $\log_{10}$ frequency scores for each word, one for each age.

Lexical diversity (LD)

Lexical diversity was computed by counting the proportion of other MCDI words with which each MCDI word co-occurred, in the CHILDES corpus for children up to a given age. This was computed in the following way. First, for each age, we constructed a 500x500 matrix, with each cell in the matrix reflecting the number of times each word co-occurred with another MCDI word in the CHILDES corpus within a 7-word (forward) window. This resulted in 15 (one for each age in months) different 500-element co-occurrence vectors for each MCDI word. For each age, we then computed the proportion of each word's vector elements that were nonzero, to obtain the proportion of MCDI word types that each word co-occurred with at that age. These computations of lexical diversity were computed across all words, without taking into consideration a word's grammatical class.

Document diversity (DD)

Document diversity was calculated by computing the proportion of the 1718 documents (number of transcripts in our CHILDES dataset) in which a word occurred, in the CHILDES corpus for children up to a given age. Each individual audio recording (document) in CHILDES captures a single event such as breakfast or bath time, so document diversity can be considered a proxy of the diversity of events in which a word occurs. This resulted in 15 document diversity scores for each MCDI word (one for each age in months).

Proportion known word co-occurrence (Pro-KWo)

Our measure of the "Proportion of Known Word Co-occurrence" (Pro-KWo), was computed as follows. We started with the co-occurrence matrix described above when computing lexical diversity. This matrix yielded counts of how many times each MCDI word co-occurred with each other MCDI word. We then took each word's 500-element co-occurrence vector and multiplied those values element-by-element by the MCDIp score for each co-occurring word. The MCDIp score, the proportion of children at that age who produced that word, thus served as a proxy measure for how likely children of that age are to have prior knowledge of that word. This yielded, for each word at each age, a 500-element vector of co-occurrence frequencies, weighted by the proportion of children who knew each of those 500 co-occurring words. Next, for each word we calculated the sum of both the original unweighted word co-occurrence vector, and the counts weighted by the MCDIp. We divided the *weighted* sum by the corresponding *unweighted* sum. The resulting scalar value is a proxy for the proportion of a word's total co-occurrences that were with known words. An example is shown in Table 1 using hypothetical but illustrative MCDIp scores and co-occurrence counts. This table shows that the words *why* and *where*, while equated in frequency, nonetheless have very different Pro-KWo scores because "where" co-occurs with more known words.

Table 1
Hypothetical Pro-KWo scores for the words *why* and *where*. Co-Occurrence values are calculated within a 7 word forward moving window.

1. Unweighted co-occurrence counts.

	ball	cup	think	did	Sum
Why	10	10	100	100	220
Where	100	100	10	10	220
MCDIp	*0.7*	*0.6*	*0.2*	*0.3*	

2. Weighted co-occurrence counts (Unweighted * MCDIp).

Why	7	6	20	30	63
Where	70	60	2	3	135

3. Unweighted Sum/ Weighted Sum			**Pro-KWo**
Why	63/220	=	0.29
Where	135/220	=	0.61

Pro-KWo Shuffle

We also created a variant of the Pro-KWo measure to deal with potential confounds in the measure. One issue of concern with our Pro-KWo measure is that we use, as a part of Pro-KWo, one MCDI-derived value (MCDIp, the proportion of children who say a word at each age). We then use the Pro-KWo score to predict MCDIp and the binary "produces vs. doesn't produce" values. This does raise the concern that Pro-KWo's potential predictive power just comes from using the MCDI to predict itself. Mathematically, the probability of this mattering is low. Consider the attempt to do a logistic regression predicting the 274 binary parental reports of whether their 20-month-old child produces *shoe*. Of these 274 events, 238 (0.867) are "yes", and 34 (0.133) are "no". Clearly, using MCDIp of *shoe* (definitionally the same, 0.867) as a predictor value in the logistic regression would be circular and pointless. But the MCDIp value is not being directly used. Instead, it, along with all 499 other MCDIp values, are all being multiplied by the corresponding co-occurrence values of each of those 499 words with *shoe*. Thus, shoe's MCDIp value is only 1 of 499*2 values going into shoe's Pro-KWo score, and thus is not likely having a large effect on its value. Nonetheless, it is important to make sure this is not driving the effect.

In order to exclude this possible circularity, we created a "Pro-KWo Shuffle" measure. For Pro-KWo Shuffle, Pro-KWo values were calculated as previously described, with the exception that the MCDIp scores used to weight word co-occurrences were randomly assigned to different words within the same age group. For the Pro-KWo Shuffle results reported below, we shuffled MCDIp scores within each age 1000 times, and used each shuffled dataset to calculate 1000 individual Pro-KWo Shuffle scores (for each word, at each age). We then correlated each word's newly created shuffled Pro-KWo score with its non-shuffled MCDIp score, for all 1000 random simulations, and averaged these correlations. Thus, in Pro-KWo Shuffle, *shoe's* MCDIp value may be randomly assigned to "daddy's" co-occurrence score, instead of being paired with *shoe's* co-occurrence score. If we see that shuffling MCDIp values within an age group still leads to Pro-KWo being strongly associated with MCDIp scores, then this may demonstrate there is a problem with using MCDI scores as part of the measure being used to predict MCDI scores. But if the predictive value of Pro-KWo Shuffle is at or close to zero, it will show that the potential confound is not a concern for our measure.

Analyses of productive vocabulary development

With these measures of children's vocabulary knowledge and key distributional measures of child-directed speech, we tested the relationships between these vocabulary measures and distributional statistics of child-directed speech. All analyses were performed in R. Mixed-effects logistic regression (glmer) analyses were performed with the lme4 package, version 1.1.26 (Bates, Maechler, Bolker, & Walker, 2015). Data and code are available at (https://github.com/AzFlores/Pro-KWo).

In our first set of analyses, we aimed to better understand the relationship between our distributional predictors and child language outcomes. We first computed correlations between our four distributional measures with each other, at each age. We then computed correlations between our four distributional measures (as well as Pro-KWo Shuffle) and MCDIp at each age. These analyses allowed us to see normative trends in the data and understand the relationships between our predictors. Pro-KWo Shuffle gives us an additional way to control for distributional effects of age. If Pro-KWo scores are changing across ages in a way that is not specific to the word-specific correspondences of co-occurrence scores and MCDI scores, then the Pro-KWo Shuffle score will also be highly correlated with MCDI. But if the correlation for Pro-KWo Shuffle is at or near zero, this will be evidence that the specific word-occurrence and MCDI correspondence was critical to Pro-KWo's high correlation with MCDI.

Our next two sets of analyses allowed us to test which distributional predictors were robust to effects of age and random effects of individual children and words. Across ages, all four (five counting Pro-KWo Shuffle) distributional predictors will be *extremely* highly correlated with MCDI, but for the uninteresting reason that they all go up with age. The MCDI scores naturally go up with age, as does a word's cumulative frequency, the proportion of words and documents with which a word co-occurs, and the proportion of its co-occurrences that are with known words. To account for and remove this age effect, we created mixed-effects logistic regression models predicting the MCDI's binary word production measure (1 = produced, 0 = did not produce), one for each of our distributional measures. Each model had the child's age and one of the four predictors as fixed factors, and child and word as random factors. To see how the measures' predictive value varied across age, we then created a separate mixed effect model for each predictor at each age, in order to test the effect of variability of our predictors within each age group.

Our third and final set of analyses mirrored our second set of analyses, but with the goal of understanding the role of grammatical class in moderating the relationship between our four distributional predictors and language outcomes. We first computed correlations between our four distributional measures separately for each of four grammatical classes (adjective, function word, noun, verb). We then computed correlations for each of our four distributional predictors with MCDIp, again, separately for each grammatical class, at each age. Finally, we created a mixed-effects logistic regression model predicting the binary word production measure with the Pro-KWo measure. Crucially, we did not compute separate models for each grammatical class. Our goal was to better understand the prediction error across words of different grammatical classes in these regression models and see if Pro-KWo is a measure that is robust to grammatical class.

Results

Correlations of distributional predictors with each other and with MCDIp

To better understand the general relationship between each statistical predictor, we first examined the correlation between each predictor with each other and with MCDIp at 24 months of age, the age where variance in MCDI is highest. Fig. 1 shows the histograms of each predictor, and the scatterplot and correlation of each predictor with each other. Within this age group, as expected from much previous language modeling work, there exists a very strong relationship (though sometimes nonlinear) between word frequency and both measures of contextual diversity. In addition, document and lexical diversity are themselves highly correlated. In contrast, our measure of Pro-KWo shows much smaller correlations with word frequency (r = -0.086), lexical diversity (r = -0.07), and document diversity (r = -0.208).

Next, to better understand whether this pattern of relationships

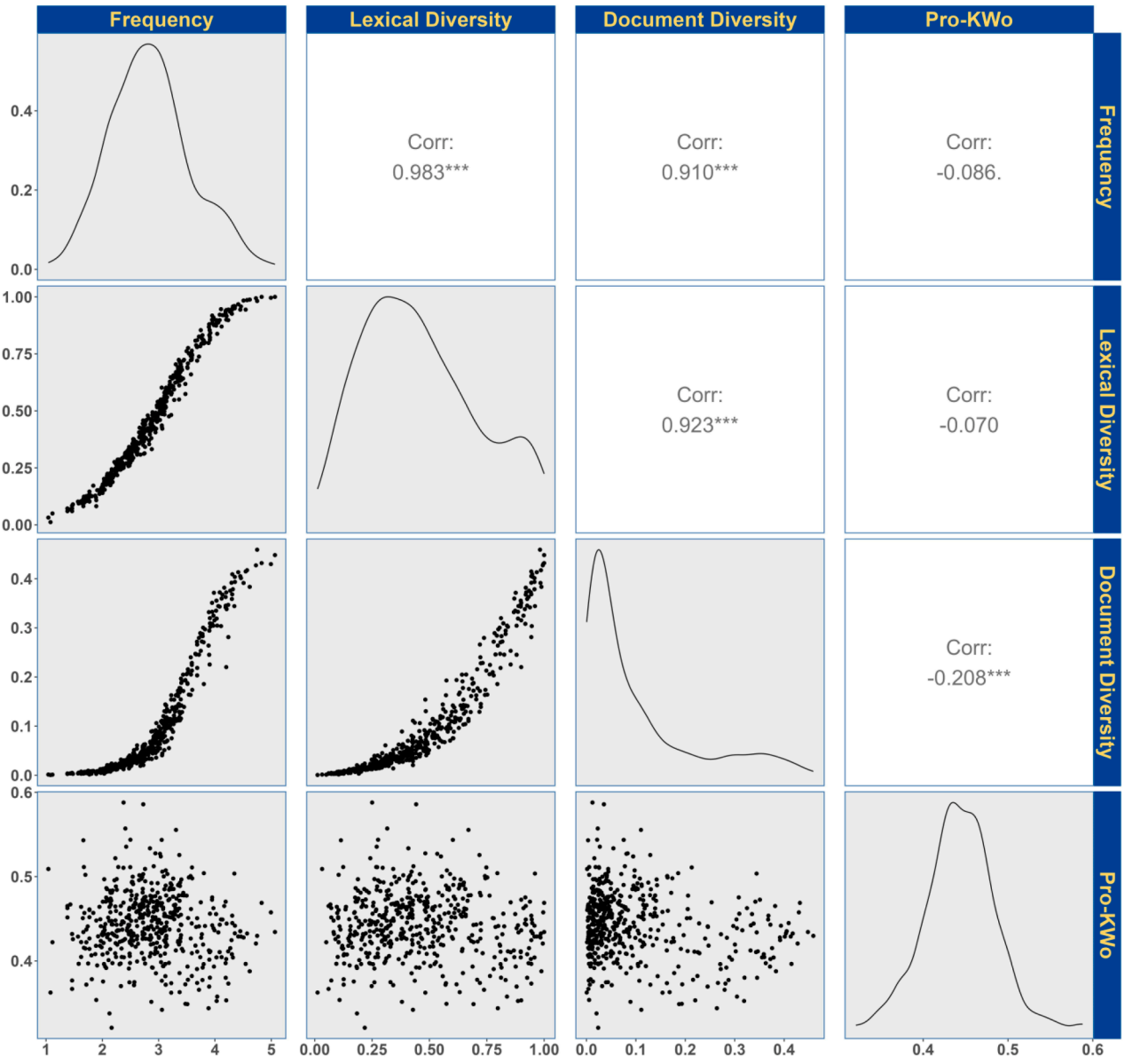

Fig. 1. Correlogram of all distributional statistics at 24 months. Frequency is the $\log_{10}$ transformed cumulative frequency. The symbols *, **, and *** in the correlogram indicate the level of statistical significance of the correlation coefficients (p-values less than 0.05, 0.01, and 0.001 respectively).

among our measures was consistent across development, we examined the correlations between our predictor variables at each individual age (16–30 months). For readability, the correlations for ages 18, 21, 24, 27, and 30 months are shown in Table 2, the rest are available in our online supplemental materials. Across age groups we found that the magnitude of correlation coefficients among Pro-KWo and each of the other predictors was small. Despite being calculated with the same language corpus (the CHILDES corpus), Pro-KWo is generally not correlated with and is therefore likely accounting for different sources of variability than the other distributional predictors.

Next, we were interested in whether the five predictors showed a relationship to the proportion of children who produced a word (MCDIp) at each age group. Scatterplots for ages 18, 21, 24, 27, and 30 months are shown in Fig. 2. Both frequency and lexical diversity showed a small correlation with MCDIp. The size of this correlation was relatively consistent largely across age. In contrast, Pro-KWo was moderately correlated with MCDIp at all age groups. Compared to other statistical predictors, Pro-KWo showed the strongest relationship to

Table 2
Correlation of all distributional statistics across five age groups, but computed within each age group. For example, the correlation of frequency at 18 months with Pro-KWo at 18 months is −0.12, at 21 months is −0.08, at 24 month is −0.09, etc. Bolded signifies the correlation is significant at the 0.01 level (2-tailed).

	Frequency					Lexical Diversity					Document Diversity				
	18	21	24	27	30	18	21	24	27	30	18	21	24	27	30
Lexical Diversity	**0.97**	**0.97**	**0.98**	**0.98**	**0.98**										
Doc. Diversity	**0.89**	**0.89**	**0.91**	**0.91**	**0.91**	**0.94**	**0.94**	**0.92**	**0.91**	**0.89**					
Pro-KWo	**−0.12**	−0.08	**−0.09**	−0.05	−0.05	**−0.12**	−0.07	−0.07	−0.04	−0.05	**−0.21**	**−0.18**	**−0.21**	**−0.17**	**−0.17**

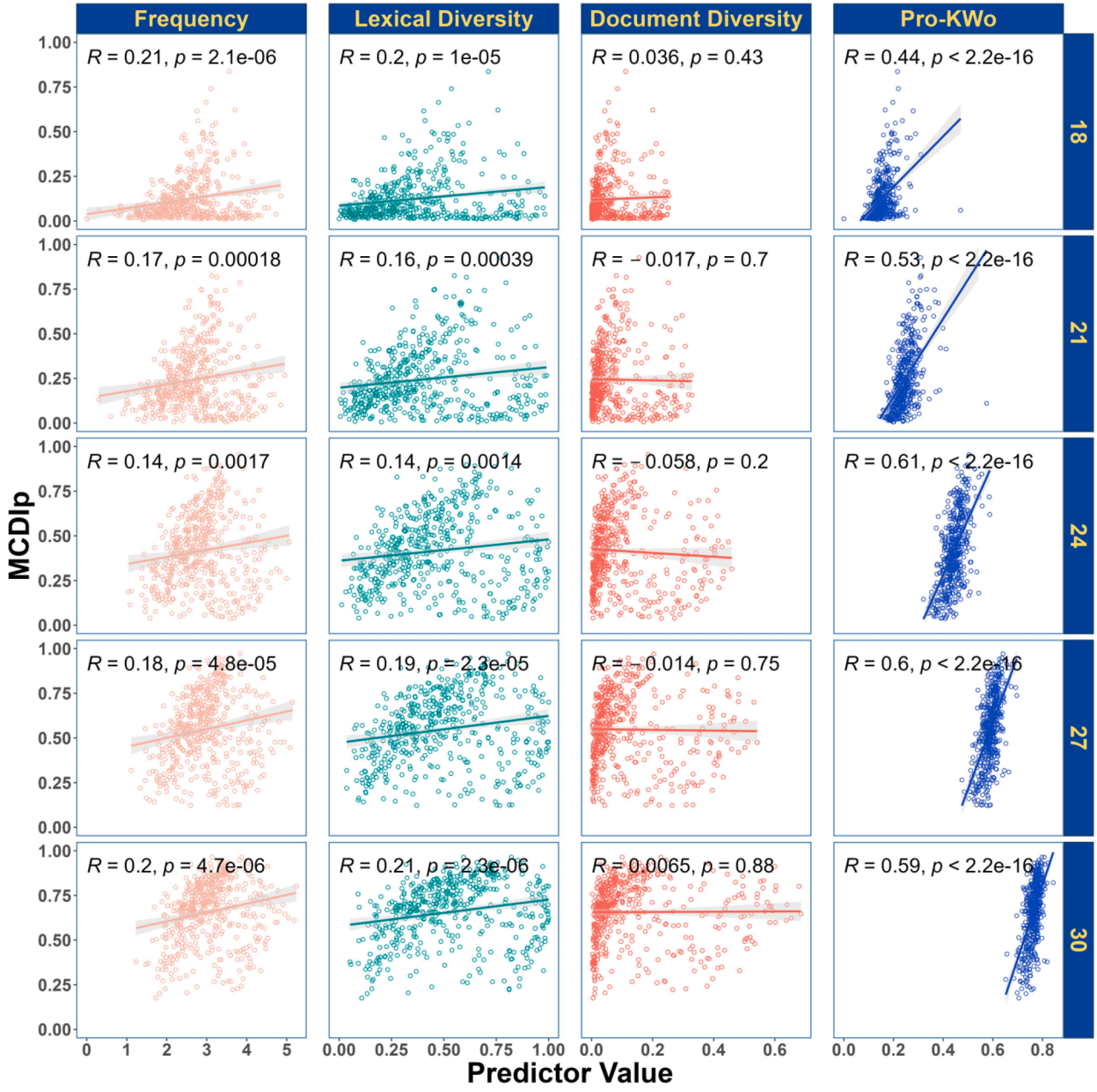

Fig. 2. Correlation between each distributional statistic and MCDIp across age groups. Frequency is the $\log_{10}$ transformed cumulative frequency. Pro-KWo shuffle is not depicted (since each 1000 random simulations created a different distribution), but the mean correlated at each age group near zero and was not significant.

MCDIp, with its effect increasing across age groups. We did not find any significant relationship between MCDIp and document diversity across any age group. Pro-KWo Shuffle correlations (-0.04, −0.07, −0.04, −0.01, and −0.03 for the ages 18, 21, 24, 27, and 30 months respectively) were not significant at any age. The nonsignificant correlation of MCDIp with Pro-KWo shuffle demonstrates that the predictive performance of Pro-KWo is not due to it using MCDI scores as a component of its score, and was not due to age-specific changes in the general distributional of the scores. Because of Pro-KWo Shuffle's effectively zero correlation with MCDIp, we did not consider it in any further analyses.

We also examined the relative stability of each predictor and MCDIp by looking at its correlation with itself across age groups. That is, is a word with a high frequency score at 18 months also high at 30 months? In order to better understand whether the predictors capture similar variance across words at each age, we examined the correlations within our predictor variables across 5 age groups (Table 3). Correlations for

Table 3
Correlation within each distributional predictor across age (beginning at 18 months). All values shown are significant at the 0.01 level (2-tailed).

Age	Frequency	Lexical Diversity	Document Diversity	Pro-KWo
18	1.00	1.00	1.00	1.00
21	0.99	0.99	1.00	0.93
24	0.98	0.98	1.00	0.84
27	0.97	0.97	0.99	0.84
30	0.97	0.96	0.99	0.71

Frequency, Lexical Diversity and Document Diversity correlations remained high across all age groups. For Pro-KWo, we observe a consistent trend such that Pro-KWo score correlations decrease as the distance between ages increases. Pro-KWo is a weighted co-occurrence measure, where the weighting value (MCDIp) reflects the proportion of children who produce a word at each age, so as the composition of

known versus unknown words changes across normative child vocabularies, Pro-KWo changes as well. This is suggestive evidence that Pro-KWo measure is picking up changes across time that are not captured, or even capturable, by frequency or contextual diversity.

Predicting MCDI "Produces" data with distributional predictors across age

Up to this point we have examined how aggregate measures of children's vocabulary knowledge (MCDIp) relate to each of our predictor variables, as well as how each predictor variable correlates with itself across ages. Next, we were interested in seeing how well each of the four distributional measures predicted individual child production scores, and how resilient these predictors were across age. To do this, first we created separate mixed effects logistic regression models for each of our four predictors with age as an additional fixed effect, and child and word as random effects. These models are shown in Table 4.

As we can see from models in Table 4, all four distributional predictors were robust to age, but also interacted with age. In other words, all predictors still significantly predicted the binary "produced" data, even after accounting for variance due to age. But all four predictors interacted with age, showing us that the effect of the different distributional predictors were different at different ages.

Predicting MCDI "Produces" data with single distributional predictors within each age

Due to the interaction of each predictor with age (and because of a pre-analysis interest in the change in the distributional statistics' predictive power at different ages), we also fit separate multilevel logistic regression model using each of our four distributional statistics as predictors (standardized and centered) at each age. As in the previous analysis, these models with each child's individual binary production data as the outcome variable, had one distributional predictor as a fixed effect, and had child and word as random effects. We fit these models for all ages from 16 to 30 months, but for brevity we show the results from months 18, 21, 24, 27, and 30. In Fig. 3 we show individual fixed effect estimates for each predictor, where each point represents a separate multilevel model. Table 5 shows the full model results for each of these five models.

Fig. 3 demonstrates the nature of the age × distributional predictor interaction found in the previous analysis. Document diversity was found to not be a significant predictor across any age group. In contrast, the effect of (log10 cumulative) frequency and lexical diversity were significant across all age groups (higher frequency and higher lexical diversity predicting higher likelihood children produce a word). The effect of Pro-KWo stands out dramatically compared to the other predictors. The effect size was positive (words that co-occurred with more already-known words were more likely to be produced), significant at all ages, and got considerably stronger as children got older. The effect of Pro-KWo was also much larger than for the other predictors, with a log-odds ratio of between 4.50 and 8.74, compared to 0.26 to 0.37 for frequency, 0.22 to 0.38 for lexical diversity, and −0.12 to −0.02 for document diversity. Translated into (slightly) more everyday language, this means that a one standard deviation increase in a word's Pro-KWo score was associated with a model's prediction of the binary "produces" variable going up by about 0.989. A one standard deviation change in a word's frequency, lexical diversity, and contextual diversity were associated with a model's prediction of the binary "produces" variable going up by about 0.565, 0.554, and 0.470, respectively.

Table 4
Parameter estimates for two predictor models, using the model: Produced ~ Age * Predictor + (1|Subject) + (1|Word) for each predictor. For all models, random intercepts of participants and words are included. Diversity and Pro-KWo were calculated using all words from the MCDI (minus the exclusions noted in the Methods), regardless of the word's grammatical class.

Factor	LogOdds	SE	z	P(>|z|)	2.5 %	97.5 %
Age	0.44	0.006	72.01	0.001	0.43	0.46
Frequency	0.26	0.017	15.32	0.001	0.22	0.29
Age * Frequency	0.02	0.000	30.00	0.001	0.01	0.02
Age	0.43	0.006	68.75	0.001	0.41	0.44
Lexical Diversity	0.47	0.017	27.04	0.001	0.44	0.51
Age * Lexical Diversity	0.01	0.000	22.32	0.001	0.01	0.01
Age	0.47	0.006	73.79	0.001	0.46	0.49
Document Diversity	−0.31	0.059	−8.57	0.001	−0.38	−0.24
Age * Document Diversity	0.02	0.001	12.89	0.001	0.01	0.02
Age	0.20	0.006	21.61	0.001	0.18	0.22
Pro-KWo	1.49	0.034	43.15	0.001	1.49	1.56
Age * Pro-KWo	−0.04	0.003	−11.53	0.001	−0.04	−0.03

Predicting MCDI "Produces" data with Pro-KWo plus the other predictors

In addition to comparing single predictor models to each other, we also fit mixed effect models that included the three other examined predictors alongside the Pro-KWo measure. Due to the high correlation among the predictor variables of frequency, lexical diversity and document diversity (as shown in Table 2), we did not fit a model with all predictors. These two-predictor models (Table 6) show that Pro-KWo is a robust predictor of word production and accounts for more and unique variability compared to frequency, lexical and document diversity .

Effects of grammatical class

In many statistical models of word learning, the effect of distributional statistics on word knowledge has differed based on grammatical class, with smaller (and even negative) effects found when examining across all words, and larger positive effect sizes when examining words within a specific grammatical class. We were interested in whether we would find similar effects within and between grammatical classes for Pro-KWo, or if this measure would be robust to the grammatical category. In computations of each measure, grammatical class was not used in any way to calculate the distributional predictors. Our approach reflects an agnostic position regarding the ways in which children categorize words (if at all) using grammatical categories. Thus, the co-occurrences between words used for Lexical Diversity and Pro-KWo were calculated using all words from the MCDI, regardless of their grammatical class.

First, we examined the same age 24-month dataset from Fig. 1, and calculated the relationships between our four predictors, both across all and within each grammatical class (for nouns, verbs, adjectives, and function words, the four grammatical classes represented from words on the MCDI). These results are shown in Fig. 4. Measures of frequency, document and lexical diversity were highly correlated with each other both across all words and within each grammatical class. This was not the case when examining the relationship between Pro-KWo and the other distributional statistics. When aggregating across all grammatical classes, as before (as in the first analysis) there was only a weak relationship between Pro-KWo and frequency (r = -0.086), lexical diversity (r = -0.070), and document diversity (r = -0.208). Within grammatical classes, these correlations tended to go up, but stayed relatively low, with the highest being between Pro-KWo and Lexical Diversity for function words (r = 0.341).

The low correlation across grammatical class suggests that whereas in aggregate Pro-KWo and the other three predictors predicted different variance in aggregate, when broken down by category, Pro-KWo and other predictors do share some sources of variance. Though nothing like the overlap of the other three predictors with each other (which were all

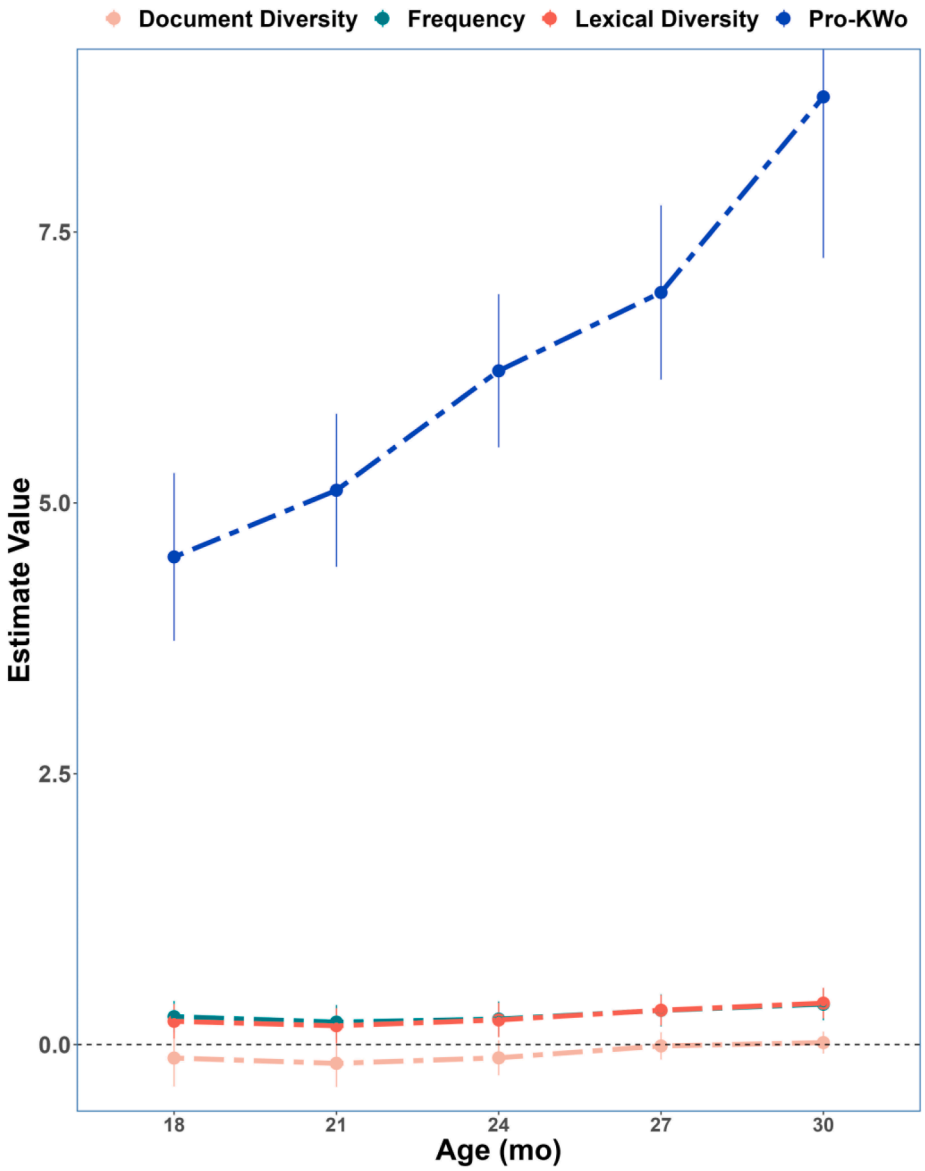

Fig. 3. Fixed effect estimates for single predictor models. Each point represents a single model with 95 % confidence intervals around each estimate.

r = 0.84 and above, both within and between grammatical classes). In essence, something about what makes nouns, verbs, adjectives, and function words different from each other such that frequency, lexical and document diversity only account for variability *within* class, is already incorporated into the Pro-KWo measure. The sets of words, especially known words, that items in different grammatical classes co-occur with seems to systematically vary by grammatical class in important ways that leads Pro-KWo, but not the other measures, to be a robust predictor across grammatical class. These findings suggest that there is still some work to do to understand how prior knowledge interacts with the learning of words of specific grammatical classes. Further it suggests that Pro-KWo as a measure of the quality of linguistic contexts that children hear may not be fully independent of other predictors, as there are small relationships between Pro-KWo and the other four distributional predictors when considering grammatical class.

Next, we examined each of our distributional predictors' relationship to MCDIp scores, both across and within grammatical category. Fig. 5 illustrates the relationship of the four statistical predictors with MCDIp at 24 months, with different colors depicting words from different grammatical categories. Fig. 6 illustrates the Pearson correlation coefficient of each of our statistical predictors with MCDIp across age groups, with different colors depicting different grammatical categories. In Fig. 5, words of the same grammatical class tend to cluster together for frequency, lexical diversity and document diversity, but not for Pro-KWo, where scores are more homogeneously distributed across a word's grammatical class. The difference in correlation by grammatical class is particularly evident in Fig. 6, where the magnitude of the correlation varies by grammatical class (and is somewhat consistent across age). However, while for frequency, lexical diversity and document diversity, correlations are substantially lower when all grammatical classes are aggregated, for Pro-KWo the correlation remains high when aggregating across grammatical class.

Across age groups, a consistent pattern emerges such that measures of frequency, document and lexical diversity show only a small correlation with MCDIp across all words. However, when considering grammatical categories, we see a marked increase in correlation values. This increase is once again most noticeable for nouns, which perform much better on all three measures when calculated within grammatical class. Notably, the correlation for document diversity demonstrates a classic Simpson's paradox, flipping from its negative correlation across all words to a positive correlation within each grammatical category. This makes sense: function words and semantically light verbs are some of

Table 5
Parameter estimates for single predictor models at each age, using the model Production ~ Predictor + (1|Subject) + (1|Word) for each predictor. For all models, random intercepts of participants and words are included.

Age	LogOdds	SE	z	P(>|z|)	2.5 %	97.5 %
Frequency						
18	0.257	0.074	3.471	0.001	0.112	0.403
21	0.207	0.080	2.567	0.010	0.049	0.365
24	0.235	0.083	2.817	0.005	0.071	0.398
27	0.316	0.078	4.181	<0.001	0.167	0.463
30	0.373	0.076	4.490	<0.001	0.224	0.523
Lexical Diversity						
18	0.215	0.081	2.655	0.008	0.056	0.374
21	0.174	0.082	2.113	0.035	0.013	0.335
24	0.228	0.081	2.815	0.005	0.069	0.386
27	0.317	0.072	4.405	<0.001	0.176	0.458
30	0.383	0.072	5.303	<0.001	0.241	0.524
Document Diversity						
18	−0.124	0.134	−0.928	0.354	−0.386	0.138
21	−0.173	0.111	−1.558	0.119	−0.391	0.045
24	−0.122	0.082	−1.481	0.139	−0.283	0.039
27	−0.013	0.064	−0.196	0.845	−0.138	0.113
30	0.019	0.052	0.359	0.720	−0.083	0.120
Pro-KWo						
18	4.501	0.395	11.395	<0.001	3.727	5.276
21	5.116	0.360	14.197	<0.001	4.410	5.823
24	6.219	0.361	17.242	<0.001	5.512	6.926
27	6.942	0.410	16.913	<0.001	6.137	7.746
30	8.747	0.758	11.533	<0.001	7.261	10.234

Table 6
Parameter estimates for two predictor models at each age, using the model Production ~ Predictor + ProKWo + (1|Subject) + (1|Word) for each predictor. For all models, random intercepts of participants and words are included.

Age	LogOdds	SE	z	P(>|z|)	2.5 %	97.5 %
18 Months						
ProKWo	4.76	0.34	13.89	<0.001	4.08	5.43
Freq	0.36	0.07	5.48	<0.001	0.23	0.49
ProKWo	4.72	0.35	13.34	<0.001	4.03	5.41
LD	0.33	0.07	3.08	<0.001	0.18	0.47
ProKWo	4.61	0.33	13.72	<0.001	3.95	5.23
DD	0.16	0.11	1.42	0.156	−0.06	0.40
21 Months						
ProKWo	5.23	0.34	15.15	<0.001	4.55	5.90
Freq	0.28	0.07	4.24	0.001	0.15	0.42
ProKWo	5.21	0.36	14.74	<0.001	4.50	5.91
LD	0.25	0.06	3.62	<0.001	0.12	0.39
ProKWo	5.17	0.36	14.49	0.000	4.46	5.86
DD	0.07	0.10	0.75	0.543	−0.11	0.26
24 Months						
ProKWo	6.37	0.33	19.09	<0.001	5.72	7.03
Freq	0.33	0.07	5.17	<0.001	0.21	0.47
ProKWo	6.33	0.31	20.18	<0.001	5.72	6.95
LD	0.31	0.06	4.86	<0.001	0.18	0.43
ProKWo	6.35	0.32	19.47	<0.001	5.71	6.99
DD	0.12	0.07	1.78	0.073	−0.01	0.25
27 Months						
ProKWo	7.07	0.38	18.57	<0.001	6.32	7.81
Freq	0.37	0.06	6.25	<0.001	0.25	0.48
ProKWo	7.01	0.39	18.11	<0.001	6.25	7.77
LD	0.34	0.06	6.14	<0.001	0.24	0.46
ProKWo	7.13	0.41	17.22	<0.001	6.32	7.95
DD	0.14	0.05	2.74	0.040	0.04	0.24
30 Months						
ProKWo	8.94	0.42	19.34	<0.001	8.03	9.85
Freq	0.43	0.06	7.06	<0.001	0.30	0.54
ProKWo	8.89	0.41	21.59	<0.001	8.08	9.69
LD	0.42	0.06	7.28	<0.001	0.31	0.53
ProKWo	9.03	0.89	10.09	<0.001	7.28	10.79
DD	0.13	0.04	3.07	0.021	0.04	0.22

the words with the highest document diversity scores as a class, and are some of the last words children say. However, of the function words, the ones with the highest document diversity scores are the ones said earlier.

In contrast, Pro-KWo shows a qualitatively different pattern than the other predictors and maintains an overall consistent relationship with MCDIp. The only exception to the consistent relationship between Pro-KWo and MCDIp within and across grammatical class is that Pro-KWo shows a smaller (though still significant) correlation when examining only verbs. In fact, verbs appear to be accounting for the increase in the predictability of ProKWo across ages. Pro-KWo's much higher consistency across grammatical classes is an important difference between Pro-KWo and other statistical predictors of word learning. Most other models of word learning find stronger effects within grammatical categories than across all grammatical categories, but Pro-KWo shows no similar boost. It is as strong a predictor within a single grammatical category and across all words.

To better understand the nature of the predictiveness of Pro-KWo across grammatical class, we dug deeper into the prediction error across items of Pro-KWo. We fit logistic regression models predicting binary (produced/not produced) vocabulary outcomes with Pro-KWo as the single predictor (Fig. 7) at 24 months. We then computed the prediction error made by the model for each item (word). Then we correlated the model's prediction error of each word with the aggregate measure of vocabulary knowledge to yield the relationship between the degree of prediction error relative to the proportion of children who produced a particular item. Fig. 7 shows that the prediction error of each grammatical class is largely overlapping, suggesting that Pro-KWo is not systematically under- or over-estimating production likelihood of grammatical classes. However, it is also clear from Fig. 7 that there is some clustering by grammatical class. Some nouns are clustered above the regression line at the top while some function words are clustered at the bottom. So Pro-KWo, while relatively more robust to grammatical class than other distributional predictors, still shows some evidence of small effects of grammatical class. Pro-KWo is slightly *under*-estimating the likelihood that children produce some nouns and *over*-estimating the likelihood that children produce some function words. Of course, it is not clear if this over and underestimation is an effect of grammatical class per se, or if Pro-KWo is broadly underestimating the most frequently produced words and overestimating the least frequently produced words, which happen to be nouns and function words respectively. Or perhaps the effect of Pro-KWo should be modeled non-linearly to best account for extreme values. Nevertheless, we find that Pro-KWo retains some degree of sensitivity to grammatical class, though far less than other distributional predictions.

General discussion

In the current work, we aimed to incorporate insights from behavioral findings into statistical models of language learning. Specifically, we aimed to incorporate the role of linguistic quality of word learning contexts, with an example of quality being defined as linguistic contexts that allow learners to leverage their existing lexical knowledge to learn new words. We also hoped to resolve issues related to the divergent

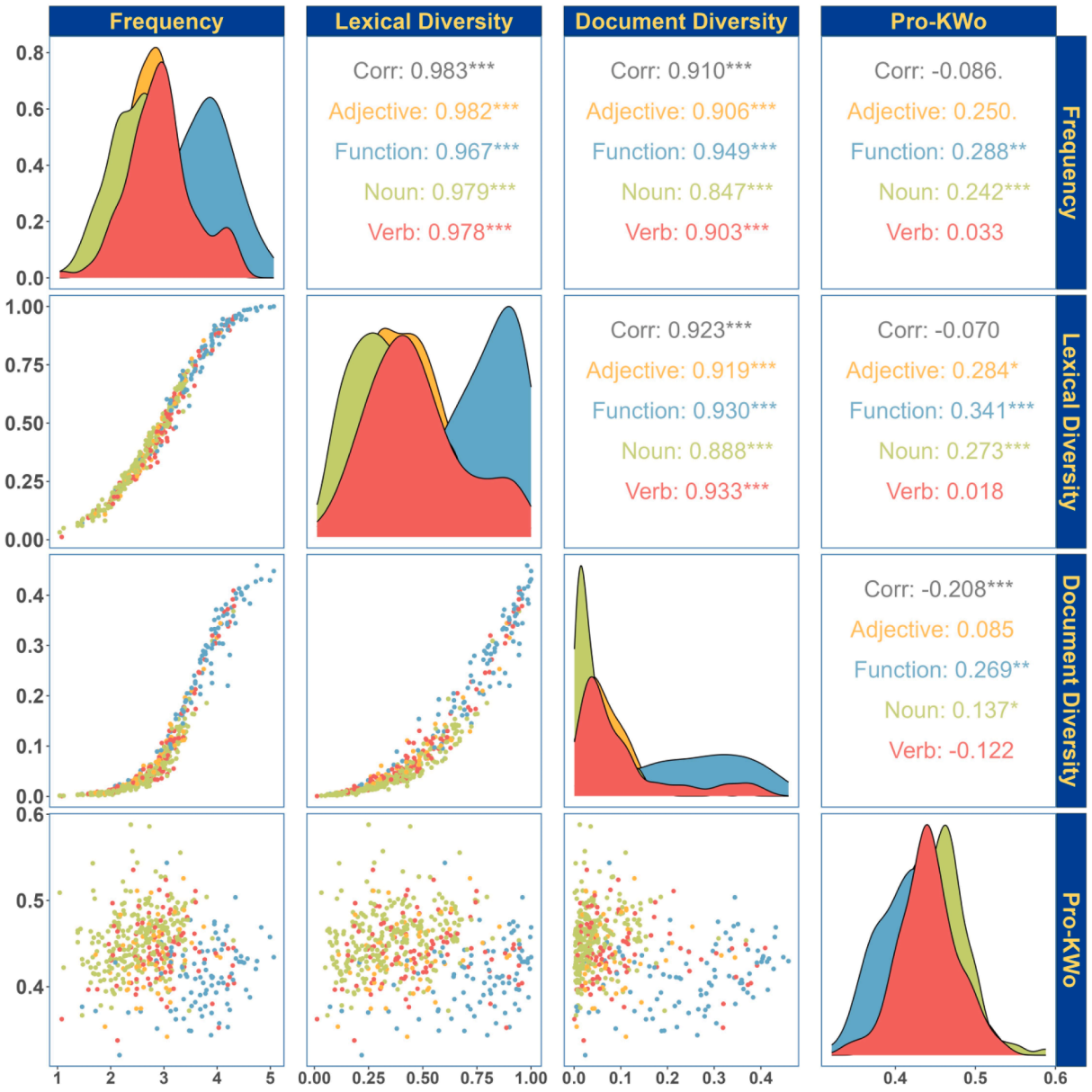

Fig. 4. Correlogram of all distributional statistics at 24 months within grammatical class. Frequency is the $\log_{10}$ transformed cumulative frequency. Note: The symbols *, **, and *** in the correlogram indicate the level of statistical significance of the correlation coefficients (p-values less than 0.05, 0.01, and 0.001 respectively).

predictions about vocabulary development from words of different grammatical classes.

To accomplish these goals, we introduced a new metric (Pro-KWo) a distributional predictor that quantifies the likelihood that children know the words with which a word co-occurs, indexing the degree the word occurs in felicitous distributional contexts. We found that this measure accounts for more variability in word learning than other distributional measures, independent variance from those measures, and crucially accounts for variability both within and across grammatical class. An important question is what exactly Pro-KWo might be capturing that allows it to robustly predict word knowledge across grammatical class. We first review how our findings compare with previous work, and then return to the literature reviewed in our introduction and discuss how our Pro-KWo measure contributes to our understanding of each.

Models of child vocabulary development

How do our findings compare with previous research modeling child vocabulary development? Hills et al. (2010) used network growth model analyses (where network connections were defined as whether words co-occurred in child-directed speech) to try to predict child vocabulary development. They tested whether newly learned words could be best understood as words that shared a lot of connections to known words in general ("lure of the associates"), words that shared a connection to specific hub words in the existing knowledge network ("preferential attachment") or words that shared a lot of connections to words in the language environment, regardless of whether they are known or not ("preferential acquisition"). They found support for the "lure of the associates" hypothesis for words overall and nouns, but found support for the "preferential acquisition" for verbs and function words, and found that none of the hypotheses predicted adjectives specifically. In general, the finding was that words that shared many co-occurrence connections to other words were more likely to be learned earlier (with some inconsistency on the importance of whether children already produced those co-occurring words).

Our results - using a very different methodology (logistic regression

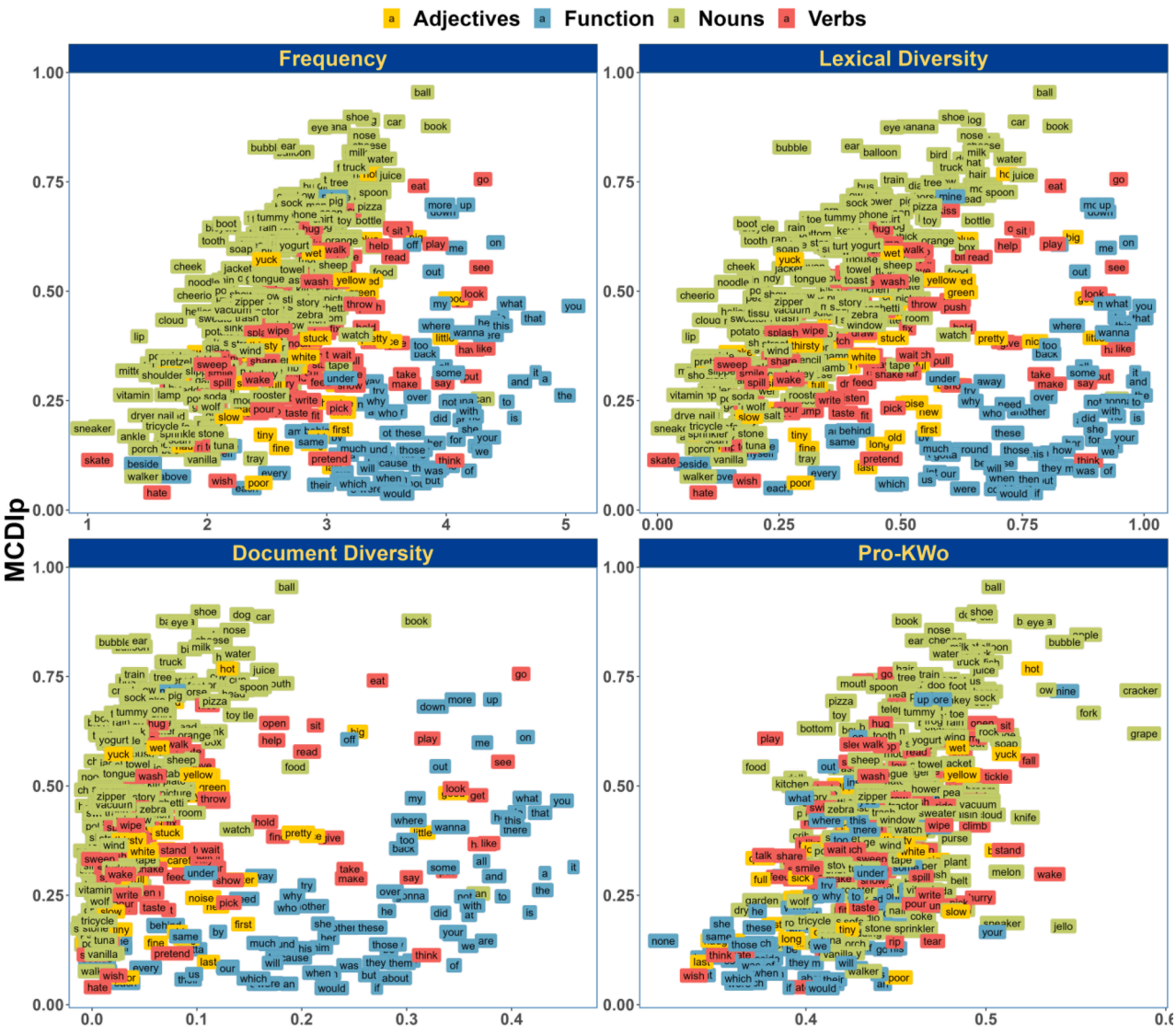

Fig. 5. Correlation of MCDIp and Distributional Statistics at 24 months.

of individual children's production responses across all ages, rather than network growth based on adding words when 50 % of children produce a word) - have some similarities but also many differences to Hills et al.'s findings. In our models, lexical diversity was most like the "lure of the associates" hypothesis. Like Hills et al., we found that lexical diversity had a small to moderate effect for predicting if children say words overall (around r = 0.20), but that this was masked by extremely high between-grammatical category variance. Lexical diversity had a much larger effect for nouns (around r = 0.70). However, it is also important to note that these results were almost exactly the same as our observed results for word frequency, which as always was extremely highly correlated with lexical diversity r = 0.983). As such it is very hard to tell how much of this effect is truly an effect of lexical diversity versus just an effect of frequency.

The results for our newly introduced Pro-KWo measure differed substantially from Hills et al. findings. We found that Pro-KWo had a high correlation with produced words across all words (about r = 0.60, with much less variation by grammatical class, though lower for verbs, r = 0.15 to r = 0.40 across ages). Why did Pro-KWo perform so differently? The Pro-KWo measure can be thought of as a descendent of the contextual diversity, in ways that make it a blend of the "lure of the associates" and "preferential acquisition" models. One way to think about it is that it is a ratio of the two: how many words does a word co-occur with that the child already knows (lure of the associates), relative to the number of words it co-occurs with overall (preferential acquisition). This may be an important difference from either of the two calculated independently. The ratio of known-word co-occurrences can be thought of as providing a confidence estimate on the meaning of a word. A child might have heard a word co-occurring with many different words. But if the child doesn't know or yet produce the words with which it is co-occurring, then the child may know that they don't yet have a good estimate of the word's meaning or proper use, and thus may be less likely to produce the word.

Pro-KWo also differs from "lure of the associates" and "preferential acquisition" in two other ways. First, it uses the words' co-occurrence counts, rather than just a binary measure of whether they co-occur at all. This means that the ratio being computed is driven more heavily by the frequent words in the learning environment. Second, it uses MCDIp measures in a quantitative-weighted manner, rather than in a binary "add them to the network or not" manner. The multiplicative nature of these two factors can mean that a word that is either low frequency (even if well known) or not well known (even if high frequency) will not contribute much to a Pro-KWo score. A word will tend to get a high Pro-KWo score to the extent that it is co-occurring with many frequent words that a child already produces (which actually gives the Pro-KWo model a little bit of the "preferential acquisition" approach as well).

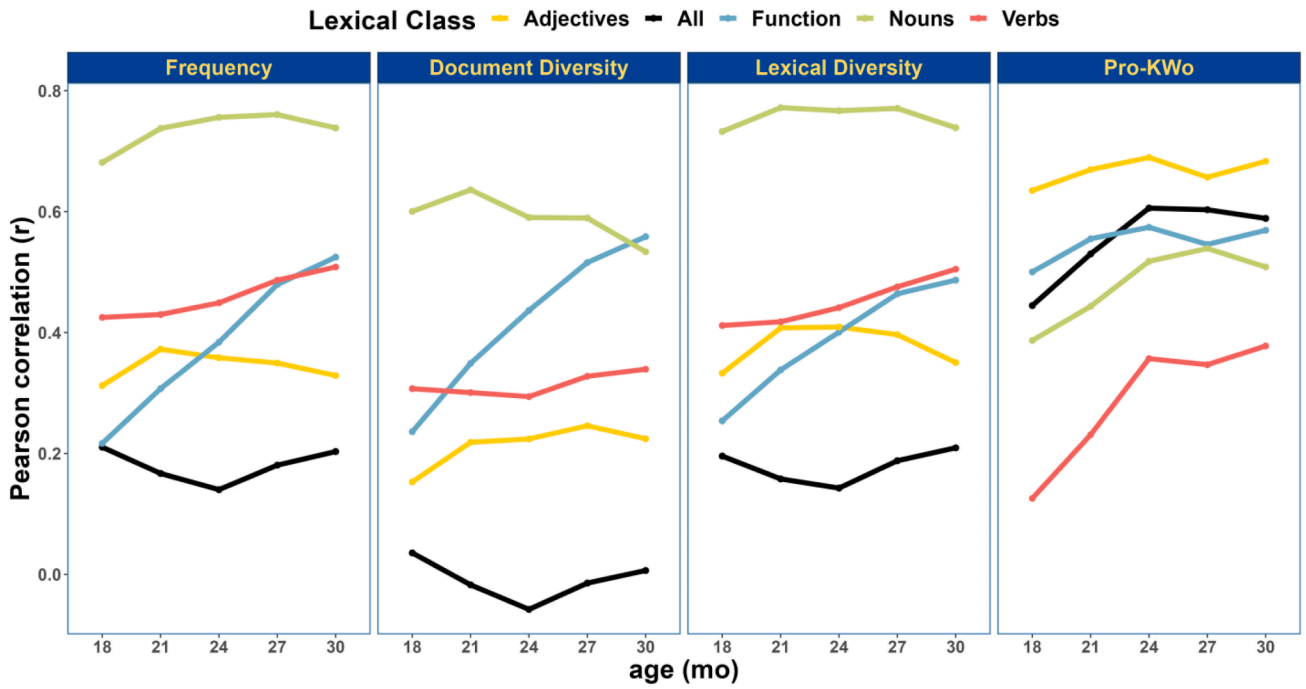

Fig. 6. Correlation of MCDIp and distributional statistics across age for each word's grammatical class.

Another interesting comparison to our work is research by Siew and Vitevich (2020). They found that sparser phonological neighborhoods are associated with earlier ages of acquisition. At first glance, this result may seem to be inconsistent with our findings, which generally support the notion that denser connectivity structure supports word learning. However, considered in the broader context of research looking at semantic versus phonological effects on lexical processing, these results are quite compatible. In adults, neighborhood phonological density is associated with an increase in recognition difficulty (Luce & Pisoni, 1998; McClelland & Elman, 1986) but a decrease in production difficulty (Dell, 1986; Vitevitch, 2002; Vitevich & Sommers, 2003; Gahl, Yao & Johnson, 2012). Likewise, other work with adult picture naming times suggests that different aspects of semantic density are associated with an increase or decrease in naming times. For example, semantic richness (a greater number of semantic features) is associated with shorter naming latencies while semantic density (degree to which features are shared across multiple entities; McRae et al., 2005) is associated with longer naming latencies (Rabovsky, Schad & Abdel Rahman, 2010). There is no single effect of neighborhood density on behavior. Instead, the content being represented (such as phonological versus semantic information) and the cognitive process being modeled (comprehension versus production) may both influence whether density is good or bad.

Quantity, Quality, and prior knowledge

The dichotomy of quantity and quality has been proposed in the behavioral word learning literature to reflect a distinction between the amount of speech children hear (Hoff, 2003; Huttenlocher et al., 1991; Huttenlocher et al., 2010; Weisleder & Fernald, 2014) and contexts that impart a particularly rich language learning opportunity (Tomasello & Todd, 1983, Tomasello, 1988; Akhtar et al., 1996; Yu, Suanda & Smith, 2019). In contrast, while statistical approaches have proposed measures of quantity of speech (i.e., word frequency), measures of the quality of speech have yielded small effects. In the current work we suggested a quality measure that links statistical approaches to well established behavioral findings.

Our proposed Pro-KWo measure leverages findings of quality language episodes that emphasize the role of prior knowledge. For instance, Pro-KWo may aid word learning through a bootstrapping process in which unknown words that co-occur with many known words within the same sentence frame (e.g., "The funny cat *plunked* the toy") are easier to learn. Such bootstrapping accounts are pervasive in the behavioral word learning literature (Fisher et al., 2010; Markman & Watchel, 1988; Yu & Smith, 2007).

Pro-KWo may also be capturing the propensity of caregivers to finely tune speech to children in a manner that is sensitive to their lexical knowledge. This has been found in studies where mother's utterances are recorded during play sessions with infants. For example, Masur (1997) found that mother's prioritize naming novel objects while in the presence of both familiar and comprehended objects. More recent investigations examine the extent to which parents attune their speech to children's vocabulary knowledge. In an experiment where parents and infants jointly engaged in a referential task, parents were shown to guide their infant to the correct referent by providing helpful information according to their estimates of the infant's vocabulary knowledge. Further, parents modulated their speech in instances where their initial assumptions of which words the infant knows were incorrect (Leung, Tunkel & Yurovsky, 2021). In both instances caregivers provide infants with language experience that is dynamically changing according to their understanding of the child's lexical knowledge. Pro-KWo provides for a proxy of such instances by utilizing aggregate measures of children's existing knowledge to differentially weight language experience, and while the above referential studies were primarily conducted with nouns (e.g., stuffed animal toys and cartoon animals) caregivers may similarly craft their speech to children according to what children already know across all grammatical classes.

Here it is important to note that in our current implementation, prior knowledge is defined as an aggregate measure of children's productive vocabulary (MCDIp). Since our measure of prior knowledge is a coarse metric, one cannot extrapolate which of the many learning mechanisms Pro-KWo may be capturing, but we do know that this initial knowledge base provides the foundations for future word learning. Thus, Pro-KWo can be compared to measures such as word frequency in that frequency captures the relevant importance of quantity of speech and Pro-KWo captures language quality. There are many other ways in which quality could be operationalized in analyses of naturalistic language data

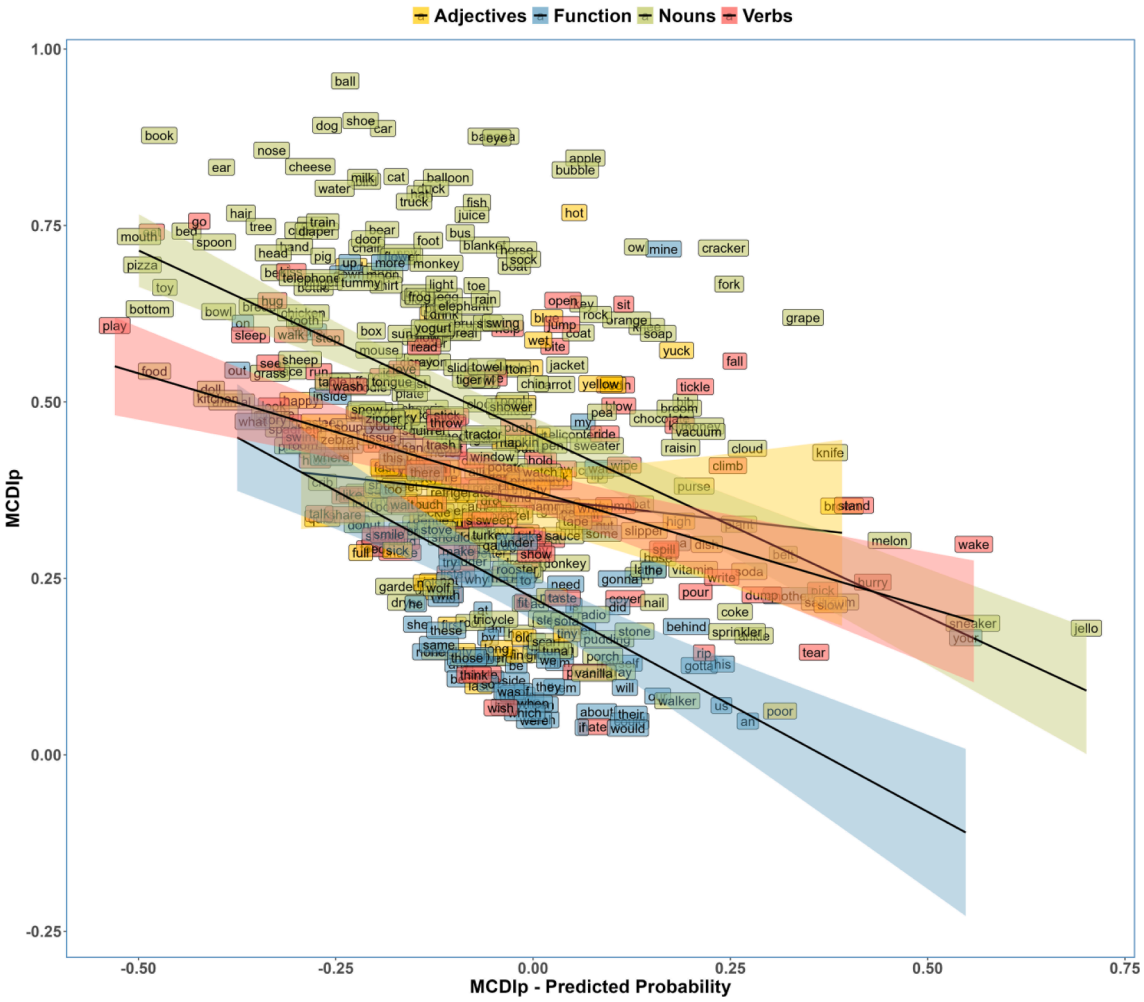

Fig. 7. Correlation of MCDIp with predicted model probabilities at 24 months. Positive values along the x-axis represent over predictions, while negative values represent under predicted values.

(Goldenberg, Repetti & Sandhofer, 2022; Hoff, 2006; Huttenlocher et al., 2010; Meredith & Catherine, 2020; Tamis-Lemonda, Kuchirko & Song, 2014) and future statistical models of word learning may define quality in other ways, inspired both by naturalistic recordings and laboratory experiments of early language learning.

Grammatical class

The behavioral word learning literature has emphasized inductive learning constraints (Markman & Wachtel, 1988; Smith et al., 2002), domain-general learning mechanisms (Carey & Bartlett, 1978; Yu & Smith, 2007; Trueswell et al., 2013) and socio-pragmatic factors (Tomasello, 1988; Akhtar et al., 1996; Yu & Ballard, 2007) that influence word learning across grammatical classes. In contrast, statistical analyses of naturalistic word learning data have found that distributional predictors of vocabulary outcomes perform best within but not across grammatical class (Goodman et al., 2008; Hills et al., 2009). The finding that knowledge of a word's grammatical class was needed in order to predict learning outcomes has implied to some researchers that frequency, lexical and document diversity exert different effects on words of different grammatical classes. This has led some to suggest potentially different mechanisms for different grammatical classes (Hills et al., 2010). Alternatively, it could mean that words from different grammatical classes might need to be represented independently, have their statistics tracked separately, or have some part of the learning system know what class a word is from when using distributional statistics.

In the current work, Pro-KWo was shown to be a robust predictor of productive vocabularies not only within each grammatical class but also across all words aggregated together. Rather than posit different effects of Pro-KWo by grammatical class, we found that this measure was uniformly predictive, suggesting that the underlying mechanism or mechanisms that are indexed by Pro-KWo may also act uniformly across grammatical class, removing the need to posit a different learning mechanism or differential representation for words from different grammatical classes. Previous differences that were attributed to grammatical class per se may indeed be an emergent property the way that different grammatical classes vary in terms of the proportion of their co-occurrences are with known words.

What exactly then, is the relationship between Pro-KWo, grammatical class, and age of acquisition for words? It has long been understood that grammatical class is itself a predictor of the order in which words are acquired, with most languages having a strong bias to learn nouns earlier, then verbs and adjectives, and last function words (Gentner, 1982). Many proposals about the semantic-conceptual nature of these differences have been suggested, which are nicely evaluated by Gentner (2006). Gentner argues that proposals involving maturational constraints on relational knowledge (Halford, Wilson, & Phillips, 1998) are not supported by data that verbs are also harder for second language learners later in life (Lennon, 1996), and that adults show a mapping advantage for nouns over verbs (Gillette et al., 1999). Gentner also argues that differential knowledge of the conceptual components of nouns

versus verbs, while possibly a partial explanation, is unlikely to be the full explanation since even the most concrete verbs like motion and causal verbs come after nouns in order of acquisition, long after children have demonstrated knowledge of the underlying concepts (Baillargeon & Wang, 2002; Childers & Tomasello, 2002; Gentner, 1975, Gentner, 1982; Golinkoff & Kerr, 1978; Pruden et al., 2004).

Gentner argues that the best explanation for verbs lagging nouns is an effect of a shift from focusing on object properties earlier in learning, and then later shifting to attend to relational properties of words (Gentner & Rattermann, 1991). This proposal is similar (though not identical) to a proposal by Gleitman et al. (2005), that the difficulty with more abstract words comes "not in overcoming conceptual difficulties with abstract word meanings but rather in mapping these meanings onto their corresponding lexical forms" (pg. 23). The success of the Pro-KWo measure is consistent with both hypotheses, and potentially adds to both explanations. One of the reasons why children may struggle early with relational mappings, or prefer attending to object properties rather than relational properties, or have difficulty mapping abstract properties to lexical forms, is that they just don't know enough of the words they would need to know to make use of the relational or abstract information they are being given.

Future directions and Limitations:

We have speculated that Pro-KWo may be capturing instances in which known words may bootstrap the learning of new words. This process may continue to play an important role across language development. Our analysis shows that Pro-KWo improves as a statistical predictor linearly with age (Fig. 3), supporting a "rich get richer" interpretation of word learning. However, our ability to accurately measure the effect of Pro-KWo at later age groups is constrained by the limited set of words within the MCDI. That is, by 30 months there are fewer words yet to be produced (MCDIp at 30 months: 0.66 across all words), it is reasonable to assume a wider range of words some of which children may normatively produce at ages beyond 30 months would further inform our characterization of Pro-KWo beyond the current analysis. Doing so would reveal whether Pro-KWo captures a dimension that is relevant early in word learning but later is reduced in importance as children begin to leverage new information that may be of greater use.

Another limitation involves our measure of word knowledge. In our current analysis we operationalize prior knowledge by using production values from a parental inventory assessment. These vocabulary questionnaires only provide us with a partial estimate of children's word knowledge. It is likely that children know a great deal more words than they say, thus one must consider how to appropriately weight production values when making assessments of children's vocabulary composition. It also means that there are surely production-side constraints affecting differences between what words children comprehend versus what words children say. Children understand function words long before they say them, and it could be that measures like Pro-KWo (and for that matter, frequency and contextual diversity) are much better predictors of comprehension than of production.

Conclusion

Overall, our results provide evidence that previously proposed learning mechanisms and biases which have historically focused on nouns, may extend to words of other grammatical classes. To our knowledge, Pro-KWo is the first statistical predictor of word learning that does not interact with a word's grammatical class. Pro-KWo then represents a first step in characterizing what kinds of information can be used to quantify which language experiences may be most useful for learning new words. And while the current implementation defines prior knowledge as an aggregate measure of children's productive vocabulary, it is a measure that closely approximates the quality of speech children hear in a way not previously reported. There is reason to believe more refined and targeted accounts of children's prior knowledge may be even more useful when incorporated into a distributional predictor of word learning. As has been noted in prior behavioral work on word learning there are additional factors beyond prior knowledge which account for vocabulary outcomes. Achieving a way of capturing these factors and subsequently incorporating them into a distributional statistic may provide more ways in which distributional statistics can be used to study word learning. For instance, a great deal of research has identified that language episodes in which the child and parent are both jointly attending to a referent are particularly informative and promote word learning. Finding ways of identifying episodes of joint attention from speech corpora may not be a straightforward process. However, such efforts may be worthwhile in that they begin to further increase the utility of large naturalistic datasets by adding important *meta*-information by which to weight language statistics. Further we contend that such approaches are necessary in order to increase the overall validity of distributional statistics of language learning.

CRediT authorship contribution statement

Andrew Z. Flores: Conceptualization, Formal analysis, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Jessica L. Montag:** Conceptualization, Funding acquisition, Methodology, Writing – original draft, Writing – review & editing. **Jon A. Willits:** Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

A link to all data is available in a public repository which we link in the manuscript.

References

Akhtar, N., Carpenter, M., & Tomasello, M. (1996). The role of discourse novelty in early word learning. *Child Development, 67*(2), 635–645. https://doi.org/10.1111/j.1467-8624.1996.tb01756.x

Ambridge, B., Kidd, E., Rowland, C. F., & Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language, 42*(2), 239–273. https://doi.org/10.1017/s030500091400049x

Anderson, N. J., Graham, S. A., Prime, H., Jenkins, J. M., & Madigan, S. (2021). Linking quality and quantity of parental linguistic input to child language skills: A meta-analysis. *Child Development, 92*(2), 484–501. https://doi.org/10.1111/cdev.13508

Arias-Trejo, N., & Alva, E. A. (2013). Early spanish grammatical gender bootstrapping: Learning nouns through adjectives. *Developmental Psychology, 49*(7), 1308. https://doi.org/10.1037/a0029621

Baillargeon, R., & Wang, S. H. (2002). Event categorization in infancy. *Trends in Cognitive Sciences, 6*(2), 85–93. https://doi.org/10.1016/s1364-6613(00)01836-2

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*. doi: 10.1016/j.jml.2017.01.001.

Beals, D. E. (1997). Sources of support for learning words in conversation: Evidence from mealtimes. *Journal of Child Language, 24*(3), 673–694. https://doi.org/10.1017/s0305000997003267

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences, 109*(9), 3253–3258. https://doi.org/10.1073/pnas.1113380109

Blackwell, A. A. (2005). Acquiring the English adjective lexicon: Relationships with input properties and adjectival semantic typology. *Journal of Child Language, 32*(3), 535. https://doi.org/10.1017/s0305000905006938

Booth, A. E., & Waxman, S. R. (2009). A horse of a different color: Specifying with precision infants' mappings of novel nouns and adjectives. *Child development, 80*(1), 15–22. https://doi.org/10.1111/j.1467-8624.2008.01242.x

Bortfeld, H., Morgan, J. L., Golinkoff, R. M., & Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech-stream segmentation. *Psychological Science, 16*(4), 298–304. https://doi.org/10.1111/j.0956-7976.2005.01531.x

Borovsky, A., Kutas, M., & Elman, J. (2010). Learning to use words: Event-related potentials index single-shot contextual word learning. *Cognition, 116*(2), 289–296. https://doi.org/10.1016/j.cognition.2010.05.004

Borovsky, A., & Peters, R. E. (2019). Vocabulary size and structure affects real-time lexical recognition in 18-month-olds. *PloS one, 14*(7), e0219290.

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. (2016). *August)*. From uh-oh to tomorrow: Predicting age of acquisition for early words across languages. In CogSci.

Braginsky, M., Sanchez, A., & Yurovsky, D. (2018). childesr: Accessing the'CHILDES'Database. *R package version*, (1).

Brandt, S., Diessel, H., & Tomasello, M. (2008). The acquisition of German relative clauses: A case study. *Journal of Child Language, 35*(2), 325–348. https://doi.org/10.1017/s0305000907008379

Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition, 81*(2), B33–B44. https://doi.org/10.1016/s0010-0277(01)00122-6

Byers-Heinlein, K., & Werker, J. F. (2013). Lexicon structure and the disambiguation of novel words: Evidence from bilingual infants. *Cognition, 128*(3), 407–416. https://doi.org/10.1016/j.cognition.2013.05.010

Cameron-Faulkner, T., Lieven, E., & Tomasello, M. (2003). A construction based analysis of child directed speech. *Cognitive Science, 27*(6), 843–873. https://doi.org/10.1207/s15516709cog2706_2

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Papers and Reports on Child Language Development, 15*, 17–29.

Cartmill, E. A., Armstrong, B. F., Gleitman, L. R., Goldin-Meadow, S., Medina, T. N., & Trueswell, J. C. (2013). Quality of early parent input predicts child vocabulary 3 years later. *Proceedings of the National Academy of Sciences, 110*(28), 11278–11283. https://doi.org/10.1073/pnas.1309518110

Chang, L. M., & Deák, G. O. (2020). Adjacent and Non-Adjacent Word Contexts Both Predict Age of Acquisition of English Words: A Distributional Corpus Analysis of Child-Directed Speech. *Cognitive Science, 44*(11), e12899.

Childers, J. B., & Tomasello, M. (2002). Two-year-olds learn novel nouns, verbs, and conventional actions from massed or distributed exposures. *Developmental psychology, 38*(6), 967. https://doi.org/10.1037/0012-1649.38.6.967

Christophe, A., Millotte, S., Bernal, S., & Lidz, J. (2008). Bootstrapping lexical and syntactic acquisition. *Language and Speech, 51*(1–2), 61–75. https://doi.org/10.1177/0023830908051001050l

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. https://doi.org/10.1111/cogs.12409

Cox, C. R., & Haebig, E. (2022). Child-oriented word associations improve models of early word learning. *Behavior research methods, 1–22*. https://doi.org/10.3758/s13428-022-01790-y

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological review, 93*(3), 283. https://doi.org/10.1037/0033-295x.93.3.283

Dickinson, D. K., & Tabors, P. O. (1991). Early literacy: Linkages between home, school and literacy achievement at age five. *Journal of Research in Childhood Education, 6*(1), 30–46. https://doi.org/10.1080/02568549109594820

Echols, C. H., Crowhurst, M. J., & Childers, J. B. (1997). The perception of rhythmic units in speech by infants and adults. *Journal of Memory and Language, 36*(2), 202–225. https://doi.org/10.1006/jmla.1996.2483

Elman, J. L. (1990). Finding structure in time. *Cognitive science, 14*(2), 179–211. https://doi.org/10.1207/s15516709cog1402_1

Estes, K. G., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychological Science, 18*(3), 254–260. https://doi.org/10.1111/j.1467-9280.2007.01885.x

Fennell, C. T., & Werker, J. F. (2003). Early word learners' ability to access phonetic detail in well-known words. *Language and Speech, 46*(2–3), 245–264. https://doi.org/10.1177/00238309030460020901

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., & Stiles, J. (1994). Variability in early communicative development. In *Monographs of the society for research in child development* (pp. i–185). https://doi.org/10.2307/1166093

Fenson, L. (2007). MacArthur-Bates communicative development inventories. *Paul H. Brookes Publishing Company*. https://doi.org/10.1037/t11538-000

Ferry, A. L., Hespos, S. J., & Waxman, S. R. (2010). Categorization in 3-and 4-month-old infants: An advantage of words over tones. *Child Development, 81*(2), 472–479. https://doi.org/10.1111/j.1467-8624.2009.01408.x

Fisher, C., Gertner, Y., Scott, R. M., & Yuan, S. (2010). Syntactic bootstrapping. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*(2), 143–149. https://doi.org/10.1002/wcs.17

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). Variability and consistency in early language learning: The Wordbank Project. *MIT Press*. https://doi.org/10.7551/mitpress/11577.001.0001

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. https://doi.org/10.1017/s0305000916000209

Ferguson, B., Graf, E., & Waxman, S. R. (2014). Infants use known verbs to learn novel nouns: Evidence from 15-and 19-month-olds. *Cognition, 131*(1), 139–146. https://doi.org/10.1016/j.cognition.2013.12.014

Gahl, S., Yao, Y., & Johnson, K. (2012). Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of memory and language, 66*(4), 789–806. https://doi.org/10.1016/j.jml.2011.11.006

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In Kuczaj, S.A. (Ed.), Language development: Vol. 2. Language, thought, and culture (pp. 301-334). Hillsdale, NJ: Erlbaum.

Gentner, D. (1989). The mechanisms of analogical learning. In Vosniadou, S., & Ortony, A., (Eds.), *Similarity and analogical reasoning (pp. 199–241)*. London: Cambrige University Press; doi: 10.1017/cbo9780511529863.011.

Gentner, D. (1975). Evidence for the psychological reality of semantic components: The verbs of possession. *Explorations in cognition, 35*, 211–246. https://doi.org/10.7551/mitpress/5237.003.0008

Gentner, D. (2006). Why verbs are hard to learn. In K. Hirsh-Pasek, & R. Golinkoff (Eds.), *Action meets word: How children learn verbs* (pp. 544–564). New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195170009.003.0022.

Gentner, D., & Namy, L. L. (2006). Analogical processes in language learning. *Current Directions in Psychological Science, 15*(6), 297–301. https://doi.org/10.1111/j.1467-8721.2006.00456.x

Gentner, D., & Rattermann, M.J. (1991). Language and the career of similarity.

Gleitman, L. (1990). The structural sources of verb meanings. *Language acquisition, 1*(1), 3–55. https://doi.org/10.1207/s15327817la0101_2

Gleitman, L. R., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. C. (2005). Hard words. *Language learning and development, 1*(1), 23–64. https://doi.org/10.1207/s15473341lld0101_4

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition, 73*(2), 135–176. https://doi.org/10.1016/s0010-0277(99)00036-0

Goldenberg, E. R., Repetti, R. L., & Sandhofer, C. M. (2022). Contextual variation in language input to children: A naturalistic approach. *Developmental psychology, 58*(6), 1051. https://doi.org/10.1037/dev0001345

Goldenberg, E. R., & Sandhofer, C. M. (2013). Same, varied, or both? Contextual support aids young children in generalizing category labels. *Journal of Experimental Child Psychology, 115*(1), 150–162. https://doi.org/10.1016/j.jecp.2012.11.011

Golinkoff, R. M., & Kerr, J. L. (1978). Infants' perception of semantically defined action role changes in filmed events. *Merrill-Palmer Quarterly, 24*(1), 53–61.

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language, 35*(3), 515. https://doi.org/10.1017/s0305000907008641

Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and brain sciences, 21*(6), 803–831. https://doi.org/10.1017/s0140525x98001769

Harris, M., Barrett, M., Jones, D., & Brookes, S. (1988). Linguistic input and early word meaning. *Journal of Child Language, 15*(1), 77–94. https://doi.org/10.1017/s030500090001206x

Harris, Z. S. (1957). Co-occurrence and transformation in linguistic structure. *Language, 33*(3), 283–340. https://doi.org/10.2307/411155

Harris, M., Jones, D., & Grant, J. (1983). The nonverbal context of mothers' speech to infants. *First Language, 4*(10), 21–30. https://doi.org/10.1177/014272378300401003

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Paul H Brookes Publishing. https://doi.org/10.1007/s00431-005-0010-2

Havron, N., Ramus, F., Heude, B., Forhan, A., Cristia, A., Peyre, H., & EDEN Mother-Child Cohort Study Group. (2019). The effect of older siblings on language development as a function of age difference and sex. *Psychological Science, 30*(9), 1333–1343. https://doi.org/10.31234/osf.io/fgpmd

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. https://doi.org/10.1016/j.cogpsych.2011.06.002

Hills, T., Maouene, J., Riordan, B., & Smith, L. B. (2009). Contextual diversity and the associative structure of adult language in early word learning. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2118–2123). Austin, TX: The Cognitive Science Society, 10.1016/j.jml.2010.06.002.

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. https://doi.org/10.1016/j.jml.2010.06.002

Hirsh-Pasek, K., Adamson, L. B., Bakeman, R., Owen, M. T., Golinkoff, R. M., Pace, A., … Suma, K. (2015). The contribution of early communication quality to low-income children's language success. *Psychological Science, 26*(7), 1071–1083. https://doi.org/10.1177/0956797615581493

Hoff, E., & Naigles, L. (2002). How children use input to acquire a lexicon. *Child Development, 73*(2), 418–433. https://doi.org/10.1111/1467-8624.00415

Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child Development, 74*(5), 1368–1378. https://doi.org/10.1111/1467-8624.00612

Hoff, E. (2006). How social contexts support and shape language development. *Developmental Review, 26*(1), 55–88. https://doi.org/10.1016/j.dr.2005.11.002

Houston, D., Santelmann, L., & Jusczyk, P. (2004). English-learning infants' segmentation of trisyllabic words from fluent speech. *Language and Cognitive Processes, 19*(1), 97–136. https://doi.org/10.1080/01690960344000143

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. https://doi.org/10.31234/osf.io/qhwv5

Hsu, N., Hadley, P. A., & Rispoli, M. (2017). Diversity matters: Parent input predicts toddler verb production. *Journal of child language, 44*(1), 63–86. https://doi.org/10.1017/s0305000915000690

Huebner, P., & Willits, J. (2021). Using lexical context to discover the noun category: Younger children have it easier. *Psychology of Learning and Motivation, 75*. https://doi.org/10.1016/bs.plm.2021.08.002

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology, 27*(2), 236. https://doi.org/10.1037/0012-1649.27.2.236

Huttenlocher, J., Waterfall, H., Vasilyeva, M., Vevea, J., & Hedges, L. V. (2010). Sources of variability in children's language growth. *Cognitive Psychology, 61*(4), 343–365. https://doi.org/10.1016/j.cogpsych.2010.08.002

Huttenlocher, J., Vasilyeva, M., Cymerman, E., & Levine, S. (2002). Language input and child syntax. *Cognitive Psychology, 45*(3), 337–374. https://doi.org/10.1016/S0010-0285(02)00500-5

Huttenlocher, J., Vasilyeva, M., Waterfall, H. R., Vevea, J. L., & Hedges, L. V. (2007). The varieties of speech to young children. *Developmental psychology, 43*(5), 1062. https://doi.org/10.1037/0012-1649.43.5.1062

Jones, M. N., & Mewhort, D. J. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological review, 114*(1), 1. https://doi.org/10.1037/0033-295x.114.1.1

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in english-learning infants. *Cognitive Psychology, 39*(3–4), 159–207. https://doi.org/10.1006/cogp.1999.0716

Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology, 29*(1), 1–23. https://doi.org/10.1006/cogp.1995.1010

Landau, B., & Gleitman, L. (1985). *Language and experience*. Cambridge, MA: Harvard University Press, 10.1002/acp.2350010109.

Lany, J., & Saffran, J. R. (2010). From statistics to meaning: Infants' acquisition of lexical categories. *Psychological science, 21*(2), 284–291. https://doi.org/10.1177/0956797609358570

Lany, J., & Saffran, J. R. (2013). Statistical learning mechanisms in infancy. *Comprehensive Developmental Neuroscience: Neural Circuit Development and Function in the Brain, 3*, 231–248. https://doi.org/10.1016/b978-0-12-397267-5.00034-0

Lany, J., Gómez, R. L., & Gerken, L. A. (2007). The role of prior experience in language acquisition. *Cognitive Science, 31*(3), 481–507. https://doi.org/10.1080/15326900701326584

Lennon, P. (1996). Getting "easy" verbs wrong at the advanced level. *International Review of Applied Linguistics in Language Teaching, 34*, 23–36. https://doi.org/10.1515/iral.1996.34.1.23

Leung, A., Tunkel, A., & Yurovsky, D. (2021). Parents fine-tune their speech to children's vocabulary knowledge. *Psychological Science, 32*(7), 975–984. https://doi.org/10.1177/0956797621993104

Lidz, J., Waxman, S., & Freedman, J. (2003). What infants know about syntax but couldn't have learned: Experimental evidence for syntactic structure at 18 months. *Cognition, 89*(3), 295–303. https://doi.org/10.1016/s0010-0277(03)00116-1

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and hearing, 19*(1), 1. https://doi.org/10.1097/00003446-199802000-00001

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior research methods, instruments, & computers, 28*(2), 203–208. https://doi.org/10.3758/bf03204766

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. Transcription format and programs* (Vol. 1). https://doi.org/10.1162/coli.2000.26.4.657

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive psychology, 20*(2), 121–157. https://doi.org/10.1016/0010-0285(88)90017-5

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82*(3), B101–B111. https://doi.org/10.1016/s0010-0277(01)00157-3

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive psychology, 18*(1), 1–86. https://doi.org/10.1016/0010-0285(86)90015-0

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. https://doi.org/10.3758/bf03192726

Meredith, L. R., & Catherine, E. S. (2020). Analyzing input quality along three dimensions: Interactive, linguistic, and conceptual. *Journal of child language, 47*(1), 5–21. https://doi.org/10.1017/s0305000919000655

Merriman, W. E., Bowman, L. L., & MacWhinney, B. (1989). The mutual exclusivity bias in children's word learning. *Monographs of the society for research in child development*, i–129. https://doi.org/10.2307/1166130

Moors, A., De Houwer, J., Hermans, D., Wanmaker, S., Van Schie, K., Van Harmelen, … Brysbaert, M. (2013). Norms of valence, arousal, dominance, and age of acquisition for 4,300 Dutch words. *Behavior research methods, 45*, 169–177. https://doi.org/10.3758/s13428-012-0243-8

Morgan, J. L., & Saffran, J. R. (1995). Emerging integration of sequential and suprasegmental information in preverbal speech segmentation. *Child Development, 66*(4), 911–936. https://doi.org/10.1111/j.1467-8624.1995.tb00913.x

Masur, E. F. (1997). Maternal labeling of novel and familiar objects: Implications for children's development of lexical constraints. *Journal of child language, 24*(2), 427–439. https://doi.org/10.1017/s0305000997003115

Naigles, L. (1990). Children use syntax to learn verb meanings. *Journal of Child Language, 17*(2), 357–374. https://doi.org/10.1017/s0305000900013817

Naigles, L. R. (1996). The use of multiple frames in verb learning via syntactic bootstrapping. *Cognition, 58*(2), 221–251. https://doi.org/10.1016/0010-0277(95)00681-8

Naigles, L. R., & Hoff-Ginsberg, E. (1998). Why are some verbs learned before other verbs? Effects of input frequency and structure on children's early verb use. *Journal of child language, 25*(1), 95–120. https://doi.org/10.1017/s0305000997003358

Nazzi, T., Dilley, L. C., Jusczyk, A. M., Shattuck-Hufnagel, S., & Jusczyk, P. W. (2005). English-learning infants' segmentation of verbs from fluent speech. *Language and Speech, 48*(3), 279–298. https://doi.org/10.1177/00238309050480030201

Onnis, L., Monaghan, P., Richmond, K., & Chater, N. (2005). Phonology impacts segmentation in online speech processing. *Journal of Memory and Language, 53*(2), 225–237. https://doi.org/10.1016/j.jml.2005.02.011

Pan, B. A., Rowe, M. L., Singer, J. D., & Snow, C. E. (2005). Maternal correlates of growth in toddler vocabulary production in low-income families. *Child Development, 76*(4), 763–782. https://doi.org/10.1111/1467-8624.00498-i1

Perry, L. K., Perlman, M., Winter, B., Massaro, D. W., & Lupyan, G. (2018). Iconicity in the speech of children and adults. *Developmental Science, 21*(3), Article e12572. https://doi.org/10.1111/desc.12572

Perry, L. K., & Saffran, J. R. (2017). Is a pink cow still a cow? Individual differences in toddlers' vocabulary knowledge and lexical representations. *Cognitive Science, 41*(4), 1090–1105. https://doi.org/10.1111/cogs.12370

Perry, L. K., & Samuelson, L. K. (2011). The shape of the vocabulary predicts the shape of the bias. *Frontiers in Psychology, 2*, 345. https://doi.org/10.3389/fpsyg.2011.00345

Perry, L. K., Axelsson, E. L., & Horst, J. S. (2016). Learning what to remember: Vocabulary knowledge and children's memory for object names and features. *Infant and Child Development, 25*(4), 247–258. https://doi.org/10.1002/icd.1933

Pruden, S. M., Hirsh-Pasek, K., Maguire, M., & Meyer, M. (2004). *Foundations of verb learning: Infants categorize path and manner in motion events, 1*, 461–472.

Rabovsky, M., Schad, D. J., & Abdel Rahman, R. (2021). Semantic richness and density effects on language production: Electrophysiological and behavioral evidence. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 47*(3), 508. https://doi.org/10.1037/xlm0000940

Rowe, M. L., Leech, K. A., & Cabrera, N. (2017). Going beyond input quantity: Wh-questions matter for toddlers' language and cognitive development. *Cognitive Science, 41*, 162–179. https://doi.org/10.1111/cogs.12349

Rowe, M. L. (2012). A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. *Child Development, 83*(5), 1762–1774. https://doi.org/10.1111/j.1467-8624.2012.01805.x

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences, 112*(41), 12663–12668. https://doi.org/10.1073/pnas.1419773112

Sadeghi, S., Scheutz, M., & Krause, E. (2017, September). An embodied incremental bayesian model of cross-situational word learning. In 2017 joint IEEE international conference on development and learning and epigenetic robotics (ICDL-EpiRob) (pp. 172-177). IEEE.Krause, E. (2017). An embodied incremental bayesian model of cross-situational word learning. The Seventh joint IEEE international conference on development and learning and on epigenetic robotics : September 18-21, 2017, Instituto Superior Técnico, Lisbon, Portugal, 172–177. Doi: 10.1109/DEVLRN.2017.8329803.

Sanchez, A., Meylan, S. C., Braginsky, M., MacDonald, K. E., Yurovsky, D., & Frank, M. C. (2019). childes-db: A flexible and reproducible interface to the child language data exchange system. *Behavior research methods, 51*, 1928–1941. https://doi.org/10.3758/s13428-018-1176-7

Schwab, J. F., & Lew-Williams, C. (2016). Language learning, socioeconomic status, and child-directed speech. *Wiley Interdisciplinary Reviews: Cognitive Science, 7*(4), 264–275. https://doi.org/10.1002/wcs.1393

Shi, R., & Lepage, M. (2008). The effect of functional morphemes on word segmentation in preverbal infants. *Developmental Science, 11*(3), 407–413. https://doi.org/10.1111/j.1467-7687.2008.00685.x

Shneidman, L. A., Arroyo, M. E., Levine, S. C., & Goldin-Meadow, S. (2013). What counts as effective input for word learning? *Journal of Child Language, 40*(3), 672. https://doi.org/10.1017/s0305000912000141

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. https://doi.org/10.1037/xge0000876

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B, 13*, 238–241. https://doi.org/10.1111/j.2517-6161.1951.tb00088.x

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science, 13*, 13–19. https://doi.org/10.1111/1467-9280.00403

Song, L., Tamis-LeMonda, C. S., Yoshikawa, H., Kahana-Kalman, R., & Wu, I. (2012). Language experiences and vocabulary development in Dominican and Mexican infants across the first 2 years. *Developmental psychology, 48*(4), 1106. https://doi.org/10.1037/a0026401

Swingley, D., & Humphrey, C. (2018). Quantitative linguistic predictors of infants' learning of specific english words. *Child Development, 89*(4), 1247–1267. https://doi.org/10.1111/cdev.12731

Tamis-LeMonda, C. S., Kuchirko, Y., & Song, L. (2014). Why is infant language learning facilitated by parental responsiveness? *Current Directions in Psychological Science, 23*(2), 121–126. https://doi.org/10.1177/0963721414522813

Tomasello, M., & Todd, J. (1983). Joint attention and lexical acquisition style. *First Language, 4*(12), 197–211. https://doi.org/10.1177/014272378300401202

Tomasello, M. (1988). The role of joint attentional processes in early language development. *Language Sciences, 10*(1), 69–88. https://doi.org/10.1016/0388-0001(88)90006-x

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology, 66*(1), 126–156. https://doi.org/10.1016/j.cogpsych.2012.10.001

Vlach, H. A., & Sandhofer, C. M. (2011). Developmental differences in children's context-dependent word learning. *Journal of Experimental Child Psychology, 108*(2), 394–401. https://doi.org/10.1016/j.jecp.2010.09.011

Vitevitch, M. S. (2002). The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(4), 735. https://doi.org/10.1037/0278-7393.28.4.735

Vitevitch, M. S., & Sommers, M. S. (2003). The facilitative influence of phonological similarity and neighborhood frequency in speech production in younger and older adults. *Memory & cognition, 31*(4), 491–504. https://doi.org/10.3758/bf03196091

Weisleder, A., & Fernald, A. (2014). Social environments shape children's language experiences, strengthening language processing and building vocabulary. Language in Interaction. Studies in honor of Eve V. Clark, 29-49. https://doi.org/10.1075/tilar.12.06wei.

Willits, J., Saffran, J., & Lany, J. (2017). Toddlers can use semantic cues to learn difficult nonadjacent dependencies. https://doi.org/10.31234/osf.io/4ca78.

Willits, J. A., Seidenberg, M. S., & Saffran, J. R. (2014). Distributional structure in language: Contributions to noun–verb difficulty differences in infant word recognition. *Cognition, 132*(3), 429–436. https://doi.org/10.1016/j.cognition.2014.05.004

Wojcik, E. H., & Saffran, J. R. (2015). Toddlers encode similarities among novel words from meaningful sentences. *Cognition, 138*, 10–20. https://doi.org/10.1016/j.cognition.2015.01.015

Wojcik, E. H., Zettersten, M., & Benitez, V. L. (2022). The map trap: Why and how word learning research should move beyond mapping. *Wiley Interdisciplinary Reviews: Cognitive Science, 13*(4), e1596.

Yu, C., & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing, 70*(13–15), 2149–2165. https://doi.org/10.1016/j.neucom.2006.01.034

Yu, C., Suanda, S. H., & Smith, L. B. (2019). Infant sustained attention but not joint attention to objects at 9 months predicts vocabulary at 12 and 15 months. *Developmental Science, 22*(1), e12735.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science, 18*(5), 414–420. https://doi.org/10.1111/j.1467-9280.2007.01915.x

Yuan, S., & Fisher, C. (2009). "Really? She blicked the baby?" Two-year-olds learn combinatorial facts about verbs by listening. *Psychological Science, 20*(5), 619–626. https://doi.org/10.1111/j.1467-9280.2009.02341.x

Zettersten, M., Potter, C. E., & Saffran, J. R. (2020). Tuning in to non-adjacencies: Exposure to learnable patterns supports discovering otherwise difficult structures. *Cognition, 202*, Article 104283. https://doi.org/10.1016/j.cognition.2020.104283