Attentional Constraints and Statistics in Toddlers' Word Learning

Sumarga H. Suanda, Seth B. Foster, Linda B. Smith, & Chen Yu
Department of Brain and Psychological Sciences
Indiana University
Bloomington, IN - USA

***Abstract*— Recent research supports the notion that word learning can be conceptualized as a statistical learning process. As many have noted however, statistical learning is constrained by processes such as attention and memory. In the current study, we observed, through toddler-perspective head cameras, toddlers' visual input as parents labeled novel objects during an object-play session. We then analyzed the co-occurrence statistics between words and objects that accumulated over the session. We also analyzed the *constrained* co-occurrence statistics which took into consideration the perceptual properties of the objects (e.g., object size) at the times words were uttered. We compared the information in these two types of statistical structures and examined which of the two best fit with the patterns of children's object-name learning. Implications of these results for statistical learning accounts of early word learning are discussed.**

Keywords—early word learning; statistical learning; language development; attention and learning

I. INTRODUCTION

Children learn words surrounded by a sea of data. The data consists of, among other things, the hundreds of words per hour to which they are exposed and the many objects with which they interact. The process of learning new words, specifically the mapping of words to their referents, can thus be viewed as a process of mining this data to figure out which words go with which objects. Multiple lines of research support this statistical learning framework for understanding early word learning. First, observational studies of children's language input have suggested that better data, as measured by statistical properties in the input such as word frequency, word contextual diversity, and word density, are correlated with better learning [1]. Second, computational models of early word learning have demonstrated that algorithms built to detect statistical regularities from learning environments similar to those of young children are capable of "learning words" and display learning signatures characteristic of children's learning [2]. Finally, a growing body of experimental research has revealed that within controlled laboratory settings, even the youngest of word learners can track the statistical regularities in their learning environment, and that they recruit this sensitivity in the service of word learning [3].

Within this statistical learning framework, the goal of the current project is to extend our understanding of children's early word learning in three ways. First, we endeavored to characterize the nature of the statistical input (i.e., word-to-referent co-occurrence patterns) available to toddlers as they engaged in a free-flowing object play session with their parents. Second, we directly examined the role these statistics played in children's learning of the object names during the play session. Third, we analyzed how these statistics interacted with attentional constraints in the learning process.

To address these issues, the current research adopts a mixture of detailed multi-sensory observations of a toddlers' audio-visual learning environment and a traditional word learning paradigm. Toddlers and their parents participated in a free play session with a set of novel objects. Toddlers were equipped with head cameras placed low on their foreheads. The camera provided a characterization of the toddlers' visual environments during the play session. Parents were given the names of the novel objects and were told to use them during the play session. Of particular interest in the current study was the information that was in the toddlers' field of view at the moment names were uttered. Following the play session, toddlers participated in a standard object name test where their learning of the new names was assessed.

This procedure allowed for direct investigation into the three research questions posed above. First, by examining head camera images and documenting which objects were in the toddlers' views during instances of parent object naming, we were able to characterize the co-occurrence patterns available in the toddlers' learning input. Second, by collecting a measure of toddlers' learning, we were able to link the co-occurrence structure of toddlers' learning environment to their object name learning. And third, by analyzing additional measures of the toddlers' experience during naming events, such as the size of objects in view and which objects were being held, we were able to consider the role such factors played in constraining learning.

The remainder of this paper is organized as follows. We first describe the methods for collecting the observational data that was the basis for our input analysis. Next, we describe the method we used for compiling and summarizing the co-occurrence structure in the input. We then report the analysis of these statistical structures, as well as how they relate to

This research was supported by National Science Foundation (NSF) grant 0924248 and Air Force Office of Scientific Research grant FA9550-09-1-0665 awarded to CY, and by NSF Postdoctoral Fellowship (2012) awarded to SHS.

toddlers' learning. We conclude with a discussion of the implications of the results for early word learning.

II. Observational Experiment

A. Participants

Thirteen parent-toddler dyads participated (*Mean toddler age* = 19.8 mos, *Range* = 15.6 - 25.7 mos); five toddlers were female. Five additional dyads participated but were excluded from the final analyses because toddlers refused to wear the head camera.

B. Experimental Environment & Stimuli

Figure 1 depicts the experimental set-up and a subset of the novel objects used. During the experiment, toddlers sat in a chair at a table across from their parents who sat on floor cushions. With this set-up, toddlers' eyes, heads and head cameras were approximately at the same level as their parents'. To assist in the automatic head-camera image processing (see *Visual Image Processing* below), the room's floor and floor-to-ceiling curtains were all white. Additionally, both parents and toddlers wore white smocks.

The stimuli for this experiment included two sets of three novel object – novel word pairings. All objects were constructed in-house, had a single main color, were similar in size, and were small enough for toddlers to handle. Each object was paired with a novel word that was bisylabic and that adhered to the phonotactic constraints of English ("*habble", "wawa", "mapoo", "zeebee", "dodi", "tema"*).

Fig 1. The experimental set up: toddler and parent equipped with head-cameras played in an all-white room with three objects at a time.

C. Apparatus

Toddlers and their parents wore identical head cameras, each embedded in a sports headband (the current results focuses solely however on information gathered from the *toddlers'* head cameras). Cameras were KPC-VSN500 square cameras that were lightweight and measured at 1 x 1 inch. The focal length was 3.6mm. The camera resolution was 550-600 lines. The camera's visual field was 90^0. Recording rate of the camera was 10 frames per second. The camera sent a video signal to a computer in an adjacent control room.

D. Procedure

After the consent process, parents were shown large laminated cards that showed pictures of each of the novel objects along with their corresponding label. Parents were told to use these names during the experimental session. However, they were not told that the goal of the study was for their toddlers to learn the object names. Instead, they were told that the goal of the study was to observe natural play patterns between parents and their toddlers. During the experiment, the object name cards were taped to the parents' side of the table (out of the toddlers' view) so that they could consult the cards if they wanted.

Once parents and their toddlers put on the white smocks, two experimenters worked together to place the head camera on the toddlers' head. One experimenter distracted the toddler with a pop-up toy while the second experimenter placed the camera low on the toddlers' forehead. The camera was positioned such that a button on the pop-up toy the toddler was pushing was centered in the camera image.

The experimental procedure consisted of two phases: a learning phase followed by a test phase. The learning phase was divided into four trials, each lasting approximately 1 to 1.5 minutes long. On each trial, dyads played freely with one of the two object sets. Object sets were interleaved. Immediately following the learning phase, infants participated in the test phase. During test, toddlers sat on their parents' lap at the same white table opposite of an experimenter. The test phase began with a familiarization procedure to ensure the toddler understood the task The familiarization consisted of a series of three-alternative forced-choice (3AFC) trials where the experimenter put three *familiar* objects on a tray (e.g., car, cup, duck) out of the toddler's reach and asked the toddler to point to a particular object, saying for example: "See these, where's the cup?" After a few familiarization trials, the experimenter proceeded to the actual testing phase, which consisted of a series of 3AFC test trials designed to test the extent to which toddlers had learned the names of the objects from the learning phase. On each trial, the experimenter asked toddlers to point to the referent of one of the novel words in the same manner as the familiarization phase above. The objects on any given trial included the novel word's referent as well as two foils randomly selected from the remaining set of novel objects. Toddlers were given neutral feedback ("thank you") regardless of which object they chose. Each novel word was tested twice, yielding a total of twelve test trials per toddler.

E. Data Processing

1) Visual Image Processing. Of particular interest in the current study were the properties of the toddler-perspective images during naming events. The specific variables of interest were the number of objects in view and the size of each object. In order to derive these variables, we employed

an in-house automated machine vision program. Briefly, this program first separates non-white, object pixels from the white background. Then, object pixels are combined into object blobs based on color similarity. Finally, each object blob is examined and given an object label based on an object recognition training procedure with the novel objects' shapes and colors (for more technical details on this program, see [4]). From this automated procedure, we derived for each of the 3500 frames per subject, the objects that were in the toddlers' field of view and the size of each of these objects. In addition to these vision variables, we manually analyzed each frame to determine which object (if any) the toddler was holding.

2) *Speech Transcription*. Transcripts of parent speech during the experiment were divided into utterances, defined as a string of speech between two periods of silence of at least 400 ms. Utterances that contained a novel word were marked as naming events. The onset and offset of these utterances was used to determine the duration of each naming event. The mean number of naming events was 68.5 events per dyad (*SD* = 37.12). The average length of naming events was 1.57s long (*SD* = .65).

3) *Forced-Choice Test Coding*. A trained coder unaware of the correct word-object pairings watched muted video clips of each trial of the testing phase and noted which object toddlers chose. For each word, we tallied the number of times toddlers selected the word's correct referent. We categorized a word as being *learned* if toddlers selected the correct referent both times the word was tested. All other words were categorized as *non-learned words*. The mean proportion of words learned across toddlers was .21 (*SD*= .19), which trended towards greater than the proportion that would be predicted by chance (probability of selecting one object at random across two 3AFC trials is .11), $t(12) = 1.74$, $p = .10$.

III. Statistical Learning Models

The goal of the current endeavor was to understand the statistical structure in toddlers' learning environments and to relate that structure to their learning. To examine the structure of each toddler's input, we constructed word-object co-occurrence matrices that represented the link between the novel words and the novel objects in a set. We then populated association matrices with the naming event data gathered from the speech transcripts of parents' input and the coding of the toddlers' head camera images. Figure 2 provides a toy example that illustrates this process. In Naming Event 1, because the parent uttered "dodi" (W1) and the red object on the right (O1) as well as the blue object on the left (O2) were visible in the toddler's view, the matrix cells representing the link between W1-O1 and W1-O2 get updated. Then on Naming Event 2, when "habble" (W2) is uttered and the blue object on the bottom (O2) and the green object at the top right (O3) were in the toddler's view, the cells representing the link between W2-O2 and W2-O3 are updated. For each dyad, we constructed association matrices from all that dyad's naming events. Because there were two sets of name-object pairings, each participant had two sets of 3x3 matrices, one for each set.

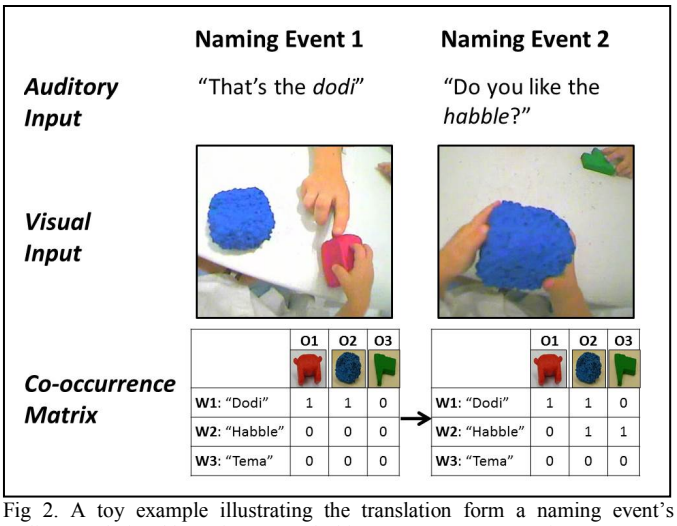

Fig 2. A toy example illustrating the translation form a naming event's auditory and visual input into a word-object co-occurrence matrix.

As is obvious in the head camera images in Figure 2, not all entities in the toddler's view are visually equal; some objects are more visually salient (e.g., closer and therefore larger in the toddler's view). And some are being actively acted on (e.g., in the toddler's hand). To understand how co-occurrence statistics interact with these visual and attentional processes, we constructed multiple types of matrices for each dyad, differing only in the way in which cells were updated. The matrices represent different conceptualizations of statistical word learning: unconstrained statistical word learning (only co-occurrence matters) versus constrained statistical word learning (co-occurrence information is modulated by visual properties or by attention). We define an unconstrained statistical word learner as a learner that simply keeps track of the co-occurrences between words and objects. When such a learner confronts a new word, the learner creates associative links of equal strength between that word and each of the objects in view. In the current context, unconstrained learning essentially acts in the same manner as the toy example illustrated in Figure 2 above with one important difference. Because naming events extend in time, and thus have multiple visual frames associated with it, the matrix gets updated using the *proportion of frames* during the naming event in which a word and an object co-occur.

We define constrained statistical learning as also a process of keeping track of the co-occurrence probability between words and objects. However, on any given naming event constrained learners might assign different associative weights to some word-object pairings over others. As listed in Table 1, we considered two types of constrained statistical learning. First, we considered a vision constrained statistical learning process. Here, for each naming event we updated matrix cells using the mean object size (proportion of pixels of image size) of each object present during the naming event. Thus, in Naming Event 2 in Figure 2 above, the association weight assigned to W2-O2 will be much larger than the association weight assigned to W2-O3. Second, we also considered a manual engagement (hand) constraint on statistical learning.

For this learner, we updated matrix cells using the proportion of time each object was in the hand of the toddler. To illustrate again in Naming Event 2 above, only the cell representing the association between W2 and O2 receives activation.

TABLE I. THE LEARNING MODELS AND THEIR UPDATING PROCEDURE

Learning Models	**Statistics Accumulated**	**Metric for Updating Association Matrix**
Unconstrained Learning Model	Co-occurrence statistics between words uttered and objects in view	Proportion of naming events in which objects were visible
Constrained Learning Models (Vision Constraint)	Co-occurrence statistics weighted by object size	Mean object sizes during naming event
Constrained Learning Models (Hand Constraint)	Co-occurrence statistics weighted by in-hand status	Proportion of naming event objects were in hand

IV. RESULTS

The analyses focus on two issues. The first is the *informativity* of the statistical structure of the input according to each of the models. That is, we ask under the assumptions of each of the learning models, was there sufficient structure in the input to enable successful word learning. The second analysis speaks more directly to the issue of whether toddlers actually utilize that structure in their input in the service of learning. We do this by examining the relation between the structure in the input and the toddlers' learning performance in the object name test. For each of these two analyses, of particular interest is the comparison between the unconstrained and constrained models of learning.

A. Informativity of the Statistical Structure in the Input

To examine the statistical structure of the toddler's inputs, from each dyad's matrices, we computed a normalized associative strength between each word and its target referent. We obtained this value by normalizing the associative strength between a word and its referent across all associative strengths associated with that particular word. Across models, we calculated a mean word-to-referent associative strength for each dyad. Figure 3 depicts these means across models, averaged across subjects. The Figure illustrates four patterns. First, the mean word-to-referent associative strength of the unconstrained statistical learning model (M = .34, SD = .02) is only marginally significantly greater than what would be predicted by chance given that there were three objects in a set (.33), $t(12) = 1.90$, $p = .08$. Second, the mean word-to-referent associative strength of the two constrained statistical learning models (M_{vision} = .39, SD_{vision} = .05, M_{hand} = .42, SD_{hand} = .14) is significantly larger than chance levels ($t_{vision}(12) = 4.33$, $p_{vision} < .001$; $t_{hand}(12) = 2.40$, $p_{hand} < .05$). Third, the mean word-to-referent associative strengths for both of the constrained models are significantly higher than the mean word-to-referent associative strength of the unconstrained model (as determined via independent samples t-tests against the unconstrained statistical learning model: $t_{vision}(24) = 3.03$, $p_{vision} < .01$, $t_{hand}(24) = 2.05$, $p_{hand} = .05$). Finally, there doesn't appear to be any differences in mean word-to-referent associative strengths across the two constrained models ($p > .10$).

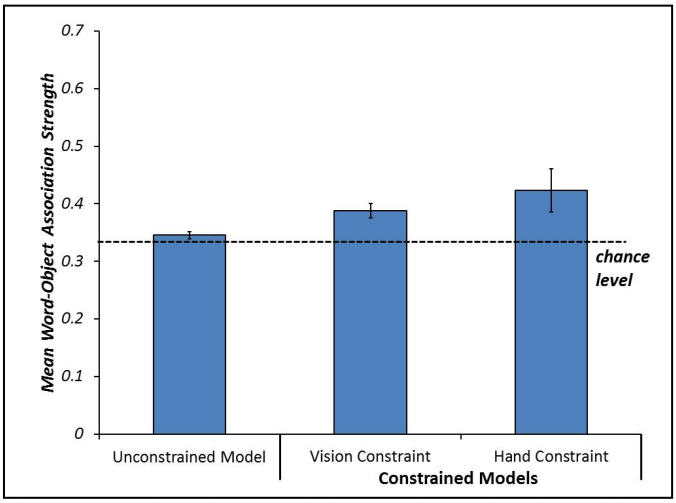

Fig 3. Mean word-to-referent association strength across learning models.

These results suggest that to the extent that word learning is viewed as a simple process of analyzing the co-occurrence information between words and candidate objects in view, learning in the current setting would be quite difficult. That is, results from the unconstrained learning model suggests that when a word is uttered, that word's referent object appears in the toddler's view at a rate that is barely more reliably than non-referent objects. However, if word learning is viewed as a statistical learning process constrained by sensori-motor processes, then there does exist sufficient structure in the environment that can be exploited for learning. Based on the visual constraint model, across a word's naming events, that word's referent tends to be larger than the other objects in view. Based on the hand constraint model, across a word's naming events, that word's referent is more likely to be in the toddlers' hands than other objects. That the analyses of the two constrained models did not appear to be different from one another is likely indicative that the two constraints (the size of an object in the toddler's view and the toddler's holding of an object) are tightly linked, providing converging, rather than unique, information about a word's likely referent. The association between the two constraints is depicted in Figure 4, which shows the association strength for each of the 78 word-referent pairings (6 word-object pairings across 13 dyads) when analyzed using the vision constraint algorithm and when analyzed using the hand constraint algorithm. The figure highlights that the longer proportion of time the referent object is held during naming events, the larger the referent object in the toddler's view. This finding and interpretation is consistent with previous work demonstrating strong coupling between a toddler's holding of an object and the visual properties of that object in the toddler's view [5].

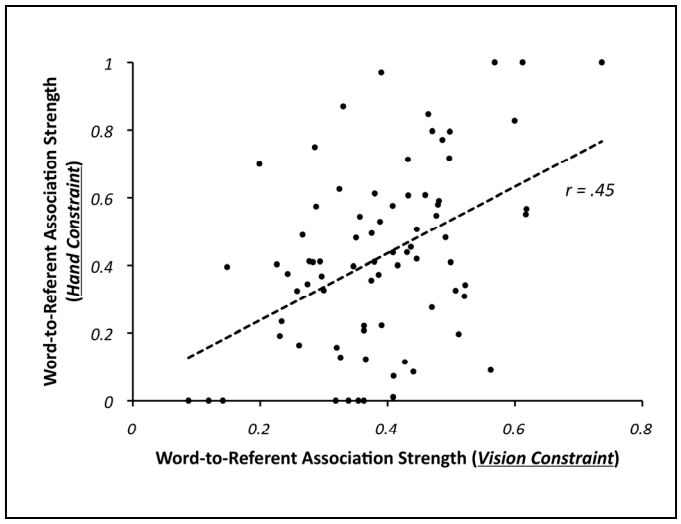

Fig 4. Relation between word-to-referent association strengths when established by the vision-constrained learning algorithm (x-axis) and when established the hand-constrained learning algorithm (y-axis).

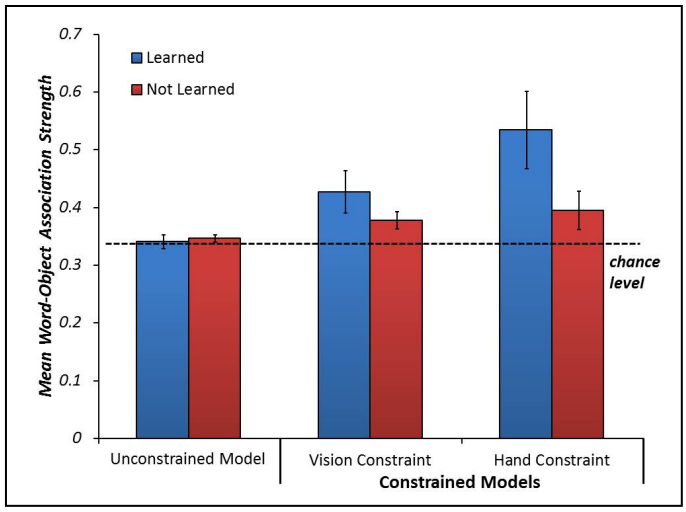

Fig 5. Mean word-object association strengths for learned and not learned words across models

B. Linking Statistical Structure in the Input to Toddler's Object Name Learning

Demonstrating that there exists sufficient structure in the toddlers' learning environment does not speak to whether the toddler participants in our study actually made use of that information. To more directly address the issue of toddler's use of the statistical structure in the learning environment, we analyzed a word's association strength in relation to whether a toddler actually learned that word. Specifically, we compared the mean association strength for words toddlers learned relative to words toddlers did not learn. As Figure 5 illustrates, for the unconstrained model, the word-to-referent mean associative strength for the learned items is no different than the associative strength for the words not learned. In contrast, for the constrained models, the word-to-referent associative strengths for the learned words are larger than the associative strength for the words not learned, raising the possibility that having more reliable input actually leads to better learning.

To address this pattern statistically, we employed mixed logit modeling [6] with the goal of trying to predict whether a word was learned using that word's mean associative strength with its referent object[1]. This method is appropriate for our analysis because (1) our dependent variable is binary (whether or not a word was learned), and (2) this method has the ability to account for random subject and item effects. For all analyses, we utilized the *lmer* function of the *lme4* package in R [7].

[1] In these models, our predictor variable was not the normalized association strength reported in the graphs above. Instead, we used the difference between the associative strength between a word and its referent and the average associative strength between that word and its non-referents. Although the normalized metric makes comparison across models easier, the raw difference is a more veridical measure of the input.

A Mixed logit model was conducted separately for each learning model. For all models, subjects and items were included in the model as random effects. Table 2 summarizes the results from the models. These results confirm the patterns observed in Figure 5. The unconstrained statistical structure of the learning environment does not have an effect on whether a word is learned or not. In contrast, co-occurrence statistics constrained by sensori-motor processes does have an effect on whether or not a word is learned. The more reliable the co-occurrence patterns between a word and its referent, taking into consideration object size or whether the object is in hand, the more likely the toddler learned that word.

TABLE 2. MODEL FIT AND COEFFICIENTS OF THE KEY PREDICTOR VARIABLE IN EACH MIXED LOGIT MODEL

Model Type	**Model Good. of Fit**	**Model Fixed Effect Parameters**			
	log-lklhd	***Coef***	***SE***	***Z***	***p***
Unconstrained Model	-39.2	.10	.20	.52	.60
Constrained Model (*Vision Constraint*)	-36.5	.04	.01	2.38	.02
Constrained Model (*Hand Constraint*)	-37.1	.14	.07	2.08	.04
Combined Model (*Vision + Hand Constraint*)	-35.8				
Obj. Size		.03	.02	1.57	.12
Obj. in Hand		.08	.07	1.15	.25

To further explore the relationship between effects of object size and object manual engagement on toddler word learning, we analyzed a combined mixed logit model that examined the independent contribution of each of these variables on learning. As reported in Table 2, when both variables were included in the logit model, neither variable remained a significant predictor of learning. This finding confirms and is consistent with our correlational analyses above (Figure 4), suggesting a tight coupling between a

toddler's manual engagement with an object and the visual features of that object from the toddler's point of view. In the context of learning the meaning of a new word, this tight coupling renders these two variables as converging rather than unique information sources to determine a word's referent.

V. Discussion

The goal of this study was to better understand the statistical structure in toddler's word learning input in a free-flowing object play context, and to examine the role of this structure on object name learning. The analysis of simple word-to-object co-occurrence patterns suggests that the input contained little reliable structure. Words and their referents were about just as likely to co-occur as words and non-referents. Thus, had the toddlers in the study relied solely on these co-occurrence patterns to acquire object labels, they would have likely learned very little, if anything at all. Although the toddlers did not actually learn many words, they did learn some, suggesting that they must have engaged in a different type of learning process. Further analyses suggest that a sensori-motor guided statistical learning process is a possible candidate mechanism. When toddler's inputs were analyzed from a view point that the input to statistical learning is constrained by basic processes such as what appeared more dominant in the toddler's visual field or what objects were in the toddler's hands, there appeared to be sufficient reliable structure in their input that could enable object name learning. Further, our analyses demonstrate that the better this attentionally-constrained statistical structure was for any given word, the greater the likelihood toddlers learned that word.

That simple word-to-object co-occurrence patterns did not play a stronger role in learning was somewhat surprising. Recent analyses of a toddler's view of the learning environment suggest that their visual environment tends to be less cluttered than that of adult learners [8,9]. We expected that this lack of clutter would have translated to a more reliable co-occurrence pattern between a word and its referent. It is possible that the current experimental context may have underestimated the utility of simple co-occurrence statistics in word learning. Given that in the current context there were only 3 objects in each set, there was little variability in the range of objects that could be in the toddler's view at the time object labels were heard. In the toddler's actual learning environment, there is likely greater diversity in the contexts in which objects are seen and object names heard. Although the small object sets used in the current experiment might be expected to ease learning for some reasons (e.g., they lessen memory demands), from a statistical point of view, these very small sets might actually make learning more difficult due to the high number of spurious correlations that could be formed (see [9] for a discussion on this point).

Although the current results suggest that tracking simple co-occurrence statistics is likely insufficient for learning, our results do suggest that the tracking of constrained co-occurrence statistics is sufficient. This finding is consistent with two lines of research. First, the results are consistent with the role of sensitivity to cross-situational statistics in learning new words [3,10]. The results are also consistent with the notion that early word learning is a constrained process [11]. Thus, the present finding that word learning involves both a mechanism for tracking statistical co-occurrences in the environment and a process of constraining the information that is tracked is likely uncontroversial. However, the simplicity of both the learning mechanism and the constraints that can capture toddler's learning patterns is surprising. The results suggest that the simple learning mechanism of accumulating word-to-referent co-occurrence patterns and the very basic attentional constraints of "attend to large things in view" or "attend to things in one's hands", can go a long way in capturing toddler's word learning patterns.

Of course, we do not suggest that these attentional constraints are the only ones that guide statistical word learning. There are likely many types of information and cues that lead learners to consider some statistics over others. Additionally, not all cues may play the same constraining roles throughout development. Thus, the goal for future research is to better understand the roles these different constraints play at different points in development, how these different constraints relate to one another, and how they interact with the statistical learning mechanism.

Acknowledgment

We thank Charlotte Wozniak, Amanda Favata, Alfredo Pereira, Amara Stuehling, and Melissa Elston for data collection. We also thank all the families that participated in this study.

References

[1] E. Hoff, and L. Naigles, "How children use input to acquire a lexicon," Child Dev., vol. 73, pp. 418-433, March/April 2002.

[2] C. Yu, "A Statistical associative account of vocabulary growth in early word learning," Lang., Learning, and Dev., vol. 4, pp. 32-62, 2008.

[3] L.B. Smith, and C. Yu, "Infants rapidly learn word-referent mappings via cross-situational statistics," Cog., vol., 125, pp. 1558-1568, 2008.

[4] C. Yu, L.B. Smith, H. Shen, A.F. Pereira, and T. Smith, "Active information selection: visual attention through the hands," IEEE Trans. on Auto. Mental Dev., vol. 1, pp. 141-151, August 2009.

[5] C. Yu, and L.B. Smith, "Embodied attention and word learning by toddlers," Cog., vol. 125, pp. 244-262, 2012.

[6] T.F. Jaeger, "Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models," J. of Mem and Lang., vol. 59, pp. 434-446, 2008.

[7] R development core team, "R: A language and environment for statistical computing," Vienna: R Foundation for Statistical Computing, http://www.R-project.org, 2005.

[8] L.B. Smith, C. Yu, and A.F. Pereira, "Not your mother's view: The dynamics of toddler visual experience," Dev. Sci. vol. 14, pp. 9-17, January 2011.

[9] D. Yurovsky, L.B. Smith, and C. Yu, "Statistical word learning at scale: The Baby's view is better," Dev. Sci., in press.

[10] C. Yu, and L.B. Smith, "Rapid word learning under uncertainty via cross-situational statistics," Psych. Sci., vol. 18, pp. 414-420, 2007.

[11] A.L. Woodward, and E.M. Markman, "Early word learning," in Handbook of child psychology: vol 2. Cognition, perception, and language, D. Kuhn and R.S. Siegler, Eds. New York: John Wiley & Sons, 1998, pp. 371-420.