UC Merced

UC Merced Previously Published Works

Title

Cross-Situational Statistics Present in an Early Language Learning Context: Evidence From Naturalistic Parent—Child Interactions

Permalink

https://escholarship.org/uc/item/3vq7v1s8

Journal

Cognitive Science, 49(6)

ISSN

0364-0213

Authors

Cain, Ellis S

Ryskin, Rachel A

Yu, Chen

Publication Date

2025-06-01

DOI

10.1111/cogs.70078

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Cross-Situational Statistics Present in an Early Language Learning Context: Evidence from Naturalistic Parent-Child Interactions

Ellis S. Cain[1], Rachel A. Ryskin[1], and Chen Yu[2]

[1] Department of Cognitive and Information Sciences, University of California, Merced
Merced, CA, United States of America

[2] Department of Psychology, Center for Perceptual Systems, University of Texas, Austin
Austin, TX, United States of America

Author Note

Correspondence concerning this article should be addressed to Ellis Cain, Department of Cognitive and Information Sciences, University of California, Merced, USA. E-mail: ecain@ucmerced.edu; Mailing address: 5200 Lake Rd, Merced, CA 95343

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

Abstract

According to the cross-situational learning account, infants aggregate statistical information from multiple parent naming events to resolve ambiguous word-referent mappings within individual naming events. While previous experimental studies have shown that infant and adult learners can build correct mappings based on statistical regularities encoded in multiple learning situations in an experiment, other studies that use more naturalistic stimuli (e.g., real-world video) reveal poor performance in adults' ability to infer the correct referent. Based on those results derived from more naturalistic stimuli, the cross-situational learning solution cannot be useful to solve the mapping problem in the real world because cross-situational statistics from the real world are much more ambiguous than those created in experimental studies. To examine the feasibility of cross-situational learning in everyday contexts, the present study aims to quantify visual-audio statistics from one of everyday activities – parent-child toy play. We analyze parent naming events in a video corpus of infant-perspective scenes during parent-child toy play in a naturalistic lab setting, where we found three distinct properties that characterize statistical regularities perceived by young learners: 1) there are a limited number of visual scene compositions perceived by young learners at the moments when they hear object names; 2) the frequencies of parent naming events are distributed in a skewed, Zipfian fashion; and 3) cross-situational statistics in naturalistic toy play are comparable to those used in laboratory experiments. Our results underscore the importance of quantifying the statistical regularities in the input from the learner's perspective in order to shed light on the mechanisms supporting early word learning.

Keywords: cross-situational word learning, naturalistic corpus study

Word count: 5753

Cross-Situational Statistics Present in an Early Language Learning Context: Evidence from Naturalistic Parent-Child Interactions

1 Introduction

One of the key challenges in word learning is that of mapping[1] words to their referents in the world. In an everyday scenario, like a child and parent playing with toys, when a parent produces an object label that is new to the child, the correct label-object mapping is not necessarily clear from the child's point of view. Many possible objects, parts of objects, events, or states could constitute what the parent is referring to (see Fig. 1A). Quine (1960) referred to this as the problem of referential uncertainty. Nonetheless, human learners, and even infants, are able to solve this problem and eventually learn correct word-referent mappings from this ambiguous input.

Examining how human learners resolve this referential uncertainty has been a focus in the literature on early word learning. Some proposed solutions rely on the ways that ambiguity is reduced within a single naming event[2]. Multiple sources of information can be brought to bear on a particular naming event to constrain the space of candidate referents in consideration. For instance, a set of cognitive biases, such as whole object bias, mutual exclusivity, taxonomy, and noun bias (Bloom et al., 1993; Markman, 1989) appear to help infants pick out the most likely referent in their environment. These biases are often thought to be learned through experience (Jara-Ettinger et al., 2022; Smith et al., 2002). Additionally, infants can infer the referential intent of the speaker in order to interpret a novel reference correctly (Baldwin et al., 1996; Bohn & Frank, 2019). For example, infants were more likely to learn a new label-object mapping when they observed an adult

---

[1] Not to be confused with *mapping* as in 'fast-mapping' from Trueswell et al. (2013), which will be discussed later.

[2] Within the scope of this paper, we will use 'naming event' to refer to an instance when a parent uses a specific word/label to refer to an object, either on its own (i.e., '*Car*!') or within an utterance ('Let's move the *car* over.'). We base this terminology on studies such as Smith and Yu (2008) and Yu et al. (2021)

producing the label while concurrently looking at the object, as opposed to only hearing the label coming from an audio speaker. Object properties might also play a role in the learning of word-object mappings, as the perceptual salience of objects has been shown to play a role in word learning (Pruden et al., 2006).

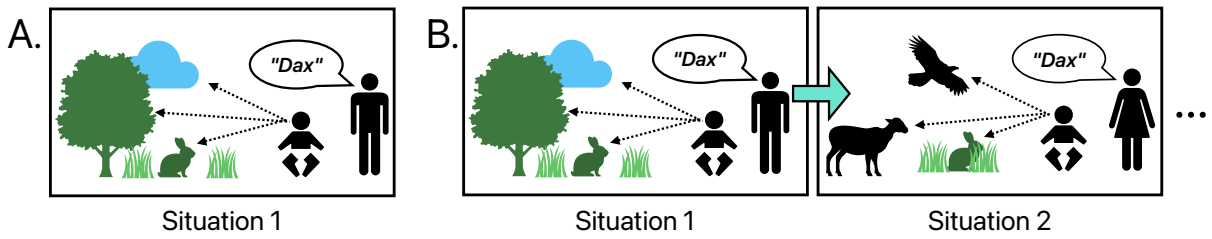


Figure 1

A) Referential ambiguity is often present within a single instance naming. In this example, there is not enough information to disambiguate what "dax" refers to. B) Given multiple exposures to learning instances in various situations, if the learner tracks the co-occurrences across each situation, they would be able to narrow down the potential referents for "dax".

Another line of work has investigated how learners reduce referential uncertainty by aggregating information over longer timescales through domain-general statistical learning mechanisms (Plunkett, 1997). Recent computational models for early language learning suggest that this type of statistical learning mechanism might underlie early word learning for nouns, whereas the cognitive biases or pragmatic factors mentioned previously would help to extend learning to different word classes and broader linguistic contexts (Meylan & Bergelson, 2022). In particular, one such mechanism for word learning is known as cross-situational learning (CSL, e.g., Fitneva & Christiansen, 2011, 2017; Smith & Yu, 2008; Trueswell et al., 2013; Yu & Smith, 2012). The logic of CSL is as follows: an infant learner encounters a to-be-learned label many times in everyday learning situations, and across those situations, the label co-occurs with the intended referent more than with other potential referents; if the infant learner tracks the co-occurrence statistics among the label

and candidate referents in each situation (i.e., what objects are in view during the naming event), regardless of the environment or the identity of the objects, then the correct label-referent pair should have the strongest aggregate statistical association relative to all the other pairings (see Fig. 1B).

To test this CSL idea, an experimental paradigm has been developed in the lab in which participants are presented with a set of ambiguous naming events – multiple auditory labels paired with multi-object visual displays – and asked to figure out which picture corresponds to each of novel words (e.g., "dax", see Figure 2A). During training, participants hear novel words/labels played in a random order, while they view a visual scene containing multiple visual objects in a random configuration (i.e., the temporal order of labels doesn't match the spatial layout of objects). Throughout the experiment, the correct label-object co-occurrences are consistent and frequent, such that whenever an object is on the screen, the label is always played (but other labels are also played). Though, within a given naming event, the mapping between label and referent is ambiguous (i.e., all mappings are equally likely), across naming events, the association between the label and correct referent is stronger than all other associations for that label. The cross-situational co-occurrence statistics have been argued to offer a solution to early word learning.

Empirical evidence from the CSL paradigm described above shows that adults, children, and even infants, succeed at learning correct label-object mappings at above chance levels across a variety of task configurations (e.g., differing numbers of objects per trial, repetitions, etc.; Fitneva & Christiansen, 2011; Smith & Yu, 2008; Yu & Smith, 2007). Moreover, a range of computational models have been developed to simulate this kind of learning (Amatuni & Yu, 2020; Bambach et al., 2018; Bhat et al., 2022; Frank et al., 2009; Kachergis et al., 2017; Stevens et al., 2017; Vogt, 2012).

1.1 Word learning in the wild

CSL assumes that infants are continuously tracking multiple word-referent associations between audio and visual stimuli, which, though trivial for a computational model, may be challenging for infants. In particular, it is unclear that the strong co-occurring regularities encoded in the well-controlled experimental paradigm are also evidenced in the real world outside of the lab.

To test CSL in more naturalistic situations, Medina et al. (2011) used third-person videos of naming events extracted from parent-child social interaction. Participants were shown 40-second short vignettes where parents uttered target words to their infants. The videos were muted except for a beep or nonce word where the target word was uttered. With these muted videos, the information available to the participants is restricted to just *when* the target word is uttered and the accompanying visual information, allowing the participants to 'simulate' being an infant learner, who wouldn't necessarily be able to utilize the surrounding linguistic context to guess the target word. Similar to the original cross-situational learning experiments, participants were asked to identify the correct referent of each word after training. Participants in this "Human Simulation Paradigm" (HSP) task performed poorly: learning accuracy was below 33% for 90% of the vignettes shown. Medina et al. concluded that there is not enough structured information in the real world to support CSL as a useful mechanism for word learning. Instead, infants in the real world have to rely on "fast mapping" of labels to referents (Trueswell et al., 2013), learning a particular mapping from a single naming event, buttressed by innate biases or pragmatics (e.g., Frank et al., 2009), and then potentially revising that mapping if it conflicts with downstream input.

In most CSL experiments, learners perceive two to four objects simultaneously and hear the corresponding words played sequentially. The results based on HSP suggested that the degrees of uncertainty in those experiments are much lower than those encountered by young learners in the real world. However, the HSP results used to support

this argument are from adult judgments of child video to estimate the degrees of referential uncertainty that young children encounter in early world learning. While the HSP approach is useful as a way to estimate child learning environments, it is an indirect way to assess referential uncertainty in the real world. Several recent studies reveal statistical regularities in the real world differ from the results based on HSP and also from those encoded in previous CSL experiments. On the visual side, studies using head-mounted cameras and head-mounted eye trackers show that the infant's view is selective, with the named object more likely to be in the center of the first-person view or generally more salient, making the referent of a novel label less uncertain (Yu et al., 2021; Yurovsky et al., 2013). This is different from the CSL experiments wherein candidate objects to a target word have an equal size. In contrast, children and caregivers in the real world typically interact with the objects and bring objects of interest close in view (see Fig. 2B). The manual actions on objects often create a non-uniform distribution of object sizes in view, with the held objects larger than others. Thus, the child's view is more structured and less ambiguous than previously thought.

On the language side, recent analyses of child-directed speech also show that children do not hear large numbers of novel words with equal frequency and at a fixed interval. Studies using natural language corpora have found that a small number of high-frequency words account for a large portion of tokens in text and speech (Piantadosi, 2014; G. K. Zipf, 1949; G. Zipf, 1935). This "Zipfian" distribution is characterized by a power law relationship between the log of word frequency and the log of the word rank. Recent CSL studies indicate that adults and children learn more quickly when exposed to Zipfian label-object frequency distributions relative to the uniform distributions that characterized earlier studies (Hendrickson & Perfors, 2019; Lavi-Rotbain & Arnon, 2022; Wolters et al., 2024).

In summary, there are contradicting views about real-world learning and the lab-based CSL experiment. According to one view, the real world is highly ambiguous, in

that the visual learning input is messy. In contrast, another view sees the real world as containing more structured learning input, whether through the (restricted) views generated by the child's interactions with objects or through the distributional structure of the language input. It remains an open question how the statistical regularities of the real world compare to the lab-based CSL experiments.

1.2 The present research

To address contradicting views on statistical regularities for cross-situational learning, we propose a direct way to measure referential uncertainty from child egocentric video. The goals of the present study are to analyze (semi-)naturalistic naming events and object presentation from child-perspective scenes[3] embedded in toy-play contexts to determine: 1) the scene composition from the child's perspective, 2) distributional properties of these naming events, and 3) whether the cross-situational learning statistics were present in their learning input. We will only be analyzing *noun* naming events, when a parent uses a specific label to refer to an object, which could be either an explicit naming event (i.e., '*Car*!') or embedded within an utterance (i.e., 'Can you move the *car* over here?'). In our analyses, we would be treating both types as naming events for *car*.

We chose toy play because it is one of everyday activities through which young children learn object-word mappings. In the present study, we first quantify the visual properties of the infant-perceived scenes that correspond to naming events. We then characterize the distributional properties of the learning input (i.e., parent naming, object appearance frequency). Lastly, we quantify the co-occurrence statistics present in this learning context and compare different ways to aggregate statistical information to build object-word mappings. If experimental tasks and assumptions about early word learning violate the properties of the audio-visual context of infant learning experiences in the real world and artificially increase statistical regularities in lab learning tasks, then the CSL mechanism, in its current form, may not be sufficient to explain how infants learn

[3] 'Scene' refers to the visual input that is paired with a naming event.

word-referent mappings in the real world. On the other hand, if there is sufficient structure in the child's first-person input to extract strong object-word associations, such results would lend further support to a CSL-based mechanism underlying early word learning.

2 Methods

2.1 Data collection

The data used in the present study were previously collected from free-flowing parent-child dyadic toy play sessions. Children ($n = 36$, $age = 15 - 25$ months old) were recruited from the local community in Monroe County, Indiana (see Yu et al., 2021 for more details). 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 3-left shows a third-person view of the study setup. At the beginning of the study, 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. The same set of 24 toys was used for all parent-child dyads. 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 3-right.

2.2 Data annotation

The YOLO object detection model (Redmon & Farhadi, 2018) was used to process child-perspective scenes to detect the presence of toy objects in the scenes. This automatic data annotation is accomplished by re-training the pre-trained YOLO model with a set of object instances of the 24 toys. Coders hand-annotated over 50,000 bounding boxes in total by hand (roughly 2000 instances per toy) for this re-training. After re-training, the algorithm automatically detected toy objects in view (with a post-training accuracy of 87%[4]), providing the sizes and locations of up to twenty-four coordinate sets per frame (x,

[4] The YOLO model performed well with the limited amount of re-training data. Alternative models like Mask R-CNN (He et al., 2017) could reach higher accuracy with the added benefit of fine-grained object masks, but would require much more hand-annotated re-training data.

y, x-length, y-length) that can be used to draw "bounding boxes" around all of the visible toys[5].

Parent speech was manually annotated at the utterance level using Audacity. There was no minimum utterance length for annotation, but utterances less than 400 ms apart were collapsed into a single utterance. In the present study, we focused on the utterances of parent speech containing object names. The object detection data and parent naming utterances were all synchronized, which allowed us to measure the composition of child-perspective scenes at the moments of parent speech. The current study includes data from 106,200 frames of egocentric videos, corresponding to 1,475 parent naming events. For each naming event, object detection data was extracted and averaged within a 3s window starting at the onset of a given naming event. This temporal window is chosen to include both the utterance itself (roughly 1 to 1.5 seconds) and also a brief period of time after the naming event. The scripts can be found at the following OSF repository: https://osf.io/bd2cs/?view_only=0d5088ae31054daab14a11f8df983d80

3 Results

3.1 Visual properties of child-perspective scenes at naming moments

To quantify the visual scene composition during naming events, YOLO object detection was used to generate bounding boxes for the toys in each frame. These bounding boxes consist of x-y coordinates with length and width in pixels, such that they indicate the area occupied by the toy in the child's view, based on the head-mounted camera. The object size data was averaged over the 3s window of data extraction. 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 $0 - 14$).

[5] These bounding boxes capture the *absolute* size of the objects in the child's field of view. The *relative* object size could be an alternative method, which could be useful for addressing environments where the objects of interest have a large range of sizes.

Figure 4 shows the overall distribution of the sizes of visible toys in a naming event, when sorted by size. Only a few toys occupy a relatively large portion of the child's field of view. The largest toy for a given scene, on average, occupies 15.11% of the child's field of view (range $0.45\% - 67.03\%$), and the second largest toy occupies 9.85% on average (range $0\% - 52.98\%$), while the sixth largest toy only occupies 4.38% on average (range $0\% - 12.12\%$).

We then used mean shift clustering to identify groups of similar scenes in the data (Comaniciu & Meer, 2002). Here, scenes were represented as a vector of the top six[6] 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 $[0.25, 0, 0, 0, 0, 0]$. These 6-object scene vectors on average cover 63.6% of the child's field of view (range $35\% - 100\%$; $SD = 12.6\%$). There were a total of 23 clusters identified. Five clusters comprised greater than 30 scenes (range $1194 - 31$). The remaining 18 clusters comprised 5 or fewer scenes. We chose to focus on the clusters that had more than 30 data points, as the remaining clusters represent largely idiosyncratic scenes. For these 5 clusters, Fig. 6 shows the centroid of each cluster along with an example scene to illustrate the visual composition of the scenes.

Cluster 1 is the most frequent ($n = 1194$ scenes) and represents scenes in which several toys are approximately equally sized, and none takes 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%) with 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-sized object and the rest in the

[6] This number was chosen to capture the general trends in toy sizes per scene. If there were too many toys included in the scene vectors, then the trends might be too specific or unable to generalize. If there were too few, then it wouldn't be able to capture the variation in number of toys per scene. Six was chosen to balance the two extremes.

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.

In summary, the majority of child-perspective scenes show a skewed distribution of toy sizes with only a few toys occupy a sizable portion of view. Moreover, the clustering results identified five common scene compositions of child-perspective scenes at naming moments. Those common scene compositions vary in the number of toys that are large in view and the total number of visible toys in the background.

3.2 Distributional properties of parent naming

We measured and tracked naming frequencies across the toys and parent-child dyads. There were 40.97 naming events on average per parent-child dyad (range $7 - 111$), with a mean of 3.09 naming events for each toy (range $1 - 19$) per parent. However, across parents, we see variation in the most frequently named toy. For example, the helmet was the most frequently named toy for one parent, whereas the rose was the most frequently named for another parent (see Fig. 7).

As seen in Fig. 7, naming frequency has a right-skewed distribution for all parents. Individually, each parent also has a right-skewed distribution, but the toy identities vary for each rank (i.e., one parent named 'helmet' the most, whereas another named 'rose' the most). For the following analyses of naming distributions, we calculated the average frequency for each rank, regardless of the toy identity. Based on a chi-square test comparing naming frequency (averaging across all parents) to a uniform distribution, the naming frequency distribution was sufficiently different from a uniform distribution ($X^2(361, N = 22) = 418, p < 0.05$).

Figure 8 shows the overall naming distribution plotted on a log-log scale, where the x-axis is the log of the rank and the y-axis is the log of the frequency. To examine how the naming frequencies were distributed in this naturalistic corpus, we followed the analyses of

Zipfian distributions in child-directed speech from Lavi-Rotbain and Arnon (2023). Namely, we fit the equation $f(r) \propto \frac{1}{(r+\beta)^{\alpha}}$ to the parent naming frequency distribution to estimate the parameters of the distribution. In the equation, toy rank in the distribution is denoted by r and frequency of the corresponding rank is $f(r)$. For the parameters that are estimated, α is the exponent of the power law and β is the correction. As with Lavi-Rotbain and Arnon (2023), we did not use the binomial distribution split method mentioned in Piantadosi (2014) due to our corpus' greatly reduced size. We used the maximum likelihood estimator to estimate the α and β parameters for our distribution, using the code implementation from Lavi-Rotbain and Arnon (2023). Lastly, to evaluate the model fit, we calculated the Pearson correlation between the original frequency distribution in the dataset and the predicted frequency distribution, using the fit parameters.

Even with such a limited frequency distribution (maximum of 22 named toys per parent), we were able to find parameter estimates that had a high correlation with the original frequency distribution ($R = 0.987$). While the α and β parameters are quite large (37.15 and 196.9 respectively), they are comparable to values estimated from child-directed speech corpora in Lavi-Rotbain and Arnon (2023), suggesting that the distribution is indeed Zipfian.

3.3 Cross-situational Statistics

Building object-name mappings in everyday contexts, like toy play, requires linking heard names with seen objects. Here we quantified cross-situational statistics by measuring word-object co-occurrences across multiple naming events. To do so, we created association matrices that accumulated co-occurrence frequencies of parent naming events with visual objects present in the child's field of view. To capture different ways to cross-situational statistics that may be accumulated over time, we developed three different ways to build association matrices: 1) basic object-name co-occurrence frequencies, 2) co-occurrence frequencies only for objects exceeding a certain size threshold ($area > 2\%$), and 3)

co-occurrences weighted by the object's size in the field of view. For the basic object-name co-occurrence matrix (1), we computed the aggregate label-object co-occurrences across subjects for each label and toy pairing. For the threshold version (2), we only computed the co-occurrences if the visual objects exceeded the threshold. For the weighted version (3), the co-occurrences were weighted by the object's size in the child's field of view. In other words, instead of adding 1 for every co-occurrence (as in the basic version), the increment was multiplied by the object size (i.e., increment $= 1 * 0.02 = 0.02$). Then, all the matrices were normalized by naming frequency to focus on the proportional difference in naming frequency. Fig. 9 shows an example schematic of how these were calculated.

3.3.1 Overall Statistics from all naming events

Figure 10 shows the three versions of association matrices. The diagonal values of these association matrices correspond to the correct word-object mappings, while the off-diagonal values correspond to spurious incorrect mappings. For the version of basic co-occurrences (Fig. 10A), the diagonal elements are more strongly associated than the off-diagonal co-occurrences. When the object size threshold ($area > 2\%$) is applied to the co-occurrences (Fig. 10B), this seems to strengthen the associations on the diagonal, consistent with the fact that parents tend to name the larger objects in view ($\rho = 0.419$). When the co-occurrences are weighted by the object size (Fig. 10C), this reduces the overall associative strength for the majority of the object-name mappings, however, the correct label-object mapping remains the strongest for a few objects. To quantify the above observation, we calculated the average associative strengths, as seen in Figure 11. On average, the correct statistics were stronger than the spurious co-occurrences. Based on a paired t-test, the mean of the correct object-name co-occurrences was statistically higher than the spurious co-occurrences ($p < 0.05$) across all three versions. Moreover, we also trained a binomial logistic regression to test if the associative strengths predict whether a label-object pairing is correct (on a per-item, per-subject level). We found that the associative strengths, in each of the the three versions, predict the correct label-object

pairing (Basic: $b = 5.19$, $SE = 1.28$, $p < 0.005$; Threshold: $b = 8.96$, $SE = 1.34$, $p < 0.005$; Weighted: $b = 280.5$, $SE = 13.8$, $p < 0.005$).

Overall, the weighted version seems to be the most effective way to build strong associations between seen objects and heard words. Thus, when hearing an object name, if a child learner pays attention to candidate objects relative to their size (i.e., larger in view), then the aggregated cross-situational statistics offer the best data for learning, compared with the versions of considering all visual objects equally, and the version of ignoring objects below a certain threshold.

In summary, with different ways to aggregate object-name co-occurrences across multiple naming events, cross-situational statistics are consistently present in the learning input, evidenced by stronger correct object-name associations over incorrect associations.

3.4 Cross-situational statistics within different visual scenes

In the above sections, the scene clustering analysis identified distinct scene compositions at naming moments. Given that two of the three versions of calculating cross-situational statistics are based on individual object sizes in view (i.e., thresholding and weighting by object size), the version that worked best at the aggregate level might not necessarily transfer to the different scene compositions. Accordingly, here we calculated cross-situational statistics in different types of visual scenes. Some of the common scene compositions may contain stronger cross-situational statistics than others. Alternatively, cross-situational statistics may be similar across different scenes.

We generated co-occurrence matrices for each of the cluster data subsets and then calculated the average associative strengths for correct and incorrect object-name mappings. Figure 12 shows the cross-situational statistics (average associative strength, etc.) in each of the clusters with the three versions of aggregation. We repeated the logistic model analysis to see if the associative strength predicts whether an object-name mapping is correct. Namely, we trained fifteen logistic regression models to predict whether a label-object mapping was correct based on a given associative strength for each aggregation

type and cluster subset. We found that for the main scene type, the correct object-name mappings are stronger than the spurious co-occurrences in all of the three versions ($p < 0.005$). For the rest of the less-frequent scene clusters, only the weighed version creates stronger association strengths of correct object-name mappings than those of incorrect mappings ($p < 0.005$) while the frequency and threshold versions could not distinguish between correct and incorrect mappings ($p > 0.05$). Table 1 shows the model results.

This is most likely due to the skewed object size distributions in the smaller scene clusters. Thus, when a visual scene in a naming event is composed of a small of objects with a similar size, any of the three ways to aggregate cross-situational statistics would lead to successful learning. However, when a visual scene is composed of a dominant object among others, the learning mechanism that aggregates co-occurrence statistics with a bias towards larger objects is the only one that would lead to successful learning.

Table 1

Estimates from each logistic model for predicting the correctness of mapping (T/F) based on associative strength values, for each calculation type and cluster.

	Basic			Threshold			Weighted by size		
Cluster	β	SE	P-value	β	SE	P-value	β	SE	P-value
1	**4.68**	1.50	1.75e-3	**8.32**	1.56	9.12e-8	**351.8**	18.4	2.33e-81
2	28.2	20.7	0.17	30.4	20.3	0.13	**664.6**	89.9	1.46e-13
3	8.10	31.5	0.80	2.70	32.1	0.99	**589.3**	84.7	3.47e-12
4	-5.98	26.2	0.82	-7.27	27.1	0.79	**578.3**	91.7	2.85e-10
5	13.1	35.8	0.71	19.2	36.5	0.60	**1143.9**	155.0	1.62e-13

4 Discussion

In the current study, we quantified the statistical regularities of parent naming in naturalistic toy play data in three ways: 1) visual properties: we demonstrated that the

majority of naming events are accompanied by visual scenes composed of a small handful of similar-size toys; 2) parent naming: we observed that parent naming events appear to be distributed in a Zipfian fashion within each parent-child dyad (i.e., some toy names are produced much more frequently than others); 3) cross-situational statistics: we found that, across multiple methods of aggregating co-occurrences of visual data and naming events, cross-situational statistics can discriminate between correct and incorrect word-referent mappings. These findings serve as first steps to connect laboratory cross-situational experiments with the learning input in the real world and provide new evidence to support cross-situational learning as a learning mechanism that underlies early word learning.

Our clustering analysis identified five common scene compositions, which the most frequent cluster likely corresponds to general play when the child is looking around and not directly manipulating objects, whereas the other clusters likely correspond to when the child is directly manipulating a specific toy (or two). These five are not exhaustive of all possible types of scene compositions, but they generally capture potential toy-play scenes, such as a scene where one object takes up a majority of the child's view with others in the background.

4.1 Comparison to lab experiments

In most cross-situational learning experiments, the visual scenes displayed on a computer screen are composed of a small number of visual objects that are equal in size and clearly separated in space (2-4 objects per training trial). The naming and appearance of toys were linked, since label-object co-occurrences were guaranteed, and both had uniform frequency distributions. Most importantly, the correct label-object co-occurrences had the highest associative strength, such that statistical learners could identify the correct mapping based on the highest associative strength.

In this naturalistic toy-play dataset, we found that the toy sizes had a right-skewed Zipfian distribution and there were a variety of scene compositions (i.e., overall organization, number of toys in view). The most common scene consisted of multiple toys

taking up similar proportions of the child's view, whereas the other common scene compositions usually had one or two toys taking up the majority of their view.

Unlike in many previous CSL experiments, in the current dataset, object names were not uniformly distributed. Rather, a small number of toys were named much more frequently than others for many of the parent-child dyads, in a Zipfian manner. There are some caveats about the Zipfian nature of the naming distribution: we focused only on one part-of-speech (nouns) during naming events, and the number of tokens was greatly limited compared to a typical corpus of adult or child-directed speech (e.g., Lavi-Rotbain & Arnon, 2023; Piantadosi, 2014). Parents had a relatively restricted vocabulary in the task (i.e., they were given specific, singular names for each of the 24 toys). Even with such a limited corpus, we were able to fit the Zipfian equation to our data. The α parameter is higher for more skewed distributions, which suggests that our limited distribution is extremely skewed, with the most named toy being named much more frequently than the second-most named toy. The β parameter is much more variable across corpora (Lavi-Rotbain & Arnon, 2023), but both of our parameter estimates ($\alpha = 37.15; \beta = 196.9$) were within the range of other analyses ($0.9 < \alpha < 130; -0.5 < \beta < 477$). It appears that, even over the course of a relatively short parent-child interaction with a restricted vocabulary, the language input is structured in a way that benefits learning.

Most importantly, as in lab CSL experiments, we found that there was a cross-situational learning signal present in the label-object co-occurrences of these naturalistic parent-child interactions, and that the correct associations were stronger than the incorrect, spurious co-occurrences. However, the strength of the signal depends on the visuo-spatial properties of the scene: for scenes with multiple objects of a similar size, the CSL signal was present no matter how associations were computed. For the scenes with skewed object size distributions, the CSL was only present when the associations were weighted by object size. This suggests that, when learners are in an environment where they hear a novel label and are faced with objects of different sizes, it is advantageous for

their word learning if they infer that the label is more likely to refer to the larger objects than the smaller ones. Whether children do in fact adopt such a bias is an open question.

While some aspects of the distributional patterns in these naturalistic parent-child interactions differed from those in the lab experiments, the presence of the cross-situational learning signal gives credence to CSL as a viable mechanism for word learning. Previous studies have demonstrated that Zipfian distributions in language learning input aid in learning (Lavi-Rotbain & Arnon, 2022; Wolters et al., 2024), but 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.

Moreover, toy play is just one potential context for early language learning. Toy play generally happens between a caretaker-child dyad, but it could also occur alone or with other children who are also in the process of acquiring language, and therefore might not intrinsically provide 'useful' statistical regularities. Toy play itself might be more common in western industrialized cultures where a caretaker is frequently present and able to engage in play with the children (Henrich et al., 2010). Other contexts may contain different statistical regularities; storybook reading often has more naming events from the parents (i.e., Zhang & Yu, 2022) and has more structured interactions.

4.2 Cross-situational learning "in the wild"

Using the Human Simulation Paradigm, which uses scenes from real-world naming events, Medina et al. argued that CSL may not be a viable mechanism for word learning in naturalistic environments. However, the visual properties of the child's field of view differ from those of other views (e.g. 3rd-person views as in HSP vs unique 1st-person views generated by children; Bambach et al., 2018; Suarez-Rivera et al., 2022; Yu & Smith, 2012). Therefore, the current data set investigates CSL using child view recordings in a (semi-)naturalistic play environment. The lab area was decorated to be similar to a home, with carpets, furniture, and stuffed animals. Parents and caretakers were instructed to play

similarly to how they would at home and allow the child to roam freely around the play area. It is possible that the different setting and recording equipment could influence how the dyads play, but this would be similar to them playing in a new environment for the first time; for example, the dyad might go to a new public library to play, which wouldn't necessarily reflect their home environment, but many aspects of their toy play and the caretaker's language usage statistics should remain consistent.

Yet, since the current study was conducted in a consistent lab environment, it affords the ability 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, and the ability to aggregate data across dyads. 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 the extent that the current data set captures naturalistic parent-child interactions, these results indicate that CSL may support word learning "in the wild."

4.3 Extensions of the current analysis

To better understand the robustness and generalizability of the patterns observed here, these analyses should be extended to similar datasets collected from other environments, during other kinds of interactions (e.g., reading). The relationship between the timing of parent naming events and the spatial properties of the visual input, as well as whether/how parent naming 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.

Our analysis also did not distinguish between the different types of naming events, whether the noun was embedded within an utterance (without clear pragmatic labeling intent; i.e., 'Let's move the *car* over.') or as a labeling event with pragmatic purpose (i.e., 'This is a *car*' or just '*Car*.'). This is not an exhaustive list of the types of naming events, but future studies could analyze the different types of naming events and explore how they could be differentially useful for learning. Related to the point above about the parent's view, the naming events with pragmatic purpose for labeling might potentially have different co-occurrence statistics or visuo-spatial properties than regular utterances (lacking the pragmatic labeling purpose).

In our analysis of the cross-situational statistics generated by these toy play interactions, object size was shown to be a potential factor that could strengthen the correct mapping. Indeed, Pruden et al. (2006) found that perceptual salience to be an important factor for learning word-object mapping in infants. Other work has suggested that semantic features might play a larger role in attention than object salience (i.e., Henderson et al., 2009, 2019), though it is unclear to what extent these findings would extend to infants who are in the process of learning language and semantic features. Future studies, therefore, could explore other visual features such as luminance, edge orientation, or color, in addition to object size (Wolfe & Horowitz, 2017), to see whether object salience predicts infant attention and if it is another signal that could improve the strength of the correct mapping in the cross-situational statistics.

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, similar to experiments that have used first-person videos as stimuli (Zhang et al., 2020) or those that use Zipfian distributions in the stimuli (Hendrickson & Perfors, 2019).

5 Conclusion

In conclusion, we analyzed the statistical regularities of a parent-child toy play in a naturalistic lab setting and found that while the distributional properties slightly differ from those found in the lab-based cross-situational learning studies, there was a cross-situational learning signal present in the learning input that could be used to identify the correct label-object mapping. This findings support cross-situational learning, in that there is enough of a signal in a quasi-naturalistic environment for CSL to be a valid learning mechanism. Future studies can explore other learning environments to see if these transfer to other contexts.

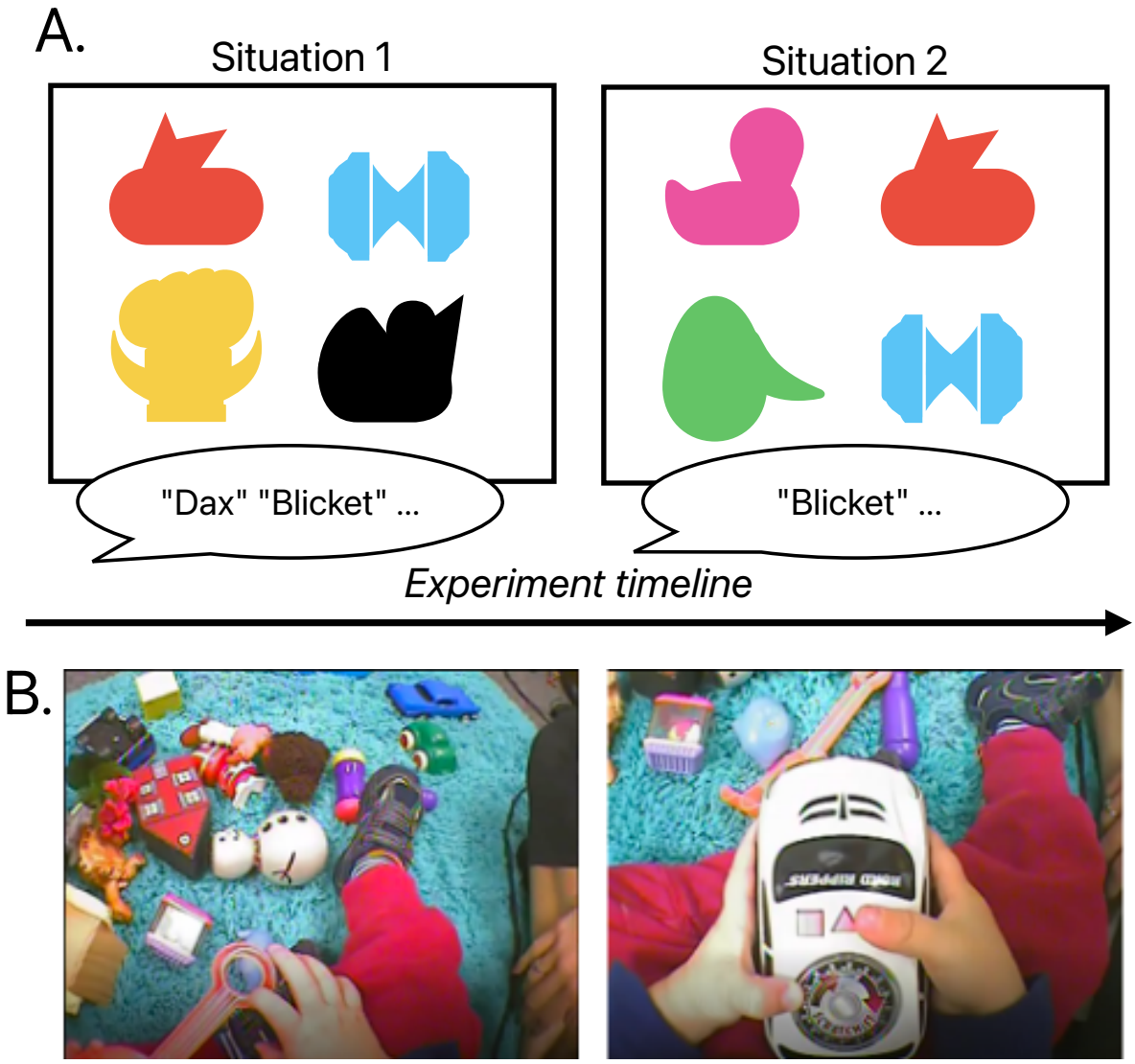


Figure 2

A) Example diagram of CSL experiment. Each trial consists of a set of novel objects, with the spoken labels presented in no particular order. Throughout the course of the experiment, the participant is proposed to track which objects co-occur with the novel labels, such as "blicket" in this example. B) Actual first-person views generated by a child during toy play.

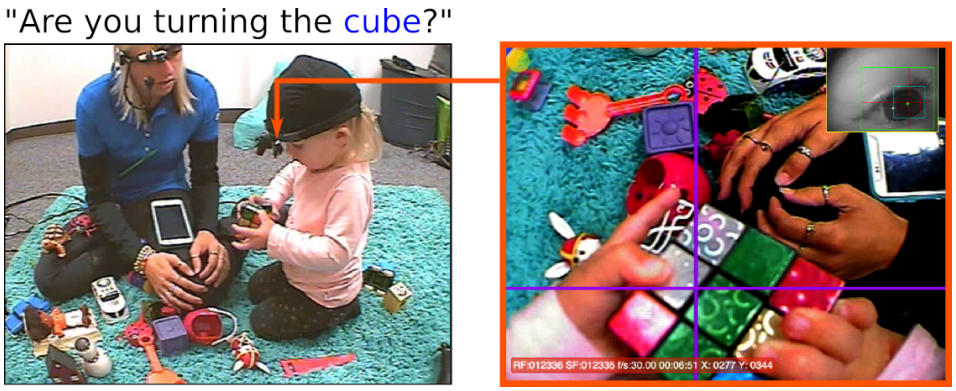


Figure 3

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

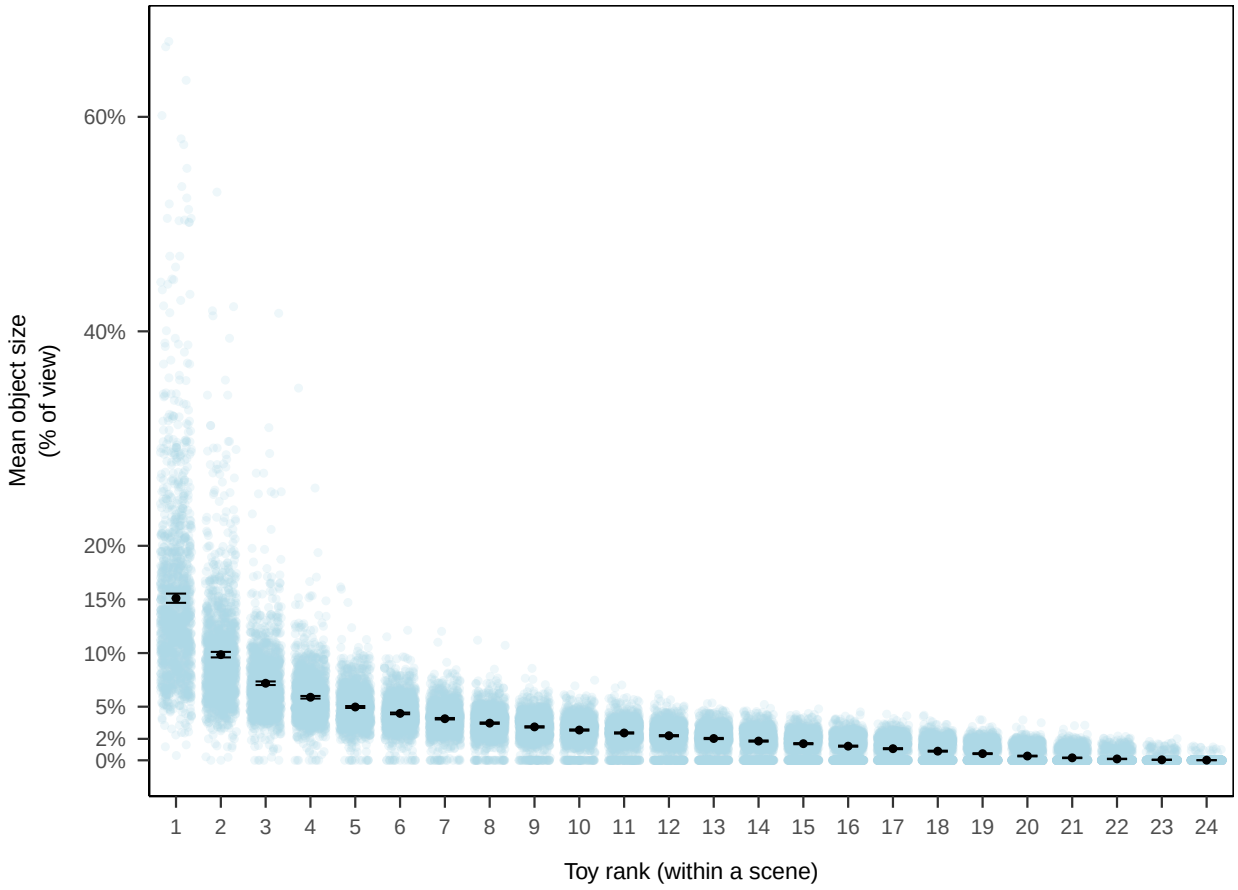


Figure 4

Mean toy sizes during a given naming event, sorted by size ranks. Error bars represent 95% *confidence interval.*

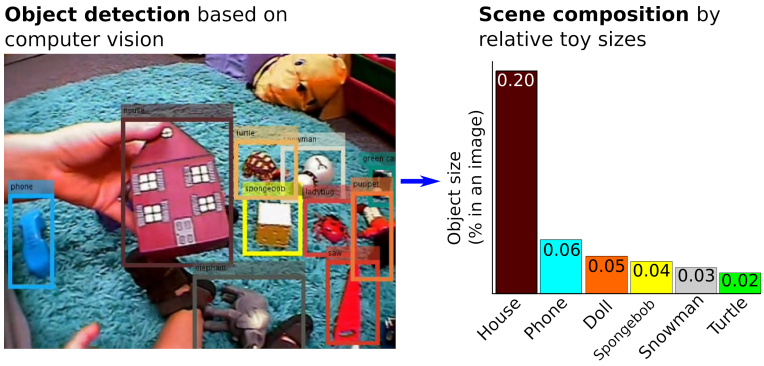


Figure 5

Example of the object detection data extracted from the child's first-person view footage (right) and the corresponding scene vector derived from that scene (left), which includes the relative toy sizes of the largest six toys.

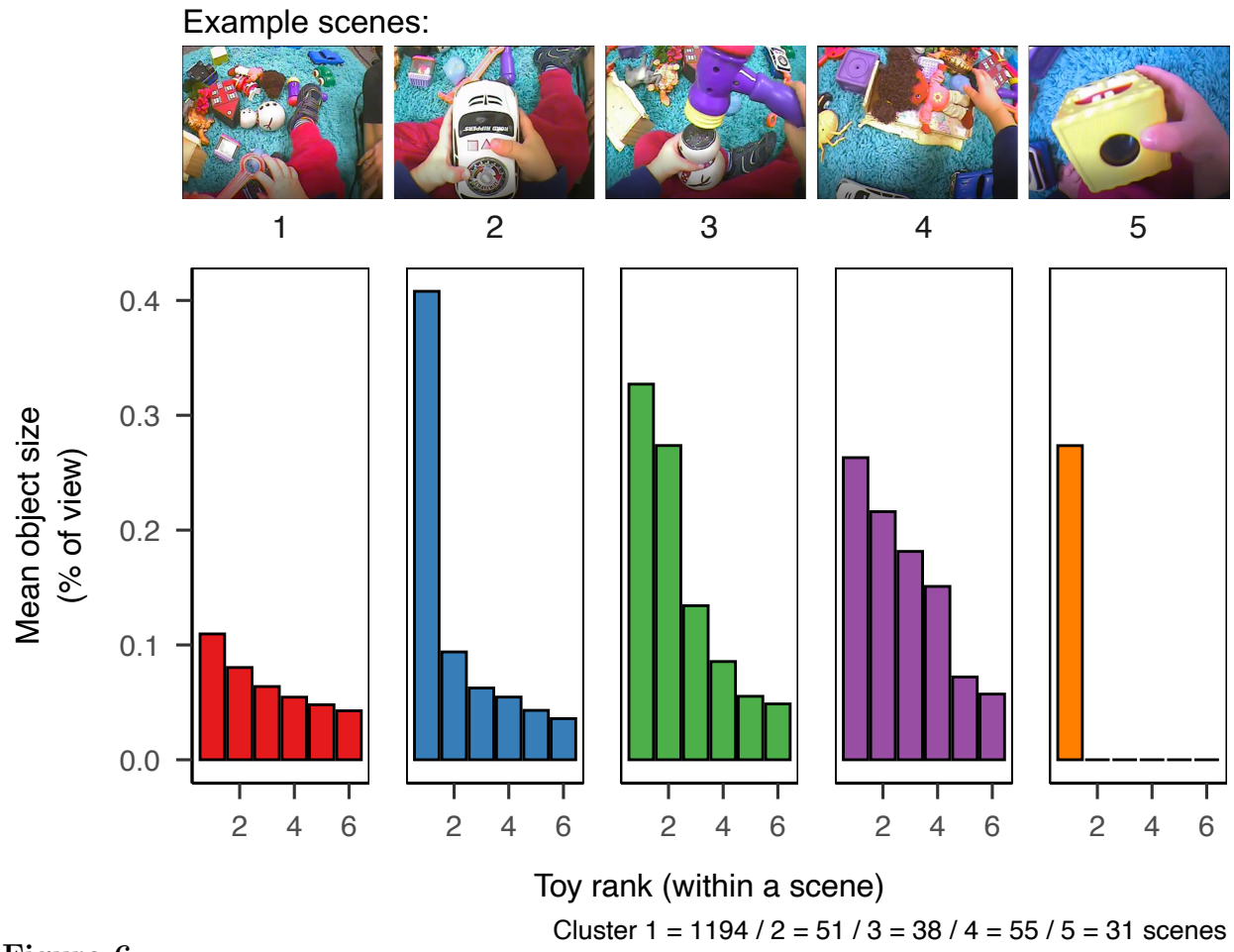


Figure 6

Five common scene compositions as identified using mean shift clustering. Each scene composition is characterized by the top six largest toys ranked by size. Scene counts for each cluster are in the bottom right. Example scenes from each of the scene clusters appear above the respective graph.

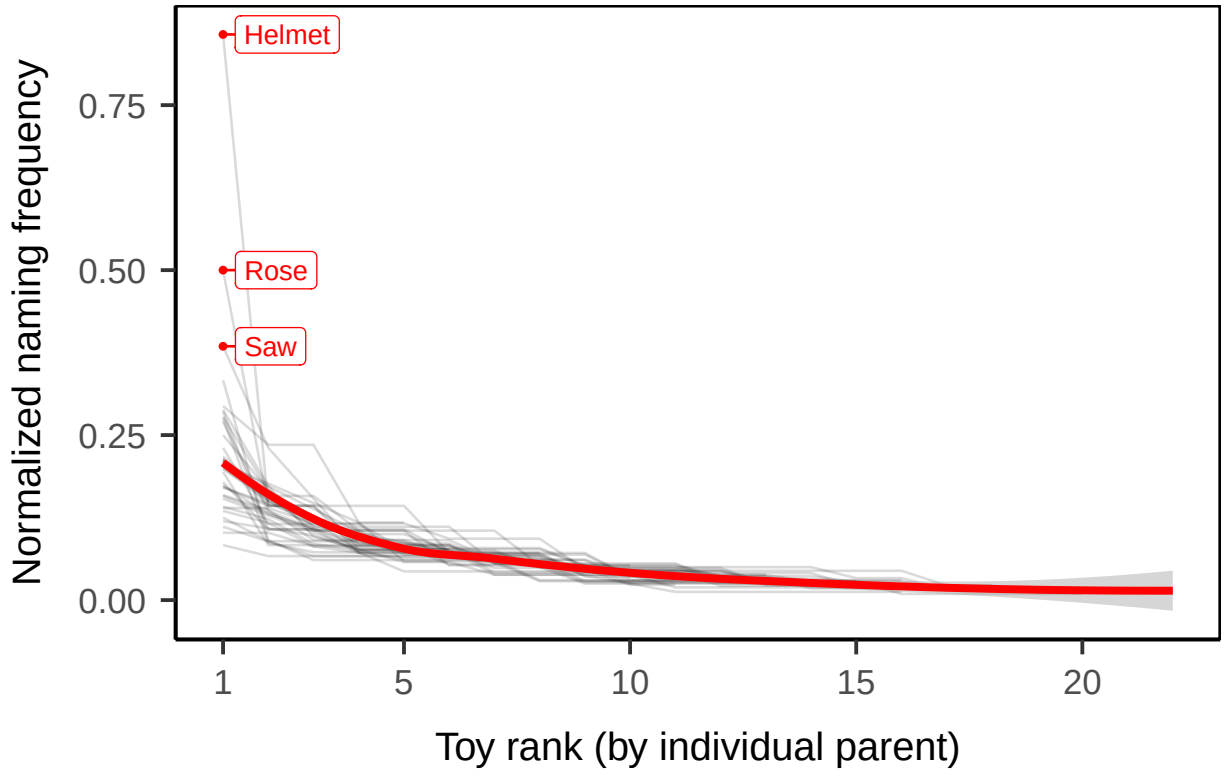


Figure 7

Frequency distribution of parent naming events, where each black line represents an individual parent. The x-axis is sorted by the frequency rank (regardless of toy identity). Frequency has been normalized for each parent, to correct for variation in overall naming frequency. The most frequently named toy has been labeled for a few parents. The LOESS fit has been added in red, highlighting the general trend across parents.

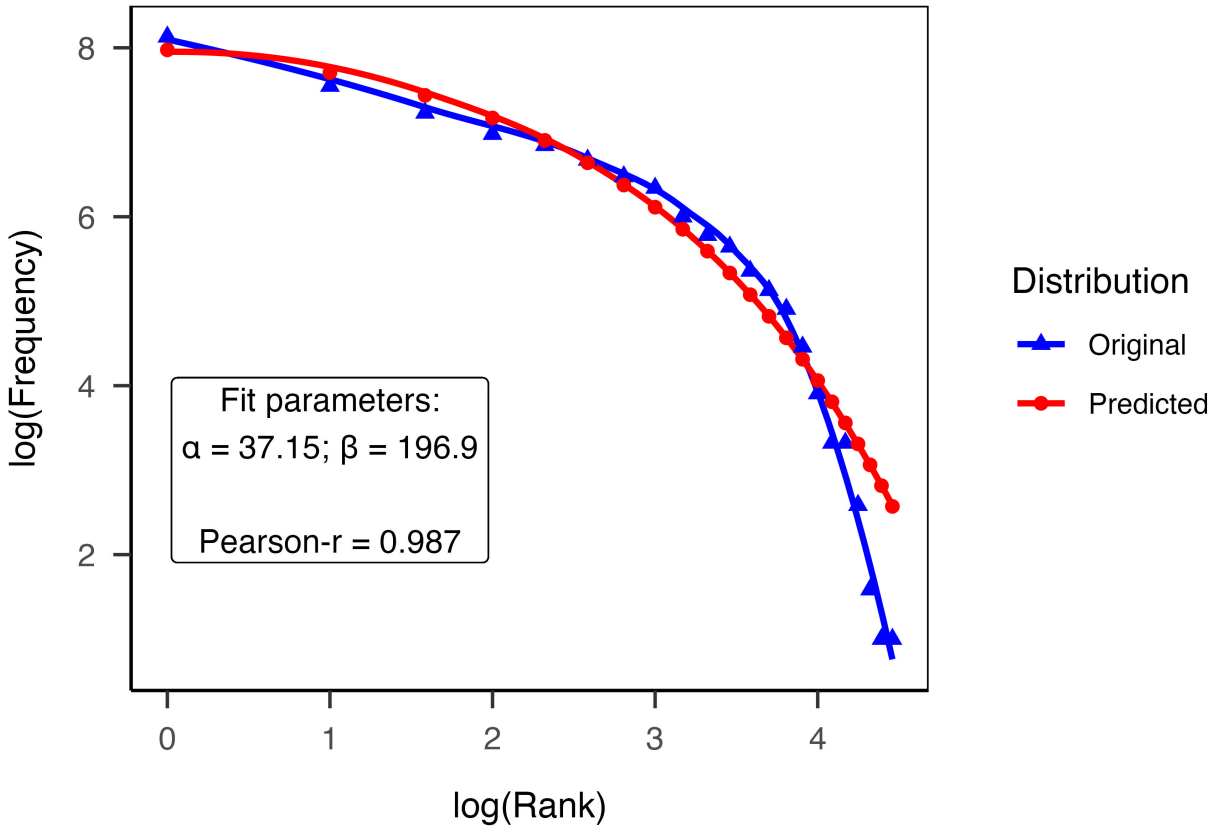


Figure 8

Log-log transform plot for the naming frequency distribution (averaged across participants). Blue represents the distribution from the original data, and red represents the Zipfian model fit based on the parameter estimates from the data. The corresponding fit parameters and Pearson-r correlation coefficient appear in the text box.

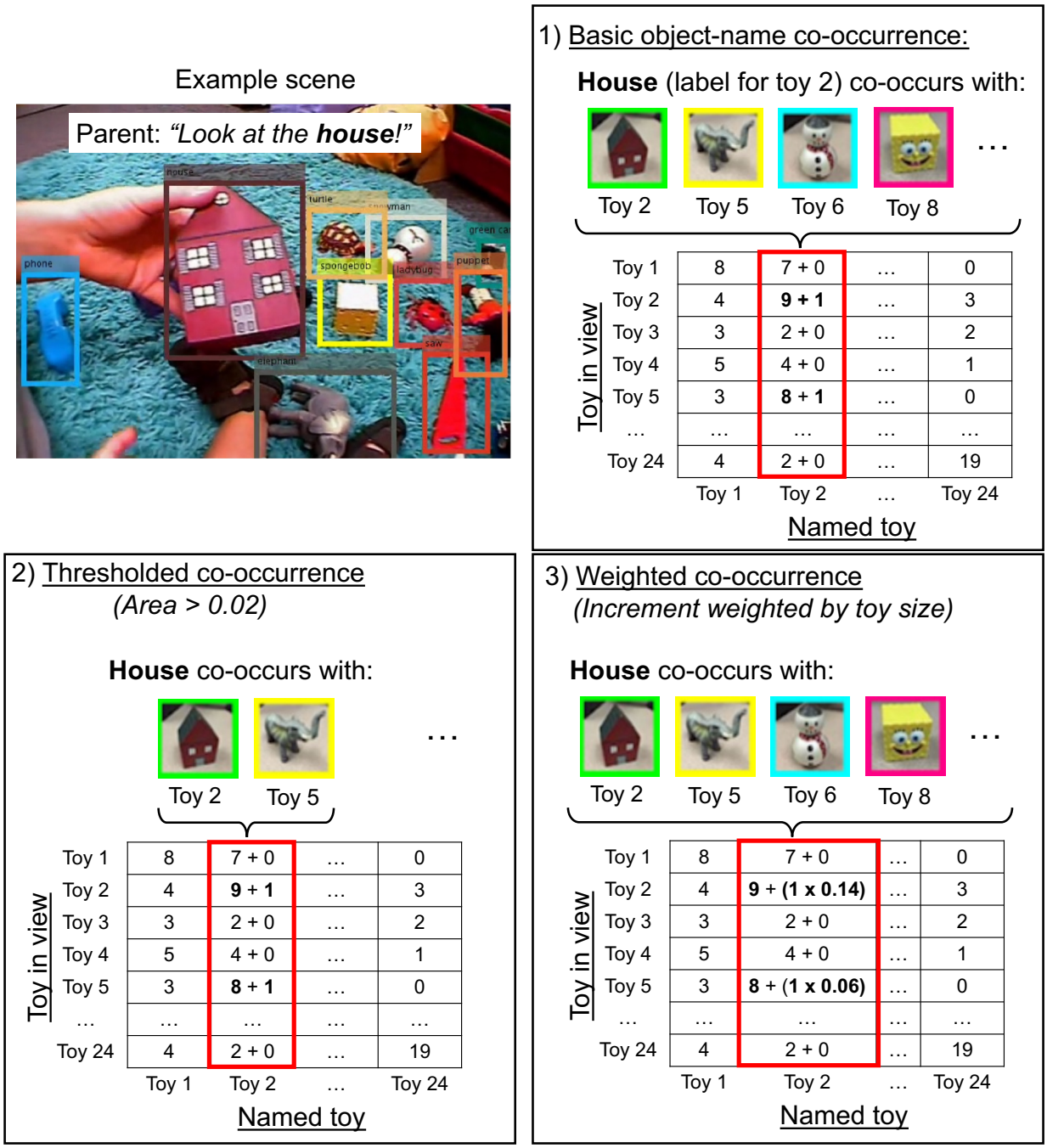


Figure 9

Schematic showing the different calculation types for creating the association matrices: 1) basic label-object co-occurrence frequencies, 2) object size threshold, and 3) weighted by size co-occurrence frequencies.

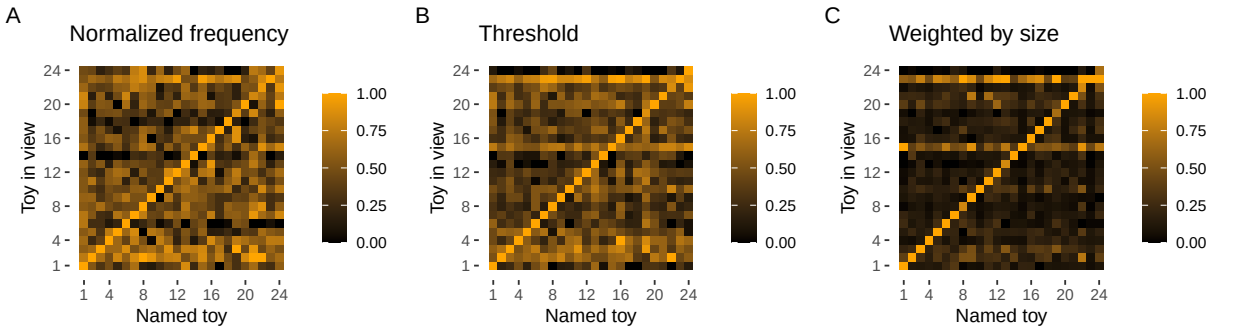


Figure 10

Associative matrices generated from the label-object co-occurrences. The three versions include A) normalized frequencies, B) co-occurrences for objects that surpass an object size threshold of 2%, and C) co-occurrences weighted by object size. Toy appearance frequencies are normalized by the naming frequency to reduce the confound of the uneven naming frequency distribution. The values in each plot have been normalized to the same range.

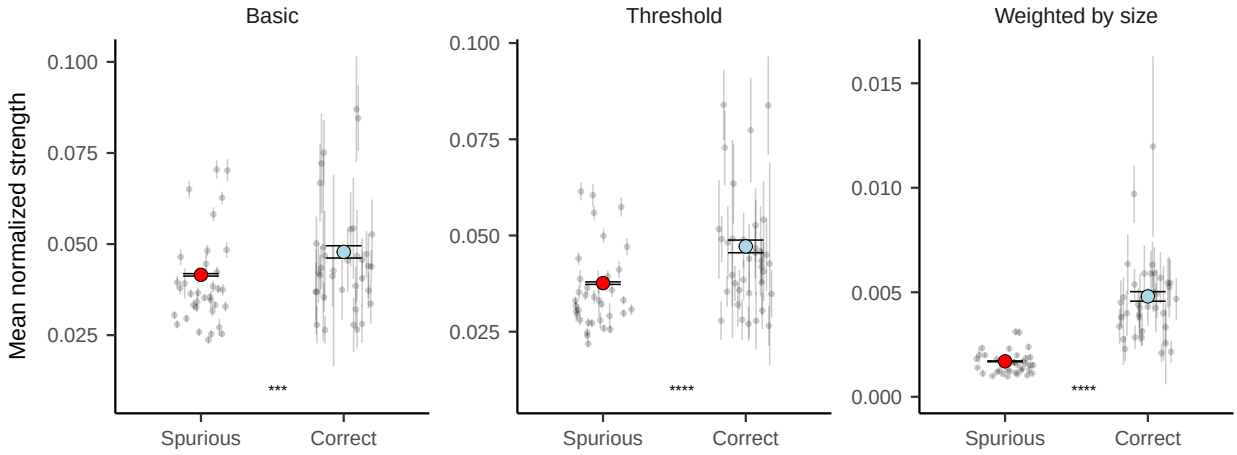


Figure 11

Comparison between the correct (diagonal) and spurious (off-diagonal) co-occurrences of the associative matrices. The colored dots represent the associative strengths averaged across the whole data set. Subject- and item-level means are in the background. Significance of difference in means indicated at the bottom of each subplot.

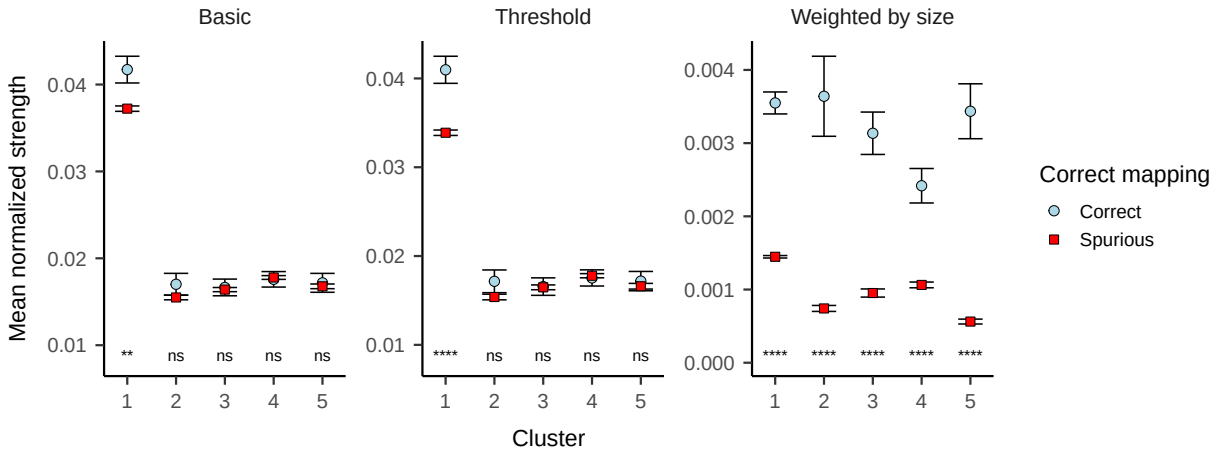


Figure 12

Average associative strength of correct and incorrect label object mappings across the different clusters and calculation types. Error bars represent 95% *confidence interval. T-test significance reported below each pair.*

References

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. B., & Yu, C. (2018). Toddler-inspired visual object learning. *Advances in Neural Information Processing Systems*, *31*, 1201–1210.

Bhat, A. A., Spencer, J. P., & Samuelson, L. K. (2022). Word-object learning via visual exploration in space (wolves): A neural process model of cross-situational word learning. *Psychological Review*, *129*(4), 640.

Bloom, L., Tinker, E., & Margulis, C. (1993). The words children learn: Evidence against a noun bias in early vocabularies. *Cognitive development*, *8*(4), 431–450.

Bohn, M., & Frank, M. C. (2019). The pervasive role of pragmatics in early language. *Annual Review of Developmental Psychology*, *1*(1), 223–249.

Comaniciu, D., & Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(5), 603–619. https://doi.org/10.1109/34.1000236

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

Fitneva, S. A., & Christiansen, M. H. (2017). Developmental changes in cross-situational word learning: The inverse effect of initial accuracy. *Cognitive Science*, *41*, 141–161.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological science*, *20*(5), 578–585.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

Henderson, J. M., Hayes, T. R., Peacock, C. E., & Rehrig, G. (2019). Meaning and attentional guidance in scenes: A review of the meaning map approach. *Vision*, *3*(2), 19.

Henderson, J. M., Malcolm, G. L., & Schandl, C. (2009). Searching in the dark: Cognitive relevance drives attention in real-world scenes. *Psychonomic bulletin & review*, *16*(5), 850–856.

Hendrickson, A. T., & Perfors, A. (2019). Cross-situational learning in a zipfian environment. *Cognition*, *189*, 11–22.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not weird. *Nature*, *466*(7302), 29–29.

Jara-Ettinger, J., Levy, R., Sakel, J., Huanca, T., & Gibson, E. (2022). The origins of the shape bias: Evidence from the tsimane'. *Journal of Experimental Psychology: General*.

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.

Lavi-Rotbain, O., & Arnon, I. (2022). The learnability consequences of zipfian distributions in language. *Cognition*, *223*, 105038.

Lavi-Rotbain, O., & Arnon, I. (2023). Zipfian distributions in child-directed speech. *Open Mind*, *7*, 1–30.

Markman, E. M. (1989). *Categorization and naming in children: Problems of induction*. mit Press.

Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. R. (2011). How words can and cannot be learned by observation. *PNAS Proceedings of the National Academy of Sciences of the United States of America*.

Meylan, S. C., & Bergelson, E. (2022). Learning through processing: Toward an integrated approach to early word learning. *Annual review of linguistics*, *8*(1), 77–99.

573 Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review
574 and future directions. *Psychonomic bulletin & review*, *21*(5), 1112–1130.
575 Plunkett, K. (1997). Theories of early language acquisition. *Trends in cognitive sciences*,
576 *1*(4), 146–153.
577 Pruden, S. M., Hirsh-Pasek, K., Golinkoff, R. M., & Hennon, E. A. (2006). The birth of
578 words: Ten-month-olds learn words through perceptual salience. *Child development*,
579 *77*(2), 266–280.
580 Quine, W. V. O. (1960). Word and object: An inquiry into the linguistic mechanisms of
581 objective reference.
582 Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv*.
583 Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object
584 name learning provides on-the-job training for attention. *Psychological science*,
585 *13*(1), 13–19.
586 Smith, L. B., & Yu, C. (2008). Infants rapidly learn word-referent mappings via
587 cross-situational statistics. *Cognition*, *106*(3), 1558–1568.
588 https://doi.org/10.1016/j.cognition.2007.06.010
589 Stevens, J. S., Gleitman, L. R., Trueswell, J. C., & Yang, C. (2017). The pursuit of word
590 meanings. *Cognitive science*, *41*, 638–676.
591 Suarez-Rivera, C., Linn, E., & Tamis-LeMonda, C. S. (2022). From play to language:
592 Infants' actions on objects cascade to word learning. *Language Learning*, *72*(4),
593 1092–1127.
594 Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify:
595 Fast mapping meets cross-situational word learning. *Cognitive psychology*, *66*(1),
596 126–156.
597 Vogt, P. (2012). Exploring the robustness of cross-situational learning under zipfian
598 distributions. *Cognitive Science*, *36*(4), 726–739.

599 Wolfe, J. M., & Horowitz, T. S. (2017). Five factors that guide attention in visual search.
600 *Nature human behaviour*, *1*(3), 0058.
601 Wolters, L., Lavi-Rotbain, O., & Arnon, I. (2024). Zipfian distributions facilitate children's
602 learning of novel word-referent mappings. *Cognition*, *253*, 105932.
603 Yu, C., & Smith, L. B. (2007). Rapid Word Learning Under Uncertainty via
604 Cross-Situational Statistics. *Psychological Science*, *18*(5), 414–420.
605 https://doi.org/10.1111/j.1467-9280.2007.01915.x
606 Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers.
607 *Cognition*, *125*(2), 244–262. https://doi.org/10.1016/j.cognition.2012.06.016
608 Yu, C., Zhang, Y., Slone, L. K., & Smith, L. B. (2021). The infant's view redefines the
609 problem of referential uncertainty in early word learning. *Proceedings of the*
610 *National Academy of Sciences*, *118*(52).
611 Yurovsky, D., Smith, L. B., & Yu, C. (2013). Statistical word learning at scale: The baby's
612 view is better. *Developmental science*, *16*(6), 959–966.
613 Zhang, Y., Amatuni, A., Cain, E., & Yu, C. (2020). Seeking meaning: Examining a
614 cross-situational solution to learn action verbs using human simulation paradigm.
615 *Proceedings of the Annual Meeting of the Cognitive Science Society*, *42*(42).
616 Zhang, Y., & Yu, C. (2022). Examining real-time attention dynamics in parent-infant
617 picture book reading. *Proceedings of the Annual Meeting of the Cognitive Science*
618 *Society*, *44*(44).
619 Zipf, G. K. (1949). Human behavior and the principle of least effort.
620 Zipf, G. (1935). The psycho-biology of language: An introduction to dynamic philology.