Quantity and diversity: Simulating early word learning environments

Jessica L. Montag[1], Michael N. Jones[2] & Linda B. Smith[2]

[1]Department of Psychology

University of California, Riverside

Riverside, California

[2]Department of Psychological and Brain Sciences

Indiana University

Bloomington, Indiana

This work was supported by NIH Grant T32 HD-07475 and by NSF grant BCS-15233982 to Smith.

Please direct all correspondence to Jessica L. Montag, Department of Psychology, University of California, Riverside, 900 University Avenue, Riverside CA, 92515

Email: Jessica.montag@ucr.edu.

Abstract

The words in children's language learning environments are strongly predictive of cognitive development and school achievement. But how do we measure language environments and do so at the scale of the many words that children hear day-in and day-out? The quantity and quality of words in a child's input is typically measured in terms of total amount of talk and the lexical diversity in that talk. There are disagreements in the literature whether amount or diversity is the more critical measure of the input. Here we analyze the properties of a large corpus (6.5 million words) of speech to children and simulate learning environments that differ in amount of talk per unit time, lexical diversity, and the contexts of talk. The central conclusion is that what researchers need to theoretically understand, measure, and change is not the total amount of words, or the diversity of words, but the function that relates total words to the diversity of words, and how that function changes across different contexts of talk.

Keywords: language development; child-directed speech; individual differences; computer simulation; linguistic quantity and quality

Early vocabulary development is characterized by marked individual differences that have significant downstream consequences for later language learning and for success in many other cognitive domains. The evidence indicates that differences in vocabulary growth among otherwise typically developing children are strongly related to differences in their language learning environments (Hoff, 2003; Huttenlocher et al., 2010; Hurtado, Marchman & Fernald, 2008; Weisleder & Fernald, 2014). In brief, some children's environments include much more talk to the child than do other environments, and those children who hear more words directed to them, not surprisingly, show more rapid and robust language learning. But what exactly is it about environments with more child-directed talk that matters? To answer that question, we need to know how to measure and compare language learning environments. This is a complex problem in part because the scale of experience is massive with the average child hearing more than 20,000 child-directed words a day or over 7 million words a year (Hart & Risley, 1995; Shneidman, Arroyo, Levin & Goldin-Meadow, 2013). The problem is also complicated by the fact that the frequency distribution of words in produced language is not normal and thus the usual assumptions about sampling from normal distributions do not apply. These issues are becoming urgent as new methods that capture language learning environments at scale (Gilkerson & Richards, 2011; Roy, et al., 2006; VanDam, et al., 2016) are outpacing our analytic and inferential methods to understand the distributions of words in the talk that we record (Greenwood, et al., 2011).

The goal of this paper is to take a step toward finding solutions by focusing on two well-used and traditional measures of the words in children's environments: Their total number and their diversity. These two measures provide an illuminating case because: (1) As is well known (Heaps 1978; Herdan 1960; more recently, Malvern et al., 2004; McKee, Malvern & Richards, 2000; Richards, 1987; Tweedie & Baayen, 1998), total words and the diversity of words are not independent and their relation changes nonlinearly with the sample size of the speech analyzed, and (2) the relation between total words and number of unique words depends on the contexts of talk. Although we concentrate on counts of words in the learning environment, it seems likely that the issues considered here – how relations among measures change as a function of sample size and context – will extend to other relevant aspects of the language learning environment.

Our approach to exploring the relation between total words in the input and their diversity is to simulate different learnings environments. Thus, we do not measure real children's learning environments nor make predictions from real or simulated environments to vocabulary development. Instead, we explore how *possible* learning environments may vary and how this affects the relation between total words and their diversity.

Background

The number of total words (the "tokens") and the diversity of those words (the number of unique "types") have played an important role in the study of language and language learning for over a century (Carroll, 1938; Estoup, 1916; Johnson, 1938; Osgood, 1952). Contemporary debates about the quantity and quality of input also often center on measures of tokens and types (Hoff & Naigles, 2002; Huttenlocher et al., 1991; 2010; Rowe, 2012; Weisleder & Fernald, 2014; c.f., Hoff, 2006; Hirsh-Pasek et al., 2015). This makes sense: More talk in the learning environment offers more repetitions, more co-occurrences, more opportunities for learning any individual word. More diversity among the words heard offers opportunities for building a bigger vocabulary and for determining the meanings of words from the larger semantic networks in which they occur.

Measures of tokens and types in the input begin with a transcription of some sample of speech directed to the child. Traditionally, researchers recorded about an hour, sometimes several hours, but with increasing frequency researchers are recording whole days and more (VanDam, et al., 2016). From a sample (say several hours long) of the words per minute in this recorded child-directed speech, one can estimate the total amount of talk that different children have heard over some more extensive period of time (such as whole days or years). Estimates made in this way indicate that the average child hears about 20,000 to 38,000 total words a day (Hart & Risley, 1995; Shneidman, et al., 2013; Weisleder & Fernald, 2014). Estimates made in this way also suggest that there is extreme individual variability in the total number of words directed to different children, ranging from as few as 2000 child-directed words a day for some children to as many as 50,000 words a day for others (Hart & Risley, 1995; Weisleder & Fernald, 2014). These differences in amount of talk to individual children are strongly predictive of the child's vocabulary size and early school achievement (Dickinson, Golinkoff & Hirsh-Pasek, 2010; Hart & Risley, 1995; Hoff, 2003; Huttenlocher et al., 2010; Rowe, 2012; Walker, Greenwood, Hart & Carta, 1994) and are also highly associated with the socio-economic standing of the families (Hoff, 2003; Huttenlocher et al., 2010; Hurtado, Marchman & Fernald, 2008; Weisleder & Fernald, 2014). Indeed, Hart and Risley (1995) projected that by the time children entered school, there was a 30-million-word gap in the cumulative number of words directed to children from poorer versus richer families. Given the predictive link between total words per unit time in child-directed speech in the home and the child's vocabulary size and school readiness, there is now a considerable public health effort directed to increasing parent talk to young children (Leffel & Suskind, 2013; Reese, Sparks & Leyva, 2010; Roberts &

Kaiser, 2011; and public health initiatives such as Providence Talks, First 5 California, and Too Small to Fail, among many others).

The total number of unique words, not just total words, is critical to building a large vocabulary. However, the opportunity to hear unique words varies with the total amount of talk and does so in a complicated way that derives from the fact that the frequency with which any individual word in a language is produced is not uniform. Instead, a very few words are very frequent but most words occur in speech quite rarely. To illustrate this point, and the problem posed in measuring total words and total unique words in a child's experience, we used the distribution of unique words in the CHILDES corpus of 6 million child-directed words. The CHILDES corpus is a compendium of many different parents talk to their children (MacWhinney, 2000). From this corpus, we created hypothetical distributions of the unique words heard in a day by children hearing on average 2,000, 20,000 or 50,000 words in a day. We did so by randomly sampling tokens from the entire corpus of 6 million words such that the hypothetical frequency distributions of individual words correspond to that of real parent talk in aggregate. The resulting distributions shown in Figure 1 plot the frequency of occurrence of individual words (the y axis) as a function of the rank of their overall frequency in the language. The figure illustrates the well-known fact that a very *few words are produced with very high frequency but that most words are produced infrequently.* This is so in all three simulated day-long environments. But critically, children who hear a greater total amount of talk will hear those highly frequent words even more frequently and, as illustrated in the long tail of infrequent words, will also hear many more unique but sparsely occurring words.

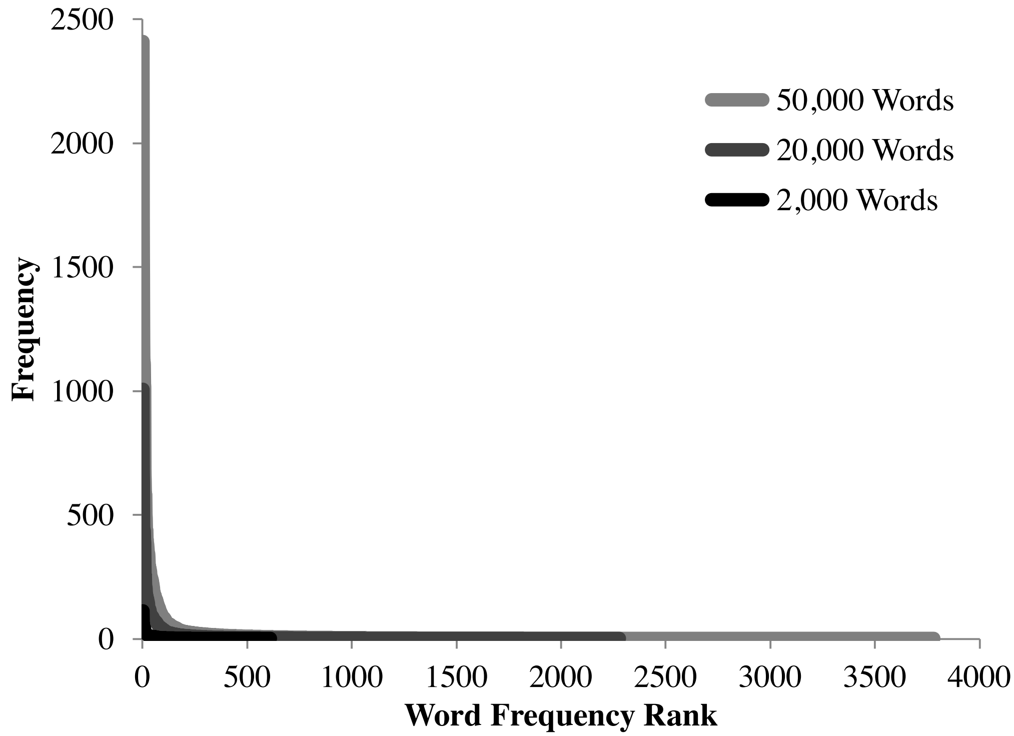
Lexical diversity, the number of unique types, in the input is positively related to the child's vocabulary size (Hart & Risley, 1995; Hoff & Naigles, 2002; Huttenlocher et al., 1991;

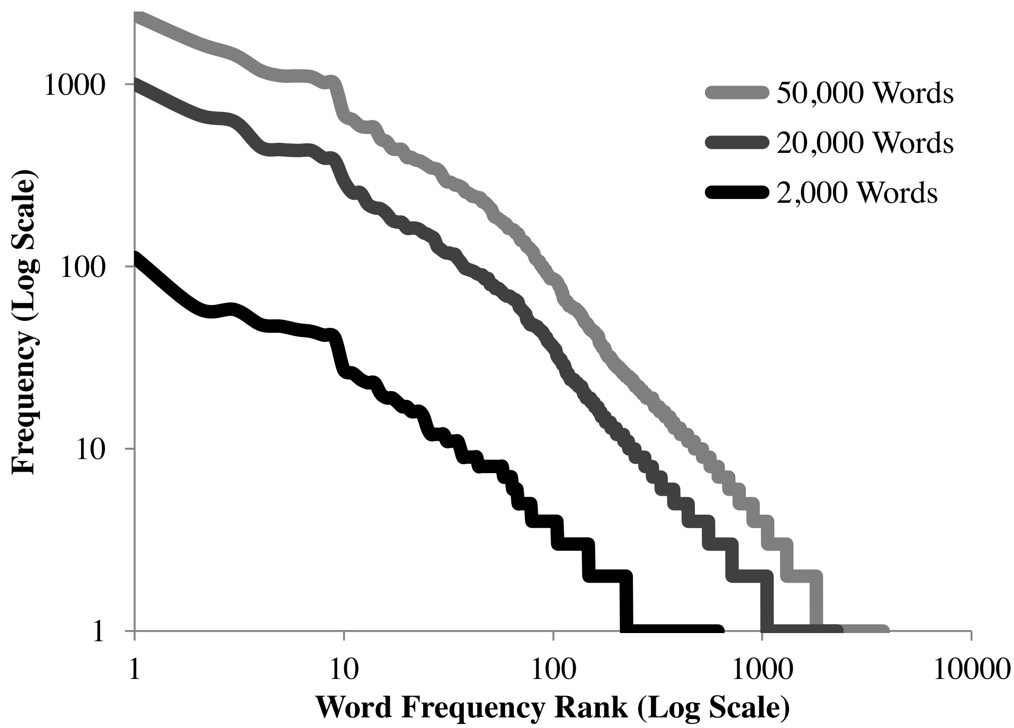
2010; Pan, Rowe, Singer & Snow, 2005; Rowe, 2008; 2012; Shneidman, et al., 2013; Weizman & Snow, 2001). In fact, many of these researchers who use type counts as an indicator of lexical diversity note a high correlation between word type and token counts. One common way to measure the overall diversity of words in some sample is to determine the number of unique word types in relation to the number of all words, or tokens. In general, more unique word types, or a high ratio between word types and token ratio (proportionally more diverse words) is considered higher quality. Although reasonable, none of this is straight forward. As is well known (Heaps 1978; Herdan 1960) and as shown in the type-token table embedded in Figure 1, the type-token ratio *decreases* as the total number of tokens sampled increase. Thus, *if parents principally differ only in how many words they sample* from the language in a unit of time, then children with a smaller day-long word count hear a higher type-token ratio, and a higher rate of *more diverse speech* but, of course, a fewer total number of unique word types, and fewer repetitions of everything. All these properties of the input are inter-related.

We ask two key questions of this interrelation. At the methodological level, the question is what constitutes a "fair" sampling of words to measure and compare learning environments? And at the theoretical level, the question is what constitutes an optimal distribution of words in the learning environment for early vocabulary development? Many of our intuitions about sampling (by researchers or by learners) are based on the properties of normal and near normal distributions. These do not apply given the frequency distributions of words in language. Normal distributions characterize such properties as the height of individuals; normal distributions are also forced on measurement systems of human traits, such as intelligence. In these distributions, scores cluster around a central tendency, making the typical value of a large enough sample representative of the population distribution. Many of the statistics we use in studies of the words children hear are based on this central tendency assumption. The words in natural language production, however, are extremely skewed with respect to their rank frequency, as in the examples in Figure 1. More formally, the distribution is characterized by a power law (Cohen, Mantegna & Havlin, 1997; Clauset, Shalizi & Newman, 2009; Ferrer-i-Cancho & Solé, 2002; Goldwater, Griffiths & Johnson, 2006; Kello, et al., 2009; Mandelbrot, 1953; Piantadosi, 2014; Simon, 1955; Zipf, 1949).

$$f(x) = ax^{-k}$$

Where the frequency of a word with rank x, is given by a power constant k which determines the steepness of the relation between a word's frequency rank and its frequency, and the scaling constant, a. These distributions lack a well-defined average value; this makes many of our usual inferences about sampling as well as statistical procedures based on central tendencies inappropriate (Clauset, Shalizi & Newman, 2009). These distributions with their few highly frequent words and the long tail of rarer words also create the complex relations between counts of types, tokens and the type-token ratio. This complexity is captured in Heaps-Herdan law (Heaps 1978; Herdan 1960): As the number of words sampled (by the researcher or by the young learner) increases, the number of unique words also increases, *but at a rate that slows as more words are added to the sample*. This presents both a measurement and conceptual problems for understanding how language environments may differ between children. Many studies indicate that the total number of types, tokens, and the ratio between the two in some sample of parent talk are positively related to child to language outcomes (Hart & Risley, 1995; Hoff & Naigles, 2002; Huttenlocher et al, 2010; Pan, et al., 2005; Rowe, 2008; 2012; Shneidman, et al., 2013; Weizman & Snow, 2001), but as Heaps-Herdan law make clear, these measures are all products of a single sample of words in a child's learning environment and we do not have a unified understanding of how these distributions of sampled words (and thus these individual measures) can vary across children's individual learning environments. This is the question we seek to understand and provide a step toward answering. Accordingly, in the simulated environments section of this report, we explore the relations among total words and the diversity of words across a set of simulated learning environments that differ in properties likely relevant to early word learning.





Tokens	Types	Type-Token Ratio
50,000	3,781	0.07562
20,000	2,277	0.11385
2,000	612	0.306

Figure 1: A word-rank by word-frequency plot that show a day's worth of speech input for hypothetical children who hear 50,000, 20,000 or 2,000 words per day (top) along with a log-log scale version of this graph (bottom). Words were randomly selected from all of CHILDES. The table shows the type counts (number of unique words) and the type-token ratio of the day's input for these three hypothetical children.

Simulated Environments

Malvern and Richard (Malvern et al., 2004; McKee, Malvern & Richards, 2000) showed how different degrees of lexical diversity can be represented in terms of the different curves relating total word tokens and type-token ratios. The solution outlined in Malvern et al. (2004)

and McKee, Malvern & Richards (2000) is the VOCD, a single value measure of lexical diversity that should be less dependent on sample size. The VOCD is quite similar to the sampling method we outline here that yields different type-token curves. Despite solving some (but not all) problems related to the size-dependence of many lexical diversity measures (McCarthy & Jarvis, 2007), the VOCD does not solve the problem we describe in Demonstration 4, that in addition to sample size, lexical diversity is dependent on how that sample was constructed, with corpora composed of small pieces of many contexts being inherently more lexically diverse than a similarly sized corpus composed of fewer, longer documents or conversations.

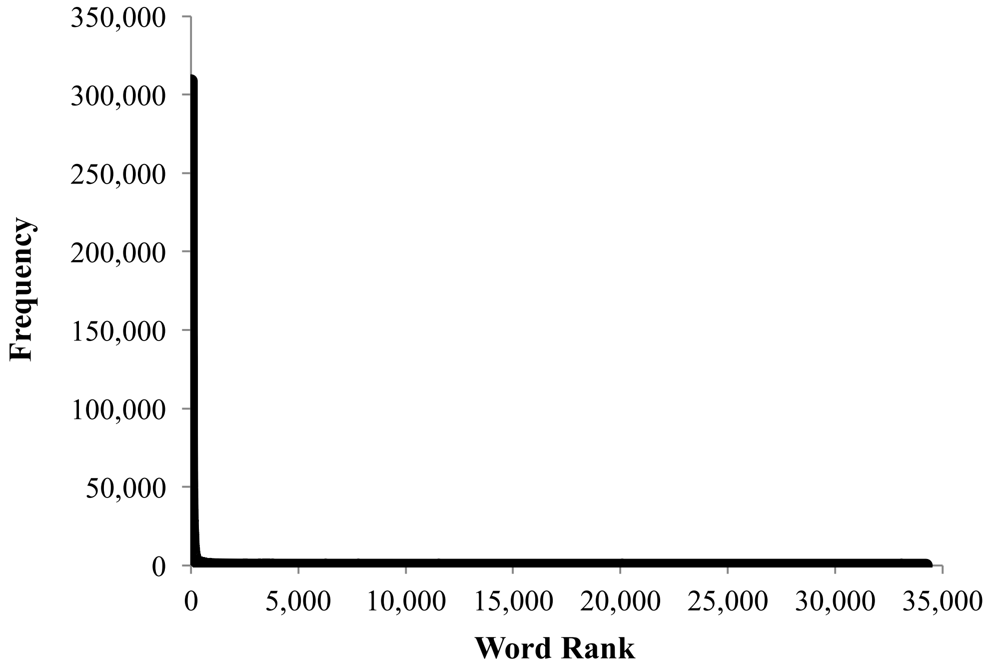
We see our work as building upon the work of Malvern et al. (2004) and McKee, Malvern & Richards (2000). Our analyses of simulated environments extend this work and lead us to this conclusion: the function that relates number of types to number tokens within a learning environment may provide the path to measuring learning environments at the scales now possible as well as provide important insights into how environments differ and how malleable the individual properties of those environments may be. Our approach was to create samples of varying sizes from different hypothetical word learning environments. All the simulated environments began with the same large corpus of caregiver speech to children, the CHILDES corpus (MacWhinney, 2000). This is a collection of transcripts of children interacting with caregivers, siblings and other adults that were collected for a variety of purposes by different language researchers in a variety of settings. Thus, this corpus of child-directed speech was created over many different parents and children. We know that the statistical analyses of words in this specific corpus capture something real about children's word-learning environments because these regularities have been repeatedly shown to predict the normative age of acquisition for words as well as a variety of linguistic devices (Diessel, 2009; Goodman, Dale & Li, 2008; Hills, Maouene, Riordan & Smith, 2010; Kidd, Leiven & Tomasello, 2006; Mintz, 2003; Ninio, 2011).

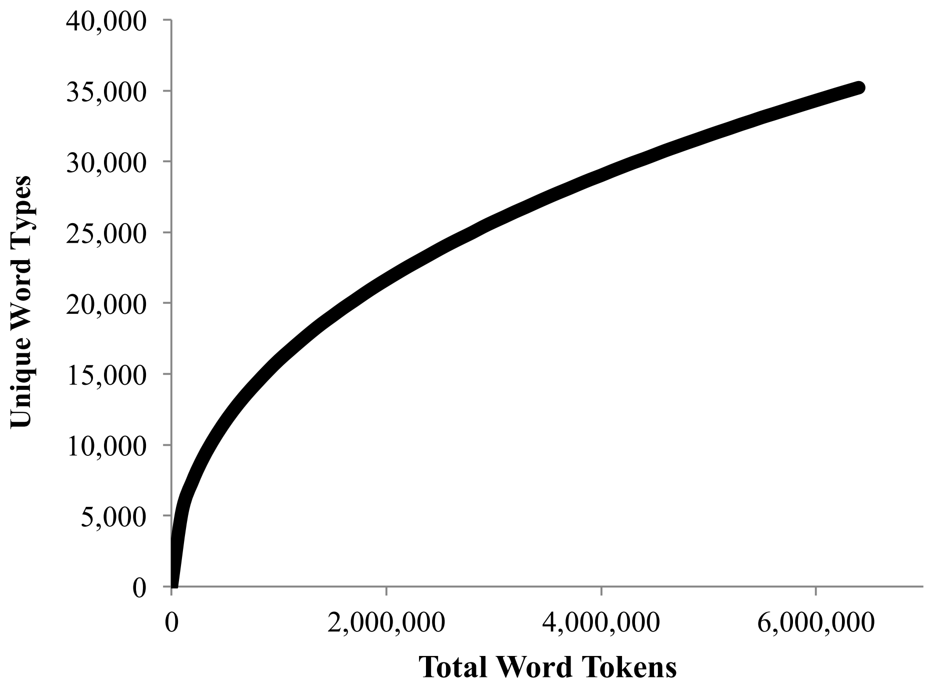
The full CHILDES corpus provides us with a baseline environment. We created this baseline "environment" using the child-directed speech from the entire American English subset of the CHILDES corpus directed at children under the age of 5 years. This corpus consists of a total of 4,432 individual conversations (contiguous recording sessions) containing a total of about 6.5 million words. We used a version of the CHILDES corpus that had been processed to (1) remove a number of the special transcription characters and other artifacts of the CHILDES coding system and (2) systematize words with idiosyncratic spellings (e.g. replace all instances of "doggy" with "doggie" to maintain consistent spelling) (Hubner & Willits, 2017). We first describe the properties of this baseline environment and then the relations between types and tokens in simulated environments derived from this baseline environment

1. The Baseline Environment

We first show that the distribution of words in the baseline corpus is characterized by a power law distribution of words and Heaps-Herdan law (Heaps 1978; Herdan 1960; see also, Malvern et al., 2004; McKee, Malvern & Richards, 2000; Richards, 1987; Tweedie & Baayen, 1998). To do this, we counted the number of times each unique word appeared in CHILDES, sorted them by frequency and plotted the subsequent frequency (number of instances in the corpus) by the frequency rank value of the words. The top panel of Figure 2 shows the result and the classic power law pattern, that a word's frequency exponentially decreases inverse to frequency rank. In other words, the corpus consists of a few very frequent words, and a large number of relatively infrequent words. For example, the left-most word in the plot is the most frequent word in the corpus, “you,” which occurs about 309,000 times in the corpus followed by “is” and “the” which occur about 218,000 and 190,000 times in the nearly 6.5-million-word corpus. The top 1% of all the word types account for 82% of the tokens; the top 5% of all words account for 95% of all tokens. Mid-frequency words, near the inflection point of the curve include words like “take” and “dough” which are the 100th and 101st most frequent words and occur about 11,500 times, and “bridge” and “quick,” the 1000 and 1001st most frequent words, which appear 443 and 442 times. Most word types are found in the long, infrequent tail, and include words that appear in a 6.5 million-word corpus only a handful of times. Some reasonably common words “stewed,” “snowboard” or “bronze,” appear only once in the corpus. In brief, as in natural language as whole, the specific words at the head of the distribution, are very frequent, but most of the words that children need to learn—the long tail of the distribution-- are infrequent. This is characteristic of power-law distributions, a strongly right-skewed shape that retains its shape regardless of the scale at which the distribution is viewed.

The relation between tokens and types, as per Heaps-Herdan law, can be captured by the curve that relates the number of types in the sample to the number of tokens. We do this at two scales: First in Figure 2, for successive samples of 100,000 words, a scale that is tractable for the scale of input that children receive in a year and second in Figure 3, for successive samples in a day for the three example children in Figure 1, a scale of samples closer to what researchers are now beginning to measure with some regularity (VanDam, et al., 2016).


350,000
300,000
250,000
200,000
150,000
100,000
50,000
0
Frequency
0
5,000
10,000
15,000
20,000
25,000
30,000
35,000
Word Rank


40,000
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
Unique Word Types
0
2,000,000
4,000,000
6,000,000
Total Word Tokens

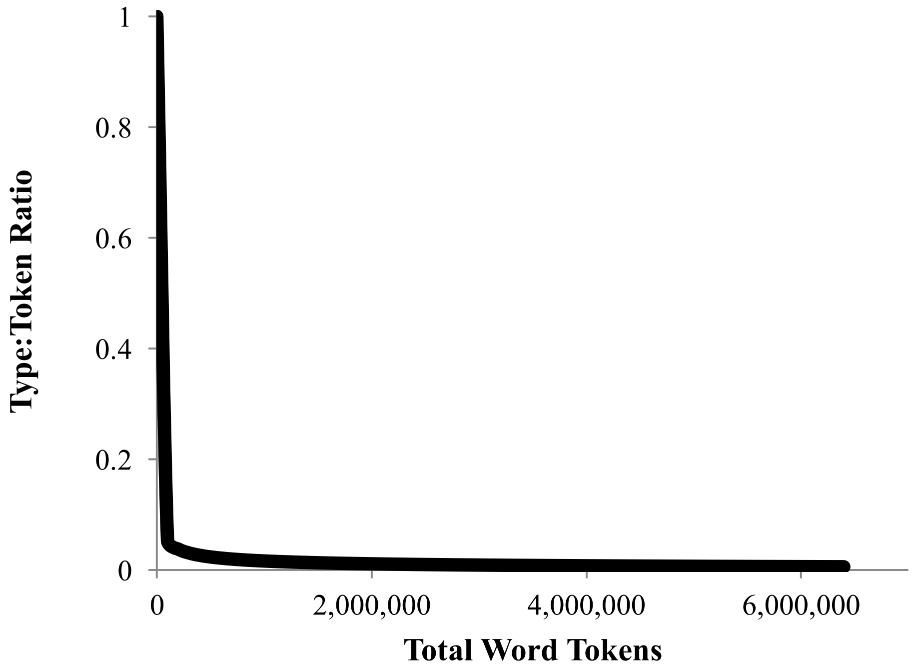


Figure 2: The frequencies of the 6.5 million words in CHILDES, sorted by frequency rank (top). The number of unique word types at given a random selection of tokens at increasing token sizes (center). The type-token ratio of unique word types to total token number at increasing token sizes (bottom).

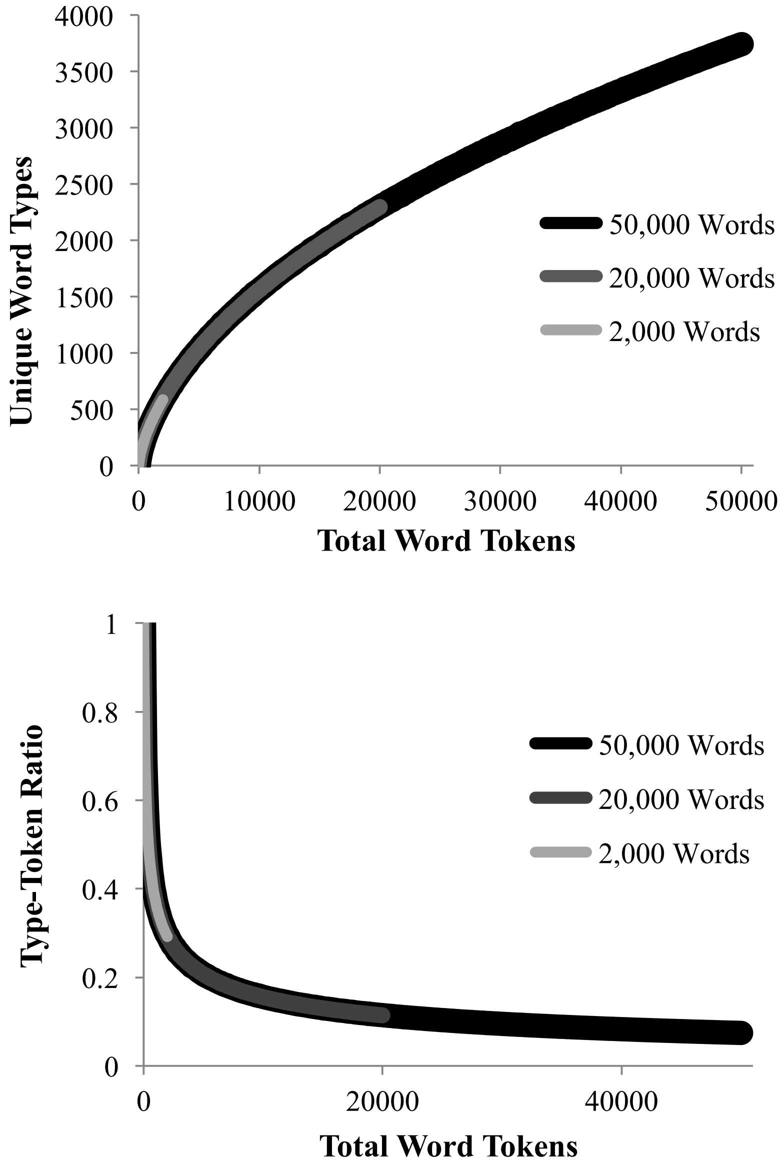


Figure 3: Total unique word type counts by total token counts (above) and type-token ratio by total word tokens (below), for 2,000, 20,000 and 50,000 words.

Figure 2 shows the function relating number of types and to number of tokens for child directed speech sampled at the larger scale. To create this figure, we randomly selected samples, with replacement, from all 6.5 million words of CHILDES that increased in increments of 100,000 words, thus collecting the types and tokens that a child who hears 20,000 words a day might hear in less than 5 days (100,000 tokens) up to about a year (6.5 million words). We then calculated the number of unique word types at each of those sample sizes, yielding counts of the number of unique words in samples of varying sizes, spanning the range from 100,000 to 6.5 million words. We repeated this sampling procedure 100 times, and calculated the average number of unique word types at each of the sample sizes. This allowed us to generate the figure shown in the middle panel of Figure 2: the number of unique word types at different token sizes. As predicted by Heaps-Herdan law, the number of unique word types increases as the total number of word tokens increases, but at a rate that slows at larger token sizes. And again, the bottom panel of Figure 3 shows the extreme sample size dependence of type-token ratios and the non-informativity of a one-time measure of types and tokens at this developmental scale.

Figure 3 depicts the function relating number of types and to number of tokens at a smaller scale, the speech heard in day by three hypothetical children whose learning environments differ only in the amount of child-directed speech in a day. To create this figure, we first randomly selected 2,000, 20,000 or 50,000 words from all 6.5 million words of childes. We then sampled, with replacement, samples that increased in increments of 100 words, and calculated the number of unique words in each of those samples. This sampling procedure was repeated 100 times and the top panel of Figure 3 shows the average type count across the 100 samples. The bottom panel was created by dividing the number of unique type counts, calculated above, by the total number of word tokens, and plotting that ratio against the total number of word tokens used to calculate that ratio.

First, these graphs show the staggering difference across children that exist in terms of the amount of spoken language children typically hear, and the unsuitability of a single type-token ratio to describe differences in the samples. All three samples were randomly drawn from the same population, so variation in type-token ratios purely reflects differences in sample size, or the rate with which children move along the type-token curve (see also Carroll, 1964; Hutchins, Brannick, Bryant & Silliman, 2005; Richards, 1987).

Because this relation between types and token counts is nonlinear, researchers in the past have often forced to-be-compared samples from different children to be the same sample size by truncating the larger sample to the length of the smallest sample in the dataset (Hoff & Naigles, 2002, c.f. Richards, 1987; Malvern et al., 2004). But as illustrated in Figures 1, 2 and 3, this is not ideal as the same measured type-token ratio can emerge from very different combinations of total tokens, total unique words and the sample size. In brief, type-token ratios from a single sample of speech can be misleading about the character of the speech as a whole. For this reason, some researchers have proposed that we abandon type-token ratios as a measure of lexical diversity (Malvern et al., 2004; McKee, Malvern & Richards, 2000). The alternative, however, is to use *the curve* relating numbers of types to numbers of tokens as the measure of the learning environment.

2. Families of Curves

Caregivers differ in the words they say to children. This may be because of differences in the words they know (see Bornstein Haynes & Painter, 1998; Rowe, 2008) or their beliefs about the words appropriate to use with children. For example, whereas one parent may label an object a “contraption,” another, may label it with the best ordinary word they can find, such as “truck” (Gleitman, Newport & Gleitman, 1984; Hayes & Ahrens, 1988; Snow, 1972). All these caregivers will generate language samples that fit Heaps-Herdan law and look like Figure 2, but the shape of the individual curves will differ. Here, we follow the lead of Malvern and Richards and colleagues (Malvern et al., 2004; McKee, Malvern & Richards, 2000) and create a family of curves that reflect these differences.

We begin with the baseline type-token curve generated from the entire CHILDES corpus and then simulate caregivers with different abilities and/or tendencies to include diverse words by randomly deleting all tokens of 10% or 20% of the types in the CHILDES corpus. In this way, we create three sets of simulated caregivers: One with 100% of CHILDES child-directed vocabulary, one with 90% of that child-directed vocabulary, and one with 80% of that child-directed vocabulary. For the 10% reduction in child-directed vocabulary, we generated a list of all the unique word types in the entire corpus and then randomly selected 10% of those unique words and eliminated all instances of those words from the CHILDES corpus. We then performed the same sampling procedure described previously, randomly selecting samples that increased in increments of 100,000 from the CHILDES corpus, and in each sample counting the number of unique word types. We repeated this procedure 100 times, each time randomly selecting a different 10% of the total unique word types to eliminate form the corpus. The procedure was identical for 20% reduction of child-directed vocabulary.

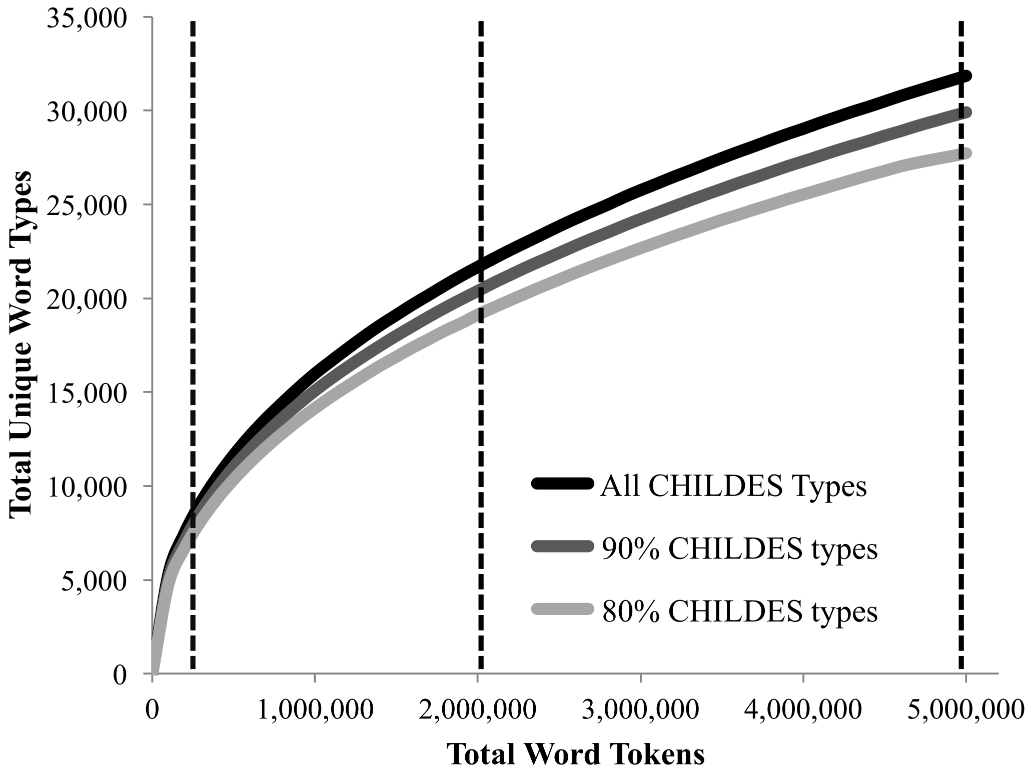


Figure 4: Type-token curves for three hypothetical children. One child's linguistic input drew from all unique word types contained in child-directed CHILDES (black line), one child's input drew from all but 10% of the unique word types in child-directed CHILDES (dark gray), and once child's linguistic input drew from all but 20% of the unique word types in child-directed CHILDES (light gray).

The results are shown in Figure 4: Learning environments with different sizes of child-directed vocabularies yield different type-token curves. Notice that although the at-scale linguistic experiences of children who learn words in the three different environments will differ – in the total tokens, in the repetitions of words, and in the diversity of those words – those differences will not be apparent in a single time-point measure of types and tokens: There are points on the 80% curve higher than those on the 100% curve. This point is tautological but is profoundly important for a unified understanding of learning environments and their malleable properties. For example: A child along the 80% curve may hear more unique words than a child along the 100% curve, if the quantity of speech they hear in a unit time is much greater and therefore the child moves faster along that lower curve. Is this a more or less optimal learning environment than moving more slowly on a higher curve? Is the number of unique words heard in a unit time the most important factor or is it the whole distribution with its repetitions of words and diversity? The answer is that we do not know.

Consider three children, one who hears 2,000 child-directed words a day and one who hears 20,000 and one who hears 50,0000. The three dashed lines in the figures show how many types and tokens these children would hear in 100 days. Along any curve, talking more yields more unique words (in a unit of time). Parents who talk more – whatever their differences in child-directed vocabulary – will move along this curve faster. In brief, children's learning environments may be characterized by (1) different type-token curves and (2) different speeds of movement along those curves. The relative contribution of these two components of the environment is not known. This would seem critical because the amount of child-directed talk (per unit time) is a malleable factor in learning environments. Thus, the proposal that amount of talk is the key to remediating individual differences in word learning may, in this way, be right, as may the general advice that parents should be encouraged to talk more to their children (e.g. Weisleder & Fernald, 2014). But, then again, it may not be right –a higher curve (or perhaps even lower type-token curve may be advantageous if there is some sweet spot of repetition and diversity) may be optimal for vocabulary growth. If the shape of the curve varies across children and if the shape of the curve matters, not just the rate of movement, can we change that shape? Is there a way for young learners to "jump" curves, to move from a language-learning environment characterized by a lower curve to one characterized by a higher one?

3: Changing the curve

Caregivers' selection of the words they say to their children is not only constrained by the caregiver's vocabulary and beliefs about child-appropriate talk, but also by context. Day in and day out, conversations about eating breakfast or getting dressed may present little diversity in the words directed to the child while a new event, such as a trip to a zoo or museum, may provide an influx of new words. Indeed, research studying how parents talk in different contexts supports this conclusion and as young children often show gains in vocabulary immediately following novel experiences such as trips to zoos (Benjamin, Haden & Wilkerson, 2010; Borun, Chambers, Dritsas & Johnson, 1997). Others have noted how picture books also provide an easy way for parents to expand contexts and topics (Snow, 1983; Massaro, 2015; Montag, Jones & Smith, 2015). Here we use picture books as our case example of how talk across varying contexts may enable parent talk to jump from one curve to another. In the simulated environments in this section, we use the text in picture books as the new-context words that can be added to the baseline environment. We chose books because we can use the text in in picture books as a sample of, albeit imperfect, (parents do not always read all words in the text; Deckner, Adamson & Bateman, 2006; Fletcher, Cross, Tanney, Schneider, & Finch, 2008; Hudson Kam & Matthewson, 2016; Whitehurst et al., 1988) source of new context words that can be added to the baseline environment. We believe this is a reasonable simulation approach because large representative surveys of parents indicate that many parents report reading books to their children at least once a week from infancy onward (Young et al., 1998). Parents chat conversationally about the contents of the book but also read the text (Deckner, Adamson &

Bakeman, 2006; Dickinson, Griffith, Golinkoff & Hirsh-Pasek, 2012; Fletcher, Cross, Tanney, Schneider & Finch, 2008; Hudson Kam & Matthewson, 2016; Mol, Bus, de Jong & Smeets, 2008; Ninio & Bruner, 1978; Whitehurst et al., 1998). Thus, the text in child-directed books provide a reasonable proxy for adding contexts to parent talk.

The starting point for the simulated environments in this section are 100% CHILDES child-directed vocabulary, the 90% child-directed vocabulary, and the 80% child-directed vocabulary of Figure 4. To each of these environments, we added the words from 100 common picture books (see Montag, Jones & Smith, 2015), thereby increasing the total words available in each environment by about 68,000 tokens. We then regenerated the type-token curves on this expanded sample. Note, this is objectively a small change. The “year-long” vocabulary for the smallest vocabulary (80% of CHILDES types) contains about 5 million tokens. One hundred books in a year is relatively small amount books in the lives of many infants and children (Bradley, et al., 2001; Deckner, Adamson & Bakerman, 2006; Young, Davis, Schoen & Parker, 1998). The 68,000 tokens are an addition of less than 1.5% of the total words.

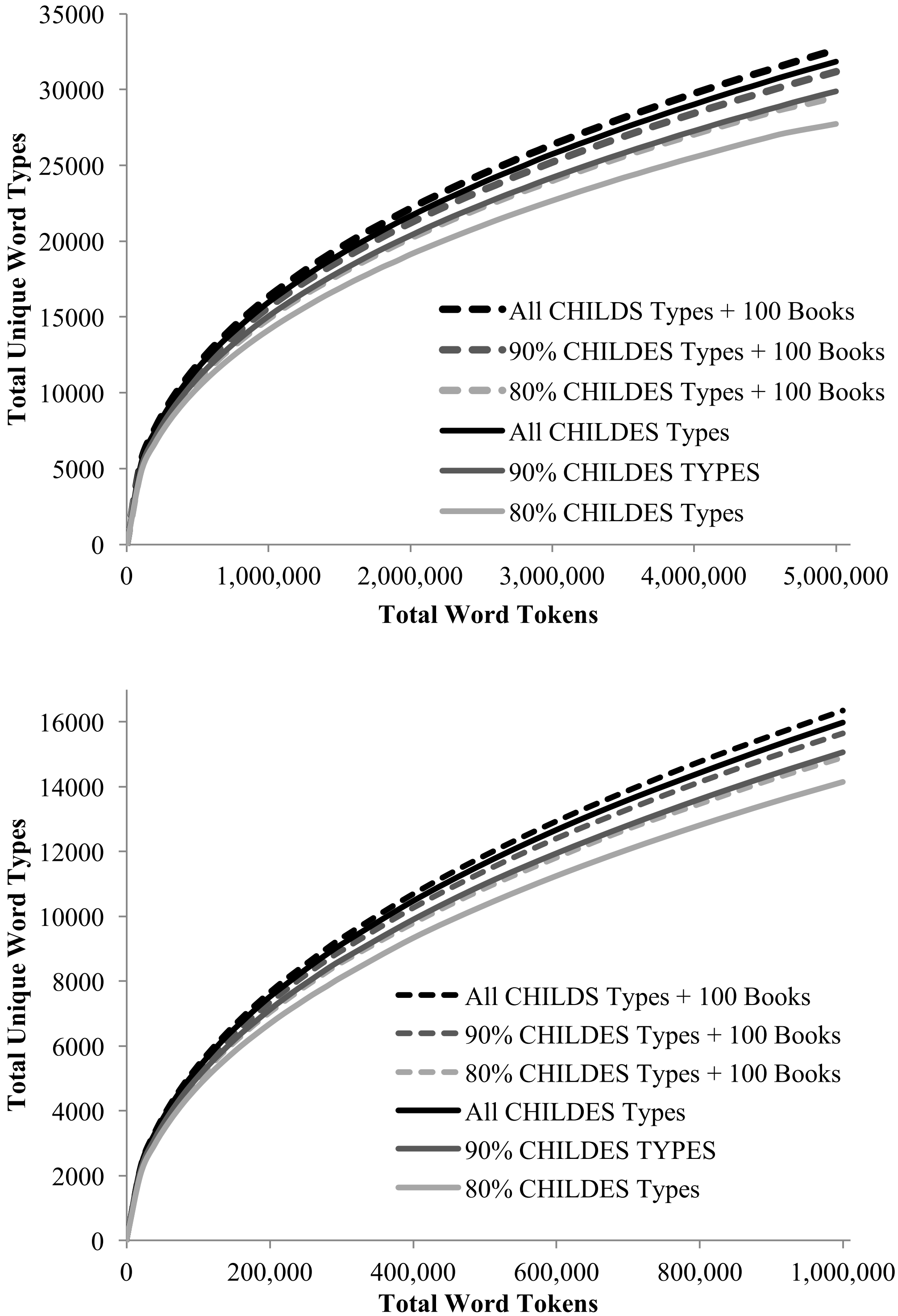


Figure 5: Type-token curves for six hypothetical children. The three solid lines refer to the same three hypothetical children plotted in Figure 4. The three dashed lines refer to what these three children's linguistic input would look like that they additionally received linguistic input in the

form of the text of 100 picture books. The bottom panel enlarges the bottom-left portion of the top panel, showing type counts for up to one million total word tokens (note the scale invariance of the curves).

However, as is evident in Figure 5, this small change in the words in the vocabulary sampled for child-directed talk changes the shape of their type-token curve. Moreover, the change is greater for the baseline environments with originally lower type-token curves—in the 80% curve, the books yield an increase of over 6% in total word types, and an increase of over 4% for the 90% curve and an increase of about 2.5% in the 100% curve (for a less than 1.5% increase in types). While all children may benefit from the more lexically diverse vocabulary in picture books, children who hear less lexically diverse spoken language from caregivers and who hear less total talk may particularly benefit from this additional source of linguistic input. The bottom panel of Figure 5 –which enlarges the early end of the type-token curve – shows that these gains emerge in smaller quantities of speech input, having effects on the early end of these curves and are not limited to large, aggregate word counts. Differences in unique token counts for smaller samples mirror those of larger samples showing 1) The scale invariance of the power-law distribution of words in a language and 2) That differences in lexical diversity, based on the range of contexts in which talk is generated, have discernible effects even at small sample sizes.

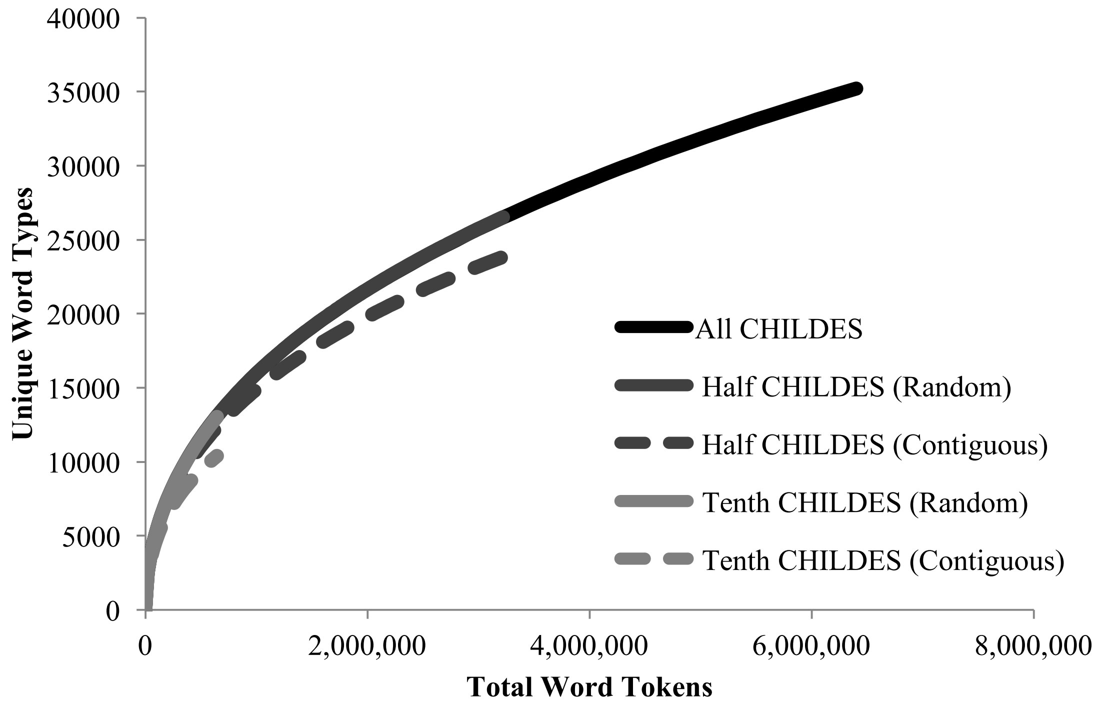
Although these findings might seem to support the idea of picture-book reading interventions to bolster language learning environments, this is not our main point. Instead, our point is that to understand the relevant properties of word learning environments, we need to understand how frequency distributions of words in the environment can vary. This simulation makes two related points: (1) Learning environments that differ in as little as the words present in children's common picture books (or the words likely to be evoked on trips to museums, zoos, and lighthouses) present fundamentally different type-token relations that may matter to language learning beyond the amount of parent talk per unit time, and (2) Relatively small differences to learning environments in terms of varying the contexts of talk may underlie observed differences in the shape of the type-token curve.

4: The distribution of contexts

Because talk is coherent and tied to the context in which it occurs, the distribution of words in time is not random (Church & Gale, 1995). For example, talk about bowls is likely to co-occur with talk about spoons and there may be many mentions of bowls close in time to each other in the morning, and few in the evening. In brief, the distribution of words in time is lumpy and bursty. They appear systematically in lumps of co-occurring words (Altman, Pierrehumbert & Motter, 2009; Firth, 1957; Landauer & Dumais, 1997; Sahlgren & Karlgren, 2005) so that the likelihood with which a word is encountered in a context is not equal to the base rate frequency of that word in the learning environment of the learner but is related to the other words uttered in this context. Individual words also appear in bursts in time (Katz, 1996; Kleinberg, 2003) and are more likely to appear at any moment if they recently appeared. Again, the likelihood with which a word is encountered at any moment is not equal to the base frequency but is conditional on whether it just appeared. These properties have been conceptualized as emerging from the same processes that generate the power-law distribution of words in speech production (Altmann, Pierrehumbert & Motter, 2009; Serrano, Flammini & Menczer, 2009).

In our previous simulations, we ignored the lumpy and bursty nature of words and treated the CHILDES as a "bag of words," randomly drawing words from the whole corpus at different sample sizes. When one samples randomly from a big bag of words, the shape of the sampled distribution is similar to the shape of the distribution for the whole bag. However, if one samples words in segments of coherent conversations, then the shape of the sample distribution is not similar to the shape of the population distribution. *This is because coherent conversations are more repetitive and less lexically diverse.* We first demonstrate this fact and then consider its broader implications, as narrative coherence is a known positive factor in early word learning (Rowe, 2012; Snow, 1983).

As in the previous simulations, we begin with the CHILDES corpus. In one set of samplings, we treat the problem, as we did in the previous demonstrations, as sampling from a big bag of words. But the CHILDES corpus is not, at its origins, a big bag of words. It is instead a series of coherent conversations, with each conversation narratively and contextually constrained in time and place. Formally, then, by taking conversations into account we shift from a conceptualization of the input as a big bag of words to a series of little –conversation sized –bags. To show the consequence of conceptualizing type-token relations within the "one big bag of words" versus of a series of conversational bags, we calculated type and token counts in subsets of CHILDES as we sampled subsets words in different ways. We used the same overall sampling procedure we used to create the plot in the center panel of Figure 1, randomly selecting samples from CHILDES that increased in increments of 20,000 words and calculating the number of unique word tokens at each of those sample sizes. However, in this simulation, we selected those words either randomly from the whole corpus (as in prior simulations) or in sets of contiguous words. We did this for smaller total samples of CHILDES (one half, one tenth and one fiftieth of the corpus) than in previous demonstrations because contiguous sampling yields the same full bag of words as random sampling when all words are sampled. The distributional properties of words in conversations –even when aggregated over many conversations – are seen in these smaller samples. The randomly selected subsets were selected by randomly selecting either one half, tenth or fiftieth of all the words (total word tokens) in child-directed CHILDES.



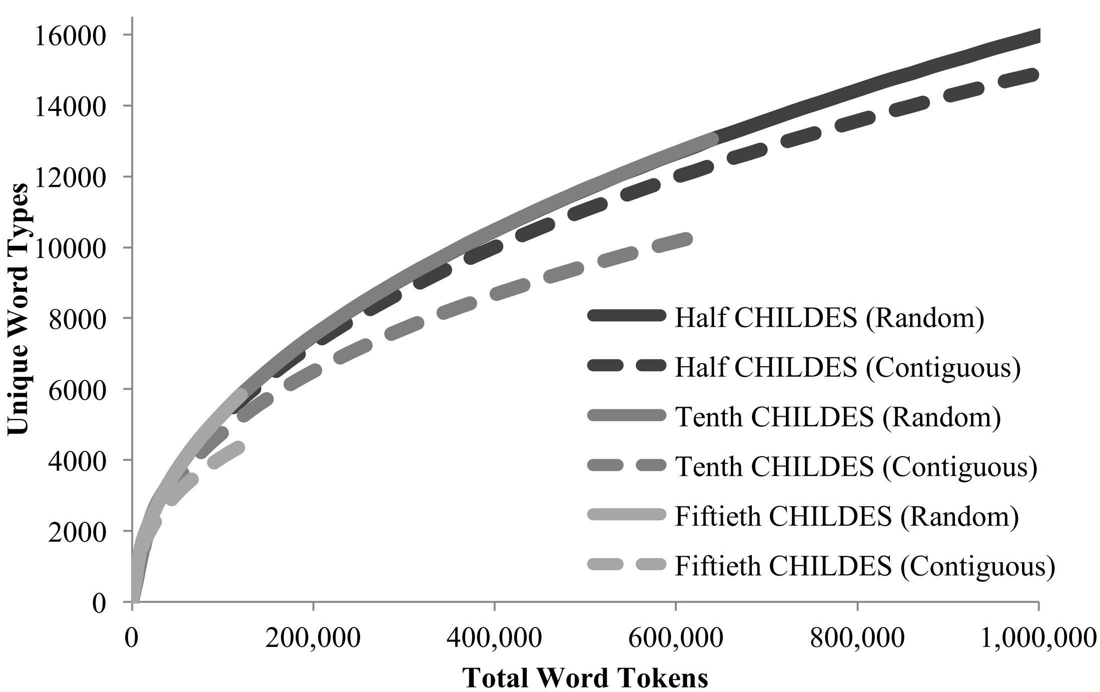


Figure 6: Type counts at different total token size in child-directed CHILDES, sampled in different ways. The black solid line is the same line presented in Figures 3-5, and refers to the total number of unique word types at increasing total token sizes. The gray solid lines refer to type counts and different token sizes selected from a random selection of child-directed CHILDES. The gray dashed lines refer to type counts and different token sizes selected from contiguous selections of child-directed CHILDES.

The results are shown in Figure 6. The solid lines were generated by calculating the number of unique word types at different sized random samples drawn from the entire child-directed CHILDES corpus and the dashed lines were generated by calculating the number of unique word types in different sized contiguous samples. The point, clear in the figure, is that word types, as a function of word tokens grows much more slowly when words are sampled as contiguous coherent samples of speech, which is of course, how children experience that speech. The shape of curve was dependent on corpus size (half, tenth or fiftieth of the whole corpus) but only for contiguously sampled CHILDES subsets, not for the bag of words sampling approach. The half of CHILDES sampled contiguously contains 10% fewer unique words than the half sampled randomly, the tenth of CHILDES sampled contiguous contains 20% fewer and the fiftieth of CHILDES contains 25% fewer unique words than the randomly sampled counterparts. This is because the smaller sample of contiguous speech means not just fewer words but *fewer conversational contexts* and thus more repetition of high frequency words. The reason that fewer conversational contexts affects lexical diversity is that when sampling randomly, any word that appears in CHILDES is as likely as any other to be selected. So, for example, if “zebra” were selected, “lion” or “dishwasher” would both be equally likely to occur. However, this assumption violates important pragmatics of language. When "zebra" occurs in conversation, perhaps at the zoo or while reading a book about animals, "lion" is far more likely to occur in the same conversation or context. Contiguous sampling of CHILDES accounts for this pragmatic fact about language. These results also show how amount of talk and contexts of talk co-vary when conversational coherence is taken into account. Limited talk and limited contexts of talk may provide a particularly poor learning environment. Relative to this point is a body of previous work showing differences in the properties of language and words generated in different contexts, for example, playtime conversation is more concrete and object focused whereas mealtime is more abstract and storytime includes more rare words and greater lexical diversity (Beals & Tabors, 1993; Hoff, 1991; Sosa, 2016; Weizman & Snow, 2001).

Beyond the unique contributions of different contexts of speech, the point we make here is that a greater diversity of contexts *itself* is associated with greater lexical diversity. In fact, the addition of even a small number of unique contexts can have consequences for overall lexical diversity. To illustrate this, to our contiguously sampled speech from childes, we added in the text of picture books, and observe a marked increase in the slope of the type-token curve. We show the resultant curves in Figure 7. In this figure, we started with a contiguously sampled tenth of child-directed CHILDES (about 650,000 words), which represents about a month of speech for the average child. To this, we added the text of either 10, 50 or 100 different picture books, numbers that are all within the range of books experienced by young children, with 100 unique books representing the higher end of distribution (Bradley, et al., 2001; Deckner et al., 2006; Young et al., 1998). From that sample of language, we then sampled, as in other simulations, samples increasing in size of 20,000 words and counted the number of unique words in each sample. We then repeated this technique 100 times, each time with a different contiguous sample from CHILDES, and with a different random sample of picture books, and plotted the mean word count of these 100 samples.

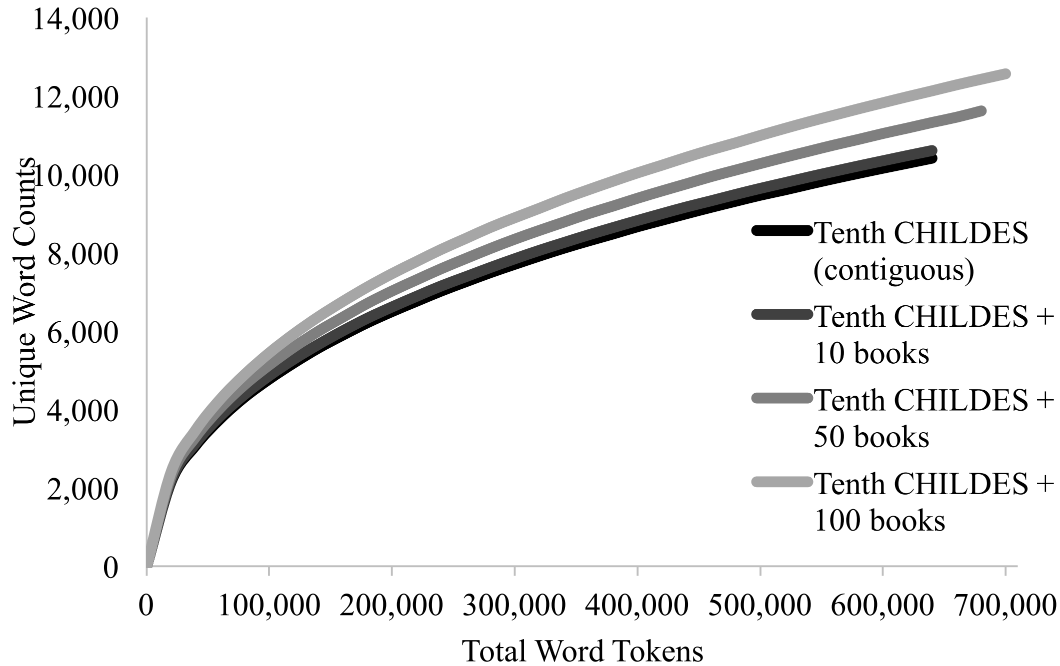


Figure 7: Type-token curves for a contiguously sampled tenth of child-directed CHILDES (about 650,000 words), plus the text of 10, 50 or 100 unique picture books.

Small additions of picture books text, which often consists of language in contexts outside those of day-to-day activities can have a profound effect on the total lexical diversity of the sample. Adding only 10 picture books yielded an increase of just under 2% unique word tokens. The average book length was 680 words, so even 10 books represents only about 1% of the total language sample. Ten picture books over the course of about a month is well within the experiences of the modal child (Bradley, et al., 2001; Young et al., 1998), though admittedly there is not existing data regarding how often books are repeated. Adding text of 50 picture books in associated with an almost 9% increase in unique word types and 100 picture books is associated with a 16.5% increase in unique word types. While 100 unique picture books a month is a very large number, given the very high number of picture books in the homes of some children (in a laboratory sample, average of 126, range of 13-1750; Deckner et al., 2006), 100 unique books may not be entirely unrealistic for a small subset of children, and 100 books of any sort is likely very realistic for some children at one end of the distribution. That said, our goal is not to literally model a month's worth of language input, but rather to illustrate the consequences of adding in language taken from a range of contexts on overall lexical diversity. Even the additions of small numbers of unique word contexts can have notable consequences for the lexical diversity of a language environment.

These observations also have implications for how we should sample the input when measuring environments. Given the contribution of conversational context on observed lexical diversity, we need to know how conversations contexts are distributed differently in different families. One possible approach is to use new wearable technology that can yield day-along or multiple day recordings. The distribution and rate of contextual diversity could be estimated by sampling parent talk at set temporal windows across the day, or in the ideal, over multiple days. We also need to understand how contextual diversity co-varies with amount of talk in real children's environments. If constrained *contexts* are the principle factor creating *less* talk and *lower* type-token curves in the input for some children, then instructing parents to talk more may not be enough to alter the input in meaningful ways.

5: Analysis of a sample child

The power-law distribution of word frequencies in language, the size-dependence of type-token ratios, and the burstiness of language as a consequence of conversational context, all matter for the analysis of naturalistic datasets. We illustrate the consequences of these principles for studying the environments of three individual children using longitudinal data—large amounts of speech directed at single children—contained in the CHILDES dataset. These sample children may not be typical in their language learning environments, but at least two of the three children we will discuss (Sarah and Adam) are not children of academics (Brown, 1973). However, they provide a way to demonstrate the applicability of the present simulations to the study of individual differences and the word learning environments of real children.

Nina is a child for whom longitudinal speech input is available in the CHILDES corpus (Suppes, 1974). She was recorded from age 1;11-3;3 and the corpus contains 52 individual sound files for a total 195,303 words. The following analyses investigate only the speech directed to Nina in the CHILDES corpus. First, Figure 8 shows the cumulative type and token counts contained in each contiguous recording in Nina's dataset.

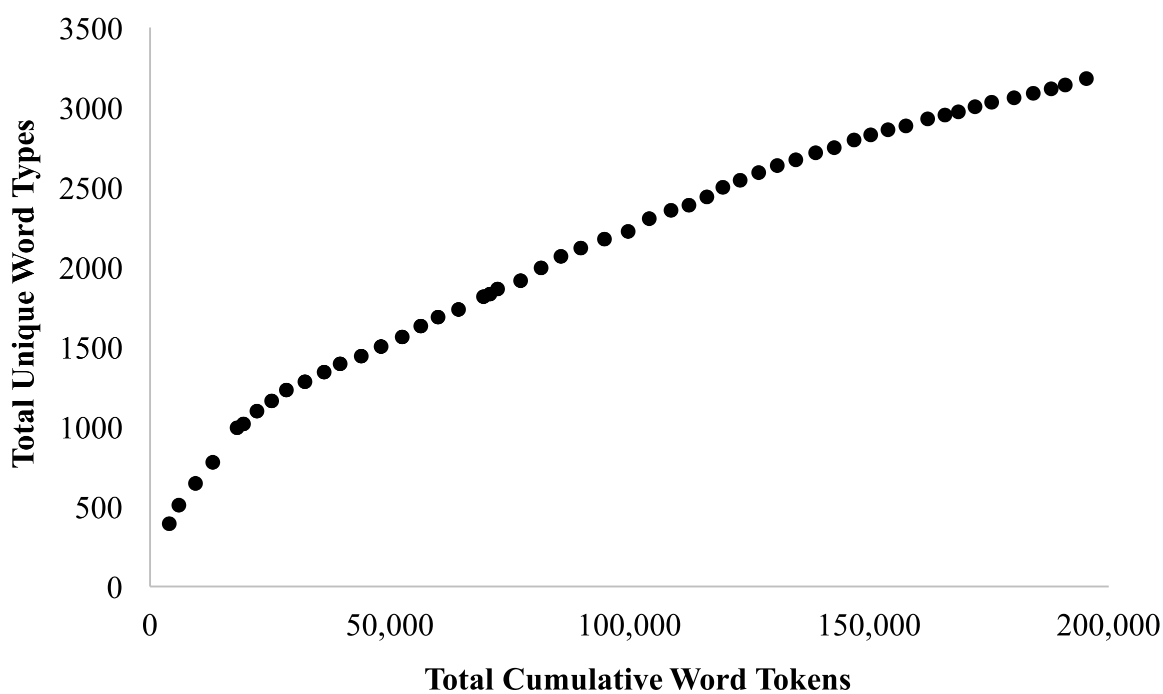


Figure 8: The cumulative number of types and tokens in Nina's language input. Each point refers to one of the 52 contiguous recordings and they are arranged chronologically (1;11-3;3)

This curve is similar in shape to the simulation data suggesting 1) The simulation data, which treats language as an unordered bag of words indeed captures something real about the shape of children's aggregated experiences data and 2) That the increase in the unique type count attributed to new word tokens decreases as the total sample size increases, the relation between types and tokens described by Heaps-Herdan law, is evident in naturalistic, longitudinal data from a single child (and at a scale of just under 200,000 words).

Next, we show hypothetical data that represents what Nina's input might look like if her caregivers used 10% or 20% fewer unique word tokens. In this analysis, like those in Simulation 2, we lumped all speech to Nina together, then removed either 10% of 20% of the unique tokens, and selected random samples that increased in increments of 10,000 words. The resultant type and token counts (mean of 100 runs with a different random sample of word types excluded each time) is plotted in Figure 9.

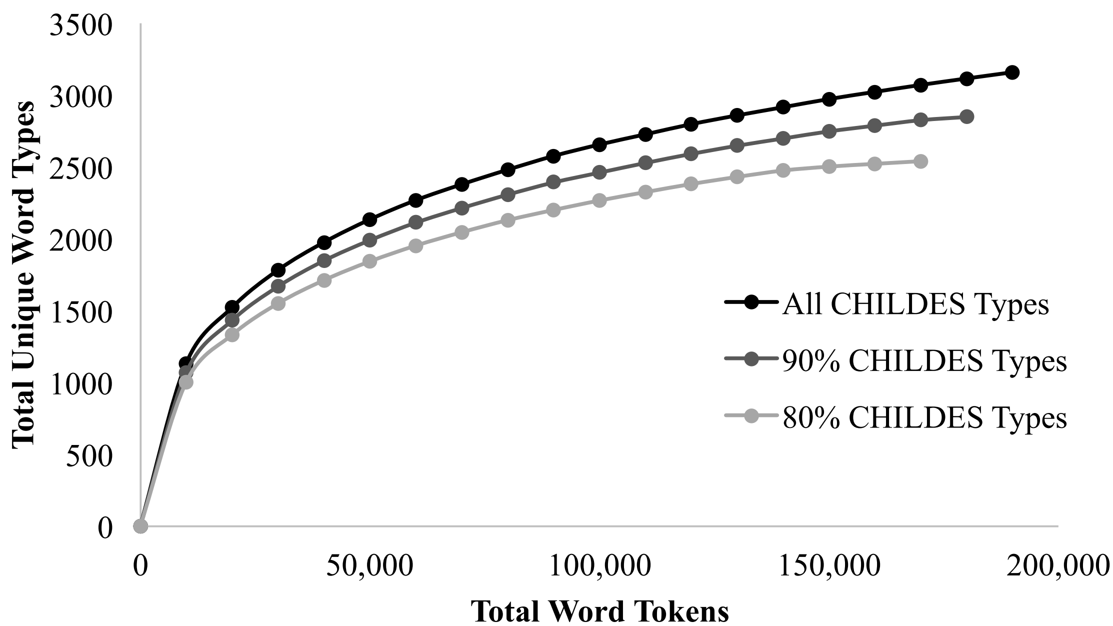


Figure 9: All 195K words of speech directed to Nina, and hypothetical data with 10% or 20% of all unique word types removed.

Again, analyzing data from a single individual yields the same pattern of results as did analyzing aggregate data from multiple individuals. As in Simulation 2, we see a family of curves that vary in slope as a consequence of lexical diversity. These curves illustrate the dissociation of the amount of speech and the lexical diversity of speech to children. The lexical diversity of caregiver speech is illustrated by the three different curves while the amount of speech is represented by location along the x-axis. These are two important parameters, diversity and amount of speech per unit time that can theoretically operate independently, and may each be important parameters to explore when measuring speech to children.

Finally, we illustrate the importance of sampling technique when estimating a child's language environment, by comparing Nina to two other children with longitudinal speech input in the CHILDES corpus, Adam and Sarah (Brown, 1973). Adam's (age 2;3-5;2) dataset consists of 55 sound files, containing a total of 123,811 words of speech directed at Adam. Sarah's (age 2;3-5;1) dataset consists of a total of 115 sound files, containing a total of 176,208 words. For reasons pertaining to analysis technique, files containing fewer than 1,000 word of speech directed at the target child were removed, which only affect 24 sound files (17,575 words) removed from Sarah's dataset.

First, Figure 10 shows the cumulative type and token counts for Nina (analogous to Figure 7), plus Adam and Sarah.

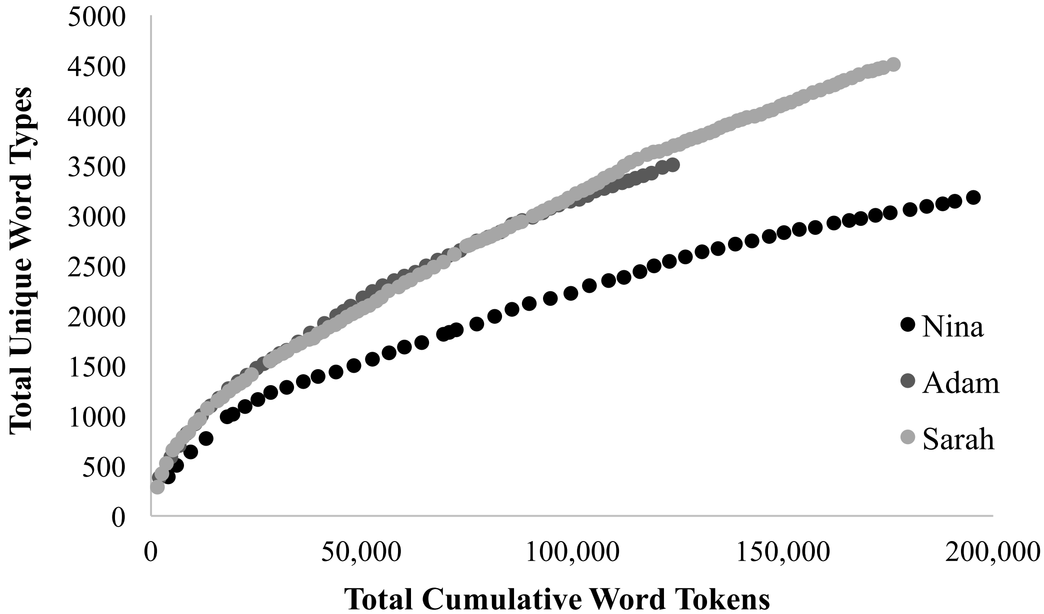


Figure 10: The cumulative number of types and tokens in Nina, Adam and Sarah's language input. Each point refers to a single contiguous sound file.

First, it is immediately obvious that Nina's curve is below the curves of Adam and Sarah, which are nearly overlapping. This may suggest Nina's language input is less lexically diverse relative to the inputs of Adam and Sarah, as illustrated with the families of curves in simulations 2 and 3. However, a second important observation is that Nina contains fewer, longer sound files than Adam and Sarah. Languages is bursty with repeated words in a context. Thus, a relevant question is, how much of the difference between Nina and Adam and Sarah could be attributed to observation that Nina's dataset may contain fewer unique conversational contexts? To answer this question, we selected only the first 1,000 words from each data file, as a rough proxy for equalizing the number of conversational contexts (we can assume that longer recordings generally contained a larger number of unique conversational contexts), so all recordings were equated for length. Figure 11 shows those curves.

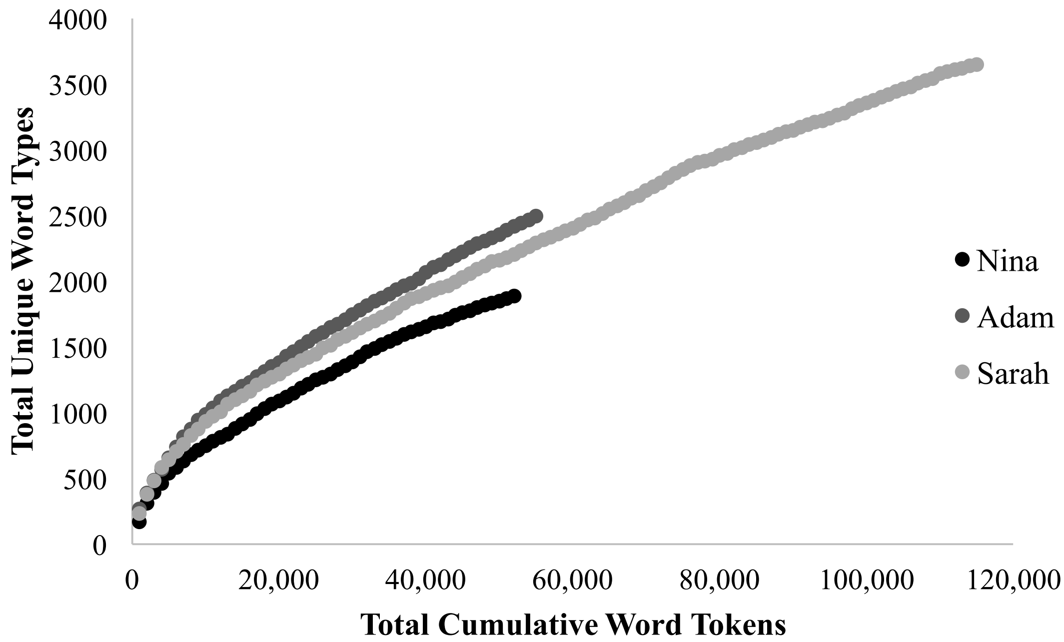


Figure 11: The cumulative number of types and tokens in Nina, Adam and Sarah's language input, when including only the first 1,000 words of each sound file. Each point refers to a single contiguous sound file.

As shown in Figure 10, after controlling for file size, the gap between Nina and the other two children has narrowed considerably. Nina's curve is now only slightly below Sarah's, which may be slightly below Adam's. In short, the qualitative pattern of curves changed dramatically as a consequence of equalizing the number of contexts from which the speech to children is obtained, suggesting that this may be a significant source of variability that is not often accounted for when comparing language input across different children, or two corpora of different sizes and construction, more broadly. This again suggests that the critical measure for analyzing contexts is the distribution of different contexts.

Finally, because Nina is younger than the other two children, equating all three children for age, and the number of recordings at each age, yields Figure 12. Age was equated by selecting the same number of sound files across the same age range (2;3-3;3), spaced approximately equally, for all three children. Now, the three children's curves are more similar, with a possible Adam-Sarah-Nina pattern of decreasing lexical diversity emerging.

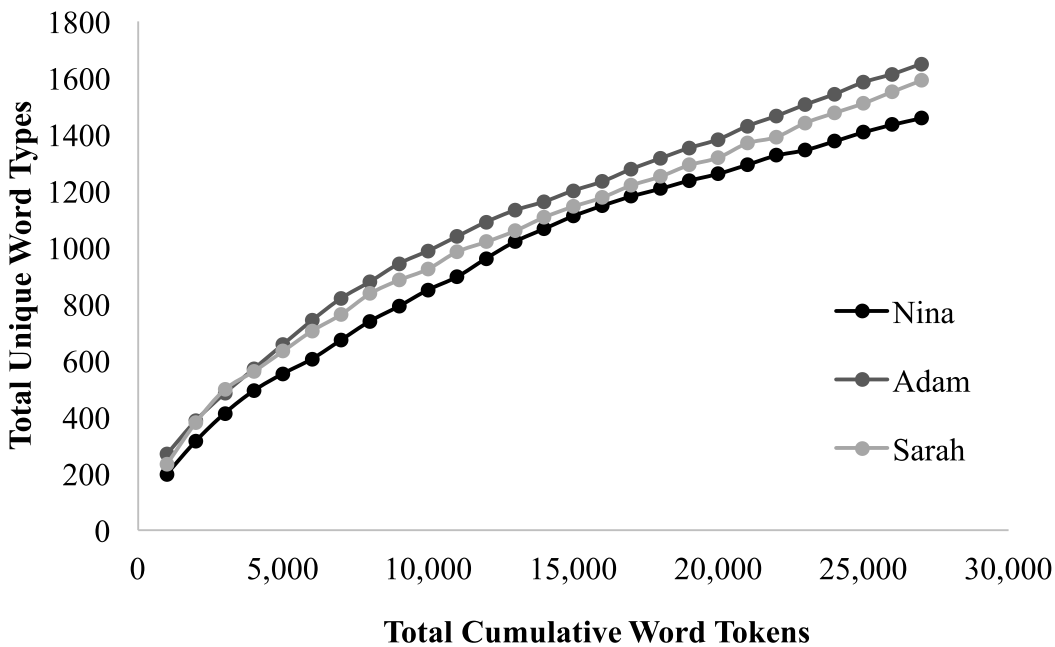


Figure 12: The cumulative number of types and tokens in Nina, Adam and Sarah's language input, when including only the first 1,000 words of each sound file. Each point refers to a single contiguous sound file

Had we only looked at Figure 10, we might have concluded that Nina encounters less lexically diverse speech than Adam or Sarah, and made predictions for Nina's vocabulary accordingly. However, when we properly control for how the language samples were collected (Figures 11 and 12), we now see that the lexical diversity in the speech countered by these three children are quite similar, and maybe we would *not* expect predictions on the basis of lexical diversity across input to be borne out in, for example, the vocabularies of these three children. Of course, an additional source of variability is the *amount* of speech each child encountered, and equating for sample size obviously ignores that potential source of variability. From these limited samples, and different procedures used to collect the data, we cannot make strong conclusions about the learning environments of these three children. Our point in this final analysis however, is three-fold: First, it is possible to determine the type-token curves for individual children. Second, the learning environment may be conceptualized as composed of the number of tokens, their diversity, which is dependent on the diversity of conversational contexts, and the rate of movement (words per unit time). Third, and relatedly, children's environments are not fixed; they need not be stuck on a single curve. The present analyses suggest that diversity in the contexts of talk may be the most effective way to alter the diversity of words in the learning environment.

Discussion

The simulations and analyses presented in this paper explore the ways in which well-known distributional properties of words in language interact to determine the word-learning environment. The field is at the edge of barrier breaking approaches (VanDam et al., 2016) that measure children's lexical learning environments at a much larger scale than has possible in the past. The specific contribution of the present study, then, is a characterization of how and why learning environments can vary, a contribution that has implications for how we think about and analyze these new larger scale measures of learning environments. These insights from these new at-scale measures, in turn, have implications for determining how differences in input environments affect the rate of children's vocabulary development, and, finally, for how we might encourage more optimal learning environments for all children.

Amount of talk

Parent talk is language, and thus the properties of parent talk and the differences between individuals must be understood within the laws of how words are distributed within language. The present analyses of simulated environments show the conceptual, methodological and ultimately practical implications of this stance. In the growing literature on the long predictive reach of early vocabulary size for developmental outcomes (Fernald, Perfors & Marchman, 2006; Hart & Risley, 1995; Huttenlocher, et al., 2010; Marchman & Fernald, 2008; Rowe, 2012; Senechal & LeFevre, 2002; Walker, Greenwood, Hart & Carta, 1994), the mounting evidence suggests that individual differences in parent talk strongly determines vocabulary size (Hart & Risley, 1995; Hirsh-Pasek, et al., 2015; Hurtado, Marchman & Fernald, 2008; Hoff, 2003; Hoff & Naigles, 2002; Huttenlocher et al, 2010; Pan, Rowe, Singer & Snow, 2005; Rowe, 2008; 2012; Shneidman, Arroyo, Levin & Goldin-Meadow, 2013; Weisleder & Fernald, 2014; Weizman & Snow, 2001). The growing public health efforts to reduce the inequalities in language learning environments makes understanding the distributional structure of words in parent talk particularly urgent. These statistical properties –in their full complexity –must be understood before we determine how parent talk influences early vocabulary development.

The simulated environments highlight three consequences of the distributional properties of words for differences in word-learning environments. First, more talk is positively but nonlinearly related to more unique words. This means that different word-learning environments need to be characterized—not by number of tokens, not by number of types, not by the ratio of types to tokens—but by the curve that relates types to tokens and by the speed with which children's aggregated word experiences move along that curve. Second, there are potentially different shaped curves relating types and tokens in the language learning environment and these shapes depend on the total vocabulary available for child-directed speech in that environment and on the distributions of contexts in which that speech is generated. Third, learning environments can shift from lower to higher curves (or higher to lower curves) with relatively small changes in the diversity of contexts of talk (e.g., with the addition of a context equivalent to reading one or two picture books a week).

These observations from *simulated* word-learning environments have consequences for how we conceptualize and measure *real-world* environments. There are many open questions. One possibility is that language-learning environments only vary minimally around a single type-token curve. This would be so if all parent sampled the words in the language in the same way and thus all converged on the same distributional properties as whole. If this were so, then the relevant differences between child-directed talk in different language learning environments would be the amount of talk. Given a common shape for the type token curve across children, amount of talk to an individual child would determine: (1) the total number of unique words the child has heard at any point in development, (2) the type-token ratio at any point in development, and, critically, (3) the speed with which the child moved along the curve aggregating life-time experiences in total words heard and in total unique words encountered. Total talk, then would be the single most important control factor in the experiential properties determining vocabulary development.

Is it really possible that learning environments all present essentially the same type-token curve and that variations in learning environments is primarily related to the rate at which words in that environment are encountered? This possibility cannot be rejected, especially since we do not know how much variation in the curve actually matters to individual learners. Further, although there is clear evidence that parents differ in amount of talk per unit time, we do not know how much the shapes of these type-token curves vary across learning environments when considered at scale. If we add in all the words that a child hears, not just words uttered by a parent, but talk with other children, teachers, shop-keepers, friends, at community events, then the type-token curves from any child might come to largely approximate some idealized distributional structure of language and thus be fundamentally the same for all children. All the relevant differences *could* be in the speed of movement along the curve of total encountered language. Although our personal views are that this is unlikely given the state of current evidence we cannot reject the idea that amount of child-directed speech is the most telling dimension of difference in learning environments.

Projections from samples of parent speech to children (Hart & Risley, 1995; Weisleder & Fernald, 2014; Shneidman, Arroyo, Levin & Goldin-Meadow, 2013), as illustrated in Figure 1, suggest extraordinary differences in the amount of child-directed speech and thus significant differences in the speed with which children move along a single universal type-token curve or any curve. Differences in speed of progression—along any one curve –is likely highly consequential for language development since the total amount of language encountered is a strong predictor of vocabulary development and because the mechanisms of learning depend on encounters with the to be learned items and their repetition. In brief, whatever else matters in the learning environment, rate of movement along the curve is likely to matter with higher rates of input leading to faster growth of the child's vocabulary. The relative size of a child's vocabulary at a given point in development predicts many other aspects language learning, including syntactic development (Bates, Betherton & Snyder, 1988; Bates & Goodman, 1997; Huttenlocher, Vasilyeva, Cymerman & Levine, 2002; Huttenlocher, et al., 2010; Marchman, Martinez-Sussman & Dale, 2004) and the speed and robustness of spoken language processing (Fernald, Marchman & Weisleder, 2013; Weisleder & Fernald, 2014). The size of a child's vocabulary also predicts (and may be a causal factor in) many realms of cognitive development, including, for example, visual object processing (Pereira & Smith, 2009), relational reasoning and problem solving (Augustine, Smith & Jones, 2011; Gentner, 2005), and working memory development (Marchman & Fernald, 2008). Other findings suggest that rate of vocabulary growth in children may be a better predictor of later language than vocabulary size at any one point in time (Rowe, Raudenbush & Goldin-Meadow, 2012). Thus, how fast children build their vocabularies along any type-token curve will have cascading consequences in many other domains. What we do not know is how the speed of movement along *the input curve* of heard words relates to the speed of movement *on the acquisition curve*. Movement along the input and learning curves need not be linearly related. This is a key open question as we move to large scale studies of parent talk and child talk.

The shape of the type-token curve

The analyses of the simulated environments strongly suggest that learning environments will vary markedly not just in the rate of movement along the type-token curve but in the shape of that of that curve. There are three potential sources of difference in the shapes of these curves. First, adults differ in their productive vocabulary sizes (Goulden, Nation & Read, 1990; Zechmeister, Chronis, Cull, D'Anna & Healy, 1995), and thus it is possible that parents with larger and smaller vocabularies will generate different input curves. Second, the words adult speakers know, however, are not the only relevant factors in determining the input (Bornstein, Haynes & Painter, 1998; Rowe, 2008). A potentially more malleable factor in determining the input to children is an adult speaker's beliefs about the appropriate words for use with children. Although this is not a topic that has been extensively studied, there are indications that this may be more critical than parent vocabulary. For example, several (small word sample) studies have reported that there is greater diversity in the words fathers as opposed to mothers use when talking to toddlers (Masur & Gleason, 1980). This mother-father difference has been linked to mothers' closer attention to and expectations about the words the child already knows (Ratner, 1988). Although the robustness and generalizability of these findings is not certain (Golinkoff & Ames, 1979; Hladik & Edwards, 1984), they highlight how different expectations concerning how one talks to a child could alter the shape of type-token curve, and by hypothesis, the rate and character of the child's vocabulary growth. If children develop in communities of adult speakers (parents, grandparents, neighbors, friends, teachers) who share similar vocabularies and similar expectations about how to talk to children, then when considered at scale, the differences in the language environments—and the shapes of the type-token curve of life-cumulative words—could be substantially different for different children.

Third, the simulated environments show how the distribution of contexts of parent talk has major effects on the shape of the type-token curve. This is because language does not just have special distributional properties with respect to the frequency of types and tokens, it also has special properties with respect to the distribution of words in time. The likelihood that someone utters a particular word depends on context (Church & Gale, 1995; Katz, 1996; Kleinberg, 2003; Firth, 1957; Landauer & Dumais, 1997; Sahlgren & Karlgren, 2005). Thus, within a context a small set of words repeat but across different contexts: the park, the store, the museum, a picture book, different words are repeated. Further, research shows that new and unusual contexts (often) yields parent talk that includes and repeats rarer and more "sophisticated" words (Weizman & Snow, 2001) and that these new contexts for talk are linked to children's addition of new words to their vocabulary (Callanan & Valle, 2008; Hoff, 2006; Weizman & Snow, 2001). The analyses of simulated environments show that adding new contexts changes the type-token curve, leading to more rapidly increasing types as a function of tokens. These simulations indicate that we do not just need to understand the distribution of words in parent talk but also the distribution of contexts in children's lives, as well as the talk that characterizes those different contexts.

However, we caution that there is no direct path from these observations about context to advice to parents, without more systematic research about the distribution of words in different learning environments. For example, several studies suggest that new and usual contexts including book reading and talk at outings such as museum trips vary with parent educational level and culture (Benjamin, Haden & Wilkerson, 2010; Dickinson & Snow, 1987; Luce, Callanan& Smilovic, 2013; Siegel, Esterly & Callanan, 2007; Tenenbaum & Callanan, 2008) leading to different words and different amounts of "rarer" words in the talk of different groups of parents. Parents for whom trips to museums are a novel or highly unusual event talk less about the exhibits than parents with more experiences in those contexts, and, as a consequence, use fewer rare words (Tenenbaum & Callanan, 2008). Note, the results may well be different if the parents for whom the museum was a never-before event for the parent took their children to and talked about a not-everyday context with which the parent was socially comfortable (see, Lee & Bowen, 2006; Sullivan, Ketende & Joshi, 2013 for perhaps related findings).

The role of contexts reminds that language learning environments have multiscale properties. The input to children is not merely a big bag of words but a sequence of small bags of words encountered in time. The consequences of the coherence of conversations and the diversity of contexts on parent talk may matter well beyond the overall type-token curve of input. A conversation about breakfast or a trip to the zoo presents the learner not just with different words but different repetitions of words close in time, repetitions we know that matter for building a narrative and for learning by the child (Horst, Parsons & Bryan, 2011). These small bags of conversation will each have their own type-token curves and these may differ in important ways for familiar contexts, for novel contexts, for book reading, at meal time versus play (Hoff, 1991; Soderstrom & Wittebolle, 2013; Sosa, 2015; Weizman & Snow, 2001). Because learning happens in real time, the type-token structure within conversations, and the distribution of smaller scale token structures that comprise the larger scale type-token curve also need to be understood. The present results strongly suggest that structure of conversations and contexts of talk are a key target for intervention.

Connecting the properties of the input to developmental outcomes

A large literature on human language processing and on early word learning suggests that the answer to the question of how the properties of input at scale relate to children's language learning outcomes will not be simple. Repetition, diversity, coherent contexts and contextual diversity have all been shown to support some aspects of lexical development (Hoff & Naigles, 2002). For example, the most frequent words in a language show marked advantages in many aspects of linguistic processing (Balota & Chumbley, 1985; Ellis, 2002; Jescheniak & Levelt, 1994; Murray & Forster, 2004; Rayner & Duffy, 1986). The words learned early by children are the ones that are common in speech to them (Goodman, Dale & Li, 2008; Hart, 1991). The co-occurrence of words, constrained by context and related meanings, builds conceptual networks of the semantic structure of language (Hills, Maouene, Riordan & Smith, 2010; Jones & Mewhort, 2007) and speeds the learning of new words when introduced in known contexts with known words (Fisher, Godwin & Matlen, 2015; Hills, Maouene, Maouene, Sheya & Smith, 2009). The contextual diversity of individual words (e.g., Adelman, Brown & Quesada, 2006; Hills et al, 2010; see Jones, Dye, & Johns, 2017 for a review) predicts both age of acquisition and the speed of adult judgments in lexical processing tasks. But at the limit, a type-token ratio of 1, diversity cannot be optimal. The open question is whether there is some ideal mix of repetition of words and contexts and of diversity of words and contexts.

This question of the relative benefits of consistency versus diversity in the training set is a subject of considerable interest in the study of human learning (e.g., Carvalho & Goldstone, 2014; 2015; Vlach & Sandhofer, 2012). In general, diversity of training instances increases generalization, but both theory and evidence suggests that for novices and early stages of learning, consistency of examples may be more important (Carvalho & Goldstone, 2014a, Gentner, 2010; Goldstein, et al., 2010; Goodman, Dale & Li, 2008). Training sets with a uniform distribution of instances are the standard in these experimental studies and thus their generalizability to training sets (the words in language) with power-law distributions may not be warranted. However, the power-law distribution itself provides a kind of “balance” between consistency and diversity. That is, the high frequency “head” provides consistency and the “long tail” provides diversity. Salkhutidinov, Torralba & Tenenbaum (2011), in a paper on the role of power-law distributions in visual object recognition, proposed that the extremely skewed distribution of visual instances and categories in the learning environment had computational benefits. That is, the power-law distribution of objects in the world may make learning easier because learning about the vast number of rare objects borrows strength (and influence on learning outcomes) from the very few high-frequency instances. In this way, the consistency of the very few high frequency items may facilitate rapid and accurate learning from the diverse and rarer instances. The power law distribution of words – and the semantic and syntactic relations among the few very high frequency words and the many much more rarely encountered words –may also play a significant role in early vocabulary and syntactic development (Goldberg, Casenhiser & Sethuraman, 2004; Naigles & Hoff-Ginsberg, 1998).

Measuring the learning environment in terms of its type-token curve provides a unified index of all these properties that may allow us to move beyond debates about quantity and quality of input (Hirsh-Pasek et al., 2015; Hoff & Naigles, 2002; Huttenlocher et al., 1991; 2010; Rowe, 2012; Weisleder & Fernald, 2014) to a better understanding of how the deeply inter-related properties of the distributions of words in human production support early vocabulary development.

Limitations

Here we concentrated on the number of words and unique words in child-directed speech. We did so because these two measures have played traditionally important roles in the study of early word learning and because their known nonlinear relation present an illustrative case of how new methods for capturing the everyday language environments of children at scale are going to expand and challenge current conceptualizations and methods. However, type and tokens are not the only relevant factors in the input. The quantity of other aspects of children's language learning environments also matter, including frequency of specific syntactic frames (Cameron-Failkner, Lieven & Tomasello, 2003; Huttenlocher et al., 2002; Huang, Leech &

Rowe, 2017; Naigles & Hoff-Ginsberg, 1998; Rowe, Leech & Cabrera, 2016) as well social behavioral factors including turn-taking, coordinated attention to the topic of speech, and parental responsivity (Hoff, 2006; Landry, Smith, Swank, Assel & Veellet, 2001; Hirsh-Pasek, et al., 2015; Ninio & Bruner, 1978; Tamis-LeMonda, Kuchirko & Sond, 2014; Tamis-LeMonda, Bornstein & Baumwell, 2001; Yu & Smith, 2012). The frequencies of these predictive properties of children's learning environments are typically estimated from observations studied at the time scales of minutes and hours. But there is persuasive evidence indicating that almost all forms of human generated behavior do not have normal or uniform distributions but instead are characterized by distributions in which a few forms of behavior are highly frequent, and most forms are rare, and in which behaviors are distributed in time in bursty bouts (Altman et al., 2009; Katz, 1996; Piantadosi, 2014). In brief, the questions that motivated the present analyses – how to measure contributions of counts of words and counts of diverse words in parent talk – comprises perhaps only a small portion of the parent behaviors that support early word learning. However, the issues and answers to how measure and understand the frequency distributions of parent-behaviors relevant to early word learning may is likely to apply to the frequency distributions of other aspects of the language learning environment (see Clerkin et al., 2017 for one example). Type-token ratios in samples of speech are also used to measure language learning in children as well as individual differences in vocabulary development (Leiven, 1978; Tardif, 1996; Templin, 1957). The issues raised here thus extend to measuring vocabulary development itself and to linking the type-token *input* curve to the type-token *acquisition* curve.

A second limitation of the present work is the use of CHILDES as the basis of the simulations since this corpus is a compendium of different conversational contexts to different children at different ages that could exaggerate, restrict or distort the amount of talk and/or lexical diversity in that talk to that which individual children hear across the daily lives. Notwithstanding these limitations, the simulated environments examined here provide us with the shape of questions we need to address as we collect and analyze multiple day-long collections of parent (and child) talk in the home.

Conclusion

In summary, the present demonstrations show how much we do not know and how much we need know about word learning environments at scale but in so doing provide a potential pathway for pursuing and for thinking about how and why word learning environments differ in the way they do. For example, rate of movement along the curve, parent vocabulary, parent expectations about how one should talk to children, the range and frequency of contexts with novel content for talk, and how parents talk in those contexts might all be in-principle be independently manipulated factors in determining the shapes of the type-token curves. But in the real world of parents and children and in the natural structure of human talk they are likely tightly inter-related in ways not yet well understood. We need to understand all of this if we are to tell parents how they should talk to their children (e.g., Leffel & Suskind, 2013; Reese, Sparks & Leyva, 2010; Roberts & Kaiser, 2011).

The words in human language have distributional properties that are well known. The causes that generate these properties are themselves not well known but characterize many natural phenomena far from language production (see Piantadosi, 2014, for a critical review). The consensus view (Goldwater, Griffiths & Johnson, 2006; Kello et al., 2009; Miller, 1957; Simon, 1955) is that power-law distributions emerge in phenomena generated by many non-independent stochastic processes and are, in fact, the mathematical marker of a phenomenon with a complex system of causes. These processes, however complex their origins, also create the data that drive word learning in children. Thus, understanding the structure of that input data is essential to a theory of early word learning. Understanding how the distributional properties of words in language to children can and do vary –and that factors responsible for that variation –are also essential to promoting healthy developmental environments for all children. Although there is much that we do not know and need to know, the positive contributions of the analyses reported here to the development of a theory of word-learning environments are these: (1) Word learning environments may be best measured in terms of the curve that relates number of types to number tokens over the months or years of cumulative input. (2) More talk in the language-learning environment may be understood in terms of the speed with which the learner moves along this curve of cumulative experienced tokens. (3) The shape of the curve relating cumulative types to cumulative tokens will vary with the size of the vocabulary from which the speakers in the learning environment draw their words for talk to children and with the diversity of contexts in which that talk occurs. (4) Relatively small changes in the diversity of contexts and topics of talk can lead to significant changes in the shape of the cumulative types-cumulative token curve.

Acknowledgements

This work was supported by NIH Grant T32 HD-07475 and by NSF grant BCS-15233982 to Smith. We thank Jon Willits for helpful discussion.

References

Adelman, J. S., Brown, G. D., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, *17*, 814-823.

Altmann, E. G., Pierrehumbert, J. B., & Motter, A. E. (2009). Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. *PLoS One*, *4*(e7678), 1-7.

Augustine, E., Smith, L. B., & Jones, S. S. (2011). Parts and relations in young children's shape-based object recognition. *Journal of Cognition and Development*, *12*, 556-572.

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance*, *10*, 340-357.

Bates, E., Bretherton, I., & Snyder, L. (1988). *From first words to grammar: Individual differences and dissociable mechanisms*. New York: Cambridge University Press.

Bates, E., & Goodman, J. (1997). On the inseparability of grammar and lexicon: evidence from acquisition, aphasia, real time processing. *Language and Cognitive Processes, 12*, 507-587.

Beals, D. E. (1997). Sources of support for learning words in conversation: Evidence from mealtimes. *Journal of Child Language*, *24*, 673-694.

Beals D. E., & Tabors P. O. (1995). Arboretum, bureaucratic and carbohydrate: preschoolers' exposure to rare vocabulary at home. *First Language,* 15, 57-76.

Benjamin, N., Haden, C. A., & Wilkerson, E. (2010). Enhancing building, conversation, and learning through caregiver–child interactions in a children’s museum. *Developmental Psychology*, *46*, 502-515.

Bornstein, M. H., Haynes, M. O., & Painter, K. M. (1998). Sources of child vocabulary competence: A multivariate model. *Journal of Child Language*, *25*, 367-393.

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*, 298-304.

Bradley, R. H., Corwyn, R. F., McAdoo, H. P., & García Coll, C. (2001). The home environments of children in the United States part I: Variations by age, ethnicity, and poverty status. *Child Development*, *72*, 1844-1867.

Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*, 33-44.

Broen, P. A. (1972). The verbal environment of the language learning child. *Monograph of the American Speech and Hearing Association, 17*, 1-103.

Brown, R. (1973). A first language: The early stages. Cambridge, MA: Harvard University Press.

Borun, M., Chambers, M. B., Dritsas, J., & Johnson, J. I. (1997). Enhancing family learning through exhibits. *Curator: The Museum Journal*, *40*, 279-295.

Callanan, M., & Valle, A. (2008). Co-constructing conceptual domains through family conversations and activities. *Psychology of Learning and Motivation*, *49*, 147-165.

Cameron-Faulkner, T., Lieven, E., & Tomasello, M. (2003). A construction based analysis of child directed speech. *Cognitive Science*, *27*, 843-873.

Carroll, J. B. (1938). Diversity of vocabulary and the harmonic series law of word-frequency distribution. *Psychological Record*, *2*, 379-386.

Carroll, J. B. (1964*). Language and thought*. Englewood Cliffs, NJ: Prentice-Hall.

Carvalho, P. F., & Goldstone, R. L. (2014a). Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Memory & Cognition*, *42*, 481-495.

Carvalho, P. F., & Goldstone, R. L. (2015). The benefits of interleaved and blocked study: different tasks benefit from different schedules of study. *Psychonomic Bulletin & Review*, *22*, 1-8.

Church, K. W., & Gale, W. A. (1995). Poisson mixtures. *Natural Language Engineering*, *1*, 163-190.

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

Clerkin, E. M., Hart, E., Rehg, J. M., Yu, C., & Smith, L. B. (2017). Real-world visual statistics and infants' first-learned object names. *Philosophy of Transactions of the Royal Society B 372: 20160055.*

Cohen, A., Mantegna, R. N., & Havlin, S. (1997). Numerical analysis of word frequencies in artificial and natural language texts. *Fractals*, *5*, 95-104.

Deckner, D. F., Adamson, L. B., & Bakeman, R. (2006). Child and maternal contributions to shared reading: Effects on language and literacy development. *Journal of Applied Developmental Psychology*, 27, 31-41.

Dickinson, D. K., Golinkoff, R. M., & Hirsh-Pasek, K. (2010). Speaking out for language why language is central to reading development. *Educational Researcher*, *39*, 305-310.

Dickinson, D. K., Griffith, J. A., Golinkoff, R. M., & Hirsh-Pasek, K. (2012). How reading books fosters language development around the world. *Child Development Research*, *2012*, 1-15.

Dickinson, D. K., & Snow, C. E. (1987). Interrelationships among prereading and oral language skills in kindergartners from two social classes. *Early Childhood Research Quarterly*, *2*, 1-25.

Diessel, H. 2009. On the role of frequency and similarity in the acquisition of subject and non-subject relative clauses. In *Syntactic Complexity: Diachrony, Acquisition, Neuro-cognition, Evolution* [Typological Studies in Language 85], T. Givón & M. Shibatani (eds), 251–276. Amsterdam: John Benjamins.

Drieman, G. H. (1962). Differences between written and spoken language: An exploratory study. *Acta Psychologica*, *20*, 36-57.

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition*, *24*, 143-188.

Estoup, J. R. 1916. Gammes Stenographiques. Institut Stenographique de France, Paris.

Fernald, A., Marchman, V. A., & Weisleder, A. (2012). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental Science*, *16*, 234-248.

Fernald, A., Perfors, A., & Marchman, V. A. (2006). Picking up speed in understanding: Speech processing efficiency and vocabulary growth across the 2nd year. *Developmental Psychology*, *42*, 98.

Ferrer-i-Cancho, R. and Solé, R.V. (2002). Zipf's law and random texts. *Advanced Complex Systems, 5*, 1–6

Firth, J. R. (1957). *A synopsis of linguistic theory. Studies in linguistic analysis*. Oxford: Blackwell.

Fisher, A. V., Godwin, K. E., & Matlen, B. J. (2015). Development of inductive generalization with familiar categories. *Psychonomic Bulletin & Review, 22*, 1149-1173.

Fletcher, K. L., Cross, J. R., Tanney, A. L., Schneider, M., & Finch, W. H. (2008). Predicting language development in children at risk: The effects of quality and frequency of caregiver reading. *Early Education and Development, 19*, 89–111.

Gentner, D. (2005). The development of relational category knowledge. In L. Gershkoff-Stowe & D. H. Rakison, (Eds.), *Building object categories in developmental time.* (pp. 245-275). Hillsdale, NJ: Erlbaum.

Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science*, *34*, 752-775.

Gilkerson, J., & Richards, J. A. (2008). *The LENA natural language study.* Boulder, CO: LENA Foundation.

Gleitman, L. R., Newport, E. L., & Gleitman, H. (1984). The current status of the motherese hypothesis. *Journal of Child Language*, *11*, 43-79.

Goldstein, M. H., Waterfall, H. R., Lotem, A., Halpern, J. Y., Schwade, J. A., Onnis, L., & Edelman, S. (2010). General cognitive principles for learning structure in time and space. *Trends in Cognitive Sciences*, *14*, 249-258.

Goldwater, S., Griffiths, T., & Johnson, M. (2006). Interpolating between types and tokens by estimating power-law generators. In Advances in neural information processing systems (Vol. 18).

Golinkoff, R. M., & Ames, G. J. (1979). A comparison of fathers' and mothers' speech with their young children. *Child Development, 50*, 28-32.

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language*, *35*, 515.

Goulden, R., Nation, P., & Read, J. (1990). How large can a receptive vocabulary be?. *Applied Linguistics*, *11*, 341-363.

Greenwood, C. R., Thiemann-Bourque, K., Walker, D., Buzhardt, J., & Gilkerson, J. (2011). Assessing children's home language environments using automatic speech recognition technology. *Communication Disorders Quarterly*, *32*, 83-92.

Hart, B. (1991). Input frequency and children's first words. *First Language*, *11*, 289-300.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experiences of young American children.* Baltimore: Brookes.

Hayes, D. P., & Ahrens, M. G. (1988). Vocabulary simplification for children: A special case of 'motherese'? *Journal of Child Language, 15*, 395-410.

Heaps, H.S. (1978). Information retrieval. Computational and theoretical aspects. New York: Academic Press.

Herdan, G. (1960). *Type-token mathematics* (Vol. 4). The Hague: Mouton.

Hess, C. W., Sefton, K. M., & Landry, R. G. (1986). Sample size and type-token ratios for oral language of preschool children. *Journal of Speech, Language, and Hearing Research*, 29, 129-134.

Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009). Longitudinal Analysis of Early Semantic Networks Preferential Attachment or Preferential Acquisition? *Psychological Science*, 20, 729-739.

Hills, T. T., Maouene, J., Riordan, B., & Smith, L. B. (2010). The associative structure of language: Contextual diversity in early word learning. *Journal of Memory and Language*, 63, 259-273.

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*, 0956797615581493.

Hladik, E. G., & Edwards, H. T. (1984). A comparative analysis of mother-father speech in the naturalistic home environment. *Journal of Psycholinguistic Research*, *13*, 321-332.

Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child development*, *74*, 1368-1378.

Hoff, E. (2006). How social contexts support and shape language development. *Developmental Review*, *26*, 55-88.

Hoff, E., & Naigles, L. (2002). How children use input to acquire a lexicon. *Child Development*, *73*, 418-433.

Horst, J. S., Parsons, K. L., & Bryan, N. M. (2011). Get the story straight: Contextual repetition promotes word learning from storybooks. *Frontiers in Psychology*, *2*, 1-11.

Huang, Y. T., Leech, K., & Rowe, M. L. (2017). Exploring socioeconomic differences in syntactic development through the lens of real-time processing. *Cognition*, *159*, 61-75.

Hudson, K. C., & Matthewson, L. (2016). Introducing the Infant Bookreading Database (IBDb). *Journal of Child Language*, 1-20.

Hurtado, N., Marchman, V. A., & Fernald, A. (2008). Does input influence uptake? Links between maternal talk, processing speed and vocabulary size in Spanish‐learning children. *Developmental Science*, *11*, F31-F39.

Hutchins, T. L., Brannick, M., Bryant, J. B., & Silliman, E. R. (2005). Methods for controlling amount of talk: Difficulties, considerations and recommendations. *First Language*, *25*, 347-363.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology*, *27*, 236.

Huttenlocher, J., Vasilyeva, M., Cymerman, E., & Levine, S. (2002). Language input and child syntax. *Cognitive psychology*, *45*, 337-374.

Huttenlocher, J., Waterfall, H., Vasilyeva, M., Vevea, J., & Hedges, L. V. (2010). Sources of variability in children's language growth. *Cognitive Psychology*, *61*, 343-365.

Jescheniak, J. D., & Levelt, W. J. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 824-843.

Johnson, W. (1939). Language and speech hygiene. Institute of General Semantics: Chicago.

Jones, M. N., Dye, M., & Johns, B. T. (2017). Context as an organizational principle of the lexicon. In B. Ross (Ed.), The Psychology of Learning and Motivation, 67:43.

Jones, M. N., & Mewhort, D. J. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*, 1-37.

Karrass, J., & Braungart-Rieker, J. M. (2005). Effects of shared parent–infant book reading on early language acquisition. *Journal of Applied Developmental Psychology*, *26*, 133-148.

Katz, S. M. (1996). Distribution of content words and phrases in text and language modelling. *Natural Language Engineering*, *2*, 15-59.

Kello, C. T., Brown, G. D., Ferrer-i-Cancho, R., Holden, J. G., Linkenkaer-Hansen, K., Rhodes, T., & Van Orden, G. C. (2010). Scaling laws in cognitive sciences. *Trends in Cognitive Sciences*, *14*, 223-232.

Kidd, E., Lieven, E., & Tomasello, M. (2006). Examining the role of lexical frequency in the acquisition and processing of sentential complements. *Cognitive Development*, *21*, 93 - 107.

Kleinberg, J. (2003). Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, *7*, 373-397.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211-240.

Landry, S. H., Smith, K. E., Swank, P. R., Assel, M. A., & Vellet, S. (2001). Does early responsive parenting have a special importance for children's development or is consistency across early childhood necessary? *Developmental psychology*, *37*, 387-403.

Leffel, K., & Suskind, D. (2013). Parent-directed approaches to enrich the early language environments of children living in poverty. *Seminars in Speech and Language, 34*, 267-277.

Lee, D. Y. (2001). Defining core vocabulary and tracking its distribution across spoken and written genres: evidence of a gradience of variation from the British National Corpus. *Journal of English Linguistics*, *29*, 250-278.

Lee, J. S., & Bowen, N. K. (2006). Parent involvement, cultural capital, and the achievement gap among elementary school children. *American Educational Research Journal*, *43*, 193-218.

Lieven, E. V . M . 1(978.) Conversations between mothers and young children: individual differences and their possible implication for the study of child language learning. In N.

Waterson & C. E. Snow. The Development of Communication, 173-187. Chichester: Wiley.

Luce, M. R., Callanan, M. A., & Smilovic, S. (2013). Links between parents' epistemological stance and children's evidence talk. *Developmental Psychology, 49*, 454-461.

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

Malvern, D., Richards, B. J., & Chipere, N. (2004). *Lexical diversity and language development: Quantification and assessment*. Basingstoke, UK: Palgrave Macmillan.

Mandelbrot, B. (1953). An informational theory of the statistical structure of language. *Communication Theory*, *84*, 486-502.

Marchman, V. A., & Fernald, A. (2008). Speed of word recognition and vocabulary knowledge in infancy predict cognitive and language outcomes in later childhood. *Developmental Science*, *11*, 9-16.

Marchman, V. A., Martínez‐Sussmann, C., & Dale, P. S. (2004). The language‐specific nature of grammatical development: evidence from bilingual language learners. *Developmental Science*, *7*, 212-224.

Markus, J., Mundy, P., Morales, M., Delgado, C. E., & Yale, M. (2000). Individual differences in infant skills as predictors of child‐caregiver joint attention and language. *Social Development*, *9*, 302-315.

Massaro, D. W. (in press). Two different communication genres and implications for vocabulary development and learning to read. *Journal of Literacy Research.*

Masur, E. F., & Gleason, J. B. (1980). Parent–child interaction and the acquisition of lexical information during play. *Developmental Psychology*, *16*(5), 404.

McCarthy, P. M., & Jarvis, S. (2007). Vocd: A theoretical and empirical evaluation. *Language Testing*, *24*, 459-488.

McKee, G., Malvern, D., & Richards, B. (2000). Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing*, *15*, 323-338.

Miller, G. A. (1957). Some effects of intermittent silence. *The American Journal of Psychology, 70*, 311-314.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*, 91-117.

Mol, S. E., Bus, A. G., de Jong, M. T., & Smeets, D. J. (2008). Added value of dialo*gic parent–child book readings: A meta-analysis. Early Education and Development, 19*, 7–26.

Montag, J. L., Jones, M. N., & Smith, L. B. (2015). The words children hear picture books and the statistics for language learning. *Psychological Science*, 26, 1489-1496.

Murray, W. S., & Forster, K. I. (2004). Serial mechanisms in lexical access: the rank hypothesis. *Psychological Review*, *111*, 721-756.

Ninio, A. (2011). *Syntactic development, its input and output*. Oxford, UK: Oxford University Press.

Ninio, A., & Bruner, J. (1978). The achievement and antecedents of labelling. *Journal of Child Language, 5*, 1–15.

Osgood, C. E. (1952). The nature and measurement of meaning. *Psychological Bulletin*, *49*, 197.

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, 763-782.

Pereira, A. F., & Smith, L. B. (2009). Developmental changes in visual object recognition between 18 and 24 months of age. *Developmental Science*, *12*, 67-80.

Phillips, J. R. (1973). Syntax and vocabulary of mothers' speech to young children: Age and sex comparisons. *Child Development*, *44*, 182-185.

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, *21*, 1112-1130.

Ratner, N. B. (1988). Patterns of parental vocabulary selection in speech to very young children. *Journal of Child Language*, *15*, 481-492.

Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, *14*, 191-201.

Reese, E., Sparks, A., & Leyva, D. (2010). A review of parent interventions for preschool children's language and emergent literacy. *Journal of Early Childhood Literacy*, *10*, 97-117.

Richards, B. (1987). Type/token ratios: what do they really tell us? *Journal of Child Language*, *14*, 201-209.

Roberts, M. Y., & Kaiser, A. P. (2011). The effectiveness of parent-implemented language interventions: A meta-analysis. *American Journal of Speech-Language Pathology*, *20*, 180-199.

Rowe, M. L. (2008). Child-directed speech: relation to socioeconomic status, knowledge of child development and child vocabulary skill. *Journal of Child Language*, *35*(01), 185-205.

Rowe, M. L. (2012). A Longitudinal Investigation of the Role of Quantity and Quality of Child Directed Speech in Vocabulary Development. *Child Development*, *83*, 1762-1774.

Rowe, M. L., Leech, K. A., & Cabrera, N. (2016). Going Beyond Input Quantity: Wh-Questions Matter for Toddlers' Language and Cognitive Development. *Cognitive Science*, 1-18.

Rowe, M. L., Raudenbush, S. W., & Goldin‐Meadow, S. (2012). The pace of vocabulary growth helps predict later vocabulary skill. *Child Development*, *83*, 508-525.

Roy, D., Patel, R., DeCamp, P., Kubat, R., Fleischman, M., Roy, B., Mavridis, M., Tellex, S., Salata, A., Guiness, J., Levit, M., Gorniak, P. (2006), 'The Human Speechome Project.' Paper Presented at the 28th Annual Conference of the Cognitive Science Society, Vancouver, Canada.

Sahlgren, M., & Karlgren, J. (2005, November). Counting lumps in word space: Density as a measure of corpus homogeneity. In *String Processing and Information Retrieval* (pp. 151-154). Springer Berlin Heidelberg.

Salakhutdinov, R., Torralba, A., & Tenenbaum, J. (2011, June). Learning to share visual appearance for multiclass object detection. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on* (pp. 1481-1488). IEEE.

Siegel, D. R., Esterly, J., Callanan, M. A., Wright, R., & Navarro, R. (2007). Conversations about science across activities in Mexican-descent families. *International Journal of Science Education*, *29*, 1447-1466.

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

Serrano, M.Á., Flammini, A., & Menczer, F. (2009). Modeling Statistical Properties of Written Text. *PLoS ONE, 4*(e5372), 1-8.

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*, 672-686.

Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika*, *42*, 425-440.

Soderstrom, M., & Wittebolle, K. (2013). When do caregivers talk? The influences of activity and time of day on caregiver speech and child vocalizations in two childcare environments. *PloS one*, *8*, e80646.

Snow, C. E. (1972). Mothers' speech to children learning language. *Child Development, 43*, 549-565.

Snow, C. E. (1983). Literacy and language: Relationships during the preschool years. *Harvard Educational Review*, *53*, 165-189.

Sullivan, A., Ketende, S., & Joshi, H. (2013). Social class and inequalities in early cognitive scores. *Sociology, 47*, 1187-1206.

Sosa, A. V. (2016). Association of the type of toy used during play with the quantity and quality of parent-infant communication. *JAMA pediatrics*, *170*(2), 132-137.

Suppes, P. (1974). The semantics of children's language. American Psychologist, 29,103– 114.

Tamis‐LeMonda, C. S., Bornstein, M. H., & Baumwell, L. (2001). Maternal responsiveness and children's achievement of language milestones. *Child Development*, *72*, 748-767.

Tamis-LeMonda, C. S., Kuchirko, Y., Lui, H. Escobar, K. & Knudson, J. (2015). Temporal Features of Infants' Real-Time Language Experiences in the Home Setting. Talk presented at biennial meeting of the Society for Research in Child Development, Philadelphia, PA.

Tamis-LeMonda, C. S., Kuchirko, Y., & Song, L. (2014). Why is infant language learning facilitated by parental responsiveness? *Current Directions in Psychological Science*, *23*, 121-126.

Tardif, T. (1996). Nouns are not always learned before verbs: Evidence from Mandarin speakers' early vocabularies. *Developmental Psychology*, *32*(3), 492.

Templin, M. C. (1957). Certain language skills in children: Their development and interrelationships. *The Institute of Child Welfare, Monograph Series No. 26.* Minneapolis: University of Minnesota.

Tenenbaum, H. R., & Callanan, M. A. (2008). Parents' science talk to their children in Mexican descent families residing in the USA. *International Journal of Behavioral Development, 32,* 1–12.

Tomasello, M. (1988). The role of joint attentional processes in early language development. *Language Sciences*, *10*, 69-88.

Tweedie, F. J., & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, *32*, 323-352.

VanDam, M., Warlaumont, A. S., Bergelson, E., Cristia, A., Soderstrom, M., De Palma, P., & MacWhinney, B. (2016). HomeBank, an online repository of daylong child-centered audio recordings. *Seminars in Speech and Language, 37*, 128‑142.

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing*, *17*, 65-83.

Vlach, H. A., & Sandhofer, C. M. (2012). Distributing learning over time: The spacing effect in children's acquisition and generalization of science concepts. *Child Development*, *83*, 1137-1144.

Walker, D., Greenwood, C., Hart, B., & Carta, J. (1994). Prediction of school outcomes based on early language production and socioeconomic factors. *Child Development*, *65*, 606-621.

Weisleder, A., & Fernald, A. (2014). Talking to children matters early language experience strengthens processing and builds vocabulary. *Psychological Science*, *24*, 2143-2152.

Weizman, Z. O., & Snow, C. E. (2001). Lexical output as related to children's vocabulary acquisition: Effects of sophisticated exposure and support for meaning. *Developmental Psychology*, *37*, 265.

Whitehurst, G. J., Falco, F. L., Lonigan, C. J., Fischel, J. E., DeBaryshe, B. D., Valdez-Menchaca, M. C., & Caulfield, M. (1988). Accelerating language development through picture book reading. *Developmental Psychology, 24*, 552–559.

Willits, J. A., Jones, M. N., & Landy, D. (2016). Learning that numbers are the same, while learning that they are different. *Proceedings of the 37th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Young, K. T., Davis, K., Schoen, C., & Parker, S. (1998). Listening to parents: a national survey of parents with young children. *Archives of Pediatrics & Adolescent Medicine*, *152*, 255-262.

Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers. *Cognition, 125*, 244-262.

Zechmeister, E. B., Chronis, A. M., Cull, W. L., D'Anna, C. A., & Healy, N. A. (1995). Growth of a functionally important lexicon. *Journal of Literacy Research*, *27*, 201-212.

Zipf, G.K. (1949). Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology. Addison-Wesley: Cambridge, MA.