How do infants start learning object names in a sea of clutter?

Hadar Karmazyn Raz (hkarmazy@iu.edu)
Drew H. Abney (dhabney@indiana.edu)
David Crandall (djcran@indiana.edu)
Chen Yu (chenyu@indiana.edu)
Linda B. Smith (smith4@indiana.edu)

Department of Psychological and Brain Sciences
Indiana University, Bloomington, IN 47405 USA

Abstract

Infants are powerful learners. A large corpus of experimental paradigms demonstrate that infants readily learn distributional cues of name-object co-occurrences. But infants' natural learning environment is cluttered: every heard word has multiple competing referents in view. Here we ask how infants start learning name-object co-occurrences in naturalistic learning environments that are cluttered and where there is much visual ambiguity. The framework presented in this paper integrates a naturalistic behavioral study and an application of a machine learning model. Our behavioral findings suggest that in order to start learning object names, infants and their parents consistently select a set of a few objects to play with during a set amount of time. What emerges is a frequency distribution of a few toys that approximates a Zipfian frequency distribution of objects for learning. We find that a machine learning model trained with a Zipf-like distribution of these object images outperformed the model trained with a uniform distribution. Overall, these findings suggest that to overcome referential ambiguity in clutter, infants may be selecting just a few toys allowing them to learn many distributional cues about a few name-object pairs.

Keywords: infancy; early word learning; machine learning; Zipfian distribution.

Introduction

The natural environment is visually cluttered with multiple namable objects in view (Clerkin, 2017). To learn their first object names, infants must link a heard object name to the referent object (Bloom, 2000). But for any heard object name, from the infant's perspective, there are multiple potential referents in view. This referential ambiguity has defined a major theoretical problem to be solved in early word learning (Quine, 1960). Despite a sea of clutter, infants already know the names of many objects by the time of their first birthday. We know this because they look to the named objects in laboratory tests (Bergelson, 2012; Swingley & Aslin, 2000) and because they begin to say object names in the contexts of those objects (Fenson et al, 1994). How does this work? The current paper integrates behavioral and modeling frameworks to explore how infants learn object names despite the referential ambiguity in their natural learning environments.

One explanation for solving referential ambiguity is the distributional cues in the language and visual input (Aslin, 2017). According to this explanation, infants track the frequencies of word-object co-occurrences to aggregate the most likely referent (Smith, Smith, & Blythe, 2011; Kachergis, Yu, & Shiffrin, 2017). A large collection of laboratory paradigms has demonstrated that infants can rapidly learn from distributional cues of visual and auditory input (e.g., Cartwright & Brent, 1997; Mintz, 2003; Mintz, Newport, & Bever, 2002; Reeder, Newport, & Aslin, 2013). However, it is still unclear how learning from distributional cues of words and objects in laboratory settings transfers to the distributional cues in the natural environment. Laboratory paradigms are typically highly controlled, presenting uniform word-object frequencies (Aslin, Saffran, & Newport, 1998; Kurumada et al., 2013). In contrast, for natural languages, word frequencies are known to follow a Zipfian distribution, in which a small number of words occur very frequently (e.g. boy, car), while many words occur rarely (Zipf, 1965). These so-called Zipfian distributions, are universal across human languages (Zipf, 1949; Piantadosi, 2014), including nouns and all words in infant-directed speech (Hendrickson & Perfors, 2018). Furthermore, recent studies show that even the distribution of objects in infants' natural visual environments follow a Zipfian distribution, where a few objects appear highly frequently and most objects are rare (Clerkin et al., 2017).

Nevertheless, the sensitivity of infants and adults to distributional cues highlights an intriguing, but as of yet untested, benefit for learning from Zipfian distributions. Theoretically, learning from a Zipfian distribution should be more difficult than a uniform distribution, as there is not enough information in a Zipfian distribution to link the referents for words that occur rarely (Blythe et al., 2010; Vogt, 2012; Reisenauer et al., 2013; Blythe et al., 2016). However, a recent adult study demonstrated that adults learn word-object links more easily from Zipfian distributions than from uniform distributions of word-object occurrences (Hendrickson & Perfors, 2018). Those results suggest that Zipfian distributions improve adults' learning by providing more statistical cues for the highly-frequent words, which in turn reduces the referential uncertainty associated with the unknown rare words. Yet, how infants learn from Zipfian distributions is still unknown.

The approach in this paper is that the natural training data for learning new object names are generated by the behaviors of the learner from the mature social partner who provides the

name. Here we demonstrate that during infant-parent interactions, objects being handled and named by the parent create Zipfian frequency distributions, in which very few events occur very frequently, forming a very small set for learning.

Recent studies of infant naturalistic environments suggest that Zipfian distributions provide a balance between *consistency* of a few high-frequent events with *diversity* of rare events (Clerkin et al., 2017; Smith & Slone, 2018; Montag, Jones, & Smith, 2017). Here we hypothesized that the parent and infant consistently select to play with a few objects, generating a training data set balanced with rare exploration of diverse objects. In other words, parents' and infants' selective and exploratory behaviors naturally generate name-object experiences that form Zipfian distributions, which is hypothesized to reduce ambiguity and optimize learning.

In this study we demonstrate infant naturalistic learning, while infants and parents were engaging with objects in a cluttered and an unstructured environment. The play sessions were recorded from the infant's egocentric perspective. From these visual experiences we report the frequency distributions of the objects with which infants and parents engaged during toy play. We subsequently applied a machine learning model to evaluate the structure of the visual "training data" produced from these play experiences. The model was tested for detecting the play objects with a training dataset of uniform distributions compared to Zipf-like distributions of infants' egocentric object views.

Behavioral Methods

To evaluate how infants learn early object names in a naturalistic environment, we conducted a toyplay experiment allowing infant-parent dyads to freely engage with object toys.

Participants

The final sample included 16 infant-parent dyads with 12 month-old infants (8 female) ranging from 12.2 to 12.5 months (*M*=12.3, *SD*=1.12) were included in the final sample.

Stimuli and Experimental Setup

Parents and their infants were invited to play in a naturalistic setting for a duration of approximately 10 minutes. Parents and infants both sat on a carpeted floor in a playroom environment. To create an unstructured environment, a random assortment of 33 toys were randomly distributed on the floor (see Figure 1). The toy objects' themes were not related in any particular way; thus, any selection and exploration of objects emerged naturally from infant and parent behaviors. The same toys were used in each session. The instructions were to play freely as they normally would at home.

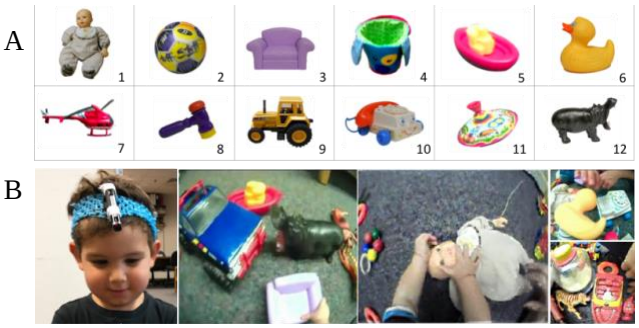

Figure 1: (A) Stimuli set. (B) Experimental setup (left to right) infant wearing a Looxcie camera, infant ego-centric views from camera during toyplay.

Egocentric View

To collect infant egocentric view, we used a commercially available, lightweight (22 g) wearable camera (Looxcie). The camera was secured to a hat that was custom fit to the infant so that when the hat was securely placed on the infant's head, the lens was centered above the nose and did not move (see Figure 1).

The head camera captured the scene in front of the viewer but did not provide direct gaze information, which in principle could be outside of the head camera image (Smith et al., 2015). However, head mounted eye-tracking studies have demonstrated that under active viewing conditions, human observers, including infants, typically turn both heads and eyes in the same direction and align heads and eyes within 500 ms of a directional shift to maintain head and eye alignment when sustaining attention (Yoshida & Smith, 2008; Smith, Yu, & Pereira, 2011). Therefore, it can be expected that a high proportion of gaze during active viewing is highly concentrated in the center of the head camera image (Yoshida & Smith, 2008).

Data Processing

The raw videos were coded using Datavyu by sampling frames at 0.2Hz (1 frame every 5 sec; 2,008 frames total). To describe the dyadic behaviors of engagement with objects, the corpus of frames was coded for (1) objects in view, (2) objects handled by the infant or parent and (3) objects named by the parent. The *objects in view* were defined as the number of objects in the field of view from the infant's perspective. The *objects handled* were coded for both the parents and infant and defined as any hand contact with objects. Finally, parents' speech was manually transcribed and then annotated for *objects naming* when objects were named explicitly (*N*=580).

We measured the statistics for objects in parents' and infants' hands and parent naming events over the entire 10-minute period. Since there were individual differences among the infants in terms of the objects they played with, we constructed rank-ordered frequency histograms (see

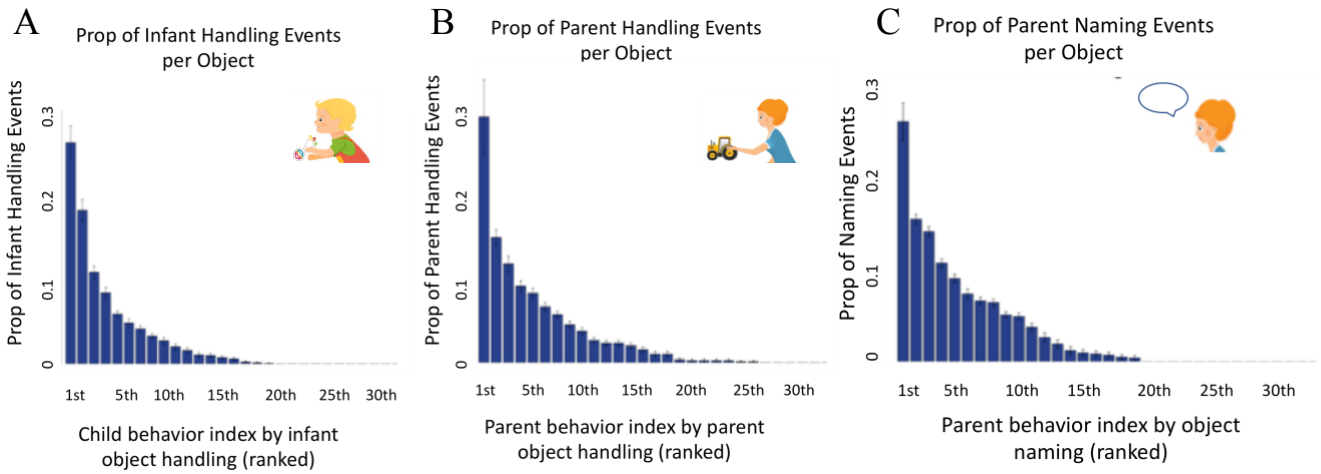

Figure 2: Ranked order histograms of the objects infant and parent handled and named (objects' ranks across these histograms are not necessarily the same). (A) histogram of infant object handling, showing the proportion of instances infants handled each object. (B) histogram of parent object handling, showing the proportion of instances parents handled each object. (C) histogram of parent object naming, showing the proportion of instances parents named each object. Error bars indicate 95% confidence intervals.

Figure 2). The rank-ordered frequency was measured independently for each infant and then combined since the most frequent object in view, handled or named differed for different infants. The objects are annotated by their rank-ordered frequency ranging from the 1st most frequent to least 33rd frequent object.

Behavioral Results

The experimental setup of dyadic play was visually cluttered as there were on average 13 objects in view every frame (*Min*=1, *Max*=30, *M*=13.18, *SD*=6.16). Figure 2 shows ranked order histograms of infant and parent object handling proportions, as well as parent naming proportions. The histograms display a Zipf-like pattern which is indicative of behavioral selectivity of objects in the scene. Specifically, these zipf-like distributions follow an approximate power-law, in which a small set of objects are handled and named very frequently and most objects are rare. Six objects (*Min*=9, *Max*=19, *M*=13.70, *SD*=3.13) account for over 80% of the total proportion of infant's object handling events. Six objects (*Min*=9, *Max*=24, *M*=15.5, *SD*=3.8) account for over 80% of the total proportion of parents' object handling events. Eight objects (*Min*=10, *Max*=48, *M*=24.20, *SD*=10.23) account for over 80% of the total proportion of parents' object naming events. These Zipf-like distributions likely reflect a balance between the parents' and infants' stability and exploration of objects, which may benefit object learning.

Modeling Methods

A machine learning model was used to test whether the distributional properties of infants' visual object experience impacted learning. In particular we wanted to understand the learning mechanism by which infants learn new object names in clutter environments.

The data

The collected corpus of 2,008 infants' egocentric views were used to construct two different toy object training sets, as detailed in Table 1. Six of the objects were selected for our machine learning study: baby doll, ball, chair, bucket, boat, and duck (see Figure 1). One of the two training datasets had a uniform frequency distribution of object images, and the other was with a Zipf-like frequency distribution. There were a few reasons for only using six specific objects for the modeling framework. First, from the raw corpus of images, only 1,200 images included at least 1 of the 6 objects for detection in the scene. Second, Bounding boxes indicating the objects' location and label were annotated for the set of 6 objects intended for detection. Note that some images were removed from the corpus due to low image quality such as high blur. The corpus was augmented by 180-degree rotation and horizontally flipped, yielding a final corpus of about 3,000 images: 2400 split into training and validation and 600 for testing.

Figure 3: Example of cluttered training images, including multiple objects for detection, labeled and annotated with a bounding box.

Each training dataset (the uniform and Zipf-like) was formed by a subset of the 2400 images for training. Due to the nature of the cluttered scenes, many images included more than 1 detectable object as seen in Figure 3. These images that included multiple objects for detection were counted toward the frequencies of more than one object when forming the datasets of uniform and Zipf-like frequency distributions (see Table 1). The final training data sets included 2,154 images each. In the *uniform* dataset, each object was present in 400 images. The *Zipfian* dataset included high frequency and low frequency images of objects. In the *Zipfian* dataset the baby doll had the highest frequency (1000 images), and the duck was the rarest (100 images).

Table 1: Distribution of object images among the uniform and right-skewed datasets for training

Images Per Object	Uniform	Zipf-like
Baby Doll	400	1000
Ball	400	600
Chair	400	320
Bucket	400	240
Boat	400	140
Duck	400	100
Total	2,154	2,154

Model Parameters

The applied machine learning model was the Faster R-CNN, Region-based Convolutional Neural Network (Ren, Girshick & Sun, 2015), a well-known, state-of-the-art machine learning model for object detection. The model is essentially a network composed of three main components: a feature extractor, a region proposal network (RPN), and a classifier. First, for the feature extraction part, we adapted a pretrained CNN VGG16 on the ImageNet data set which includes approximately 1.2 million images (Russakovsky et al., 2015). The model has 16 layers and classifies images into 1000 object categories (e.g. keyboard, mouse, coffee mug, pencil). The input images (size 224X224) are inputted into the VGG16 network. The network evaluates the distinctive visual features for the whole image, which allows us to detect multiple objects in each image. Second, after feature extraction the regions are proposed, therefore only running one CNN over the entire image instead of multiple CNN's for each proposed region. Finally, a single softmax layer, outputs the class probabilities directly for each region. The last fully connected layer and classification layer were adjusted for the number of classes in the data set applied in this framework (N classes= 7, including 'background').

Training

The Fast-RCNNs were trained with two different datasets that varied in the frequency distributions of object images: the Zipf-like and uniform distributions. The network was trained for 1000 epochs (iterations).

Modeling Results

To determine whether infants' selective behavior benefits learning, we applied a machine learning model trained with Infants' egocentric views. We compared the model's performance of object recognition when trained with a Zipf-like vs. a uniform distribution of infants' egocentric views as seen in Table 2. Overall, the model trained with the Zipf-like dataset had a significantly higher (t=-3.35, p<0.05) mean average precision (mAP=40%) compared to the model trained with the uniform dataset (mAP=23%). This pattern of results suggests that a Zipf-like distribution of data yields higher accuracy and benefits learning.

To further evaluate whether the Zipf-like distribution of objects in infants' egocentric views reduce ambiguity we evaluated the average precision of each object (see Table 2). For the baby doll and the ball there were more images in the Zipf-like dataset relative to the uniform distribution and there was accordingly a higher average precision for these objects in the Zipf-like trained model (AP=56% and 48%, respectively). The chair had a similar number of images in both datasets and had a similar average precision in the Zipf-like (AP=23%) and uniform datasets (AP=21%). For the rest of the objects (the bucket, boat and the duck), there were less images in the Zipf-like dataset, yet a higher average precision in the Zipf-like trained model (AP= 29% and 38% and 43%, respectively) compared with the model trained on the uniform distribution. These patterns of results, where there is higher precision despite less training images suggests that information has been shared among objects reducing likely competing objects and reducing ambiguity.

Discussion

In the real world, the sea of visual clutter provides multiple competing referents for every heard object name. This paper explored how infants solve this referent ambiguity in a cluttered environment to learn first object names. Here we presented a behavioral study of infants and their parents playing freely with objects and applied a learning model to explore the 'behind the scene' learning machinery.

To observe how infants learn object names in a cluttered environment, we recorded the play from infants' egocentric view. In order to describe experiences relevant for object name learning, we reported the frequency distributions of the

Object AP	Baby Doll	Ball	Chair	Bucket	Boat	Duck	mAP
Uniform	0.25	0.23	0.21	0.19	0.28	0.24	0.23
Zipf-Like	0.56	0.48	0.25	0.29	0.38	0.43	0.40

Table 2: Model test results: average precision per object for the uniform and Zipf-like datasets

objects infants and parents handled, as well as the objects parents named. We found that the frequency distributions of objects handled and named approximated a right-skewed Zipf-like distribution with few highly frequent objects along with many low frequency objects. This finding suggests that in a cluttered environment, infants and parents consistently select a set of a few objects for learning and rarely explored the other objects. The consistent object handling and naming behaviors during early word learning offers repetition, a key component for learning (Hintzman & Block, 1971, Vlach, 2014).

The infant-parent dyads consistently created datasets that were highly selective and focused on just a few objects. These dynamic patterns of selection may be due to the influence of other systems such as human memory or attention which decays in a power-law pattern (e.g., Wixted & Ebbesen, 1991; Wixted, 2004; Baronchelli, Ferrer-i-Cancho, Pastor-Satorras, Chater, & Christiansen, 2013). These non-uniform distributions have been shown as optimal conditions for adults and may help solve learning problems across many domains (Schuler, Reeder, Newport & Aslin, 2017; Hendrickson & Perfors, 2018; Caron & Vincent, 2002; Salakhutdinov, Torralba, & Tenenbaum, 2011).

As we could not directly observe infants' learning machinery, we applied a machine learning model to explore how infants may be learning from a Zipfian distribution. The application of the model was weaved with the behavioral study by using infants' egocentric object views from the behavioral study of play as the training images for the learning model. The learning machinery from Zipfian distributions was evaluated by comparing a training apparatus of a Zipf-like and a uniform frequency distribution of object images. The testing demonstrated that the training using a Zipf-like distribution yielded higher accuracy than a uniform distribution. Interestingly, the testing also demonstrated that low frequency objects were learned at higher rates when trained in the Zipf-like distribution.

The Zipf-like model's patterns of results were consistent with machine learning and adult studies of Zipfian learning (Schuler, Reeder, Newport & Aslin, 2017; Hendrickson & Perfors, 2018; Caron & Vincent, 2002; Salakhutdinov, Torralba, & Tenenbaum, 2011). These studies suggested that the learned features of highly frequent items are shared with the low frequency items to reduce referent ambiguity. For example, when learning to recognize a rare vehicle such as 'bus', the exemplar shares features of wheels and window shields from an already learned 'car', a highly frequent vehicle. It has also been suggested that low frequency items such as 'napkin' may benefit from co-occurrences with high frequency objects such as 'bowl'. These model's results coincide with infants laboratory studies demonstrating that infant early word learning is tuned to statistical cues of word-object co-occurrences. These findings also suggest that infants may be able to learn object names not only from uniform distributions of word-object occurrences but also from a Zipfian distributions.

Beyond previous early word learning studies, the findings in this paper suggest that infants solve referential ambiguity in a sea of clutter by consistently selecting a few objects and rarely exploring a large subset of objects. This behavior may benefit learning and reduce ambiguity in clutter by allowing to learn a lot of statistical cues about a few objects. Finally, this paper offers a methodological framework of incorporating behavioral paradigms with computational modeling to stretch our understanding of cognition.

Acknowledgements

This research was supported in part by NSF grant BCS-1523982, NICHD T32HD007475-22 and F32HD093280, and by Indiana University through the Emerging Area of Research Initiative – Learning: Brains, Machines, and Children.

References

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. Psychological Science, 9, 321–324. doi:10.1111/1467-9280.00063

Aslin, R. N. (2017). Statistical learning: a powerful mechanism that operates by mere exposure. *Wiley Interdisciplinary Reviews: Cognitive Science, 8*(1-2), e1373.

Baronchelli, A., Ferrer-i-Cancho, R., Pastor-Satorras, R., Chater, N., & Christiansen, M. H. (2013). Networks

in cognitive science. *Trends in Cognitive Sciences*, *13*(7), 348–360.

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.

Blythe, R., Smith, A., & Smith, K. (2016). Word learning under infinite uncertainty. *Cognition*, *151*, 18-27.

Blythe, R., Smith, K., & Smith, A. (2010). Learning times for large lexicons through cross-situational learning. *Cognitive Science*, *34*(4), 620–642.

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

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.

Caron, Y., Makris, P., & Vincent, N. (2002). *A method for detecting artificial objects in natural environments.* Paper presented at the Pattern Recognition, 2002. Proceedings. 16th International Conference on.

Cartwright, T. A., & Brent, M. R. (1997). Syntactic categorization in early language acquisition: Formalizing the role of distributional analysis. Cognition, 63, 121–170. doi:10.1016/S0010-0277(96)00793-7

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., ... & Stiles, J. (1994). Variability in early

communicative development. *Monographs of the society for research in child development*, i-185.

Hendrickson, A., & Perfors, A. (2018). Cross-situational learning in a Zipfian environment.

Hintzman, D. L., & Block, R. A. (1971). Repetition and memory: Evidence for a multiple-trace hypothesis. *Journal of Experimental Psychology*, *88*(3), 297.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2017). A bootstrapping model of frequency and context effects in word learning. *Cognitive science*, *41*(3), 590-622.

Kurumada, C., Meylan, S., & Frank, M. C. (2013). Zipfian frequency distributions facilitate word segmenta-

tion in context. *Cognition*, *127*, 439–453.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. Cognition, 90, 91– 117. doi:10.1016/S0010-0277(03)00140-9

Mintz, T. H., Newport, E. L., & Bever, T. G. (2002). The distributional structure of grammatical categories in speech to young children. Cognitive Science, 26, 393–424. doi:10.1207/s15516709cog2604_1

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

Piantadosi, S. (2014). Zipf's word frequency law in natural language: A critical review and future directions.

Psychonomic Bulletin and Review, *21*, 1112–1130.

Quine, W. (1960). *Word and object*. Cambridge, MA: MIT Press.

Reisenauer, R., Smith, K., & Blythe, R. (2013). Stochastic dynamics of lexicon learning in an uncertain and

nonuniform world. *Physics Review Letters*, *110*(258701).

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems* (pp. 91-99).

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Berg, A. C. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211-252.

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.

Schuler, K. D., Reeder, P. A., Newport, E. L., & Aslin, R. N. (2017). The effect of Zipfian frequency variations on category formation in adult artificial language learning. *Language Learning and Development*, *13*(4), 357-374.

Smith, L. B., Yu, C., & Pereira, A. F. (2011). Not your mother's view: The dynamics of toddler visual experience. *Developmental science*, *14*(1), 9-17.

Smith, L. B., & Slone, L. K. (2017). A developmental approach to machine learning?. *Frontiers in psychology*, *8*, 2124.

Smith, K., Smith, A. D., & Blythe, R. A. (2011). Cross-situational learning: An experimental study of word-learning mechanisms. *Cognitive Science*, *35*(3), 480-498.

Smith, L. B., Yu, C., Yoshida, H., & Fausey, C. M. (2015). Contributions of head-mounted cameras to studying the visual environments of infants and young children. *Journal of Cognition and Development*, *16*(3), 407-419.

Swingley, D., & Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition*, *76*(2), 147-166.

Vlach, H. A. (2014). The spacing effect in children's generalization of knowledge: allowing children time to forget promotes their ability to learn. *Child Development Perspectives*, *8*(3), 163-168.

Vogt, P. (2012). Exploring the robustness of cross-situational learning under Zipfian distributions. *Cognitive*

Science, *36*, 726–739.

Wixted, J. (2004). The psychology and neuroscience of forgetting. *Annual Review of Psychology*, *55*, 235-269.

Wixted, J., & Ebbesen, E. (1991). On the form of forgetting. *Psychological Science*, *2*, 409–415.

Yoshida, H., & Smith, L. B. (2008). What's in view for toddlers? Using a head camera to study visual experience. *Infancy*, *13*(3), 229-248.

Zipf, G. (1949). *Human behavior and the principle of least effort*. New York, NY: Addison-Wesley.

Zipf, G. (1965). Human behavior and the principle of least effort: An introduction to human ecology. New York, NY: Hafner.