Quantifying cross-situational statistics during parent-child toy play

Ellis Cain[1], **Rachel Ryskin**[1], **Chen Yu**[2]

ecain@ucmerced.edu, rryskin@ucmerced.edu, chen.yu@austin.utexas.edu

[1]Cognitive and Information Science, University of California - Merced, USA
[2]Department of Psychology, The University of Texas at Austin, USA

Abstract

According to the cross-situational word-referent learning account, infants aggregate statistical information from multiple parent naming events to resolve ambiguous word-referent mappings within individual events. While some experimental studies have shown that infants and adult learners are sensitive to these statistical regularities, other studies that use naturalistic stimuli (e.g., real-world scenes with toys) reveal poor performance in adults' ability to infer the correct referent. In the current study, we examined whether the properties of young learners' input "in the wild" may differ from those found in laboratory experiments. We analyzed the temporal and spatial regularities of parent naming events from a naturalistic data set of video recordings and eye-tracking collected while parents and children played with toys. We also examined how these regularities affected infants' visual selection of information through attention. Overall, we found that parents were less likely to name the same toy twice than to name two different toys in sequence, except at short lags ($0s > t > 5s$). Most of the visual scenes accompanying naming events were composed of several toys of approximately equal (and small) size. Child attention to the target toy appeared to be modulated primarily by object size. These results underscore the importance of quantifying the regularities found in naturalistic data in order to shed light on the type of mechanism used in word learning.

Keywords: word learning; cross-situational learning; vision and attention; exploratory data analysis

Introduction

One of the key challenges in word learning is that of mapping words to their referents in the world. In an everyday learning context such as toy play, when a parent produces an object label that is new to the child, the correct label-object mapping is not necessarily clear or explicitly stated from the child's point of view. Quine (1960) referred to this as *referential uncertainty*. Nonetheless, human learners, and even infants, are able to solve this problem by building correct word-referent mappings from this ambiguous input.

Several theoretical accounts have been proposed to explain how human learners solve the mapping problem. On one account, learners reduce in-the-moment ambiguity using social (Baldwin et al., 1996) and linguistic cues (Abend, Kwiatkowski, Smith, Goldwater, & Steedman, 2017). On another account, learners aggregate statistical information about word-object mappings across individually ambiguous learning situations; this mechanism is termed cross-situational learning (Yu & Smith, 2007; Smith & Yu, 2008; Fitneva & Christiansen, 2011). The current paper focuses on examining the cross-situational learning solution for early word learning.

Cross-situational word-referent learning was proposed to provide a statistical solution for initial mappings of words to referents. All varieties of computational models succeed at this kind of learning (Amatuni & Yu, 2020; Bhat, Spencer, & Samuelson, 2018; Bambach, Crandall, Smith, & Yu, 2018). Laboratory experiments show that adults, older children, and toddlers are quite good at this kind of learning (Yu & Smith, 2007; Smith & Yu, 2008; Fitneva & Christiansen, 2011). Following the general design principle in statistical language learning, the experimental task of cross-situational learning was explicitly invented to demonstrate that learning could emerge from the aggregation of experiences with individually ambiguous naming events. On each trial, the learner heard multiple words and saw multiple objects with no information about which word went with which object. Across trials, each word always co-occurred with just one referent; thus there was cross-trial certainty despite within-trial uncertainty. But this cross-trial certainty only holds if the learner samples, remembers, and aggregates word-referent co-occurrences across trials.

However, one critique that has been leveraged against the cross-situational learning solution is that it may not work with real-world data. Trueswell, Gleitman, and colleagues (Medina, Snedeker, Trueswell, & Gleitman, 2011) asked adults to guess the intended referent when presented with 3rd-person video (no audio) of parent naming events. In this Human Simulation Paradigm (HSP), adults were very poor at guessing referents and further showed no ability to aggregate information about word-referent correspondences across these highly cluttered visual scenes. If everyday scenes paired with words are insufficient to support statistical learning in adults, is it still possible that infants learn their first object names by aggregating heard words and seen things across multiple naming experiences in the clutter of the real world?

The key to answering this empirical question is to measure and quantify the statistics of the input that young learners perceive in the real world. Toward this goal, the infant-perspective scenes that coincide with parent

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society*.

naming events are the input for learning. Hence, theories and research on early object name learning need to work with these scenes (Yu, Zhang, Slone, & Smith, 2021). Further, the relevant data for learning concerns the scene elements visually selected by the active learner. Thus, we need to know the properties of the infant-perspective scenes that coincide with parent naming and we also need to know how infants visually select information in those scenes.

The goals of the present study are to analyze real-world scenes that capture parent naming moments from the infant's perspective in order to determine: 1) what temporal regularities are present in parent naming during naturalistic toy play; 2) what complexity and composition characterizes visual scenes during parent naming, and 3) how infants allocate their attention (as measured by eye-gaze) and thereby constrain their input. We first describe the temporal properties of the auditory input (i.e., parent naming) and how it impacts child attention. We then characterize the visuo-spatial properties of the infant-perceived scenes corresponding to naming events and how they affect child attention. Since we do not have a direct measure of learning as in the lab-based cross-situational learning experiments, we measure infant gaze since visual attention is necessary (but not sufficient) for learning. If laboratory tasks and assumptions about early word learning violate the properties of the audio-visual context of infant learning experiences in the real world, then the cross-situational learning mechanism, in its current form, may not be sufficient to explain how infants learn word-referent mappings.

Data and Data Processing

The data used in this analysis were collected from free-flowing parent-child dyadic toy play sessions (n = 36, age = 15−25 mo), each involving the same set of 24 toys. Each session lasted an average of 8.11 minutes (range 0.95−16.17 min), with 291.79 minutes of video data in total. Figure 1-left shows a third-person view of the experiment setup. At the beginning of the experiment, the 24 toys were randomly spread on the floor. The parents were asked to play as they would at home and to keep their child engaged with those toys. During the play session, the parent and child each wore a head-mounted eye-tracker with a front-facing camera capturing what was in their field of view (sampling rate of 30Hz for both). An example of the child's view can be seen in Figure 1-right.

The videos from the eye-trackers and front-facing cameras were synchronized and calibrated using the Yarbus program (Positive Science LLC) to generate a gaze cross-hair in each image frame indicating the child's gaze location (Fig. 1 right). The field-of-view videos were then processed using YOLO object detection (Redmon & Farhadi, 2018), which was trained on an annotated

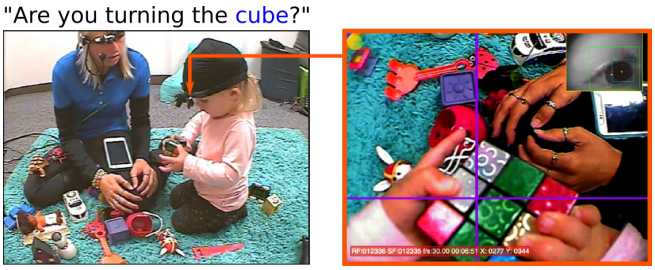

Figure 1: Third-person view of the experiment setup (left) and the child egocentric view (right), showing the gaze cross-hair.

sub-sample of frames to detect toys present in the video. After training, the algorithm automatically detects toy objects in view, providing up to twenty-four coordinate sets per frame (x, y, x-length, y-length) that are used to draw "bounding boxes" around all of the visible toys. Parent speech and child vocalizations were manually annotated at the utterance level using Audacity. There was no minimum utterance length for annotation, but utterances less than 400ms apart were collapsed into a single utterance. The gaze data, object detection data, and transcriptions were all synchronized, which allowed us to measure children's visual attention and scene composition temporally aligned to individual spoken utterances.

The current study includes data from 106,200 frames of egocentric videos from the child's front-facing camera, corresponding to 1,475 parent naming events. For each naming event, object detection data and gaze fixations were extracted within a 3s window starting at the onset of a given naming event.

Study 1: Temporal Continuity of Parent Naming Events

Previous lab experiments showed that word learning is facilitated when human learners are exposed to repeated naming of the same object within a short period of time (Kachergis, Yu, & Shiffrin, 2009), compared with when naming frequently switches between different objects. The goals of Study 1 were to examine whether such temporal regularities can be observed in parent naming during free-flowing toy play and to compare child attention to named (target) objects when hearing repeated names and when hearing different names.

The lag between any two naming events was determined by their inter-onset interval (IOI), as seen in Figure 2. Parent naming utterances were categorized into two groups: 1) **same** when the parents named the same toy repeatedly (e.g., parent names "doll", then 5s later names "doll" again); or 2) **different** when the parents named one toy first and then switched to name a different toy (e.g., parent names "doll", then 3s later names

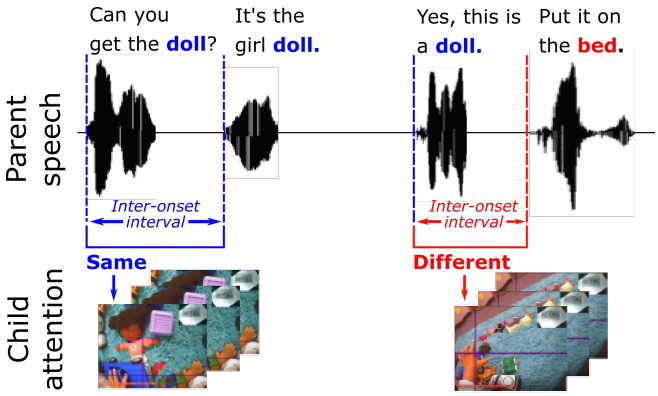

Figure 2: Temporal continuity of parent naming. The graph shows examples of the three different events of interest: the same toy being repeated, an isolated naming, and two different toys being named.

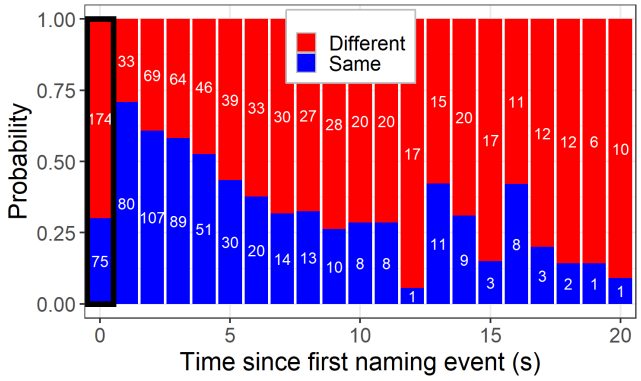

Figure 3: Likelihood of the parent naming the same or different toy based on the time since the first naming event (IOI). The black box demarcates within-utterance naming pairs (t=0s).

"bed"). Due to the annotation scheme, naming event pairs within the same utterance had an IOI of t=0s.

First, we examined the overall distribution of IOIs and the likelihood of repeated parent naming based on these IOIs. Then, we measured child attention during these naming events (from onset of utterance to 3s after) and modeled the relationship using a logistic mixed effects model.

Results

There were 40.97 naming events on average per subject (range 7−111), with a mean of 3.09 naming events for each toy (range 1−19) per subject. Different naming pairs were more frequent (n = 869) than same naming pairs (n = 570).

Using the inter-onset interval between any two naming events to quantify the temporal aspect of parent naming behavior, we found that the overall distribution of the IOIs, regardless of toy identity, was right skewed; the mean time lag was 10.925s, while the median was 4.25s (SD=25.56s, range 0−570.39s). The IOIs for same naming events were on average shorter (M=5.87s) and less variable (SD=9.31s, range 0−110.68s) when compared to those for different naming events (M=14.24s, SD=31.59s, range 0−570.39s). Since the IOIs for both same and different naming event pairs are not normally distributed, we used a Wilcox rank sum test on their distributions, which showed that the difference is statistically significant (W=3.08e5, p-value < 3.33e-15).

We grouped the inter-onset intervals by second and calculated the conditional probability of the child hearing the same or different toy as the second label given the IOI (Fig. 3). The graph includes those naming event pairs that are less than 20s apart to focus on the naming events that are temporally close together (same = 544, different = 703). Within a single utterance (t=0s), children are more likely to hear the parent name two different toys. However, for separate naming events (t>0s), children are more likely to hear the same label repeated within four seconds after the first naming, but this probability decreases as the inter-onset interval increases.

In order to quantify child attention based on the inter-onset interval, we calculated the proportion of time that the child fixated the target during the second naming event. Here we excluded naming event pairs that are within a single utterance (t=0s), since there is not enough resolution in the data to extract the child's attention proportion during each individual word. Figure 4 shows attention to target proportions from the second naming event across lags. The mean attention proportions and line of best fit are plotted over the individual proportions (for a given naming event) binned by second. With this new set of temporally close naming pairs (0<IOI<=20s), we found that children spend similar proportions of time fixating the target object during the second naming event for repeated toys (attention proportion: M=0.41, SD=0.39) and for two different toys (M=0.38, SD=0.35). Attention proportion did not appear to change as IOI increased either for same or different naming pairs.

We used a logistic mixed effects model to test the relationship between (centered) IOI, naming pair type (same/different), and their interaction and the probability that the child looked at the target (a binarized version of the fixation proportion plotted in Figure 4). Parent-child dyads were entered as random intercepts. Neither the IOI nor the naming pair type had a significant relationship to the probability that the child looks at the named toy (IOI: $b = 0.001$, $SE = 0.02$, $p = 0.66$; Pair type: $b = 0.08$, $SE = 0.15$, $p = 0.58$). The interaction between the two predictors was also not significant ($p = 0.36$).

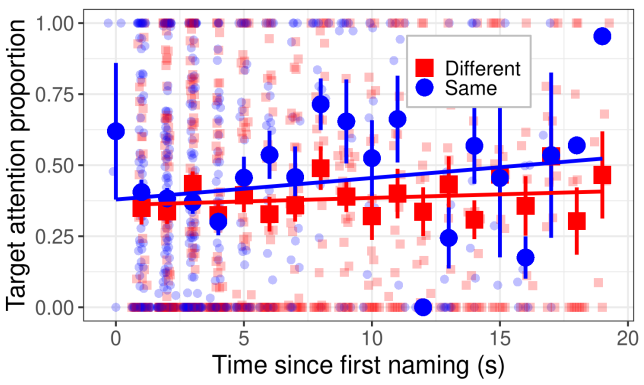

Figure 4: Proportion of time spent looking at the target for (second) naming events that are preceded by the naming of the same toy or different toy. Time since first naming was binned for each second.

In sum, parent naming events seem to be organized such that repetitions of the same toy name mostly occur within a short period of time, while sequential naming of different toys occurs at longer lags. However, within a single utterance, parents are most likely to name two different toys. The time between two naming events and whether they refer to the same object or a different object do not appear to influence how much attention the child pays to the referent of the second naming event.

Study 2: Spatial Properties of Visual Input During Naming Events

When hearing a toy name, what was the infant's visual input and how did they allocate attention? Study 2 aimed to answer these questions by analyzing the spatial regularities of scenes from the infant's viewpoint and their impact on child attention. Taking a statistical learning perspective, the child is a statistical learner that extracts a subset of the information that they are presented with. To that end, we aim to quantify two aspects of scene composition during naming events; the first aspect is the distribution of toy sizes in the child's field of view, and the second aspect is the combination of these toys in view to comprise a visual scene (Fig. 5).

For the distribution of object sizes, we ranked the objects in the child's view during a naming event according to size. We then calculated the average object sizes for each size rank per scene, regardless of identity, and represented each scene as a vector of the top six object sizes. Mean shift clustering, an unsupervised clustering algorithm, was used to identify patterns in scene composition and extract prototypical scenes from these scene vectors. Similar to Study 1, we quantified child attention based on these prototypical scenes and then modeled their relationship using a mixed effects model.

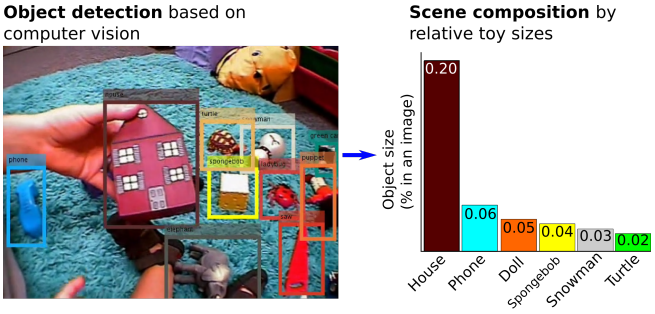

Figure 5: An example of the object detection data after processing (left) and the corresponding extracted toy sizes (right).

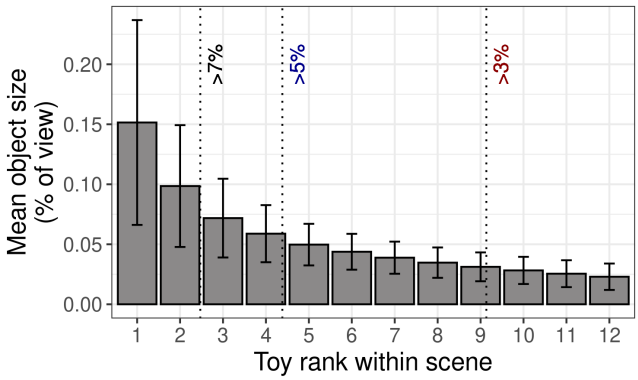

Figure 6: Mean toy sizes during a given naming event, sorted by size ranks. Dashed lines indicate the average number of toys in the child's field of view during a naming event, based on toy size cutoff.

Results

During naming events, there were an average of 17.05 toys in view (SD=4.08, range 1−24). However, when we filtered the data to only include toys in the foreground (toy size > 5% of image), there was an average of 4.38 toys in view (SD=2.29, range 0−14).

To get a more detailed look at regularities of toy sizes, we calculated the average size of toys by rank within a scene (Fig. 6). The bottom half of largest toy sizes were dropped in order to focus on the more noticeable objects in the child's field of view. Only a few toys occupy a relatively large portion of the child's field of view; the largest toy during the naming event occupies about 15% (SD=8.53%). We tried different potential cut-offs for identifying toys in the foreground: those taking up >3%, >5%, and >7% of the image. With the first, most generous definition (toy size >3%), there was an average of 9.134 toys per scene (SD=3.39, range 0−21). The second foreground cut-off value (toy size >5%) gave an average of 4.382 toys (SD=2.288, range 0−14), which most closely matches the original cross-situational learning studies. With a stricter definition of the foreground

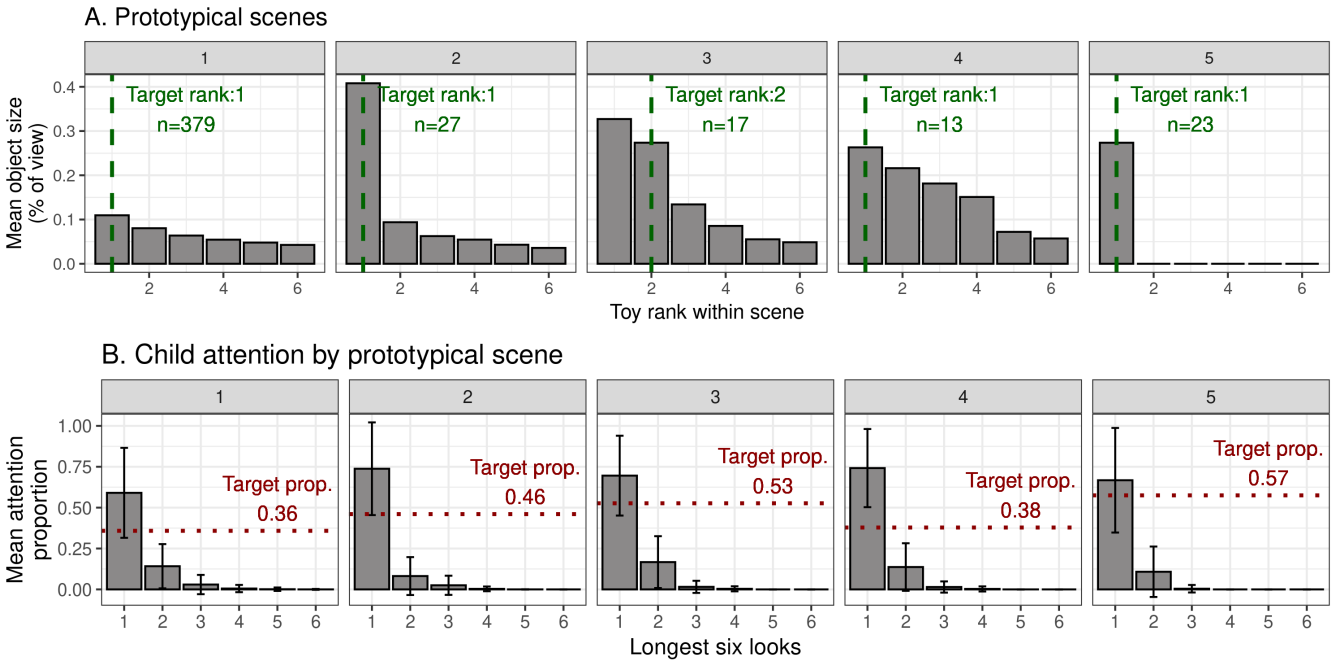

Figure 7: Five major clusters as identified by the mean shift clustering results. A) Each graph represents a prototypical scene composition during a naming instance, with each bar being the nth largest toy. Dashed lines indicate the modal target toy rank and the corresponding count. B) Each graph represents the proportion of time the child fixated on any object (for the longest six looks) for scenes belonging to each of the 5 clusters. Dotted lines indicate the mean proportion of looks to the target object.

(toy size >7%), the average number of toys per scene dropped from 4.38 toys to 2.47 toys (SD=1.65, range 0−11).

We used mean shift clustering to identify groups of similar scenes in the data. Here, scenes are represented as a vector of the top six largest toy sizes. For example, the scene in Fig.5 would be coded as [0.20, 0.06, 0.05, 0.04, 0.03, 0.02], while a scene with only one toy in view would be [0.25, 0, 0, 0, 0, 0]. There were 23 clusters identified, though the majority of the clusters were outlying single scenes (18 had less than 10 scenes per cluster, 4 had less than 60 scenes, and one included 1194 scenes). Due to being too idiosyncratic, those outlying clusters with less than 10 instances were not included in the following analyses.

From the clusters, we then extracted "prototypical" scene compositions of naming events. Figure 7A shows the prototypical scenes as identified by the mean shift clustering, where each subplot is an example scene from the data which is closest to its cluster's center point. Cluster 1 had the most instances (n=1194) and represents scenes in which several toys are approximately equally-sized and none take up more than 12% of the child's visual field. Cluster 2 (n=51) represents scenes in which one object takes up a large proportion of the visual field (near 40%) and there are several other toys in the background (<10% of view). Cluster 3 (n=38) represents scenes in which two toys take up a large proportion of the visual field, with another medium size object and the rest in the background. Cluster 4 is the second most common (n=55) and represents scenes in which there is more variability in the sizes of toys in the child's view, most of which are of medium size. Finally, cluster 5 (n=31) represents naming instances with a large, single toy in the child's visual field.

The dashed lines in Fig. 7 indicate the most frequent rank of the toy labelled by the parent among the sorted toy sizes, along with their count. With the exception of cluster 3, the labelled toy was most frequently the largest of the six toys in view (1: 379/1194, 2: 27/51, 4: 13/55, 5: 23/31). For scenes with one toy in focus (clusters 2 and 5), the labelled toy is the largest for just over half of the instances.

We quantified child attention regarding their overall gaze behavior (regardless of toy identity) and their attention on the named toy. For their overall gaze behavior, we sorted their gaze proportions (of the longest six looks) for each scene and calculate the mean proportions by rank. Figure 7B shows these distributions of gaze proportions based on each cluster. Even though the prototypical scenes are quite different in their scene composition, the children's gaze patterns are quite similar; one long look, with a few short looks. Likewise, the mean proportions of target attention are similar as

well (range 0.36−0.57). For all clusters, the mean proportions of gaze to the target are lower than the longest look proportion suggesting that children did not always look longest at the target during a naming event.

Target size rank significantly predicted whether the child looked at the target ($b = -0.24$, $SE = 0.017$, $p <$2e-16) such that the probability of fixation was larger for objects which were relatively larger. Children were also more likely to fixate the target when the scene composition belonged to cluster 3 relative to cluster 1 ($b = 1.09$, $SE = 0.54$, $p = 0.045$).

In sum, during naming events, toys are distributed in the scenes such that only a few toys occupy the foreground of the scene (M=4.38 toys/scene) when we use a size cut-off of >5%. Through mean shift clustering, we identified five prototypical scene compositions. For most of these prototypical scenes, the target was frequently the largest toy, however, only in those with one object in focus (clusters 2 & 5) did it occur for more than half of the naming events. Children's overall looking behavior during a naming event was similar across clusters. Children were more likely to look at the named object when it was large relative to the other objects in the scene and slightly more so when the composition of the scene consisted of two larger toys with other toys in the background.

General Discussion

In the current study, we quantified the statistical regularities of parent naming and child attention in naturalistic toy play data. We demonstrated that parent naming tendencies (repeat same vs. name different toy) change as a function of the time between naming events. We also observed that the majority of naming events were accompanied by scenes in which several toys only occupied an equal, small proportion. Furthermore, we found that the child's gaze distribution is consistent based on these regularities and, based on our models, identified the size rank of the target toy as the most significant predictor of child attention. These findings extend previous work on learners' sensitivity to co-occurrence statistics of visual and auditory input (Kachergis et al., 2009).

Our results reveal statistical regularities that are similar to those found in cross-situational learning studies (Yu & Smith, 2007; Smith & Yu, 2008; Fitneva & Christiansen, 2011; Kachergis et al., 2009). For example, the average naming frequencies of each toy reflected some of the conditions typical of the lab studies (3.09 vs 3/6/9), but the range was much broader (1−19). Likewise, with two of the toy size cut-offs (>5%, >7%), the average number of toys in the foreground of the child's view matched the typical number presented on screen (4.38, 2.47 vs. 4, 2). However, the mean shift clustering analysis identified prototypical scenes with either one or two large toys with a gradient of sizes, as opposed to the n equal-sized objects typical of CSL experiments. The overall input statistics in the real world are similar to those found in the CSL studies, giving credence to their derived accounts of statistical learning. Whether learning in CSL studies could be improved by adapting the exposure to more closely match properties of the child's real-world experience (e.g., scene compositions where candidate referents are not all equally-sized and targets are more likely to be relatively big) is an open question.

Trueswell, Gleitman, and colleagues' HSP work did use real-world data. However, previous research (Yu & Smith, 2012; Bambach et al., 2018) showed differences in the visual information generated by the child's field of view compared to other views (e.g. 3rd-person views). Therefore, we specifically chose the current data set for it's child view recordings in a naturalistic play environment. The lab area was decorated to be similar to a home, with carpets, furniture, and stuffed animals. Based on the instructions, the parents and caretakers played similarly to how they would at home. Yet, since it was a controlled lab environment, researchers were able to record much more information than would be available at the participants' actual homes through the use of multiple room cams, head cams, and eye-trackers. While certain aspects of the child's everyday play context cannot be wholly replicated (i.e., it is not their home) and the presence of recording equipment may impact their behavior in some way, these data balance ecological validity with dense, high-quality, multi-modal measurements which uniquely afford the type of analysis presented here.

To better understand the robustness and generalizability of the patterns observed here, these analyses should be extended to similar datasets collected from other environments. The relationship between the timing of parent naming events and the spatial properties of the visual input, as well as whether/how it is linked causally to child attention, also remains unclear. Based on the current data, we cannot conclude whether it is the parent or the child leading the interaction to generate these regularities. For instance, it may be the case that the parent chooses to name the object which they think appears largest to the child in that moment. In contrast, parents may be more focused on what is in their own field of view and the naming instance may draw the child's attention to particular objects. Future analyses could integrate the parent's view and explore the relationship between the spatial and temporal regularities. Lastly, when designing CSL experiments, the supporting regularities can be informed by this type of analysis, such that they incorporate more variations in the naming frequency and toy sizes or scene composition, or even use the first-person videos as stimuli (Zhang, Amatuni, Cain, & Yu, 2020; Zhang et al., 2021).

Acknowledgements

This research used data that was funded by the National Institute of Child Health and Human Development grant R01HD093792 to CY.

References

Abend, O., Kwiatkowski, T., Smith, N. J., Goldwater, S., & Steedman, M. (2017). Bootstrapping language acquisition. *Cognition*, *164*, 116–143.

Amatuni, A., & Yu, C. (2020). Decoding eye movements in cross-situational word learning via tensor component analysis. , *42*(42).

Baldwin, D. A., Markman, E. M., Bill, B., Desjardins, R. N., Irwin, J. M., & Tidball, G. (1996). Infants' reliance on a social criterion for establishing word-object relations. *Child development*, *67*(6), 3135–3153.

Bambach, S., Crandall, D., Smith, L., & Yu, C. (2018). Toddler-inspired visual object learning. *Advances in Neural Information Processing Systems*, *31*, 1201–1210.

Bhat, A., Spencer, J. P., & Samuelson, L. K. (2018). A dynamic neural field model of memory, attention and cross-situational word learning. In *Cogsci*.

Fitneva, S. A., & Christiansen, M. H. (2011). Looking in the wrong direction correlates with more accurate word learning. *Cognitive Science*, *35*(2), 367–380.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2009). Frequency and Contextual Diversity Effects in Cross-Situational Word Learning. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 31, p. 7).

Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. R. (2011). How words can and cannot be learned by observation. In *Pnas proceedings of the national academy of sciences of the united states of america*.

Quine, W., & Van, O. (1960). Word and object: An inquiry into the linguistic mechanisms of objective reference.

Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv*.

Smith, L., & Yu, C. (2008, March). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568. doi: 10.1016/j.cognition.2007.06.010

Yu, C., & Smith, L. B. (2007, May). Rapid Word Learning Under Uncertainty via Cross-Situational Statistics. *Psychological Science*, *18*(5), 414–420. doi: 10.1111/j.1467-9280.2007.01915.x

Yu, C., & Smith, L. B. (2012, November). Embodied attention and word learning by toddlers. *Cognition*, *125*(2), 244–262. doi: 10.1016/j.cognition.2012.06.016

Yu, C., Zhang, Y., Slone, L. K., & Smith, L. B. (2021). The infant's view redefines the problem of referential uncertainty in early word learning. *Proceedings of the National Academy of Sciences*, *118*(52).

Zhang, Y., Amatuni, A., Cain, E., Wang, X., Crandall, D., & Yu, C. (2021). Human learners integrate visual and linguistic information cross-situational verb learning. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43).

Zhang, Y., Amatuni, A., Cain, E., & Yu, C. (2020). Seeking meaning: Examining a cross-situational solution to learn action verbs using human simulation paradigm. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 42).