PHILOSOPHICAL TRANSACTIONS B

rstb.royalsocietypublishing.org

Research

Cite this article: Clerkin EM, Hart E, Rehg JM, Yu C, Smith LB. 2017 Real-world visual statistics and infants' first-learned object names. *Phil. Trans. R. Soc. B* **372**: 20160055. http://dx.doi.org/10.1098/rstb.2016.0055

Accepted: 29 July 2016

One contribution of 13 to a theme issue 'New frontiers for statistical learning in the cognitive sciences'.

Subject Areas:
cognition

Keywords:
statistical learning, infants, word learning, visual statistics, egocentric vision

Author for correspondence:
Linda B. Smith
e-mail: smith4@indiana.edu

Real-world visual statistics and infants' first-learned object names

Elizabeth M. Clerkin[1], Elizabeth Hart[1], James M. Rehg[2], Chen Yu[1] and Linda B. Smith[1]

[1]Department of Psychological and Brain Science, Indiana University, Bloomington, IN 47203, USA
[2]Interactive Computing, Georgia Institute of Technology, Atlanta, GA 30332, USA

LBS, 0000-0001-7163-8181

We offer a new solution to the unsolved problem of how infants break into word learning based on the visual statistics of everyday infant-perspective scenes. Images from head camera video captured by 8 1/2 to 10 1/2 month-old infants at 147 at-home mealtime events were analysed for the objects in view. The images were found to be highly cluttered with many different objects in view. However, the frequency distribution of object categories was extremely right skewed such that a very small set of objects was pervasively present—a fact that may substantially reduce the problem of referential ambiguity. The statistical structure of objects in these infant egocentric scenes differs markedly from that in the training sets used in computational models and in experiments on statistical word-referent learning. Therefore, the results also indicate a need to re-examine current explanations of how infants break into word learning.

This article is part of the themed issue 'New frontiers for statistical learning in the cognitive sciences'.

1. Introduction

Despite 30 years of intensive study, we do not know how infants learn their first words. The core theoretical problem facing researchers in this field is referential ambiguity. Novice learners must acquire language by linking heard words to perceived scenes, but everyday scenes are highly cluttered. For any heard word, there are many potential referents, and thus within any learning moment, there is much uncertainty as to the relevant scene elements for determining the meaning of an unknown word [1,2]. By the time children are 2 years old, they have many resources at their disposal to resolve referential ambiguity; a large experimental literature [3–6] documents their skill in using social cues, known words, category knowledge and pragmatic context to determine the referent of a novel word. However, that same literature shows that this knowledge develops incrementally during earlier stages of word learning (e.g. [7–9]) and reflects the specific cultural and language context in which learning takes place (e.g. [10–13]). Thus, the prowess of 2-year-old children depends on processes and knowledge not available to 1-year-old infants, yet these young infants have already begun learning object names [14–16]. How, given the clutter in everyday scenes, have 1-year-old infants managed to map heard names to their referents?

(a) The unrealized promise of statistical word-referent learning

Recent theory and experiments have offered a promising solution to the beginning of object name learning, one that does not require the infant to resolve referential ambiguity within a single situational encounter [17]. Instead, young learners could keep track of the potential referents that co-occur with a word across different situations and use that aggregated data to statistically determine the likely referents of to-be-learned words [18]. This form of bottom-up learning could explain how infants acquire their first object names and how they get on the

THE ROYAL SOCIETY PUBLISHING

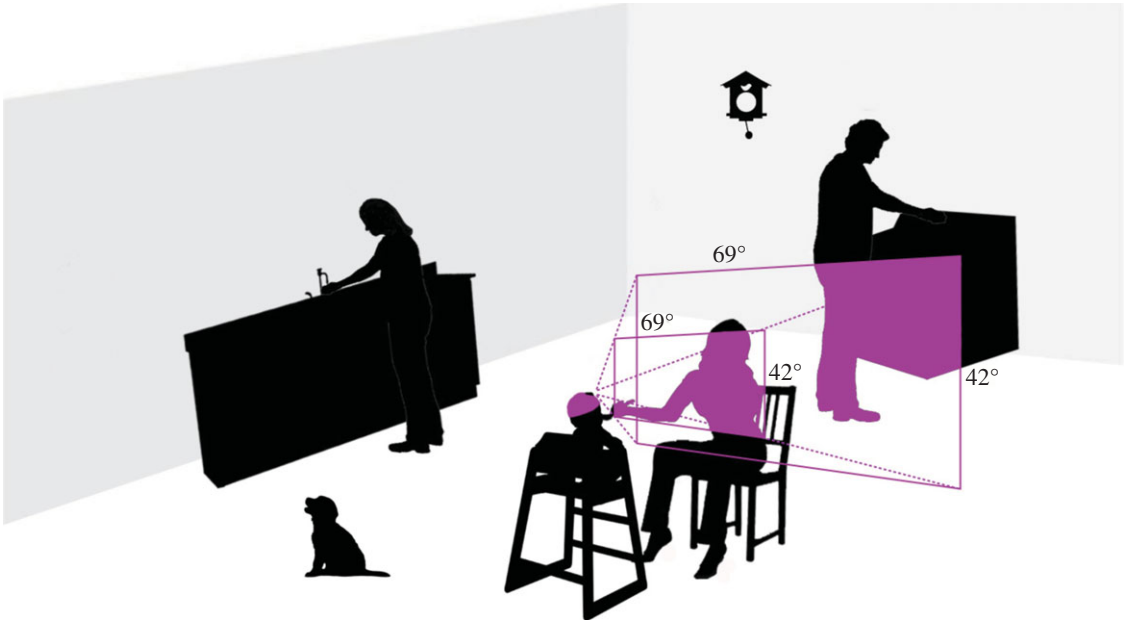

Figure 1. The selective nature of egocentric views. The field of view indicated with shading corresponds to the field of view of the head camera used in the study. The rectangles illustrate the frustrum of the camera field of view, which is 69° in the horizontal and 42° in the vertical.

path to becoming 2-year olds who have the knowledge and skills to infer the likely referent of a novel name within a single coherent context.

The problem with this statistical learning solution is that it may be beyond the abilities of 1-year-old infants [17]. Considerable evidence shows that adults and a whole variety of computational models are very good at cross-situational word-referent learning even in contexts of high uncertainty (e.g. [7,9,19–22]). Several studies have also shown that 1-year-old infants can aggregate word-referent co-occurrence data in simple laboratory experiments with minimal visual clutter and short temporal lags between repeated naming events (e.g. [18,23,24]). However, other studies show that even minor increases in task complexity disrupt infant learning [17,23,24]. To emphasize the non-scalability of infant statistical word-referent learning to the real world, Trueswell, Gleitman and coworkers [25,26] asked adults to guess the intended referent when presented with the video (but not the audio) of parents naming objects for their toddlers. Adults proved to be very poor at this and showed no ability to aggregate information about word-referent correspondences across these highly cluttered visual scenes. In brief, 1-year-old infants' perceptual, attentional and memory systems may be insufficiently robust to handle the clutter of everyday scenes. Thus, the question of how infants learn their first object names remains unanswered.

(b) Visual statistics as the proposed solution

Here we provide evidence that the visual ambiguity problem may be solved, at least in part, by the frequency distribution of objects in infants' visual worlds. Research in vision indicates that long-term experience with specific visual object categories markedly increases their recognition in suboptimal visual conditions of clutter, partial occlusion and unusual views (e.g. [27–29]); i.e. the viewing conditions of everyday life. Experiments on novel word learning with infants in the laboratory also indicate that visual familiarity with objects (prior to naming) enhances learning and retention of the name-object link [30–32]. Thus, by supporting perceptual and memory processes, visual familiarity could be critical to statistical learning. Further, if only a relatively small set of recurring objects in the lives of infants are highly frequent, these high-frequency objects could comprise a privileged class of candidates as the referents of heard object names. By this *Pervasiveness Hypothesis*, some object categories are naturally much more prevalent than others and the familiarity this prevalence brings enables infants to track objects and their co-occurrences with heard words across scenes.

Consistent with this proposal, the frequency distribution of object categories in large corpora of visual scenes are characterized by a few objects that are very frequent and many objects that occur much more rarely [33,34]. Like the distribution of elements in many naturally occurring phenomena [35–37], the frequency distribution of visual objects in these corpora of visual objects are extremely right skewed and best described by a power law. However, the data to date on the distribution of object categories come from analyses of natural scene databases [33,34]. These scenes are photographs purposefully taken by adults and thus potentially biased in their content by the visual and cognitive systems of the photographer [38–40]. More critically, they are not the scenes experienced by infants in the early stages of word learning.

Accordingly, as a first step in testing the Pervasiveness Hypothesis, we sought to determine the frequency distribution of visual objects in infant egocentric scenes. Egocentric vision is an emerging field that studies vision from the individual perceiver's point of view; a viewpoint dependent on momentary location and bodily orientation [38–41]. Research with infants, toddlers and adults (e.g. [39,40,42–47]) using head cameras and head-mounted eye trackers shows that egocentric views are highly selective (see [38] for review). Figure 1 provides an illustration; the environment near the infant contains many objects: the sink, the father, the woman at the sink, the clock and the dog. These are all in the same vicinity as the infant and could be seen by the infant—if the infant were located in the room differently or if the infant turned his or her head and eyes to those objects. But at the moment captured in the image, none of these

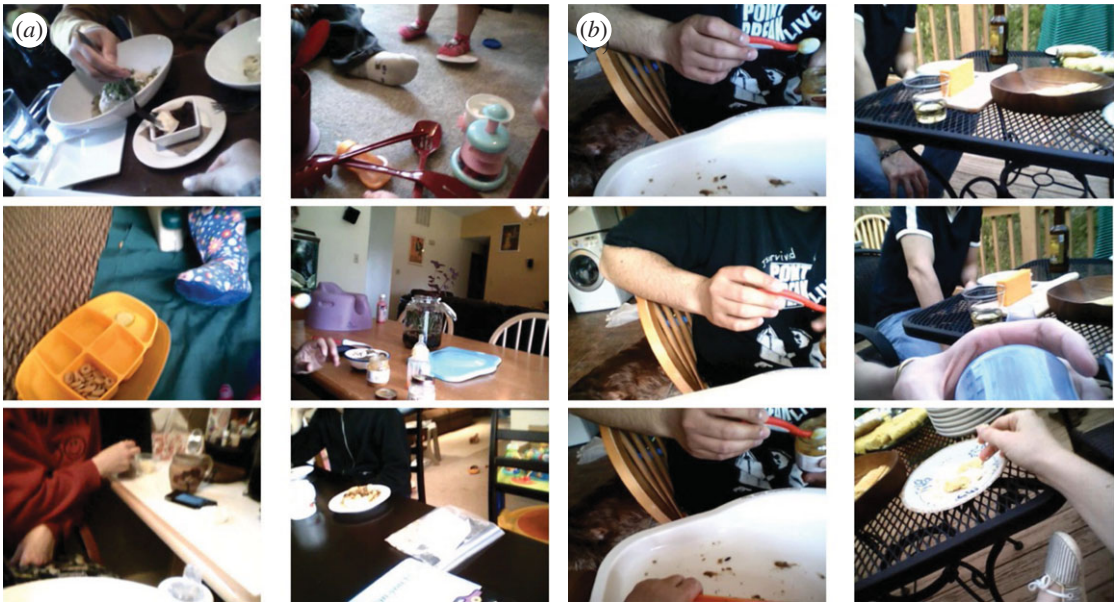

Figure 2. Infant-perspective head camera images. The left panel of six images labelled (*a*) shows the variety of 'mealtime' contexts examined in this study. The panel labelled (*b*) shows two sequences (one per column) of images sampled at 0.2 Hz.

objects is in the infant's view; instead, the momentary view is much narrower. This momentary selectivity of some scenes and their contents over others—if governed by principled internal and external constraints on infant location and infant bodily orientation—has the potential to systematically, day in and day out, privilege some object categories over others. By the Pervasiveness Hypothesis, these high-frequency objects should coincide with the object names normatively learned first by infants.

2. Material and methods

(a) Collection of the corpus

The corpus of scenes was collected from eight infants (three males) who were all between the ages of 8 1/2 and 10 1/2 months. We focused on this age because it is just prior to the traditional milestone (the first birthday) of first words [2] and because recent laboratory studies indicate that infants this age may already know some name-object mappings [14–16]. To collect the infant-perspective scenes, we used a commercially available, wearable camera (Looxcie) that was easy for parents to operate, safe (did not heat up) and very lightweight (22 g). The camera was secured to a hat that was custom fit to the infant so that when the hat was securely placed on the infant, the lens was centred above the nose and did not move. Parents were given the hat with the camera and instructed how to use it at an initial meeting; they were asked to collect videos throughout the daily activities of their infant. Parents were told that we were interested in their infant's everyday activities and that they were free to choose to record whenever it suited their family's schedule. The average amount of video collected per infant was 4.4 h (s.d. = 1.4). The diagonal field of view (FOV) of the camera was 75°, vertical FOV was 42° and horizontal FOV was 69° with a 2″ to infinity depth of focus. The camera recorded at 30 Hz, and the battery life of each camera was approximately two continuous hours. Video was stored on the camera until parents had completed their recording and then transferred to laboratory computers for storage and processing.

Head cameras measure the scene in front of the viewer but do not provide direct information as to momentary gaze, which in principle could be outside of the head camera image [38]. However, head mounted eye-tracking studies show that under active viewing conditions, human observers including infants typically turn both heads and eyes in the same direction, align heads and eyes within 500 ms of a directional shift, and maintain head and eye alignment when sustaining attention [44,46,48–54]. The result is that the distribution of gaze in active viewing (not watching screens) is highly concentrated in the centre of the head camera image [44].

(b) Selection of video segments for coding

We chose a single activity context for analysis because of the likely sparseness of individual objects across contexts (see [55] for discussion). We chose mealtime as the at-home activity because it occurs multiple times per day every day, and for infants this age, does so in various contexts and postures. Thus mealtime should yield a large number of distinct types of objects – but by hypothesis, a relatively small set of high-frequency object categories. Mealtime was defined very broadly as including any eating behaviour by anyone wherever it occurred as well as closely related activities such as preparation of and cleaning up after meals. A total of 8.5 h of video of 147 individual events (with a mean duration of 3.5 min, s.d. = 7.2) met this definition. Of the 147 mealtimes, 16 conformed to the image in figure 1 of the infant sitting in a high chair; the rest—as illustrated in the head camera images in figure 2*a*—occurred in a variety of contexts, including on the floor, at restaurants and while the infant was being carried. The total number of mealtime frames was 917 207 (mean per infant = 114 651, s.d. = 57 785). These scenes comprise a normative mealtime corpus for 8 1/2 to 10 1/2 month-old infants and thus reflect the aggregated object frequencies across contributing infants.

(c) Coding of head camera images

The 917 207 frames in the mealtime corpus were sampled at 0.2 Hz (one image every 5 s) for coding as illustrated in figure 2*b*, which yielded a total of 5775 coded scenes. Sampling at 0.2 Hz should not be biased in any way to particular objects and appears to be sufficiently dense to capture major regularities (see [56] for relevant data).

Scene-to-text coding is an approach that has been well used in previous research on scene regularities with respect to visual object recognition (e.g. [33,34]). For this study, adults (through Amazon Turk) were paid to label the objects in each image. Each

coding set consisted of 20 sequentially ordered images (sampled from 100 s of the video). For each image in a coding set, four coders were asked to provide labels for five in-view objects. The instructions to coders were pre-tested on Amazon Turk and then hand-checked for the effectiveness, frame-by-frame for a set of 200 frames and 15 Turk coders. The final instructions consisted of a set of eight training scenes and feedback that coders completed prior to each of the 20 sequential scenes in each coding set. The training scenes were structured to clarify the following instructions: to exclude body parts but not clothing, to name objects with everyday nouns ('spoon' not 'soup spoon' or 'silverware'), to label images on objects (e.g. the pictured elephant on a child's cup), to prioritize foreground not background objects (the toy on the floor not the floor) and to not repeat names in an image (if there were three cups in view, the word 'cup' could be provided just once in the list of five objects for that scene). We used this conservative approach to multiple instances of a category, one that underestimates the frequency of individual objects, because of the difficulty of counting individual objects within groups (baskets of balls, cupboards of cups, drawers of spoons). Coders were asked to try to supply five unique object labels per image but could supply fewer if the image was sparse (e.g. only one cup was in view).

The labels supplied by the coders were 'cleaned' with respect to the following properties: spelling errors were corrected, plural forms were changed to singular (e.g. dogs to dog), abbreviations were changed to the whole word (e.g. fridge to refrigerator) and adjectives were removed (e.g. redbird to bird). However, distinct names—even if semantically close, such as *mug* and *cup*, or *couch* and *sofa*, were not collapsed. We made this decision on four grounds: first, closely related words such as *mug* and *cup* are not collapsed on the normative child vocabulary inventories [57]. Second, for the full variety of objects and object names that might be supplied by coders, there is no rigorous system for determining which names are close enough to comprise a single object category versus which are not. Third, basic level categories are defined by the common nouns used by people to label objects, and thus the data from the coders themselves seems the most defensible approach. Fourth, this approach seemed least likely to systematically bias certain words over others.

The main dependent measure was the unique objects, or Types, listed by the four coders of each image. Because these scenes contain potentially many more objects that the five listed by any individual coder, and because variability among coders in the five objects listed is a potential measure of the clutter in the scene, we included all object labels offered by any coder as present in a scene. Given four coders and five potential object labels by each coder, the maximum number of unique Types per image was 20.

(d) Age of acquisition categories

We defined three age of acquisition (AoA) categories for the names of the objects provided by the coders. *First Nouns* were defined from the receptive vocabulary norms of the Bates-MacArthur Infant Communicative Developmental Inventory. This is a widely used parent-report checklist with considerable reliability and validity [57]. This inventory—designed for 8- to 16-month olds—contains 396 words, 172 of these are names for concrete objects (excluding body parts), and all these names were in the receptive vocabulary of at least 50% of 16-month-old infants in a large normative study [57]. *Early Nouns* were defined from the productive vocabulary norms (there are no receptive norms) for the Bates-MacArthur Toddler Communicative Developmental Inventory [57]. This parent checklist, again with considerable reliability and validity, consists of object names that were in the productive vocabularies of at least 50% of 30-month-old children in a large normative study [57]. We designated as Early Nouns only the 105 names for concrete things that were not also on the infant form (that is, on our First Noun list). These Early Nouns name very common everyday objects and thus one might expect these to be objects to be visually frequent in households with children. However, the critical key prediction is that the objects named by these Early Nouns, unlike objects named by First Nouns, will *not* be pervasively frequent in the egocentric views of 8- to 10-month-old infants. *Later Nouns* consisted of all other labels supplied by the coders.

Table 1. The 30 most frequent object Types named by nouns on the First Nouns list, the Early Nouns list and the Later Nouns list.

30 most frequent object Types by AoA category		
First Nouns	**Early Nouns**	**Later Nouns**
table	tray	shelf
shirt	jar	container
chair	sofa	bag
bowl	tissue	counter
cup	napkin	lid
bottle	basket	curtain
food	washing machine	tablecloth
window	knife	bin
pants	dryer	cabinet
spoon	bench	seat
toy	can	painting
plate	yogurt	handle
door	bucket	wood
picture	sauce	fireplace
couch	walker	cloth
box	sandwich	cushion
glasses	belt	straw
telephone	grass	mug
glass	scarf	outlet
light	closet	cord
book	pretzel	letters
sweater	soda	frame
paper	sidewalk	sweatshirt
refrigerator	ladder	dresser
blanket	potato	railing
jeans	stick	ring
pillow	stone	tub
lamp	strawberry	vase
plant	popcorn	desk
fork	garbage	trim

3. Results

The coders labelled 745 unique object Types, 133 of which were labelled by nouns on the First Nouns list (77% of the possible First Nouns) and 59 of which were labelled by Early Nouns (55% of the possible Early Nouns); 553 other object Types were reported as in view and thus make up the Later Nouns comparison group. By adult subjective judgement norms [58], the average AoA for the object names in this category was 6.15 years (s.d. = 1.56). Table 1 shows the 30 most frequent

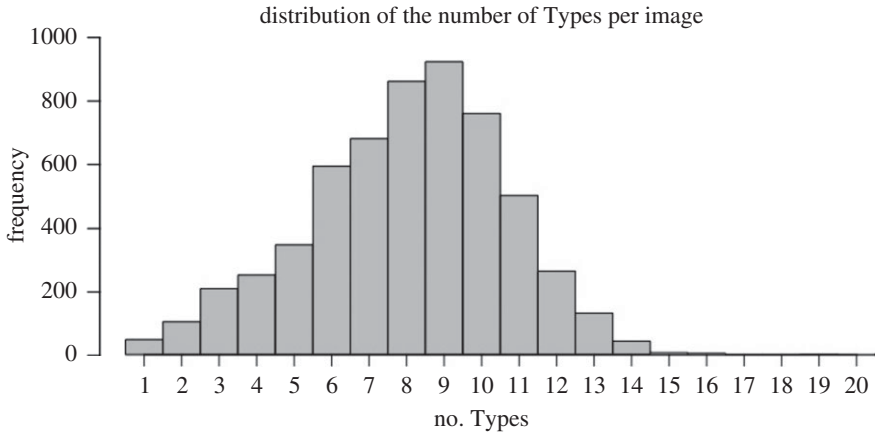

Figure 3. Histogram of the number of reported object Types per image, showing the clutter (i.e. many objects present) characteristic of these infant-perspective views.

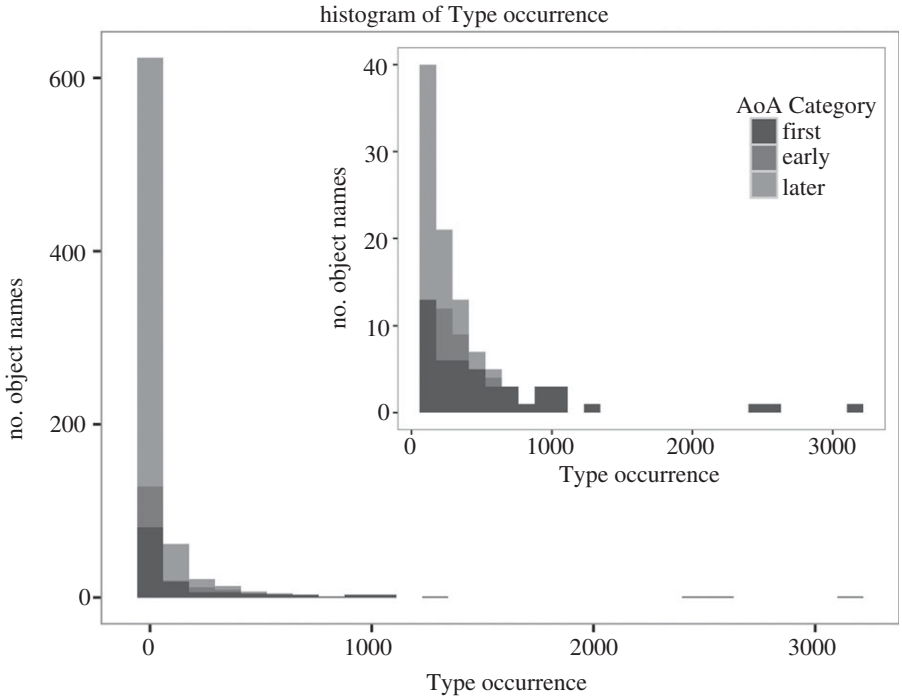

Figure 4. The frequency distribution of unique object Types occurring in the scenes. Inset is the frequency distribution of the 100 most frequently occurring object Types, omitting the 645 least frequent objects. Even among the 100 most frequent objects, the distribution is extremely right skewed.

object Types in the three AoA categories. In brief, many different common objects were present in these scenes, and these are objects named by nouns learned in childhood.

Figure 3 shows a histogram of the number of unique Types per image. The maximum, if each coder reported five different objects, is 20. The minimum, if all coders reported just one object and all reported the same object, is one. The histogram shows that most images were cluttered (unique Types reported per image, Mdn = 8, $M = 7.94$, s.d. = 2.73). Types labelled by First Nouns appeared in 97% of the coded images and thus show the same distribution with respect to clutter as shown in figure 3. In sum, the individual scenes in this corpus show the clutter and referential ambiguity assumed by theorists of statistical word-referent learning.

(a) Frequency distribution

Figure 4 shows that the frequency distribution of unique object Types is extremely right skewed. Most of the Types are very infrequent; over 75% of all types occur in 25 images or fewer. However, a small number of types are pervasively present. Seven items, less than 1% of the 745 unique object Types, occur in more than 1000 of the 5775 images, accounting for 33% of all reported object instances. An inset shows the distribution of the 100 most frequently reported Types, which includes those Types most pervasively present in the scenes. The three different shades of grey in figure 4 show the AoA categories for the names of the reported objects. As is evident in the figure, the very high-frequency objects—the tail of words that occur more than 1000 times in these scenes—are all named by nouns in the First Words category.

A power law was fitted to our data with the following estimated parameters: $\alpha = 2.44$, $x_{\min} = 238$. A Kolmogorov–Smirnov test was performed to test the power law's goodness of fit, $D = 0.07$, $p = 0.96$. The large p-value provides reasonable confidence that the observed distribution of objects in these egocentric scenes—like many natural distributions [36,59]—is well described by a power law.

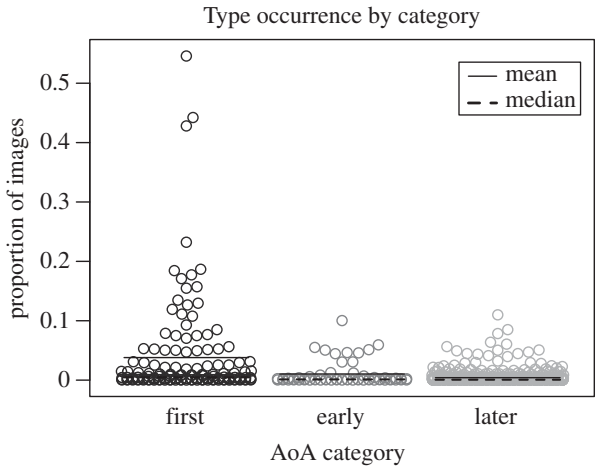

Figure 5. The proportion of images in which each object Type occurred and the mean and median for each AoA category. The most frequent Types for each AoA category are listed in table 1.

(b) Object frequency and age of acquisition

Figure 5 shows the proportion of images in which each object Type occurred as well as the median and mean proportion of images for each AoA category. As is evident, objects named by First Nouns were more frequent than objects named by Early or Later Nouns. The comparison of First and Early Nouns is particularly striking as both sets of object names are acquired quite early in childhood and refer to objects common in households with infants. We used Mann–Whitney–Wilcoxon tests to compare these frequencies due to the non-normality of the data. Object Types named by First Nouns (Mdn = 0.47% of images) were more frequent than Early Nouns (Mdn = 0.12% of images), $U = 5145$, $p < 0.001$, and they were also more frequent than Later Nouns (Mdn = 0.07% of images), $U = 54\,610.5$, $p < 0.0001$. The 15 most frequent objects in these scenes were all on the First Noun list. These findings provide direct evidence for two predictions by the Pervasiveness Hypothesis: the infant's egocentric views during mealtime persistently present a small set of objects, and those high-frequency object categories have names that are among the very first object names normatively learned first by infants.

The theoretical import of these results is highlighted by considering their day in and day out consequences. We used two sources of information to estimate the total amount of mealtime experiences infants would have between 8 and 10 months of age: (i) the frequency and duration of eating events per hour of collected video for each infant in this sample, and (ii) the frequency of eating events from a time-sampling study of daily activities of infants this age [60]. By both measures, 8- to 10-month-old infants engage in about five eating events a day, and from the present head camera recordings, those events each last about 3.5 min. Images (frames) are a measure of time. Accordingly, we combined this information to estimate the cumulative number of hours of visual experience of each object Type that would be expected over the two-month period from 8 to 10 months. Figure 6 shows that the estimated cumulative frequency of the 10 most frequent object categories—all named by nouns in the First Words category; over this two-month period, the total estimated experience markedly outpaces all other reported object categories. In brief, from their day in day out mealtime experiences, some object categories would be visually very

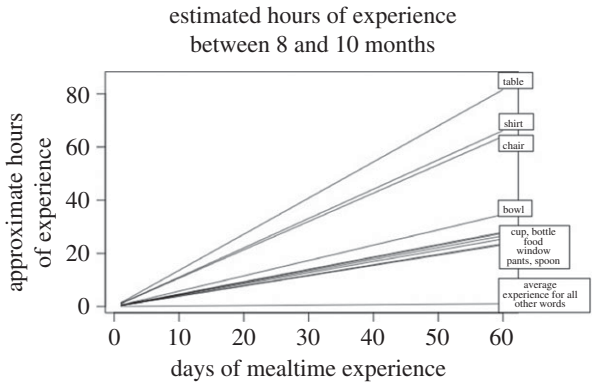

Figure 6. Projected visual experience (in hours) over the period between 8 and 10 months for the 10 most frequent objects and the average for all other objects.

well known to 10-month old infants; the many other visual objects that comprise the visual clutter in everyday scenes would be experienced much more rarely.

4. General discussion

From the infant perspective, individual mealtime scenes are highly cluttered with many different objects in view. However, across the scenes in the corpus, there is a small set of objects that are repeatedly present. This statistical fact about the distribution of visual objects suggests that not all the objects in the clutter are equal contenders as referents for heard words. By the Pervasiveness Hypothesis, the prevalence of a few object categories biases the candidate referents that are tracked and linked to heard words, a proposal that is supported by the observed correspondence between the set of highly frequent visual objects and normatively first-learned object names. These findings provide new insights about how truly novice learners may break into object name learning. They also illustrate how a developmentally informed conceptualization of statistical learning may emerge as the plausible mechanism through which infants learn their first object names.

(a) The visual side of name-object learning

The general consensus is that infants learn their first object names through ostensive definition, by linking heard words to seen objects. There is a large literature documenting infants' abilities to find words in the speech stream and on the perceptual and statistical learning that supports that necessary step to word-referent learning (see [61]). By contrast, there has been little study of the visual side of the problem. In part, this neglect derives from the oft-cited assumption that basic level categories—the object categories named by First Nouns—'carve nature at its joints' and are 'given' to young learners by the structure in the world and by their visual systems (e.g. [62,63]). However, this characterization does not fit contemporary understanding of visual object recognition [28,64]. Certainly, our adult ability to visually recognize objects is rapid, robust and accurate [64]. But the developmental evidence shows that this is hard won, through a protracted, multi-pathed, learning-dependent set of processes and that object recognition does not show its full adult prowess until adolescence [65–68]. Although, there is much that we do not know about the development of these processes, we do know

that recognition in clutter, under partial occlusion, from multiple views, and across different instances of the same category are all challenging—and are particularly so for infants and toddlers [69–72]. These challenging visual contexts are the setting of everyday learning. However, everything we know about perceptual learning indicates that extensive visual experience with specific categories leads to their more robust detection and recognition in challenging contexts [73].

Infants learning their first object names are in the early stages of building the internal mechanisms for processing and representing visual objects [66]. Their greater visual familiarity with some objects over others may yield a reasonably small set of visual object categories that can be recognized, detected and robustly remembered across the clutter and complexity of everyday scenes. If these ideas are near right, then the referential ambiguity of these scenes for young learners may be much reduced relative to judgements of these same scenes by adults. This hypothesis, in need of direct experimental test, is a form of the 'less as more' hypothesis [74,75]: infants who lack the perceptual and cognitive power of adults may be solving simpler and more tractable tasks than one would think from the adult perspective.

When does the visual learning that, by hypothesis, privileges high-frequency objects occur—prior to or simultaneous with learning the name? The correspondence of high-frequency visual objects in 8- to 10-month-old infants' egocentric views with first-learned object names—names normatively learned after the first birthday—may indicate that critical visual learning occurs developmentally prior to word-referent learning. This sequence of learning, first about visual objects and then about the links between objects and names, fits a theoretical idea prevalent in deep-learning solutions to visual object recognition. Successful visual object recognition is often achieved via an unsupervised phase of visual learning that precedes and speeds a supervised phase in which object categories are labelled [76]. However, it may also be the case that visual learning and object name learning proceed in tandem. Recent studies measuring 6- to 12-month-old infants' receptive knowledge of object names suggest that normative measures (based on parent report) may overestimate the age of acquisition of early learned names as these young infants consistently looked to named referents when tested in the laboratory, showing at the very least partial knowledge of name-object mappings [4]. Further, although the current analyses only considered the visual statistics, it is likely that the infants are also hearing the spoken names of the high-frequency objects.

All this is relevant to the plausibility of cross-situational word-referent learning as an account of novice word learning. Computational models and analyses of the statistical learning problem have made it clear that even small decreases in ambiguity facilitate statistical word-referent learning [77]. The visual prevalence of a select set of objects and the more robust visual processing likely associated with that prevalence simplifies the computational problem. Further, the source of simplification—visual familiarity—seems likely to directly counter the limitations on attention and memory that have been observed with novel objects in laboratory studies of infant cross-situational word-referent learning [23,24].

(b) Right-skewed distributions and statistical learning

The entire set of visual objects identified in infants' head camera images includes both a smaller set of high-frequency items *and* a much larger set of infrequent items. One possibility is that only the few high-frequency objects contribute to learning and that the many more numerous but infrequent objects simply do not occur enough to be registered and, therefore, are irrelevant to the account of how infants learn their first object names. This idea is consistent with both theory and evidence indicating that for novices in the early stages of learning, consistency and a small set of learning targets is optimal [78,79]. By this idea, a world that presented young learners with only the high-frequency objects and did not include the many low frequency objects at all would be ideal. There are several reasons, however, to believe that this might be the wrong conclusion and that the right-skewed distribution itself is relevant to visual learning about the high-frequency objects.

The experience of high-frequency objects in the clutter of many rarer objects may create what has been called a 'desirable difficulty' [80], by forcing the learning system to define appropriate category boundaries and by preventing recognition solutions that are over-fit to just a few experienced targets (e.g. [33,78,79]). Further, power-law distributions present other regularities that are likely relevant to learning. These distributions are believed to be the product of the multi-scale dynamics of the processes that generate the distribution and other inter-related regularities ([35–37]; see also [81]). One property is scale invariance: the right-skewed shape of the frequency distribution characterizes the distribution at different scales. For example, the distribution of objects within one meal, within all the mealtime clips, within a larger corpus of activities that include mealtime, play and dressing would retain the same shape. The object categories at the heads and tails may be different, but there would be many rare objects and a few highly frequent objects. These multi-scale non-uniform distributions are a likely product of the coherent, non-random, structure of the physical world. For example, within infants' everyday environments, bowls are likely to be in scenes with tables, bottles, cups and spoons; moreover, scenes with these co-occurring objects are likely to occur close in time. When these co-occurrences, in space (same scene) and/or time (adjacent scenes), are represented as networks, they typically exhibit a scale-free or small-world structure, showing clusters of inter-related items as well as sparser links among the clusters [82–84]. The structure of these small-world networks, in turn, are relevant to how new items are added to the network with the addition of new items predicted by their connectivity to already required items (see [85]). Past research shows that by the time children know 50 words, the network structure of their vocabularies shows the characteristic properties of small-world networks [86]. Small-world patterns in the statistics of *visual* objects could developmentally precede those in vocabulary and be central not just to understanding how infants acquire first words, but to how their vocabularies grow.

Because this is the first study to examine the frequency distribution of objects in infant-perspective views, there are many questions that remain to be answered about the natural statistics of objects in infant egocentric scenes. However, we do know that infants learn visual object categories and they learn object names. A reasonable assumption is that the learning mechanisms that accomplish this are well fit to the natural distributional statistics of words and objects in infant lives. Although non-uniform distributions have been shown to make statistical word-referent learning more difficult for some models and for adult learners ([87,88] but see [9]), in the natural learning contexts of infants, these non-uniform

distributions may be part of the solution to the problem of referential ambiguity.

(c) Developmental changes in visual environments

Infants are not stationary learners; motor, perceptual, attentional and cognitive processes all change dramatically over the first 2 years of life. As a consequence, children's visual environments also change [42–46,56]. Previous studies examining egocentric vision in older word learners (16- to 18-month-olds) have shown that these older infants often create—through their own manual actions—views in which a single object is large, centred, un-occluded and visually dominant [46,47]. Parents often name objects at these moments and when they do, toddlers are highly likely to learn and remember those object names [22]. All current indications are that younger infants, prior to their first birthday, do not create these uncluttered one-object-in-view scenes because they are not able to stably hold and sustain manual actions on objects [22]. The within-scene clutter observed in this study of 8- to 10-month-olds fits this pattern and suggests that the learning of *first* object names may take place in a different visual environment than the learning of early and later object names. If this is so, visual pervasiveness and the resultant visual familiarity with particular objects may be most critical to object name learning only at the start of word learning. In sum, the visual environment for learning first object names may have unique properties that differ substantially from those several months forward.

Ethics. All experimental protocols and consent materials were approved by the Indiana University Institutional Review Board. Parents of all participating infants provided written informed consent prior to the experiment.

Data accessibility. The text object labels provided by all coders for all images and the metadata associating images with specific individuals and the 147 mealtime events will be deposited to the Databrary repository at https://databrary.org/volume/268. Sample images that do not contain faces will also be deposited.

Authors' contributions. L.B.S., C.Y. and J.M.R. conceived the larger research program of which this study is a part. L.B.S. supervised the collection of the data. E.M.C., E.H. and L.B.S. formulated the specific research question and the analysis plan. E.H. supervised the coding of the data. E.M.C., E.H. and L.B.S. analysed the data. E.M.C. and L.B.S. wrote the paper. All authors gave final approval for submission.

Competing interests. We have no competing interests.

Funding. This article was funded by NSF grant BCS-15233982 and NIH grants R01HD 074601, R01HD 28675, T32HD007475.

Acknowledgements. We thank the families who participated in the study and our colleagues who provided additional assistance with this work, especially Ariel La, Swapnaa Jayaraman, Caitlin Fausey, Jaclyn Berning and Amanda Essex.

References

1. Quine WV. 1960 *Word and object*. Cambridge, MA: MIT Press.
2. Bloom P. 2000 *How children learn the meanings of words*. Cambridge, MA: MIT Press.
3. Tomasello M, Tomasello M. 2009 *Constructing a language: a usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.
4. Smith LB, Colunga E, Yoshida H. 2010 Knowledge as process: contextually cued attention and early word learning. *Cognit. Sci.* **34**, 1287–1314. (doi:10.1111/j.1551-6709.2010.01130.x)
5. Waxman SR, Gelman SA. 2009 Early word-learning entails reference, not merely associations. *Trends Cogn. Sci.* **13**, 258–263. (doi:10.1016/j.tics.2009.03.006)
6. Bion RA, Borovsky A, Fernald A. 2013 Fast mapping, slow learning: disambiguation of novel word–object mappings in relation to vocabulary learning at 18, 24, and 30 months. *Cognition* **126**, 39–53. (doi:10.1016/j.cognition.2012.08.008)
7. Smith K, Smith AD, Blythe RA. 2011 Cross-situational learning: an experimental study of word-learning mechanisms. *Cognit. Sci.* **35**, 480–498. (doi:10.1111/j.1551-6709.2010.01158.x)
8. Yu C, Smith LB. 2012 Modeling cross-situational word–referent learning: prior questions. *Psychol. Rev.* **119**, 21. (doi:10.1037/a0026182)
9. Kachergis G, Yu C, Shiffrin RM. 2016 A bootstrapping model of frequency and context effects in word learning. *Cognit. Sci.* (doi:10.1111/cogs.12353)
10. Imai M, Gentner D. 1997 A cross-linguistic study of early word meaning: universal ontology and linguistic influence. *Cognition* **62**, 169–200. (doi:10.1016/S0010-0277(96)00784-6)
11. Waxman SR, Senghas A, Benveniste S. 1997 A cross-linguistic examination of the noun-category bias: its existence and specificity in French- and Spanish-speaking preschool-aged children. *Cognit. Psychol.* **32**, 183–218. (doi:10.1006/cogp.1997.0650)
12. Yoshida H, Smith LB. 2005 Linguistic cues enhance the learning of perceptual cues. *Psychol. Sci.* **16**, 90–95. (doi:10.1111/j.0956-7976.2005.00787.x)
13. Fernald A, Morikawa H. 1993 Common themes and cultural variations in Japanese and American mothers' speech to infants. *Child Dev.* **64**, 637–656. (doi:10.2307/1131208)
14. Bergelson E, Swingley D. 2012 At 6–9 months, human infants know the meanings of many common nouns. *Proc. Natl Acad. Sci. USA* **109**, 3253–3258. (doi:10.1073/pnas.1113380109)
15. Tincoff R, Jusczyk PW. 2012 Six-month-olds comprehend words that refer to parts of the body. *Infancy* **17**, 432–444. (doi:10.1111/j.1532-7078.2011.00084.x)
16. Mani N, Plunkett K. 2010 Twelve-month-olds know their cups from their keps and tups. *Infancy* **15**, 445–470. (doi:10.1111/j.1532-7078.2009.00027.x)
17. Smith LB, Suanda SH, Yu C. 2014 The unrealized promise of infant statistical word–referent learning. *Trends Cogn. Sci.* **18**, 251–258. (doi:10.1016/j.tics.2014.02.007)
18. Smith L, Yu C. 2008 Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition* **106**, 1558–1568. (doi:10.1016/j.cognition.2007.06.010)
19. Frank MC, Goodman ND, Tenenbaum JB. 2009 Using speakers' referential intentions to model early cross-situational word learning. *Psychol. Sci.* **20**, 578–585. (doi:10.1111/j.1467-9280.2009.02335.x)
20. Yu C. 2008 A statistical associative account of vocabulary growth in early word learning. *Lang. Learn. Dev.* **4**, 32–62. (doi:10.1080/15475440701739353)
21. Yu C, Ballard DH. 2007 A unified model of early word learning: integrating statistical and social cues. *Neurocomputing* **70**, 2149–2165. (doi:10.1016/j.neucom.2006.01.034)
22. Yu C, Smith LB. 2012 Embodied attention and word learning by toddlers. *Cognition* **125**, 244–262. (doi:10.1016/j.cognition.2012.06.016)
23. Smith LB, Yu C. 2013 Visual attention is not enough: individual differences in statistical word-referent learning in infants. *Lang. Learn. Dev.* **9**, 25–49. (doi:10.1080/15475441.2012.707104)
24. Vlach HA, Johnson SP. 2013 Memory constraints on infants' cross-situational statistical learning. *Cognition* **127**, 375–382. (doi:10.1016/j.cognition.2013.02.015)
25. Trueswell JC, Medina TN, Hafri A, Gleitman LR. 2013 Propose but verify: fast mapping meets cross-situational word learning. *Cognit. Psychol.* **66**, 126–156. (doi:10.1016/j.cogpsych.2012.10.001)
26. Medina TN, Snedeker J, Trueswell JC, Gleitman LR. 2011 How words can and cannot be learned by observation. *Proc. Natl Acad. Sci. USA* **108**, 9014–9019. (doi:10.1073/pnas.1105040108)
27. Sun GJ, Chung ST, Tjan BS. 2010 Ideal observer analysis of crowding and the reduction of

crowding through learning. *J. Vis.* **10**, 16. (doi:10.1167/10.5.16)

28. Pinto N, Cox DD, DiCarlo JJ. 2008 Why is real-world visual object recognition hard? *PLoS Comput. Biol.* **4**, e27. (doi:10.1371/journal.pcbi.0040027)
29. Curby KM, Glazek K, Gauthier I. 2009 A visual short-term memory advantage for objects of expertise. *J. Exp. Psychol. Hum. Percept. Perform.* **35**, 94. (doi:10.1037/0096-1523.35.1.94)
30. Kucker SC, Samuelson LK. 2012 The first slow step: differential effects of object and word-form familiarization on retention of fast-mapped words. *Infancy* **17**, 295–323. (doi:10.1111/j.1532-7078.2011.00081.x)
31. Fennell CT. 2012 Object familiarity enhances infants' use of phonetic detail in novel words. *Infancy* **17**, 339–353. (doi:10.1111/j.1532-7078.2011.00080.x)
32. Graham SA, Turner JN, Henderson AM. 2005 The influence of object pre-exposure on two-year-olds' disambiguation of novel labels. *J. Child Lang.* **32**, 207–222. (doi:10.1017/S030500090400666X)
33. Salakhutdinov R, Torralba A, Tenenbaum J. 2011 Learning to share visual appearance for multiclass object detection. In *IEEE Conf. on Computer Vision and Pattern Recognition* (*CVPR*), *2011*. New York, NY: IEEE.
34. Yuen J, Russell B, Liu C, Torralba A. 2009 LabelMe video: building a video database with human annotations. In *IEEE 12th Int. Conf. on Computer Vision, (ICCV), 2009*. New York, NY: IEEE.
35. Clauset A, Shalizi CR, Newman ME. 2009 Power-law distributions in empirical data. *SIAM Rev.* **51**, 661–703. (doi:10.1137/070710111)
36. Piantadosi ST. 2014 Zipf's word frequency law in natural language: a critical review and future directions. *Psychon. Bull. Rev.* **21**, 1112–1130. (doi:10.3758/s13423-014-0585-6)
37. Kello CT *et al.* 2010 Scaling laws in cognitive sciences. *Trends Cogn. Sci.* **14**, 223–232. (doi:10.1016/j.tics.2010.02.005)
38. Smith LB, Yu C, Yoshida H, Fausey CM. 2015 Contributions of head-mounted cameras to studying the visual environments of infants and young children. *J. Cognit. Dev.* **16**, 407–419. (doi:10.1080/15248372.2014.933430)
39. Fathi A, Ren X, Rehg JM. 2011 Learning to recognize objects in egocentric activities. In *IEEE Conf. on Computer Vision and Pattern Recognition* (*CVPR*), 2011. New York, NY: IEEE.
40. Foulsham T, Walker E, Kingstone A. 2011 The where, what and when of gaze allocation in the lab and the natural environment. *Vision Res.* **51**, 1920–1931. (doi:10.1016/j.visres.2011.07.002)
41. Fausey CM, Jayaraman S, Smith LB. 2016 From faces to hands: changing visual input in the first two years. *Cognition* **152**, 101–107. (doi:10.1016/j.cognition.2016.03.005)
42. Kretch KS, Franchak JM, Adolph KE. 2014 Crawling and walking infants see the world differently. *Child Dev.* **85**, 1503–1518. (doi:10.1111/cdev.12206)
43. Frank MC, Simmons K, Yurovsky D, Pusiol G. (eds). 2013 Developmental and postural changes in children's visual access to faces. In *Proc. of the 35th annual meeting of the Cognitive Science Society* (eds M Knauff, M Pauen, N Sebanz, I Wachsmuth), pp. 454–459. Austin, TX: Cognitive Science Society.
44. Yoshida H, Smith LB. 2008 What's in view for toddlers? Using a head camera to study visual experience. *Infancy* **13**, 229–248. (doi:10.1080/15250000802004437)
45. Sugden NA, Mohamed-Ali MI, Moulson MC. 2014 I spy with my little eye: typical, daily exposure to faces documented from a first-person infant perspective. *Dev. Psychobiol.* **56**, 249–261. (doi:10.1002/dev.21183)
46. Pereira AF, Smith LB, Yu C. 2014 A bottom-up view of toddler word learning. *Psychon. Bull. Rev.* **21**, 178–185. (doi:10.3758/s13423-013-0466-4)
47. Smith LB, Yu C, Pereira AF. 2011 Not your mother's view: the dynamics of toddler visual experience. *Dev. Sci.* **14**, 9–17. (doi:10.1111/j.1467-7687.2009.00947.x)
48. Schmitow C, Stenberg G. 2015 What aspects of others' behaviors do infants attend to in live situations? *Infant Behav. Dev.* **40**, 173–182. (doi:10.1016/j.infbeh.2015.04.002)
49. Bloch H, Carchon I. 1992 On the onset of eye-head coordination in infants. *Behav. Brain Res.* **49**, 85. (doi:10.1016/S0166-4328(05)80197-4)
50. Daniel BM, Lee DN. 1990 Development of looking with head and eyes. *J. Exp. Child Psychol.* **50**, 200–216. (doi:10.1016/0022-0965(90)90039-B)
51. Ballard DH, Hayhoe MM, Li F, Whitehead SD, Frisby J, Taylor J, Fisher RB. 1992 Hand-eye coordination during sequential tasks [and discussion]. *Phil. Trans. R. Soc. Lond. B* **337**, 331–339. (doi:10.1098/rstb.1992.0111)
52. Bambach S, Crandall DJ, Yu C. 2013 Understanding embodied visual attention in child-parent interaction. In *IEEE Third Joint Int. Conf. on Development and Learning and Epigenetic Robotics* (*ICDL*), *2013*. New York, NY: IEEE.
53. Ruff HA, Lawson KR. 1990 Development of sustained, focused attention in young children during free play. *Dev. Psychol.* **26**, 85. (doi:10.1037/0012-1649.26.1.85)
54. Schmitow C, Stenberg G, Billard A, von Hoffsten C. 2013 Measuring direction of looking with a head-mounted camera. *Int. J. Behav. Dev.* **37**, 471–477. (doi:10.1177/0165025413495749)
55. Dagan I, Marcus S, Markovitch S. 1993 Contextual word similarity and estimation from sparse data. In *Proc. 31st Annu. Meet. Association for Computational Linguistics, 22–26 June 1993, Columbus, OH* (eds ACL), pp. 164–171. Morristown, NJ: Association for Computational Linguistics.
56. Jayaraman S, Fausey CM, Smith LB. 2015 The faces in infant-perspective scenes change over the first year of life. *PLoS ONE* **10**, e0123780. (doi:10.1371/journal.pone.0123780)
57. Fenson L, Dale PS, Reznick JS, Bates E, Thal DJ, Pethick SJ *et al.* 1994 Variability in early communicative development. *Monogr. Soc. Res. Child Dev.* **59**, 1–185. (doi:10.2307/1166093)
58. Kuperman V, Stadthagen-Gonzalez H, Brysbaert M. 2012 Age-of-acquisition ratings for 30,000 English words. *Behav. Res. Methods*. **44**, 978–990. (doi:10.3758/s13428-012-0210-4)
59. Newman ME. 2005 Power laws, Pareto distributions and Zipf's law. *Contemp. Phys.* **46**, 323–351. (doi:10.1080/00107510500052444)
60. Fausey CM, Smith LB. In preparation. Time-sampling study of daily events in infants from 1 to 24 months age.
61. Romberg AR, Saffran JR. 2010 Statistical learning and language acquisition. *Wiley Interdiscip. Rev. Cognit. Sci.* **1**, 906–914. (doi:10.1002/wcs.78)
62. Rosch E, Mervis CB, Gray WD, Johnson DM, Boyes-Braem P. 1976 Basic objects in natural categories. *Cognit. Psychol.* **8**, 382–439. (doi:10.1016/0010-0285(76)90013-X)
63. Gentner D. 1982 Why nouns are learned before verbs: linguistic relativity versus natural partitioning. Technical Report no. 257. Urbana, IL: University of Illinois, Center for the Study of Reading.
64. Kourtzi Z, Connor CE. 2011 Neural representations for object perception: structure, category, and adaptive coding. *Annu. Rev. Neurosci.* **34**, 45–67. (doi:10.1146/annurev-neuro-060909-153218)
65. Nishimura M, Scherf S, Behrmann M. 2009 Development of object recognition in humans. *F1000 Biol. Rep.* **1**, 56. (doi:10.3410/B1-56)
66. Smith LB. 2009 From fragments to geometric shape changes in visual object recognition between 18 and 24 months. *Curr. Direct. Psychol. Sci.* **18**, 290–294. (doi:10.1111/j.1467-8721.2009.01654.x)
67. Jüttner M, Müller A, Rentschler I. 2006 A developmental dissociation of view-dependent and view-invariant object recognition in adolescence. *Behav. Brain Res.* **175**, 420–424. (doi:10.1016/j.bbr.2006.09.005)
68. Rentschler I, Jüttner M, Osman E, Müller A, Caelli T. 2004 Development of configural 3D object recognition. *Behav. Brain Res.* **149**, 107–111. (doi:10.1016/S0166-4328(03)00194-3)
69. Kwon MK, Luck SJ, Oakes LM. 2014 Visual short-term memory for complex objects in 6- and 8-month-old infants. *Child Dev.* **85**, 564–577. (doi:10.1111/cdev.12161)
70. Kraebel KS, West RN, Gerhardstein P. 2007 The influence of training views on infants' long-term memory for simple 3D shapes. *Dev. Psychobiol.* **49**, 406–420. (doi:10.1002/dev.20222)
71. Gerhardstein P, Schroff G, Dickerson K, Adler SA. 2009 The development of object recognition through infancy. In *New directions in developmental psychobiology* (eds BC Glenyn, RP Zini), pp. 79–115. New York, NY: Nova Science Publishers.
72. Farzin F, Rivera SM, Whitney D. 2010 Spatial resolution of conscious visual perception in infants. *Psychol. Sci.* **21**, 1502–1509. (doi:10.1177/0956797610382787)
73. Logothetis NK, Sheinberg DL. 1996 Visual object recognition. *Annu. Rev. Neurosci.* **19**, 577–621. (doi:10.1146/annurev.ne.19.030196.003045)
74. Elman JL. 1993 Learning and development in neural networks: the importance of starting small.

Cognition **48**, 71–99. (doi:10.1016/0010-0277(93)90058-4)

75. Newport EL. 1990 Maturational constraints on language learning. *Cognit. Sci*. **14**, 11–28. (doi:10.1207/s15516709cog1401_2)
76. Erhan D, Bengio Y, Courville A, Manzagol P-A, Vincent P, Bengio S. 2010 Why does unsupervised pre-training help deep learning? *J. Mach. Learn. Res*. **11**, 625–660.
77. Blythe RA, Smith K, Smith AD. 2010 Learning times for large lexicons through cross-situational learning. *Cognit. Sci*. **34**, 620–642. (doi:10.1111/j.1551-6709.2009.01089.x)
78. Carvalho PF, Goldstone RL. 2014 Putting category learning in order: category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Mem. Cognit*. **42**, 481–495. (doi:10.3758/s13421-013-0371-0)
79. Carvalho PF, Goldstone RL. 2015 The benefits of interleaved and blocked study: different tasks benefit from different schedules of study. *Psychon. Bull. Rev*. **22**, 281–288. (doi:10.3758/s13423-014-0676-4)
80. Bjork EL, Bjork RA. 2009 Making things hard on yourself, but in a good way: creating desirable difficulties to enhance learning. In *Psychology and the real world* (eds MA Gernsbacher, RW Pew, LM Hough), pp. 56–64. Gordonsville, VA: Worth Publishers/Macmillan Higher Education.
81. Blythe RA, Smith AD, Smith K. 2015 Word learning under infinite uncertainty. *Cognition* **151**, 18–27. (doi:10.1016/j.cognition.2016.02.017)
82. Sadeghi Z, McClelland JL, Hoffman P. 2015 You shall know an object by the company it keeps: an investigation of semantic representations derived from object co-occurrence in visual scenes. *Neuropsychologia* **76**, 52–61. (doi:10.1016/j.neuropsychologia.2014.08.031)
83. Serrano MÁ, Flammini A, Menczer F. 2009 Modeling statistical properties of written text. *PLoS ONE* **4**, e5372. (doi:10.1371/journal.pone.0005372)
84. Baronchelli A, Ferrer-i-Cancho R, Pastor-Satorras R, Chater N, Christiansen MH. 2013 Networks in cognitive science. *Trends Cogn. Sci*. **17**, 348–360. (doi:10.1016/j.tics.2013.04.010)
85. Hills TT, Maouene M, Maouene J, Sheya A, Smith L. 2009 Longitudinal analysis of early semantic networks preferential attachment or preferential acquisition? *Psychol. Sci*. **20**, 729–739. (doi:10.1111/j.1467-9280.2009.02365.x)
86. Beckage N, Smith L, Hills T. 2011 Small worlds and semantic network growth in typical and late talkers. *PLoS ONE* **6**, e19348. (doi:10.1371/journal.pone.0019348)
87. Vogt P. 2012 Exploring the robustness of cross-situational learning under Zipfian distributions. *Cognit. Sci*. **36**, 726–739. (doi:10.1111/j.1551-6709.2011.1226.x)
88. Reisenauer R, Smith K, Blythe RA. 2013 Stochastic dynamics of lexicon learning in an uncertain and nonuniform world. *Phys. Rev. Lett*. **110**, 258701. (doi:10.1103/PhysRevLett.110.258701)