# The everyday statistics of objects and their names: How word learning gets its start

**Elizabeth M. Clerkin (emclerki@indiana.edu)**
**Linda B. Smith (smith4@indiana.edu)**
Department of Psychological and Brain Sciences, Indiana University
1101 E. 10th Street Bloomington, IN 47405

**Abstract**

A key question in early word learning is how infants learn their first object names despite a natural environment thought to provide messy data for linking object names to their referents. Using head cameras worn by 7 to 11-month-old infants in the home, we document the statistics of visual objects, spoken object names, and their co-occurrence in everyday meal time events. We show that the extremely right skewed frequency distribution of visual objects underlies word-referent co-occurrence statistics that set up a clear signal in the noise upon which infants could capitalize to learn their first object names.

**Keywords:** word learning; natural statistics; egocentric vision

## Introduction

Infants begin learning object names before their first birthday. We know they can do this because infants look to the pictures of an object upon hearing the name (Bergelson & Swingley, 2012). By these measures, individual infants do not know many object names, and their knowledge is fragile. Still, it is clear that the start of learning object names begins well before infants produce words. By consensus, these novice learners must begin learning object names by linking the heard word to visually present objects. The ability to do this has been demonstrated in experimental studies (e.g., Smith & Yu, 2008). The problem is that the everyday visual world is much noisier and cluttered than the learning tasks presented in the laboratory (Clerkin, Hart, Rehg, Yu, & Smith, 2017). Laboratory studies also show that in the period just prior to the first birthday, infants have limited attention skills and quite limited memories for the learned object-word pairings taught in a single laboratory session (Vlach & Johnson, 2013). Accordingly, the field lacks a complete understanding of how object name learning gets its early start.

Learning depends on both the internal learning mechanisms and the data for learning. There are critical gaps in current knowledge about the everyday experiences that comprise the data for early object name learning. We know that parent-naming events are often ambiguous as the visual world is cluttered (Cartmill et al., 2013), parents often do not talk to the child in the home during natural activities (Tamis-LeMonda, Custode, Kuchirko, Escobar, & Lo, 2018), and parents only sometimes name the objects in the child's view during naturalistic play (Yurovsky, Smith, & Yu, 2013). Still, we know little about the statistical structure of everyday experiences across multiple naming events (but see Bergelson & Aslin, 2017 for recent work on this topic). Here we provide evidence-based estimates on three key statistical properties of the learning environment: the frequency distribution of heard object names, of seen visual objects, and their co-occurrence.

### Rationale

The frequency distributions of words in parent talk to children are known to be extremely skewed with a small set of extremely frequent words and a much larger set of very rare words (Montag, Jones, & Smith, 2018). A small set of words that are heard pervasively – day in and day out – might define a constrained set upon which object name learning could get its start. Analyses of one large corpus of child-directed talk, however, suggests that the frequency distribution for object names in parent talk is not as skewed as other grammatical classes such that there are less dramatic differences between the most and least frequent object names (Sandhofer, Smith, & Luo, 2000). However, these analyses considered all parent talk – not talk within a particular context. Parent talk, and the words infants hear, are context bound (Montag et al., 2018). The child should be much more likely to hear the words "spoon" and "table" at mealtime than to hear the words "bat" or "ball." Thus, the key question for the role of very high frequency objects names at the start of object name learning may lie in the pervasiveness of a select set of objects names within a context.

There is very little evidence on the frequency distribution of visual objects in the natural environment generally or in infant everyday experiences in which these objects that must be linked to heard names. The evidence that does exist about the natural visual environment – from analyses of large corpora of photographs (Salakhutdinov, Torralba, & Tenenbaum, 2011) and from one analysis of head camera images collected by infants in their home (Clerkin et al., 2017) – suggests that the frequency distribution of object categories will be extremely skewed. The latter evidence further suggests that the very high frequency categories will correspond to the object names that are learned early by infants. Common sense and extant evidence from photography corpora (Sadeghi, McClelland, & Hoffman, 2015) also suggests that visual objects will be context dependent, with spoons and tables more likely in the immediate visual scene at mealtime than bats and balls.

For novice learners to learn object names, heard names must co-occur with referents in their experience. If a few object categories and their names are concurrently pervasive in infant everyday experiences, then there is a clear statistical solution to how object name learning starts – with the learning of the names of those few pervasive objects in infant experiences. Here we provide direct evidence on this possibility and show that the pervasive objects and pervasive

Figure 1. Example (non-consecutive) images from the videos recorded during infant mealtimes.

names in infant language learning environments *do not correspond* well, but that the learning environment offers a different statistical solution to the start of object name learning based on the 1) the skewed distributions of visual objects and 2) the quantity and quality of word-referent co-occurrences.

## Method

### The Corpus

We chose the mealtime context for three reasons: it is frequent, occurring on average 5 times a day for infants in this age group, the names for objects likely present at mealtimes are among the earliest learned concrete nouns by normative age of acquisition, and it is a potentially challenging context for learning given the sparsity of parent talk (Tamis-LeMonda et al., 2018) – very unlike contrived play contexts in laboratories. These mealtime events were selected from head cameras[1] embedded in hats worn by 14 infants aged 7 to 11 months at home as they went about their daily activities with no experimenters present (see Clerkin et al., 2017; Jayaraman, Fausey, & Smith, 2015 for details). Parents were not told specific activities to record and were told to record any and all activities during the times their infants were awake over a period of several days.

Figure 1 shows example images extracted from the video. Critically, the video collected from the head cameras is from the infants' ego-centric view. Thus, we have captured the visual environment directly in front of the infants' faces and the objects in it to which infants could be attending in any given moment. This ego-centric perspective is highly dependent on the infants' motor abilities, their interests, and their location and posture in any given moment. In sum, not only are we studying the natural word learning context at scale, but we are doing so with reference to the infants' own point of view.

Any video that included eating or meal preparation was included in the mealtime corpus which totaled 16.99 hours of footage and consisted of 344 mealtime events with 24.57 per subject on average (SD = 20.02).

### Coding

**Visual Objects** Still images were down-sampled from the video recordings at a rate of 0.2hz (1 image every 5 seconds). The 11,549 down-sampled images were then coded by naïve adult coders for the 5 most obvious objects in the scene using basic level nouns; (see Clerkin et al., 2017 for more details). Each image was coded by 4 coders. These adult judgements of objects that are in view do not necessarily align with what the target infant's visual attention in the moment; however, we use these adult judgements as a way of describing the clutter of the natural environment from which infants are presumably visually sampling.

We chose to keep the coders' responses as intact as possible to avoid biasing the data; however, we did clean the data in the following ways. First, extraneous adjectives were removed (e.g., "baby spoon" was reduced to "spoon"); however, if an adjective-word combination was listed in the dictionary (e.g., "high chair"), it remained as a unique object. Also, different forms of the same object name were collapsed (e.g., "cup" and "cups" were both counted as instances of "cup"). Finally, words that were overly general (e.g., "food") or clearly did not refer to a concrete object (e.g., "color") were removed entirely. The frequencies of visual object categories are reported as the proportion of frames in which the object category occurred.

**Object Names** All speech in the target infants' environment was transcribed for each mealtime using Datatyvu (Datavyu Team, 2014). The audio data was broken down into 5 second intervals for ease of coding and to have an appropriate comparison to the visual data coded at 1 image every 5 seconds. It should be noted that infants this age do not talk, and thus none of the transcribed speech is the target infants' own vocalizations. Naming events (defined as any moment an object name was said) were extracted from the speech stream for object names that referred to objects which were reported as occurring at least once in the visual scenes. The speech transcripts were cleaned as described above for visual objects. The frequencies of object names are reported as the number of naming instances for each name across the corpus

[1] The Looxcie 2 weighs 22 grams and has a 75° diagonal field of view.

as a proportion of the number of 5 second intervals containing any speech.

**Age of Acquisition Categories** In order to understand how the statistics of the natural learning environment relate to learning first words, objects were broken down into two age of acquisition (AoA) categories. Objects in the First category were those named by nouns on the MacArthur Bates Communicative Developmental Inventory – MCDI (Fenson et al., 2007) and are present in the receptive vocabulary of 50% of 18-month-old children in the Wordbank repository of thousands of MCDI administrations (Frank, Braginsky, Yurovsky, & Marchman, 2016). Later objects were all other objects given by the coders.

**Co-Occurrence** Co-occurrence was coded by three trained raters in the laboratory. Each naming instance was located in the video, and if during the 5 second interval surrounding the naming instance an object which could be called by that name was visually present, then the coder recorded there was a co-occurrence. Approximately 20% of the naming instances were coded by all 3 coders. The final judgment for those instances was the response recorded by at least 2 of the 3 coders. The overall percentage agreement between the coders was 76.2%, but there were no naming instances on which at least 2 coders did not agree. Co-occurrence is reported as the proportion of naming instances during which a corresponding object was visually present.

Table 1: Summary of data coded.

| | Num Frames | Num Speech Intervals |
|---|---|---|
| *Total* | 11549 | 12237 |
| *With Talk* | – | 6833 |
| *Without Talk* | – | 5404 |

Table 2: Object and object name counts

| | Num Unique Objects | Num Unique Object Names |
|---|---|---|
| *Total* | 1095 | 350 |
| *First* | 118 | 97 |
| *Later* | 977 | 253 |

## Results

Table 1 provides the number of fames coded for visual object and the total number of 5 second speech intervals, the number containing any speech, and the number containing no speech. Table 2 provides the number of unique visual objects and unique object names. As is apparent, there are many more objects than object names, showing considerable selectivity in parent talk relative to the wide variety of objects in view.

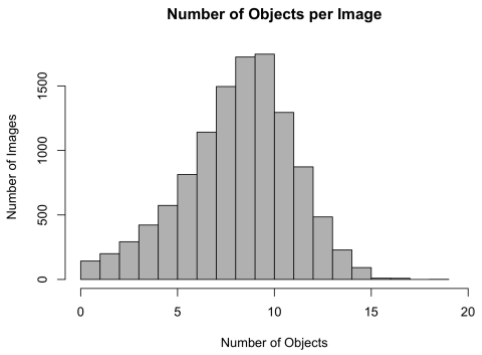


Figure 2. The frequency distribution of the number of objects per image across the corpus.

We consider the statistical regularities characterizing the visual objects, then the object names, and finally, their co-occurrence.

### Visual Objects

The number of objects coded in each scene is an indication of the clutter present in the natural visual environment. Because 4 coders named a maximum of 5 objects each per image, the number of possible objects recorded as visually present in a scene ranged from 1 to 20. Figure 2 shows the frequency distribution of the number of objects per image. On average, images contained 8.63 objects (median = 9), which supports the long-held idea that the visual world is cluttered and that for most naming events there are multiple possible referents that a novice learner could consider.

In total, coders recorded 1,095 unique objects with a total of 97,407 object instances. Only 351 of these visual objects also occurred as object names in speech, and the reported analyses focus on these 351 objects that occurred in both modalities[2]. There were 72,446 total object instances for this smaller set. Figure 3a shows the proportion of images in which each object category appeared plotted against its rank frequency. As in Clerkin et al. (2017), visual objects occur in these natural scenes with a right skewed frequency distribution[3]. A small number of objects were pervasively present and a large number of objects occurred rarely with the 20 most frequent object categories (see table 3) accounting for 65.47% of all object tokens and the 37 most frequent object categories (that is, 10.5% of the 351 objects) accounting for 80.18% of all object tokens (see Figure 4).

Further, the AoA category of an object name is significantly related to the frequency of its corresponding visual object in the corpus. 97 of the visual objects reported by coders (that also appeared in the speech modality) were First objects and 253 were Later objects. Mann-Whitney-Wilcoxon tests were used to compare the frequencies of objects in these categories due to the non-normality of the

[2] All analyses follow the same statistics pattern when all 1,095 visual objects are analyzed.

[3] The distribution is referred to as right-skewed based on a histogram of the frequency distribution in which the placement of the most frequent objects is reversed on the x-axis as compared to Figure 3. We find the rank order plots better visualizations for our purposes.

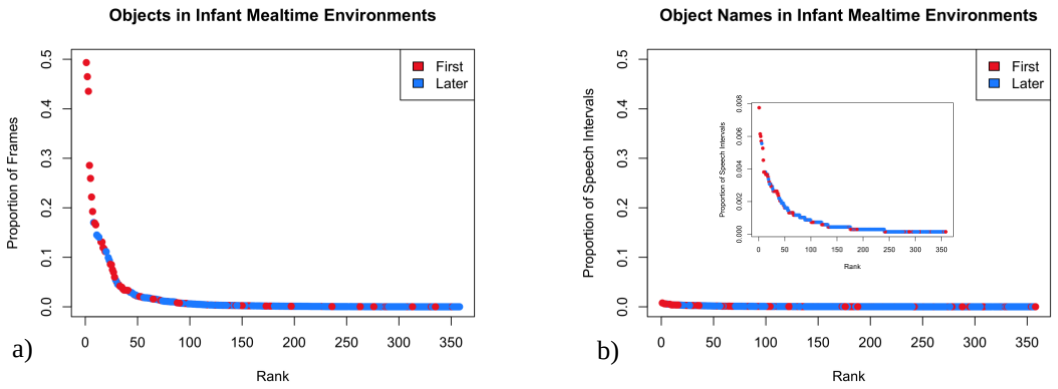


Figure 3. a) The proportion of images in which each visual object category appeared against its rank. b) The number of naming events for each object name as a proportion of the number of 5 second intervals containing any speech. Inset is the same information plotted with a smaller y axis.

data. Objects named by words in the First category (mean = 4.50%; Mdn = 0.55% of images) were significantly more frequent than objects named by Later words (mean = 0.75%; Mdn = 0.10% of images), U = 17609.5, $p < 0.0001$[4]. 11 of the 15 most frequent objects belonged to the First category. In sum, infants' visual experience during mealtime is dominated by a small set of objects named by very early learned words. These results suggest that day-in and day-out experience with these visual objects may be important for learning their names.

## Object Names

Talk overall was extremely sparse in these mealtime scenes. Any speech, not just speech including object names, only occurred in 55.83% of the total video time (see table 1).

Table 1: Top visual objects, object names, and their AoAs.

| Visual Objects | AoA |
|---|---|
| table | First |
| shirt | First |
| chair | First |
| window | First |
| bowl | First |
| cup | First |
| bottle | First |
| cabinet | Later |
| door | First |
| pants | First |
| picture | Later |
| counter | Later |
| tray | Later |
| spoon | First |
| toy | First |

| Object Names | AoA |
|---|---|
| egg | First |
| cheese | First |
| paper | First |
| book | First |
| camera | Later |
| water | First |
| juice | First |
| milk | First |
| paint | Later |
| spoon | First |
| table | First |
| dog | First |
| page | Later |
| plate | First |
| watch | First |

Object names in speech occurred even more rarely; 117 mealtime events contained some speech but none of the target object names. The overall lack of talk and object names appears quite ordinary and typical when watching and listening to content these natural videos. These infants do not yet talk themselves, and the speech stream thus often contains terms of endearment and comments directed to the baby, talk between adults, and periods of silence as the parents and their infants go about their daily lives.

Nonetheless, 351 unique object names were said during mealtime activities across 1,941 naming events. It should be noted that only a small number of object names were said – about a third of those possible based on the list of visual objects. Figure 3b shows the number of naming instances as a proportion of the 5 second intervals containing any talk for each object name plotted against its rank frequency. Though the distribution of object talk is not uniform, it does not follow the pattern of extreme skewness as does the objects or might be predicted by the statistics of natural language more generally. The 40 most frequent object categories accounted for only 48.79% of all object name tokens, and the 123 most frequent object names (that is, 35.04% of all object names) were required to account for 80.06% of all object names tokens. Though object names do not appear equally frequently, there is not a clear set of object names that dominate talk about objects in this natural mealtime context. Note in Figure 4 the difference in the shapes of the curves for the proportions of unique visual objects and object names that account for all tokens.

A large proportion (97 out of 118) of the First words whose visual objects appeared in the images occurred in the auditory domain as well. Proportionally fewer of the possible Later objects had names that were said during mealtime; only 253 of the 977 Later object names were spoken in the corpus. As with the visual objects, object name frequency is significantly related to AoA. Object names from the First category (mean

[4] All reported p-values have been corrected for multiple comparisons using the Holm correction.

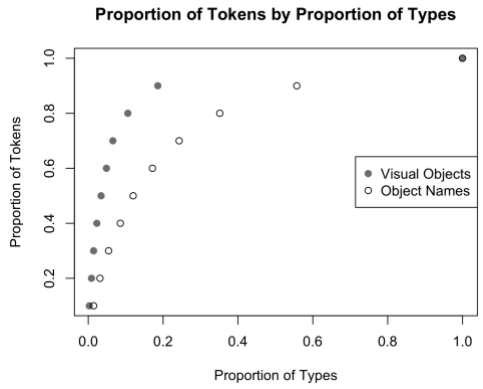


Figure 4. The proportion of types accounting for the proportion of tokens at intervals of 0.1.

= 0.14%; Mdn = 0.07% of speech intervals) were spoken more frequently than object names from the Later category (mean = 0.06%; Mdn = 0.03% of speech intervals), U = 16765.5, $p < 0.0001$. 12 of the 15 most frequent object names belonged to the First category. This, unsurprisingly, supports the idea that hearing objects names is important for learning them.

## Correspondence and Co-occurrence

If the objects present most frequently in the visual environment were those whose corresponding names occur frequently in the environment, it would seem that the problem of breaking into learning first object names is solved. However, while there is a highly significant positive relationship between visual frequency and spoken frequency for object-name pairs, the relationship is very weak[5], $\tau_B$ = 0.17, $p < 0.0001$. As a demonstration, the 40 most frequent objects and the 40 most frequent object names only have 11

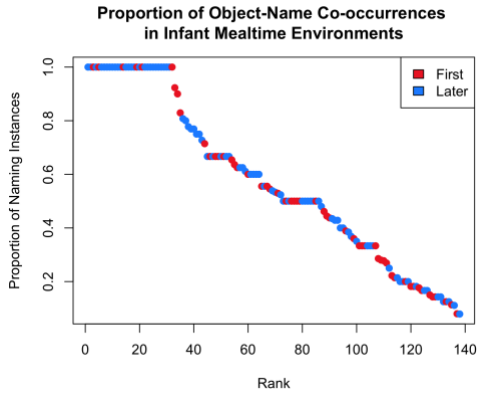


Figure 5. The proportion of naming instances in which a corresponding object was visually present. Only object-name pairs which ever co-occurred are shown.

items in common, and only 1 object name appears in the top 10 for frequency in both modalities. In sum, the pervasive visual objects are not named by words that are especially frequent in this context.

The number of co-occurrences between objects and their names even by our generous measure was very low. 213 of the 351 object-name pairs never occurred in the same 5 second interval. For the 138 that co-occurred at least once, the maximum number of co-occurrences was 34 (mean = 4.43; Mdn = 2). Because raw co-occurrence is so rare and the timescales of visual objects and spoken words are so different, we turned co-occurrence, reported here as the proportion of naming instances in which a corresponding object was visually present.

Figure 5 shows the co-occurrence proportion by rank order for the 138 object-name pairs that ever co-occurred. For co-occurrence, we do not find a right skewed frequency distribution but rather one that is bimodal. Further, co-occurrence proportion shows a statistical relationship with AoA that is opposite to those found for visual objects and object names individually. The co-occurrence proportion of the Later category (mean = 60.55%, Mdn = 62.95%) was significantly higher than that of the First category (mean = 48.83%, Mdn = 50%), U = 1662, $p < 0.01$. In fact, 26 Later objects co-occurred with their corresponding names 100% of the time whereas only 6 First objects did so. This result on its face seems surprising because there is a strong theoretical and empirical basis for the idea that co-occurrence is key to learning object-name mappings.

However, when the frequency of object names in the corpus are considered, it becomes clear why co-occurrence proportion was related to AoA in this direction. Co-occurrence proportion is in fact negatively correlated with word frequency, $\tau_B = -0.41$, $p < 0.0001$. This means that for many object names which were said perhaps only once, the corresponding visual objects were likely to be present during that naming instance. It makes sense that objects that are unusual in the context would be more likely to be present in the moment when those objects' names are said. For example, "fire extinguisher" (which is logically an unusual item for the mealtime context) was named once and the object was present, giving it a co-occurrence proportion of 1. This result suggests that it is important to consider not only the quantity of the co-occurrences (frequency) but also the quality of co-occurrence between object-names pairs (co-occurrence proportion) as it is unclear how much very young infants could learn from a single co-occurrence.

## Quantity and Quality: Strength

To assess the quantity and quality of the co-occurrence of object-name pairs, we created a new compound measure of co-occurrence strength which was the proportion of naming instances during which the visual object was present - multiplied by the number of mealtime events in which both

[5] Kendall's rank correlation used instead of Pearson's product moment correlation due to the non-normality of the data.

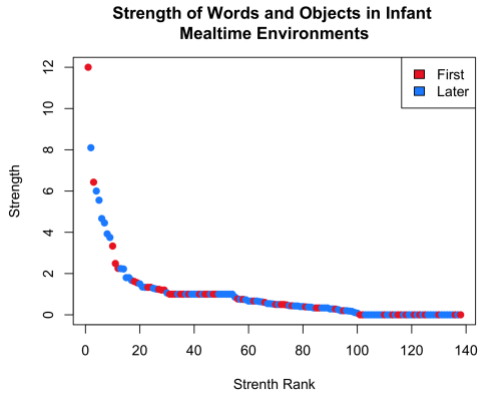


Figure 6. The strength of the co-occurrence of the object-word pair against its rank. Only object-name pairs which ever co-occurred are shown.

the visual object and the object name ever occurred. This measure allows us to approximate the potential learnability of an object-name pair.

Figure 6 shows the strength of the object-name pairing plotted against rank strength. Here we see again the skewed distribution with a small number of object-name pairings with relatively high strength values and a large number of object-name pairings with low strength values. We also again find a significant difference between strength of First object-name pairs (mean = 1.69, Mdn = 1) and Later object-name pairs (mean = 0.51, Mdn = 0.35), U = 3203, $p < 0.0001$. 13 of the 15 highest strength values belong to object-naming pairing for the First AoA category. This result suggests that the quality and quantity of co-occurrences between object-name pairs may be important for infants breaking into object name learning. Critically, the strength of the co-occurrence is underlain by the skewed frequency of visual objects.

## Discussion

The results taken together do not support the hypothesis that many may have theorized: object names that occur frequently have pervasive referents, and these word-referent pairs occur frequently and simultaneously, thus providing a simple statistical solution for how infants can learn their first object names. Instead, the evidence of the present study suggests a different solution underlain by the pervasiveness of a few object categories in the visual environment. Object names that refer to visually pervasive objects may be said relatively rarely, but because the objects are visually pervasive, whenever the object name is said, the object is likely in the infants' view. The extremely skewed frequency distribution of objects in view in the mealtime context thus makes each naming event for those objects count – as demonstrated by our measure of co-occurrence strength.

Studies of the word-learning environment for children have typically focused on the frequency and diversity of the words (Hart, 1991; Montag et al., 2018). However, investigations of the natural environment including the visual domain are taking off with the advent of small, wearable cameras. Another recent at-home study which also examined the frequency of objects and their names in the natural environment found that the overall proportion of object-name co-presence predicts 6-month-olds' performance in an in-laboratory word comprehension task (Bergelson & Aslin, 2017). This result supports the idea that the statistical structure of the learning environment is directly related to word learning. Our results further suggest that the visual side of the learning problem specifically may be critical to the start of object name learning because it sets up the opportunities for learning moments.

The frequency distribution of visual objects during a particular context (mealtime in the present case) partitions potential referents into two potential classes for young learners –those that are typically present in this context and those that are not; classes that will be different for each context. Those that are persistently present provide a *selective* visual foundation to linking the objects to their referents.

The foundation for the *early* learning of object names may be contexts – such as mealtime, dressing, getting into the car – that occur day-in and day-out and are characterized by the same object categories repeatedly and pervasively present. These routines may bias the linking of even sparse naming events to those visually pervasive objects. Contexts that repeat in this way, along with the statistical structure of visual objects in those contexts, may be a critical contributing factor for early learning. This idea is consistent with the evidence on the value of repeatedly reading infants their favorite pictures books in supporting word learning (Horst, Parsons, & Bryan, 2011). For older children, the diversity of words in the learning environment may matter most for vocabulary development (Montag et al., 2018), but for the earliest learners, consistency of the visual content of repeated contexts may be the key.

In sum, the co-occurrence statistics of object names and their referents in the contexts comprising the early natural learning environment, as underlain by the extremely right skewed frequency distribution of visual objects, set up a clear signal in the noise which infants may use to learn their first object names.

## Acknowledgements

We thank the families who participated in this study. We also thank Teagan Wilson, Remi Reich, Bryce Hockman, and Baker Nasser for in laboratory coding and our colleagues in the Cognitive Development Lab at Indiana University for helpful comments. This article was funded by NSF grant BCS-15233982, by NIH grants R01HD 074601, R01HD 28675, T32HD007475, and by Indiana University through the Emerging Area of Research Initiative - Learning: Brains, Machines, and Children.

## References

Bergelson, E., & Aslin, R. N. (2017). Nature and origins of the lexicon in 6-mo-olds. *Proceedings of the National Academy of Sciences*, *114*(49), 12916-12921.

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences, 109*(9), 3253-3258.

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

Clerkin, E. M., Hart, E., Rehg, J. M., Yu, C., & Smith, L. B. (2017). Real-world visual statistics and infants' first-learned object names. *Phil. Trans. R. Soc. B, 372*(1711), 20160055.

Fenson, L., Bates, E., Dale, P. S., Marchman, V. A., Reznick, J. S., & Thal, D. J. (2007). *MacArthur-Bates communicative development inventories*.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2016). Wordbank: An open repository for developmental vocabulary data. *Journal of child language*, 1-18.

Hart, B. (1991). Input frequency and children's first words. *First Language, 11*(32), 289-300.

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*, 17.

Jayaraman, S., Fausey, C. M., & Smith, L. B. (2015). The faces in infant-perspective scenes change over the first year of life. *PloS one, 10*(5), e0123780.

Montag, J. L., Jones, M. N., & Smith, L. B. (2018). Quantity and diversity: Simulating early word learning environments. *Cognitive Science, 42*, 375-412.

Sadeghi, Z., McClelland, J. L., & Hoffman, P. (2015). You shall know an object by the company it keeps: An investigation of semantic representations derived from object co-occurrence in visual scenes. *Neuropsychologia, 76*, 52-61.

Salakhutdinov, R., Torralba, A., & Tenenbaum, J. (2011). *Learning to share visual appearance for multiclass object detection.* Paper presented at the Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on.

Sandhofer, C. M., Smith, L. B., & Luo, J. (2000). Counting nouns and verbs in the input: Differential frequencies, different kinds of learning? *Journal of child language, 27*(3), 561-585.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition, 106*(3), 1558-1568.

Tamis-LeMonda, C. S., Custode, S., Kuchirko, Y., Escobar, K., & Lo, T. (2018). Routine Language: Speech Directed to Infants During Home Activities. *Child development*.

Datavyu Team. (2014). Datavyu: A video coding tool. Databrary Project, New York University. URL http://datavyu.org.

Vlach, H. A., & Johnson, S. P. (2013). Memory constraints on infants' cross-situational statistical learning. *Cognition, 127*(3), 375-382.

Yurovsky, D., Smith, L. B., & Yu, C. (2013). Statistical word learning at scale: The baby's view is better. *Developmental Science, 16*(6), 959-966.